

Searching for Saddle Points by Using the Nudged Elastic Band Method: An Implementation for Gas-Phase Systems

Núria González-García,^{†‡} Jingzhi Pu,[†] Àngels González-Lafont,[‡] José M. Lluch,[‡] and Donald G. Truhlar^{*,†}

Departament de Química, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain, and Department of Chemistry and Supercomputing Institute, University of Minnesota, 207 Pleasant Street Southeast, Minneapolis, Minnesota 55455-0431

Received January 27, 2006

Abstract: A new implementation of the Nudged Elastic Band (NEB) optimization method is presented. This approach uses a global procedure that yields the whole reaction path, and thus it provides an alternative to the sequential optimization of the transition state and consequent calculation of the minimum energy path. Furthermore the algorithm is very useful when one is not sure if a saddle point exists, because it can be used to eliminate the possibility of a saddle point when one does not exist. Three different versions of the NEB algorithm have been implemented. The influences of various parameters and methodological choices on the performance of the method have been studied, and the quality of the results is assessed by comparison with the saddle point and minimum energy path calculations sequential method. Recommendations are made for algorithmic choices and default parameters.

1. Introduction

Characterization of the potential energy surface (PES) is a key step in the study of any reaction. In most cases, all detailed information about the PES is obtained from electronic structure calculations.^{1,2} The steps that are usually followed to characterize a PES are, first, locating the minima and saddle points on the PES and, second, calculating the reaction paths connecting those stationary points.^{3–6} Finding the saddle points (SPs) can be especially difficult for large systems, even in the gas phase. In many cases the decisive factor in this search is to start with a good guess, which can be obtained by many different procedures such as by analogy to previously studied systems, by finding the maximum-energy structure along an approximate reaction coordinate, by performing partial optimizations, or by carrying out a full optimization at a lower level of theory where one can afford to calculate multiple Hessians to guide the search. But even

using multiple Hessians or a good guess, this search can fail. Once the SP is optimized and characterized, the minimum energy path (MEP) can be calculated.

Many different methods have been presented for finding saddle points and reaction paths.^{3–24} Some reaction paths just provide a good guess to start a search for the saddle point. One of these methods is the distinguished reaction coordinate method,⁹ where one degree of freedom, called the distinguished coordinate, is chosen and kept fixed at a sequence of values, while all the other coordinates are relaxed for each of these values. The value of the distinguished coordinate is incremented in a stepwise fashion, and the system is dragged from reactants to products. The maximum-energy geometry along the path is taken as the initial guess for the saddle-point search. The intuitively assumed reaction coordinate can turn out to be a bad one, although some authors have developed methods that overcome some of the disadvantages of the distinguished reaction coordinate method.^{25,26} Another useful algorithm for finding saddle points is the eigenvector following (EF) method.^{10–14} This algorithm starts from a local minimum and follows the most

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

[†] University of Minnesota.

[‡] Universitat Autònoma de Barcelona.

gradually ascending generalized-normal-mode eigenvector, step by step, until reaching the saddle point. Once the saddle point is characterized, the EF algorithm can be used to trace the MEP from the saddle point to the minima it connects; the first step in such a process is to follow the unique descending eigenvector at the saddle point. Unfortunately, obtaining the generalized-normal-mode eigenvectors by diagonalizing the Hessian matrix requires expensive computations so that this method is only viable for small systems or low levels of electronic structure theory.

A promising alternative to these traditional methods is provided by a group of methods that may be classified as chain-of-states methods. In these methods, a path is represented by a set of discrete structures forming a chain of replicas of the system. The structures, called images, are then optimized to try to make the entire chain lie on the MEP. Since the MEP passes through the saddle point, these methods simultaneously locate the SP and the MEP. The Nudged Elastic Band (NEB) method^{15–24} is an example of this chain-of-states approach. The NEB method can be used either as an alternative to traditional methods when they fail or as an inexpensive way to characterize the PES. A key advantage of the NEB is that it provides a global search, whereas many traditional methods only converge in the vicinity of a good initial guess.

The present article presents some illustrations of the NEB method based on a new implementation in an electronic structure code, namely MULTILEVEL.²⁷ The implementation requires only two or three initial geometries, which are provided by the user, and after generating an initial path, the program will optimize it to the MEP. Alternatively the initial steps of the NEB minimization can be used to provide a good initial guess for a more traditional TS search.

Section 2 describes the NEB method and reviews some of the previous implementations. Section 3 tests the new implementation for several reactions and recommends a version that performs quite well along with a set of default values for several of the parameters of the method. The systems used are all related to dimethyl sulfide (DMS) degradation in the atmosphere. DMS is thought to be the major biogenic component of the global atmospheric budget. The potential role of DMS in global climate change has been a subject of considerable controversy; it has been suggested that a biological/chemical cycle based on DMS could form the basis of an efficient method of climate regulation.²⁸ Previous work on DMS and the hydroxyl radical has been published,^{29,30} and here we apply the NEB technique to some unsolved questions on reactions in that degradation scheme that had not been studied yet.

2. The Nudged Elastic Band Method

2.1. Theory. In the Nudged Elastic Band (NEB) method,^{15,16,19} the reaction path is described by a discrete sequence of images consisting of two fixed end points (\vec{R}_0 and \vec{R}_{n+1}) and n intermediate movable images ($\vec{R}_1, \vec{R}_2, \dots, \vec{R}_n$). This sequence is called the chain or the elastic band. Spring interactions are added between adjacent images. The total force (also called the adjusted force) acting on each image is the sum

of the spring force \vec{F}_i^s and the force \vec{F}_i^t from the potential energy surface (which will be called the true force). The band is optimized, minimizing the total force acting on each image. During this process, the true force tends to pull the images toward the end points, giving the lowest resolution in the region nearest to the saddle point. This behavior is known as the sliding-down phenomenon. On the other hand, corner cutting is induced by the spring force pulling the sequence of images to the concave side of the MEP in the regions where it is curved. These two problems are solved by projecting out the component of the true force parallel to the chain of images and the component of the spring force perpendicular to the chain.

Then the adjusted force acting on an image, i , is given by

$$\vec{F}_i = \vec{F}_{\parallel}^s + \vec{F}_{\perp}^t \quad (1)$$

which is the sum of the spring force along the tangent to the chain and the true force perpendicular to the chain. The parallel component of the spring force, in the first version of the method,^{15,16,19} is calculated as

$$\vec{F}_{\parallel}^s = \{k[(\vec{R}_{i+1} - \vec{R}_i) - (\vec{R}_i - \vec{R}_{i-1})] \cdot \hat{\tau}_i\} \hat{\tau}_i \quad (2)$$

where k is the spring constant and $\hat{\tau}_i$ is the unit tangent vector at an image i . Furthermore, in the original version, the tangent is estimated by using the normalized line segment between two nonadjacent images along the path, \vec{R}_{i+1} and \vec{R}_{i-1}

$$\hat{\tau}_i = \frac{\vec{R}_{i+1} - \vec{R}_{i-1}}{|\vec{R}_{i+1} - \vec{R}_{i-1}|} \quad (3)$$

but a slightly better way^{15,19,20} is to bisect the two unit vectors

$$\tau_i^{(0)} = \frac{\vec{R}_i - \vec{R}_{i-1}}{|\vec{R}_i - \vec{R}_{i-1}|} + \frac{\vec{R}_{i+1} - \vec{R}_i}{|\vec{R}_{i+1} - \vec{R}_i|} \quad (4a)$$

and then normalize so that

$$\hat{\tau}_i = \tau_i^{(0)} / |\tau_i^{(0)}| \quad (4b)$$

Using eqs 4a and 4b to define the tangent ensures that the images are equally spaced (when the spring constant k is the same for each adjacent pair) even in regions of large curvature of the path. This last way of estimating the tangent will be called the bisection NEB or B-NEB version of the NEB algorithm in this work. The tangent vector is also used to obtain the perpendicular component of the true force

$$\vec{F}_{\perp}^t = \vec{F}_i^t - (\vec{F}_i^t \cdot \hat{\tau}_i) \hat{\tau}_i \quad (5)$$

In systems where the force along the minimum energy path is large compared to the restoring force perpendicular to the path, the system can develop kinks, preventing the band from converging to the MEP. At kinks the angle between the vectors $\vec{R}_i - \vec{R}_{i-1}$ and $\vec{R}_{i+1} - \vec{R}_i$ is large, so that including some fraction of the perpendicular component of the spring force may tend to straighten the elastic band.¹⁹

An improved estimation of the tangent was proposed²⁰ to eliminate kinks. Instead of using both adjacent images, $i + 1$ and $i - 1$, just the image with the highest energy is used for the estimation of the tangent at image i . The new tangent,

which replaces eq 4a, is

$$\tau_i^{(0)} = \begin{cases} \tau_i^+ & \text{if } V_{i+1} > V_i > V_{i-1} \\ \tau_i^- & \text{if } V_{i+1} < V_i < V_{i-1} \end{cases} \quad (6a)$$

where

$$\tau_i^+ = \vec{R}_{i+1} - \vec{R}_i \text{ and } \tau_i^- = \vec{R}_i - \vec{R}_{i-1} \quad (6b)$$

and V_i is the potential energy, $V(\vec{R}_i)$, of image i . If both adjacent images are either lower or higher in energy than image i , then the tangent is taken to be a weighted average of the vectors to the two neighboring images. This weight is determined from the energy. The weighted average only plays a role at extrema along the MEP, and it serves to smoothly switch between the two possible tangents $\hat{\tau}_i^+$ and $\hat{\tau}_i^-$; otherwise, there is an abrupt change in the tangent as one image becomes higher in energy than another, and this can lead to convergence problems. If image i is at a minimum ($V_{i+1} > V_i < V_{i-1}$) or at a maximum ($V_{i+1} < V_i > V_{i-1}$), then the tangent estimate becomes

$$\tau_i^{(0)} = \begin{cases} \tau_i^+ \Delta V_i^{\max} + \tau_i^- \Delta V_i^{\min} & \text{if } V_{i+1} > V_{i-1} \\ \tau_i^+ \Delta V_i^{\min} + \tau_i^- \Delta V_i^{\max} & \text{if } V_{i+1} < V_{i-1} \end{cases} \quad (7a)$$

where

$$\Delta V_i^{\max} = \max(|V_{i+1} - V_i|, |V_{i-1} - V_i|) \quad (7b)$$

and

$$\Delta V_i^{\min} = \min(|V_{i+1} - V_i|, |V_{i-1} - V_i|) \quad (7c)$$

Finally, the tangent vector needs to be normalized, using eq 4b. With this modified tangent, the elastic band is well behaved and should converge rigorously to the MEP if a sufficient number of images are included in the band. Another modification included in the same version^{20,18} of the NEB method that introduced eqs 6 and 7 is the evaluation of an improved spring force:

$$\vec{F}_{\text{spring}}^s = k(|\vec{R}_{i+1} - \vec{R}_i| - |\vec{R}_i - \vec{R}_{i-1}|)\hat{\tau}_i \quad (8)$$

This new definition of the spring force, when used instead of eq 2, ensures equal spacing of the images when the same spring constant, k , is used for the springs even in regions of high curvature where the angle between $\vec{R}_i - \vec{R}_{i-1}$ and $\vec{R}_{i+1} - \vec{R}_i$ is large. This version of the NEB defined by eqs 6–8 is called improved tangent NEB or IT-NEB.

Another modification of the NEB, called Climbing Image NEB: CI-NEB,¹⁸ has been introduced for the purpose of using an NEB calculation to converge a saddle point. This new method modifies the definition of the total force on the highest-energy image after a few iterations. After identifying this image as point i_{\max} , the force on i_{\max} is given not by eq 1, but rather by

$$\vec{F}_{i_{\max}} = -\nabla V(\vec{R}_i) + 2\nabla V(\vec{R}_i)|_{\parallel} = -\nabla V(\vec{R}_i) + 2(\nabla V(\vec{R}_i) \cdot \hat{\tau}_i)\hat{\tau}_i \quad (9)$$

The highest energy image, i_{\max} , is not affected by the spring forces at all. The total force acting on all the other images

is still defined by eq 1 (as in B-NEB and IT-NEB), and the definition of the tangent and the spring force are still given by eqs 6–8 (as in IT-NEB). The CI-NEB algorithm should converge to the saddle point more efficiently than either B-NEB or IT-NEB.

2.2. Implementations. 2.2.1. G98+NEB. As far as we know, the only NEB implementation that has been made available in a distributed computer program for gas-phase systems is the one implemented by Alfonso and Jordan.²² The driver is called *G98+NEB*, and it performs NEB calculations using energies and forces obtained from the *Gaussian98* package.³¹ The driver consists of the NEB code and several script files that mediate the information flow between the NEB code and the *Gaussian98* program. In addition, it includes utility codes to generate the initial points via linear interpolation. The basic *G98+NEB* procedure can be summarized as follows:

(i) By using a utility code, an N -point approximation of the path is generated using linear interpolation in Cartesian coordinates between the two end points.

(ii) The energy and force for each movable image as well as at the end points are computed by calling *Gaussian98*.

(iii) Spring interactions with spring constant k between adjacent images are added, and the tangent is computed by the IT-NEB algorithm.

(iv) The corresponding projections of the true and parallel forces are computed by using the previously calculated tangent vectors, using eq 8.

(v) The points in the elastic band are brought to the nearest MEP via minimization of the adjusted NEB forces. This can be done using conventional minimization techniques such as the steepest descent or modified Broyden³² method or by using damped dynamics procedures.

2.2.2. MULTILEVEL4.1. Our implementation of the NEB method has been incorporated in the MULTILEVEL program.²⁷ This implementation follows the global scheme depicted in Figure 1. This diagram can be compared with the one used in the G98+NEB program depicted in ref 22. The first important difference between our implementation and the one previously described is the way in which the initial set of images is generated. The *Interpolation* utility in the *G98+NEB* package can lead to an unphysical chain if the user does not start with physically consistent orientations for the two end points. In MULTILEVEL4.1 an initial reorientation of the two end points is performed by Chen's algorithm,³³ adopting the implementation from the *PolyRate* package.³⁴ After this reorientation, the initial set of images is generated by linear interpolation in Cartesian coordinates. However, the user can supply an external file containing all n movable images plus the two end points, and the program will read it and avoid the two previous steps.

Once the initial set of images is generated, the program enters subroutine NEBGHK. First, the potential energies V_i are calculated for all images, and the corresponding gradients are calculated for all images except the two end points. After this, the local tangent is evaluated, and the projections of the true and spring forces are carried out. Both the B-NEB¹⁵ and the newer²⁰ (IT-NEB) definitions for these two variables (the tangent and the spring force) have been implemented.

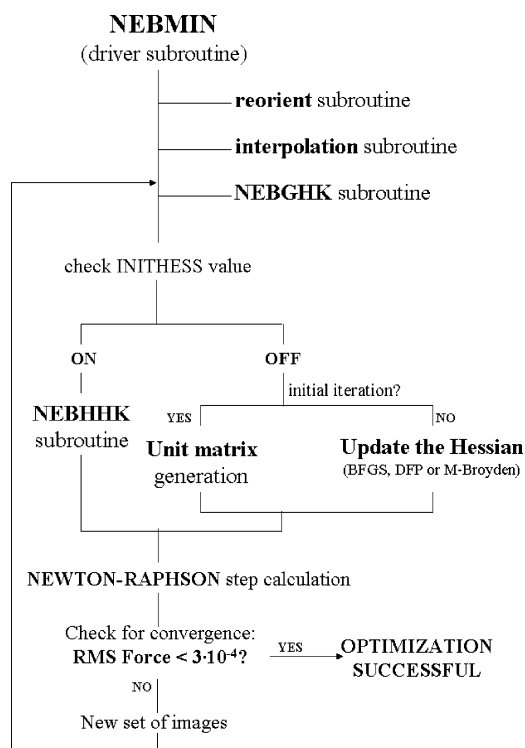


Figure 1. Flow diagram of the implementation of the NEB in the MULTILEVEL package. Note that INITHESS can be switched on or off depending on the way that the initial Hessian is obtained. If it is on, the initial Hessian will be calculated, while if is off, the initial Hessian will be approximated by a scaled unit matrix.

However, the default options correspond to eqs 6–8 (IT version) because it was demonstrated by Jónsson and co-workers²⁰ that the previous ones can promote the formation of kinks along the band. The CI algorithm is also implemented, and it uses the same formulation for these variables (tangent and spring forces) as the IT option. It just modifies the total force acting on the highest energy image after a few iterations, which is taken as five iterations in our implementation. This number was chosen after verifying that the highest energy image at this iteration remains the same along all the following minimization cycles in the tests we have made.

Finally subroutine NEBGHK calculates the adjusted force vector \vec{F} (see eq 1) which collects the $M \vec{F}_i$, where M is the number of images. The components of the resulting vector \vec{F} , of order $M \times 3N$ (where N is the number of atoms of the system), will be minimized using one of the available quasi-Newton methods.

The quasi-Newton optimization methods available in MULTILEVEL-v4.1 are variations of the Newton–Raphson method where an approximate Hessian matrix (or its inverse) is gradually updated using the gradient and displacement vectors of the previous steps.^{35–42} The displacement that is performed to move toward a stationary point is given by

$$\Delta \vec{q}_k = -H_k^{-1} \vec{g}_k \quad (10)$$

where \vec{g}_k is minus the adjusted force vector corresponding to iteration k , and the elements in H_k^{-1} should be obtained

from the derivatives of the gradient components. The possibility of obtaining the exact inverse Hessian H_k^{-1} corresponding to the adjusted force vector is implemented in our program only for the B version of the NEB algorithm. However, this process is computationally very expensive, and to avoid it another strategy will be followed.

The first option that must be set by the user is the choice of methods for calculating the initial Hessian. The user can choose between three different possibilities: using a scaled unit matrix, a low-level initial Hessian, or a high-level initial Hessian. A low-level Hessian means that the force constant matrix is evaluated at a lower electronic level than the gradients derived from the potential energy V_i . A high-level Hessian implies that the Hessian would be calculated at the same electronic level as the energies and gradients previously obtained.

During the minimization cycles, the Hessian can be recalculated or updated. Furthermore, the number of cycles after which the Hessian is to be recalculated must be defined by the user. Between two recalculations, the Hessian H_k is kept frozen. There are various possibilities for updating the Hessian H_k during the minimization cycles: the BFGS algorithm,^{36–39} the Davidson-Fletcher-Powell (DFP) algorithm,⁴³ or the Modified Broyden method,³² as described by Alfonso and Jordan.²² The quasi-Newton methods generate new geometries (by eq 10) that form the initial chain to start another minimization cycle. This process is repeated until the convergence criterion is satisfied.

These procedures are the same for the CI-NEB¹⁸ option except that the definition of the total force acting on the highest energy image which will be given by eq 9 after five iterations.

The default convergence criterion is based on the root-mean-square of the components of the true force acting on the whole band. This value is required to be smaller than or equal to $3 \cdot 10^{-4}$ hartrees per bohr ($E_h a_0^{-1}$), which is the same default convergence criterion used by *Gaussian03*⁴⁴ in optimizations.

Another key parameter which may be monitored along the optimization is the maximum component of the adjusted force at the highest energy image. As this image is supposed to converge to the saddle point, the true force should approach zero. This feature will be used as the criterion to determine the “best” parameters for each option that can be controlled by the user.

3. Calculations, Results, and Discussion

3.1. Testing Various Options. With the aim of determining good default values for the various options of the program, we tested them for the H-abstraction from CH_4 by the hydroxyl radical. To build the initial sequence of images, we used three structures: the ones corresponding to the van der Waals complexes in the entrance and exit channels (named *wellR* and *wellP*, respectively) and an intermediate structure, where the distance between the oxygen in the OH and one of the hydrogens in the methane is 1.2 Å. This distance would be a reasonable starting value for saddle points where an OH is abstracting a hydrogen. An initial

reorientation of the three structures (*wellR*, intermediate structure, and *wellP*) is carried out followed by two linear interpolations: from *wellR* to the intermediate point and from the intermediate point to *wellP*. The new structures obtained from each interpolation plus the intermediate structure, included as another movable image, formed the initial guess of movable images.

The program does not automatically carry out the interpolations mentioned in the previous paragraph; they must be done manually, and the initial set of movable images is then supplied by the user in an external file.

The electronic-structure level chosen for these tests was density functional theory with the modified Perdew-Wang 1-parameter functional for kinetics: MPW1K.⁴⁵ (This functional was optimized⁴⁵ to a database of barrier heights and reaction energies. Several studies have demonstrated that the MPW1K functional gives good performance for kinetics.^{46–50} However, the increased percentage of Hartree–Fock exchange in MPW1K deteriorates the atomization energy calculation.⁴⁵) The 6-31+G(d,p) basis set^{51,52} was chosen as a good compromise between cost and efficiency for the system studied. To compare the performance of the various options, we always performed 40 iterations, and then the maximum component of adjusted force at the highest energy image was checked. As the NEB algorithm should converge to the SP, the forces at the highest energy image should go to zero.

The parameters we will test here are: the number of images, the choice of B, IT, or CI for the NEB algorithm, the way of generating the initial Hessian to obtain the displacement during the quasi-Newton minimization, how to update this Hessian as the optimization proceeds, and the spring constant. Other parameters that can be controlled by the user are not tested here. Examples would be inclusion of some intermediate points (apart from the two fixed end points) in order to generate the initial chain, calculation of a new full Hessian along the minimization, etc.

We started our study using some reference values of the parameters taken from previous work. For example, we used a value for $k_{\text{spring}} = 0.02 \text{ E}_{\text{h}}\text{a}_0^{-2}$ as recommended by Alfonso and Jordan,²² and we used a scaled unit matrix as initial Hessian with $\text{HSCALE} = 100 \text{ E}_{\text{h}}\text{a}_0^{-2}$ (see below for the definition of HSCALE). Due to the arbitrariness of these parameters, we performed an iterative determination of the optimum parameters; that is, once we obtained the best value for one parameter, we reoptimized all the others. In this way, we could obtain best parameters even if their optimum values are coupled. The intermediate results of this iterative optimization are not discussed, and we will just discuss the final results. Figures 2–5 show the results after 40 steps of path minimization for various values of each parameter, and Table 1 gives the corresponding values for the maximum component of the adjusted force at the highest energy image (the criterion followed to decide the best parameters).

3.1.3. Number of Images. The number of images to be used in carrying out the NEB minimization should depend on the objectives of the user. For example, if one is using the NEB method just to obtain the qualitative nature of the path from reactants to products and to check if there is an energy maximum along this path, a small number of images

Table 1: Influence of the Various Parameters on the NEB Minimization^a

variable	variable value	maximum component ($\text{E}_{\text{h}}\text{a}_0^{-1}$)	
		iteration 20	iteration 40
n^b	11	0.0280	0.0219
	21	0.0282	0.0229
	41	0.0277	0.0228
HSCALE ($\text{E}_{\text{h}}\text{a}_0^{-2}$)	1	0.0048	0.0048
	10	0.0229	0.0056
	10 ²	0.0283	0.0222
NEB algorithm	10 ³	0.0302	0.0258
	B-NEB	0.0373	0.0341
	IT-NEB	0.0255	0.0202
update scheme	CI-NEB	0.0229	0.0056
	BFGS	0.0229	0.0056
	DFP	0.0297	0.0288
k_{spring} ($\text{E}_{\text{h}}\text{a}_0^{-2}$)	mBroyden	0.0301	0.0271
	1	0.0305	0.0162
	10 ⁻¹	0.0195	0.0142
	10 ⁻²	0.0229	0.0056
	10 ⁻³	0.0234	0.0059
	10 ⁻⁴	0.0234	0.0059

^a The absolute value of the maximum component of the adjusted force at the highest energy image (in atomic units: $\text{E}_{\text{h}}\text{a}_0^{-1}$) at iterations 20 and 40, respectively, is used as a test for the convergence of the method. The rows corresponding to recommended values are in bold. In each case, when one parameter or choice is varied, the other four are fixed at their recommended value, except for the first 3 rows in the table. That row has $n = 11$, $\text{HSCALE} = 100 \text{ E}_{\text{h}}\text{a}_0^{-2}$, NEB choice = CI-NEB, update scheme = BFGS, and $k_{\text{spring}} = 0.02 \text{ E}_{\text{h}}\text{a}_0^{-2}$. ^b n is the number of movable images.

can be used. However, if the user wants to tightly optimize the saddle point, more images should be included. However, the most efficient way to get an accurate saddle point would be to carry out the NEB minimization in more than one cycle: one could start using a few images just to locate the maximum energy region. Then, with that region defined, we would run a second NEB minimization concentrating all the images in the area surrounding the saddle point. In this way, the highest-energy image can be made to approach closer and closer to the saddle point. Various kinds of interpolation could be done between the three highest-energy images in order to obtain a better geometry for the saddle point; for example, the user can apply a quadratic interpolation.

We tested the performance for $n = 11, 21$, and 41 movable images. These numbers include 5, 10, or 20 interpolated images plus the intermediate image that is included. All of the images were obtained by using the same initial points, as explained above; we just increased the number of points generated between them. In Table 1 the maximum component of the adjusted force at the highest energy image are shown. As can be seen, the results do not depend strongly on the number of images used. Therefore, for the H-abstraction in the $\text{CH}_4 + \text{OH}$ system, 11 movable images provide a good compromise between cost and accuracy. This option was not reoptimized with the best values for the other parameters, and in Table 1 the results shown correspond to those obtained with the values recommended by Alfonso and Jordan.²²

For all the tests presented in the rest of section 3.1, we set $n = 11$ except in Table 1, which also shows results for $n = 21$ and 41.

3.1.2. Scaled Unit Matrix. Using a unit matrix as an approximate initial Hessian is well-known to be an efficient

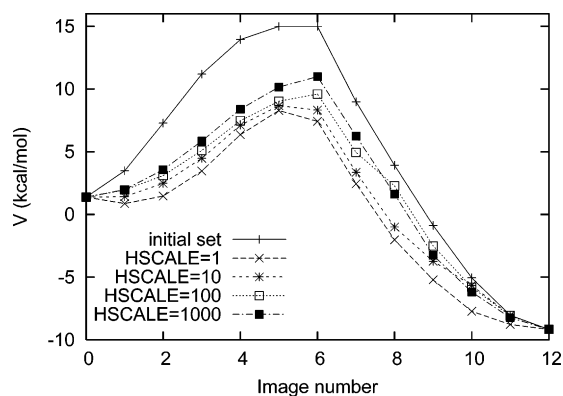


Figure 2. Influence of the HSCALE ($E_h a_0^{-2}$) value used in the minimization for $\text{CH}_4 + \text{OH}$ system. All other variables were kept unchanged: CI-NEB algorithm, $k_{\text{spring}} = 0.01 E_h a_0^{-2}$, and BFGS update scheme. The relative potential energy (vs reactants) is depicted vs the image number at iteration 40. The “initial set” stands for the energy at the initial chain of images.

way to obtain the displacement in quasi-Newton minimizations. This unit matrix can be scaled in order to obtain a matrix closer to the real Hessian. The value used will be named HSCALE (as the unit matrix has no units, the HSCALE gives dimension to that matrix in order to adjust it to approximate a Hessian). We tested several values for this variable, and the results are shown in Figure 2. The values for the maximum component of the adjusted force at the highest energy image can be seen in Table 1. In all the cases, the other three parameters were kept fixed at the values: CI-NEB algorithm, BFGS update scheme, and $k_{\text{spring}} = 0.01 E_h a_0^{-2}$. Although the options $\text{HSCALE} = 10^{-2} E_h a_0^{-2}$ and $\text{HSCALE} = 10^4 E_h a_0^{-2}$ were also tested, we do not show the results; both values yield a band that does not minimize correctly. The smallest value ($\text{HSCALE} = 10^{-2} E_h a_0^{-2}$) gave an initial quasi-Newton step (see eq 10) so large that the displacement moved the new set of images very far away from the initial one, and the energies did not minimize. On the other hand, the largest value ($\text{HSCALE} = 10^4 E_h a_0^{-2}$) gave a very small initial quasi-Newton step so that the next set of images was almost equal to the initial one. The intermediate values for HSCALE perform quite well, but the best option is HSCALE equal to $1 E_h a_0^{-2}$.

One could choose other options for the initial Hessian. For example, the program could compute the “true” force constant matrix (at the actual electronic level or lower) for each image in the set of images. The calculated Hessian corresponds to the “true” Hessian of the system without spring forces; i.e., without the terms representing the interactions between images; recall that the gradient in eq 10 is the “total” gradient (also called “adjusted” gradient), not the “true” gradient. In fact, and probably for this reason, this option did not work better than the scaled unit matrix.

The third option, the use of exact Hessians (as defined in eq 10), cannot be recommended either; especially when one takes into account the computational cost to compute the Hessian, which is $(M \times 3N) \times (M \times 3N)$, where N is the number of atoms of the system and M is the number of

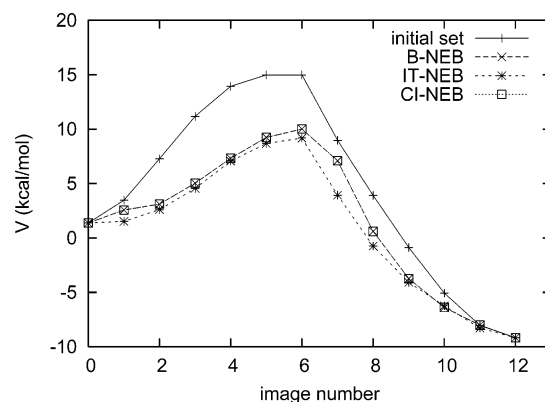


Figure 3. Influence of the iterative process chosen in the minimization for $\text{CH}_4 + \text{OH}$ system. All other variables were kept unchanged: BFGS update scheme, $k_{\text{spring}} = 0.01 E_h a_0^{-2}$, and $\text{HSCALE} = 10 E_h a_0^{-2}$. The relative potential energy (vs reactants) is depicted vs the image number at iteration 40. The “initial set” stands for the energy at the initial chain of images.

images. Nevertheless, MULTILEVEL4.1 does have an option to use the exact Hessian if the user chooses the B-NEB algorithm. In this case, the Hessian also contains the off-diagonal terms corresponding to the spring forces. These terms involve the tangent vector, and they are easier to derive for B-NEB than for IT-NEB or CI-NEB. This option, however, is computationally more expensive than the scaled unit matrix and hence is not recommended either.

To summarize, we recommend using a scaled unit matrix with $\text{HSCALE} = 1-10 E_h a_0^{-2}$ in order to start the NEB minimization.

3.1.3. Iterative Process. Three different versions of the iterative scheme in the NEB algorithm were implemented: B-NEB,¹⁵ IT-NEB,²⁰ and the CI-NEB.¹⁸ All the other options were fixed for the three tests by employing k_{spring} equal to $0.01 E_h a_0^{-2}$ and a scaled unit matrix as a initial Hessian, with $\text{HSCALE} = 10 E_h a_0^{-2}$. The results are depicted in Figure 3. The values of the maximum component of the adjusted force at the highest energy image listed in Table 1 show small differences between the three options. Although these differences are small, the CI-NEB performed better than the other ones, as was expected due to its special design for SP optimizations.

3.1.4. Update Scheme. The initial Hessian matrix was approximated by a scaled unit matrix. However, during the optimization of the path, this matrix can be and is updated by using one of the three possible schemes: BFGS,³⁶⁻³⁹ DFP,⁴³ or modified Broyden.³² We compared the performance of these three algorithms; in all cases we used a scaled unit matrix as an initial Hessian with $\text{HSCALE} = 10 E_h a_0^{-2}$, the CI-NEB algorithm was used, and k_{spring} was set equal to $0.01 E_h a_0^{-2}$. The results obtained are depicted in Figure 4. As can be seen, after 40 iterations, the BFGS algorithm gives the best converged path, followed by the modified Broyden algorithm. Since the computational costs for these three schemes are very similar, we recommend using BFGS. Table 1 shows that the maximum components of the adjusted force at the highest energy image confirm the better performance of the BFGS update scheme.

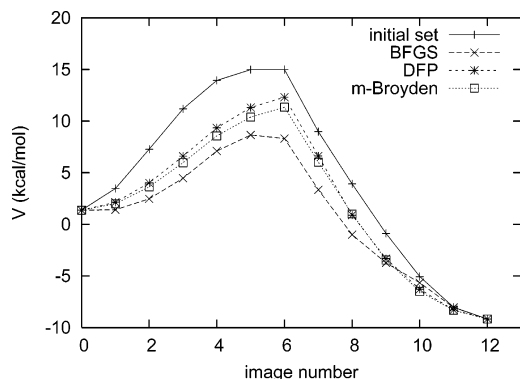


Figure 4. Influence of the update scheme used in the minimization for $\text{CH}_4 + \text{OH}$ system. All other variables were kept unchanged: CI-NEB algorithm, $k_{\text{spring}} = 0.01 \text{ E}_{\text{h}a_0^{-2}}$, and $\text{HSCALE} = 10 \text{ E}_{\text{h}a_0^{-2}}$. The relative potential energy (vs reactants) is depicted vs the image number at iteration 40. The “initial set” stands for the energy at the initial chain of images.

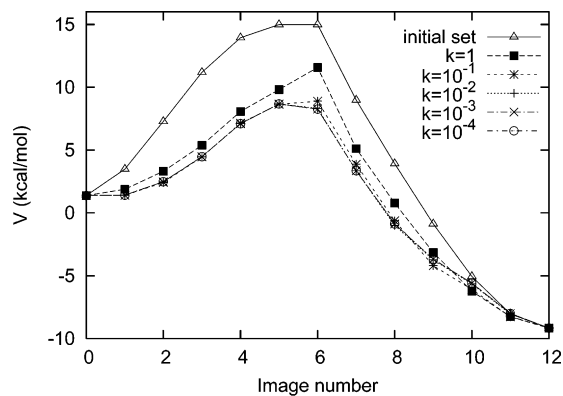


Figure 5. Influence of the k_{spring} ($\text{E}_{\text{h}a_0^{-2}}$) value used in the minimization for $\text{CH}_4 + \text{OH}$ system. All other variables were kept unchanged: CI-NEB algorithm, $\text{HSCALE} = 10 \text{ E}_{\text{h}a_0^{-2}}$, and BFGS update scheme. The relative potential energy (vs reactants) is depicted vs the image number at iteration 40. The “initial set” stands for the energy at the initial chain of images.

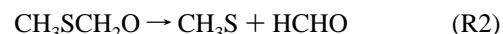
Note that the update scheme needed in the NEB is qualitatively different from the update schemes needed in searching for saddle points in a quasi-Newton procedure. In a quasi-Newton procedure, one can use special update schemes for Hessians that contain a negative eigenvalue.⁴¹ However, for the NEB calculation, we are minimizing all the components of the total force, and the Hessian should not contain any negative eigenvalue.

3.1.5. Spring Constant. To test the influence of the value of the spring constant on the performance of the NEB minimization, we used many different values. For these tests, the CI-NEB and the BFGS choices were used as the iterative process and update scheme, respectively, and a scaled unit matrix with $\text{HSCALE} = 10 \text{ E}_{\text{h}a_0^{-2}}$ was used as initial Hessian. The results obtained for NEB minimization after 40 iterations are depicted in Figure 5, and the maximum components of the adjusted force at the highest energy image are shown in Table 1. As can be seen, using a large value for the k_{spring} results in very poor performance of the

algorithm. This occurs because the global force on each image is almost totally due to the spring force. The best value for k_{spring} is $0.01 \text{ E}_{\text{h}a_0^{-2}}$.

3.2. Tests. In this section, we will present some tests of the method on various systems. Based on the studies presented in section 3.1, we used the following options to characterize the paths in these test applications: (i) spring constant: $0.01\text{--}0.001 \text{ E}_{\text{h}a_0^{-2}}$; (ii) scaled unit matrix as initial Hessian ($\text{HSCALE} = 1\text{--}10 \text{ E}_{\text{h}a_0^{-2}}$); (iii) update scheme: BFGS; (iv) NEB algorithm: CI-NEB; and (v) n (number of movable images): depending on the purpose of the user ($10\text{--}15$ in order to obtain a good initial guess). When the points in the NEB profile are closer, the reaction path is smoother.

The radicals CH_3SO_2 and $\text{CH}_3\text{SCH}_2\text{O}$ are key species formed in the addition and abstraction mechanisms, respectively, of the $\text{DMS} + \text{OH}$ reaction.⁵³ There are no theoretical studies of the pathways of their dissociations. However, these dissociations are potentially important because the final products are directly linked to SO_4^{2-} formation⁵³



The other reaction chosen is also part of the abstraction mechanism, and it involves a heavy-atom transfer:



It has been proposed that this reaction takes place via an intermediate:⁵⁴



The main aim in studying these reactions was to obtain a global knowledge of the path: does the reaction take place via a saddle point or does the energy along the MEP change monotonically. One possibility to answer those questions would be to perform an NEB minimization. Furthermore, if a maximum along the NEB profile appeared, we could then refine that NEB minimization in order to get closer to the saddle point. The electronic-structure level chosen was again the MPW1K density functional method⁴⁵ for the reasons explained above. Moreover, it has been shown that this functional works well for these kind of reactions.²⁹

3.2.1. CH_3SO_2 and $\text{CH}_3\text{SCH}_2\text{O}$ Dissociations. The initial guess of images for both reactions R1 and R2 was built by direct interpolation from the minima (CH_3SO_2 and $\text{CH}_3\text{SCH}_2\text{O}$, respectively) to a “product-like” structure. For reaction R1, we used an end point where the C–S distance was 3 \AA , while for reaction R2, that distance was 4.5 \AA . The end points with stretched bonds were not optimized. The results for both reactions are depicted in Figure 6.

The top half of Figure 6 shows that the CH_3SO_2 dissociation does not present any maximum of energy along the NEB. It is reasonable to conclude that this dissociation takes place by a monotonic increase in the energy, so searching for a saddle point would be fruitless. The same conclusion could be reached by computing a distinguished reaction-

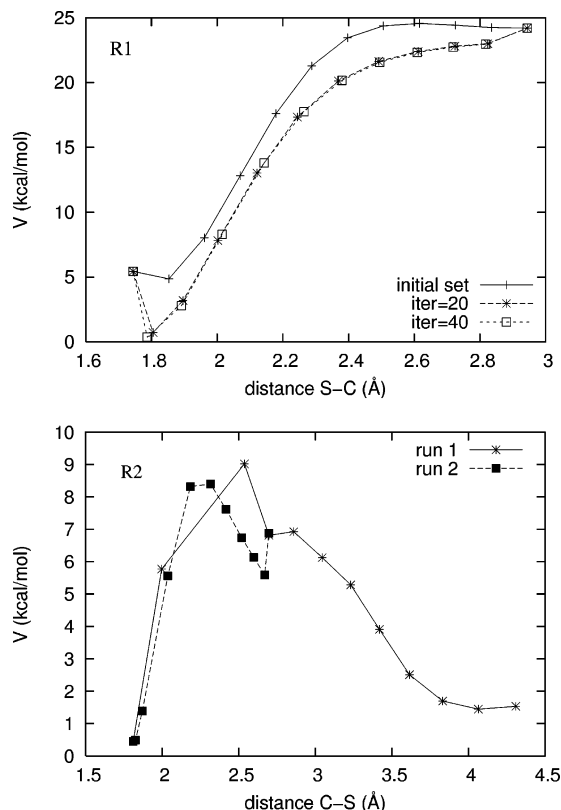


Figure 6. NEB minimization for CH_3SO_2 (top) and CH_3SCH_2O (bottom) systems. Top: Relative potential energy (vs reactant, CH_3SO_2) vs breaking bond distance (d_{S-C}) after 20 and 40 iterations (just one run was performed). Bottom: Relative potential energy (vs reactant, CH_3SCH_2O) vs breaking bond distance (d_{S-C}), after 40 iterations for run 1 and run 2. Run 2 was built by constraining the area to the maximum of energy in run 1.

coordinate path. Although in this case choosing a coordinate would be very easy (because the reaction is the breaking of one bond), there are many cases where selecting one distinguished coordinate would be a difficult choice, and the coordinate chosen could turn to be unrepresentative of the MEP. This is one of the advantages of the NEB minimization: the path is minimized without any restriction on the coordinates of the images. It is worth noting that the location of the saddle point (if it exists) should be independent of the particular system of coordinates chosen. However, this is not true for the reaction path; the NEB path should converge to the steepest descents path in whatever coordinate system is used; the user should keep in mind that only in an iso-inertial coordinate system is the steepest-descents path equal to the intrinsic reaction path, which is the MEP in the notation we usually use.

In contrast with the results for reaction R1, reaction R2 showed a maximum of energy along the NEB path, as can be seen in Figure 6. This suggests although it does not prove, that the dissociation of CH_3SCH_2O takes place via a saddle point. To characterize this saddle point, we performed a second NEB minimization by limiting the images to the area around that energy maximum. In particular, we built a second initial path using two images from the first run as the final end points of the second run. This second run yielded a

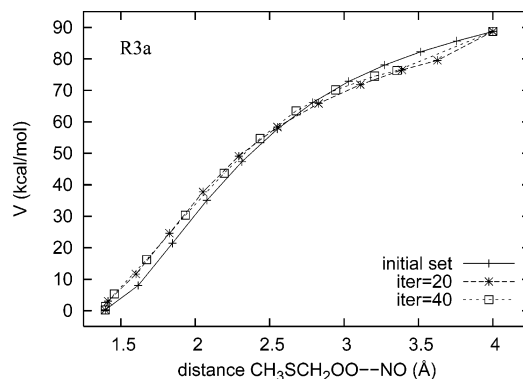


Figure 7. NEB minimization for NO addition to CH_3SCH_2OO radical (reaction R3a). The relative potential energy (vs the addition minimum, CH_3SCH_2OONO) as a function of the forming bond O–N distance is depicted. Three different stages of the minimization are shown: the initial guess, the energies after 20 iterations, and the final energies obtained after 40 iterations.

maximum of energy very close to the true saddle point. Using the geometry of the maximum-energy image as a starting point, we performed a transition-state (TS) search with the Gonzalez-Schlegel algorithm⁵⁵ as implemented in *Gaussian03*,⁴⁴ and it converged to the saddle point after only three cycles. The maximum-energy point in the second NEB run was 8.4 kcal/mol above reactants, while the real saddle point is at 7.8 kcal/mol; the normal mode associated with the reaction coordinate (C–S stretching) has a frequency of 430i cm^{-1} , while at the saddle point it is 364i cm^{-1} .

3.2.2. $CH_3SCH_2OO + NO \rightarrow CH_3SCH_2O + NO_2$. We studied reaction R3 by assuming that it takes place in two steps: (1) the NO addition to the radical to form a stable complex and (2) the CH_3SCH_2OONO dissociation to form NO_2 and CH_3SCH_2O , whose dissociation is discussed above. The initial guess of images for both reaction R3a and reaction R3b were built by linear interpolation between the stationary point representing the addition complex (CH_3SCH_2OO-NO) and a “reactant-like” and “product-like” geometry, respectively. For the NO addition we used a “reactant-like” structure where the forming O–N bond had a length of 4 Å; the “product-like” geometry for the CH_3SCH_2OONO dissociation was represented by a structure with a breaking O–O bond of length 4 Å. These two end points were not optimized.

In Figure 7 the NEB minimization for the NO addition is shown. It can be observed that NO addition takes place via a barrierless association. We did not perform any refinements for this path because no saddle point is expected when the potential energy profile along the reaction path is monotonic. Figure 8 shows the successive NEB minimizations for reaction R3b. As the first run suggested that this dissociation takes place via a saddle point, we ran more NEB cycles in order to get closer to that point. Each successive run was performed by constraining the search to the maximum-energy area. After 4 runs, we used the highest energy image as the initial guess for a conventional TS search. By using the Gonzalez-Schlegel algorithm⁵⁵ as implemented in *Gaussian03*⁴⁴ we did succeed in finding the saddle point for reaction R3b. Comparing the highest-energy image in the

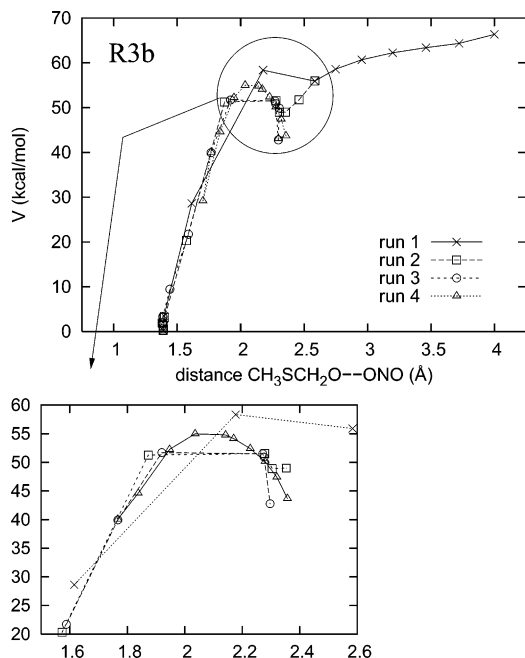


Figure 8. NEB minimization for $\text{CH}_3\text{SCH}_2\text{OONO}$ dissociation (reaction R3b). The relative potential energy (vs the minimum, $\text{CH}_3\text{SCH}_2\text{OONO}$) as a function of the breaking bond O–O distance is depicted. The results for four consecutive runs are shown. A zoom in the highest-energy area is also depicted.

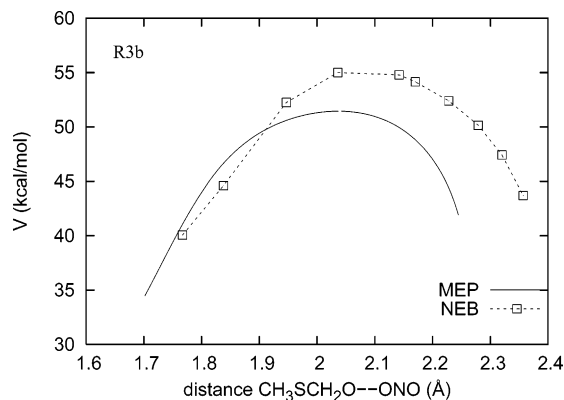


Figure 9. A comparison between the energy along the NEB and the energy along the MEP for $\text{CH}_3\text{SCH}_2\text{OONO}$ dissociation (reaction R3b) is shown. The relative potential energy along the path (vs the minimum, $\text{CH}_3\text{SCH}_2\text{OONO}$) is depicted vs the breaking bond O–O distance.

fourth NEB run with the saddle point shows that the point was geometrically very close to the saddle point. For example, the breaking O–O bond was 2.04 Å in both cases. However, the highest-energy image had two additional imaginary frequencies, besides the one associated with the reaction coordinate. Those additional frequencies were eliminated during the Gonzalez-Schlegel minimization. The value of the imaginary frequency for the mode associated with the reaction coordinate went from 367i to 189i cm^{-1} . After characterizing the saddle point we computed the MEP in order to compare the potential energy along the MEP to that along the NEB path. This comparison is shown in Figure 9. The differences between these two paths show that the NEB minimizations did not converge to the MEP mainly because the minimization was not completed. It is important

to remark here that the MEP could be computed because we previously obtained the saddle point by using the NEB algorithm.

In contrast, when using a distinguished reaction coordinate, we could not characterize the saddle point. Since a saddle point is needed to start the MEP calculation, using the NEB algorithm was a necessary starting point.

The main aim of the studies reported in this section was to obtain a global view for each reaction studied and to characterize the saddle points (if they exist). If we wanted to converge the NEB to the MEP we would need to use more images along the NEB and more cycles for the minimization (all our minimizations were stopped after 40 cycles).

4. Conclusions

The Nudged Elastic Band method was originally proposed for condensed-phase systems and is particularly useful for such cases. We have shown here that it can also be very useful for gas-phase reactions, and we have implemented it for this kind of application in the MULTILEVEL²⁷ program. Three options are included for the tangent vector, and users can choose among several options for other variables as well. On the basis of our studies we recommend default values that can guide the user in preliminary searches.

In some cases the optimum parameters to be used depend on the purpose of the calculation. For example, Figures 2–5 are not smooth due to the small number of images used. We did not employ more images because in those cases we were carrying out exploratory work. Later, in Figures 6 (bottom) and 8, we increased the number of images used (we concentrated additional images in the more interesting region). In this way, we obtained a smoother path.

The performance of the code has been illustrated for reactions involved in the DMS chemistry of the atmosphere. We characterized the paths for those reactions to show that one can elucidate whether they take place via a saddle point. For reactions where the profile showed a maximum in the potential energy profile, we characterize the saddle point by performing an iterative NEB minimization followed by a saddle-point search. One can characterize an unknown reaction in a systematic fashion by first doing a general NEB minimization for a broad whole range of the reaction coordinate and then constraining the search to the most interesting regions.

Acknowledgment. This work was supported in part by the U.S. Department of Energy, Office of Basic Energy Sciences. N.G.G. would like to acknowledge the Generalitat de Catalunya for the BE fellowship received.

References

- (1) Truhlar, D. G.; Steckler, R.; Gordon, M. S. *Chem. Rev.* **1987**, *87*, 217.
- (2) *Supercomputer Algorithms for Dynamics and Kinetics of Small Molecules*; Laganà, A., Ed.; Kluwer: Dordrecht, 1989; p 105.
- (3) McKee, M. L.; Page, M. *Rev. Comput. Chem.* **1993**, *IV*, 35.
- (4) Page, M. *Computer Phys. Commun.* **1994**, *84*, 115.

- (5) *The Reaction Path in Chemistry: Current Approach and Perspectives*; Heidrich, D., Ed.; Kluwer: Dordrecht, 1995.
- (6) Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1515.
- (7) Halgren, T. A.; Lipscomb, W. N. *Chem. Phys. Lett.* **1977**, *49*, 225.
- (8) Melissas, V. S.; Truhlar, D. G. *J. Chem. Phys.* **1992**, *96*, 5758.
- (9) Rothman, M. J.; Lohr, J. L.; Ewig, C. S.; Wazer, van J. R. In *Potential Energy Surfaces and Dynamical Calculations*; Truhlar, D. G., Ed.; Plenum: New York, 1979; pp 653–660.
- (10) Cerjan, C. J.; Miller, W. H. *J. Chem. Phys.* **1981**, *75*, 2800.
- (11) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. *J. Phys. Chem.* **1985**, *89*, 52.
- (12) Baker, J. *J. Comput. Chem.* **1985**, *7*, 385.
- (13) Culot, P.; Dive, G.; Nguyen, V. H.; Ghuysen, J. M. *Theor. Chim. Acta* **1992**, *82*, 189.
- (14) Quapp, W. *Chem. Phys. Lett.* **1996**, *253*, 286.
- (15) Mills, G.; Jónsson, H. *Phys. Rev. Lett.* **1994**, *72*, 1124.
- (16) Mills, G.; Jónsson, H.; Schenter, G. K. *Surf. Sci.* **1995**, *324*, 305.
- (17) Gillilan, R. E.; Wilson, K. R. *J. Chem. Phys.* **1992**, *97*, 1757.
- (18) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9901.
- (19) Jónsson, H.; Mills, G.; Jacobsen, K. W. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B. J., Cicotti, G., Coker, D. F., Ed.; World Scientific: Singapore, 1998; pp 385–404.
- (20) Henkelman, G.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9978.
- (21) Crehuet, R.; Field, M. *J. Chem. Phys.* **2003**, *118*, 9563.
- (22) Alfonso, D. R.; Jordan, K. D. *J. Comput. Chem.* **2003**, *24*, 990.
- (23) Chu, J.-W.; Trout, B.; Brooks, B. *J. Chem. Phys.* **2003**, *119*, 12708.
- (24) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. *J. Chem. Phys.* **2004**, *120*, 7877.
- (25) Besal, E.; Bofill, J. M. *Theor. Chem. Acc.* **1998**, *100*, 265.
- (26) Quapp, W.; Hirsch, M.; Imig, O.; Heidrich, D. *J. Comput. Chem.* **1998**, *19*, 1087.
- (27) Zhao, Y.; Rodgers, J.; Lynch, B.; González-García, N.; Fast, P.; Pu, J.; Chuang, Y.; Truhlar, D. *MultiLevel4.1*; University of Minnesota: Minneapolis, MN, 2005.
- (28) Charlson, R. J.; Lovelock, J. E.; Andreae, M. O.; Warren, S. G. *Nature* **1987**, *326*, 655.
- (29) González-García, N.; González-Lafont, A.; Lluch, J. M. *J. Comput. Chem.* **2005**, *26*, 569.
- (30) González-García, N.; González-Lafont, A.; Lluch, J. M. *J. Phys. Chem. A* **2006**, *110*, 788.
- (31) *Gaussian 98, revision A.6 and A.7*; Gaussian, Inc.: Pittsburgh, PA, 1998.
- (32) Johnson, D. *Phys. Rev. B* **1988**, *38*, 12807.
- (33) Chen, Z. *Theor. Chim. Acta* **1989**, *75*, 481.
- (34) Corchado, J. C.; Chuang, Y.-Y.; Fast, P. L.; Villà, J.; Hu, W.-P.; Liu, Y.-P.; Lynch, G. C.; Nguyen, K. A.; Jackels, C. F.; Melissas, V. S.; Lynch, B. L.; Rossi, I.; Coitiño, E. L.; Fernández-Ramos, A.; Pu, J.; Albu, T. V.; Steckler, R.; Garret, B. C.; Isaacson, A. D.; Truhlar, D. G. *PolyRate9.3*; University of Minnesota: Minneapolis, MN, 2003.
- (35) Fletcher, R. *Practical Methods of Optimization*; Wiley: Chichester, 1991.
- (36) Broyden, C. *J. Inst. Math. App.* **1970**, *6*, 76.
- (37) Fletcher, R. *Comput. J. (UK)* **1970**, *13*, 317.
- (38) Goldfarb, D. *Math. Comput.* **1970**, *24*, 23.
- (39) Shannon, D. *Math. Comput.* **1970**, *24*, 647.
- (40) Wittbrodt, J.; Schlegel, H. *J. Mol. Struct. (THEOCHEM)* **1997**, *398*, 55.
- (41) Bofill, J. *J. Comput. Chem.* **1994**, *15*, 1.
- (42) Anglada, J.; Bofill, J. *J. Comput. Chem.* **1998**, *19*, 349.
- (43) Powell, M. *Math. Program* **1971**, *1*, 26.
- (44) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.
- (45) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811.
- (46) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2001**, *105*, 2936.
- (47) Parthiban, S.; Oliveira, G. D.; Martin, J. M. L. *J. Phys. Chem. A* **2001**, *105*, 895.
- (48) Claes, L.; François, J. P. F.; Deleuze, M. S. *J. Am. Chem. Soc.* **2002**, *124*, 7563.
- (49) Claes, L.; François, J. P. F.; Deleuze, M. S. *J. Am. Chem. Soc.* **2003**, *125*, 7129.
- (50) Iron, M. A.; Lo, H. C.; Martin, J. M. L.; Keinan, E. *J. Am. Chem. Soc.* **2002**, *124*, 7041.
- (51) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.
- (52) Hehre, W. J.; Radom, L.; R. Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.
- (53) Wayne, R. P. *Chemistry of Atmospheres*, 3rd ed.; Oxford University Press: New York, 2000.
- (54) Resende, S. M.; Almeida, W. B. D. *Phys. Chem. Chem. Phys.* **1999**, *1*, 2953.
- (55) González, C.; Schlegel, H. B. *J. Phys. Chem.* **1990**, *94*, 5523.

JCTC Journal of Chemical Theory and Computation

Empirical Valence-Bond Models for Reactive Potential Energy Surfaces Using Distributed Gaussians

H. Bernhard Schlegel* and Jason L. Sonnenberg

Department of Chemistry, Wayne State University, Detroit, Michigan 48202

Received March 3, 2006

Abstract: A new method for constructing empirical valence bond potential energy surfaces for reactions is presented. Building on the generalized Gaussian approach of Chang–Miller, $V_{12}^2(\mathbf{q})$ is represented by a Gaussian times a polynomial at the transition state and generalized to handle any number of data points on the potential energy surface. The method is applied to two model surfaces and the HCN isomerization reaction. The applications demonstrate that the present method overcomes the divergence problems encountered in some other approaches. The use of Cartesian versus internal or redundant internal coordinates is discussed.

Introduction

In the empirical valence bond (EVB) approach, the potential energy surface (PES) for a reaction in solution is modeled as an interaction between a reactant and a product PES.¹ The interaction between surfaces results in an avoided crossing and yields a smooth function describing the reaction on the ground-state potential energy surface. Good empirical approximations for the noninteracting potential energy surfaces of reactants and products are available from molecular mechanics methods. To obtain a reliable model of a PES for a reaction, a suitable form of the interaction matrix element or resonance integral, $V_{12}(\mathbf{q})$, is needed.

For a two-state system, the interaction between reactant and product surfaces is taken as a modified Morse function in Warshel and Weiss' original multistate EVB method.² The function is adjusted to reproduce barrier heights gleaned from experiments or high-level ab initio calculations, but the form of the surface is not flexible enough to fit frequencies at the transition state (TS). Chang and Miller represented the square of the resonance integral, V_{12}^2 , with a generalized Gaussian.³ The exponents of the Gaussian are chosen to fit the structure and vibrational frequencies of the TS from electronic structure calculations. This form of the EVB surface is sufficiently accurate for molecular dynamics.^{4–12} The Chang–Miller model has also been applied by Jensen^{13,14} and Anglada et al.¹⁵ to transition-state optimizations.

More elaborate functions of the interaction matrix elements were used in the molecular mechanics/valence bond model developed by Bernardi et al. for exploring photochemical reaction potential energy surfaces.¹⁶ Minichino and Voth

generalized the Chang–Miller method³ for N-state systems and provided a scheme to correct gas-phase ab initio data for solutions.¹⁷ Truhlar and co-workers employed a generalized EVB approach by using distance-weighted interpolants to model the interaction matrix elements in their multiconfiguration molecular mechanics method.^{18–20}

The simplicity of the Chang–Miller resonance integral formulation is appealing, but certain difficulties must be overcome to provide the greater flexibility required to model more complex chemical reactions using molecular dynamics. The present article explores two possibilities for improving the representation of the interaction matrix elements. In particular, the generalized Gaussian utilized in the Chang–Miller approach is replaced with a quadratic polynomial times a spherical Gaussian. This avoids the well-known problem caused by negative exponents that may arise in practice.^{12,15,18} Second, to improve the accuracy of the fit, a linear combination of Gaussians times quadratic polynomials placed at suitable locations on the potential energy surface is employed.

Model Description

The EVB model describes a reactive PES in terms of a linear combination of reactant and product wave functions. The coefficients are obtained by solving a simple 2×2 Hamiltonian for the lowest energy.

$$\Psi = c_1\psi_1 + c_2\psi_2 \quad (1)$$

$$\mathbf{H} = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \quad (2)$$

* Corresponding author. E-mail: hbs@chem.wayne.edu.

$$V_{11} = \langle \psi_1 | \hat{H} | \psi_1 \rangle, \quad V_{12} = V_{21} = \langle \psi_1 | \hat{H} | \psi_2 \rangle, \\ V_{22} = \langle \psi_2 | \hat{H} | \psi_2 \rangle \quad (3)$$

$$V = \frac{1}{2}(V_{11} + V_{22}) - \sqrt{[\frac{1}{2}(V_{11} - V_{22})]^2 + V_{12}^2} \quad (4)$$

Each matrix element is a function of molecular geometry, \mathbf{q} . Good approximations for V_{11} and V_{22} are available from molecular mechanics. However, much less is known about the functional form of the interaction matrix element, V_{12} .

In Warshel and Weiss's approach,² the interaction matrix element V_{12} is chosen to reproduce the barrier height (obtained from experiments or calculations). For cases where greater accuracy is required, it is also desirable to match the position and vibrational frequencies of the TS in addition to the barrier height. The Chang–Miller approach³ describes the interaction matrix element by a generalized Gaussian positioned at or near the transition state

$$V_{12}^2(\mathbf{q}) = A \exp[\mathbf{B}^T \cdot \Delta \mathbf{q} - \frac{1}{2} \Delta \mathbf{q}^T \cdot \mathbf{C} \cdot \Delta \mathbf{q}], \quad \Delta \mathbf{q} = \mathbf{q} - \mathbf{q}_{\text{TS}} \quad (5)$$

where \mathbf{q}_{TS} is the transition-state geometry. The coefficients are chosen so that the energy, gradient, and second derivatives of the EVB surface match ab initio calculations at the TS. Following Chang–Miller's notation,³ this yields simple, closed-form equations for parameters A , \mathbf{B} (a vector), and \mathbf{C} (a matrix).

$$A = [V_{11}(\mathbf{q}_{\text{TS}}) - V(\mathbf{q}_{\text{TS}})][V_{22}(\mathbf{q}_{\text{TS}}) - V(\mathbf{q}_{\text{TS}})] \quad (6a)$$

$$\mathbf{B} = \frac{\mathbf{D}_1}{[V_{11}(\mathbf{q}_{\text{TS}}) - V(\mathbf{q}_{\text{TS}})]} + \frac{\mathbf{D}_2}{[V_{22}(\mathbf{q}_{\text{TS}}) - V(\mathbf{q}_{\text{TS}})]} \text{ and} \\ \mathbf{D}_n = \frac{\partial V_m(\mathbf{q})}{\partial \mathbf{q}} \Big|_{\mathbf{q}=\mathbf{q}_{\text{TS}}} - \frac{\partial V(\mathbf{q})}{\partial \mathbf{q}} \Big|_{\mathbf{q}=\mathbf{q}_{\text{TS}}} \quad (6b)$$

$$\mathbf{C} = \frac{\mathbf{D}_1 \mathbf{D}_1^T}{[V_{11}(\mathbf{q}_{\text{TS}}) - V(\mathbf{q}_{\text{TS}})]^2} + \frac{\mathbf{D}_2 \mathbf{D}_2^T}{[V_{22}(\mathbf{q}_{\text{TS}}) - V(\mathbf{q}_{\text{TS}})]^2} - \\ \frac{\mathbf{K}_1}{V_{11}(\mathbf{q}_{\text{TS}}) - V(\mathbf{q}_{\text{TS}})} - \frac{\mathbf{K}_2}{V_{22}(\mathbf{q}_{\text{TS}}) - V(\mathbf{q}_{\text{TS}})} \text{ and} \\ \mathbf{K}_n = \frac{\partial^2 V_m(\mathbf{q})}{\partial \mathbf{q}^2} \Big|_{\mathbf{q}=\mathbf{q}_{\text{TS}}} - \frac{\partial^2 V(\mathbf{q})}{\partial \mathbf{q}^2} \Big|_{\mathbf{q}=\mathbf{q}_{\text{TS}}} \quad (6c)$$

The original version of the Chang–Miller method runs into difficulties when \mathbf{C} has one or more negative eigenvalues.^{12,15,18} In these cases, the form of V_{12}^2 in eq 5 diverges for large $\Delta \mathbf{q}$ values. The simplest solution to this problem switches the interaction matrix element to zero in regions where the unmodified V_{12}^2 is negative or divergent.¹⁵ Another approach is to include suitable cubic and quartic terms in the Gaussian to control asymptotic behavior.¹²

In the present article, an alternative form for V_{12}^2 is proposed. Instead of using a generalized Gaussian as in eq 5, a quadratic polynomial times a spherical Gaussian is employed.

$$V_{12}^2(\mathbf{q}) = A[1 + \mathbf{B}^T \cdot \Delta \mathbf{q} + \frac{1}{2} \Delta \mathbf{q}^T \cdot (\mathbf{C} + \alpha \mathbf{I}) \cdot \Delta \mathbf{q}] \\ \exp[-\frac{1}{2} \alpha |\Delta \mathbf{q}|^2] \quad (7)$$

Fitting to the energy, gradient, and Hessian at the transition state yields the same formulas for A and \mathbf{B} as those in the

Chang–Miller case; the expression for \mathbf{C} is slightly different.

$$\mathbf{C} = \frac{\mathbf{D}_1 \mathbf{D}_1^T + \mathbf{D}_2 \mathbf{D}_2^T}{A} + \frac{\mathbf{K}_1}{V_{11}(\mathbf{q}_{\text{TS}}) - V(\mathbf{q}_{\text{TS}})} + \\ \frac{\mathbf{K}_2}{V_{22}(\mathbf{q}_{\text{TS}}) - V(\mathbf{q}_{\text{TS}})} \quad (8)$$

The exponent α is chosen to be small enough so that the PES is smooth but not so small that the reactant and product energies are affected significantly. One approach is to choose α to give a good fit for the energies along the reaction path. The form of V_{12}^2 in eq 7 can also be viewed as expanding V_{12}^2 as a linear combination of s-, p-, and d-type Gaussians.

$$g(\mathbf{q}, \mathbf{q}_K, 0, 0, \alpha) = \exp[-\frac{1}{2} \alpha |\mathbf{q} - \mathbf{q}_K|^2]$$

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

$$g(\mathbf{q}, \mathbf{q}_K, i, j, \alpha) = (\mathbf{q} - \mathbf{q}_K)_i (\mathbf{q} - \mathbf{q}_K)_j \exp[-\frac{1}{2} \alpha |\mathbf{q} - \mathbf{q}_K|^2] \quad (9)$$

Because the coefficients in eq 7 are linear, the procedure can be readily generalized to include Gaussians at multiple centers, \mathbf{q}_K . For example, one could choose to place the Gaussian centers at the TS, reactant minimum, product minimum, and a few points along the reaction path to either side of the transition state. The generalized form of V_{12}^2 can be written as

$$V_{12}^2(\mathbf{q}) = \sum_K \sum_{i \geq j \geq 0}^{\text{NDim}} B_{ijk} g(\mathbf{q}, \mathbf{q}_K, i, j, \alpha) \quad (10)$$

where NDIM is 3 times the number of atoms for a Cartesian coordinate system or the number of coordinates if internal or redundant-internal coordinates are utilized. The Gaussian exponents are chosen such that the fit is sufficiently smooth for energies along the reaction path and V_{12}^2 is acceptably small at the reactants and products, if these are not already included in \mathbf{q}_K . In the simplest approach, the exponents are all equal; alternatively, if suitable criteria exist, they may be different for different centers, or even for different directions. The B_{ijk} coefficients are obtained by fitting to V_{12}^2 and its first and second derivatives at a number of points, \mathbf{q}_L , which can conveniently be the same as \mathbf{q}_K .

$$V_{12}^2(\mathbf{q}_L) = \sum_K \sum_{i \geq j \geq 0}^{\text{NDim}} B_{ijk} g(\mathbf{q}_L, \mathbf{q}_K, i, j, \alpha) \\ \frac{\partial V_{12}^2(\mathbf{q})}{\partial \mathbf{q}} \Big|_{\mathbf{q}=\mathbf{q}_L} = \sum_K \sum_{i \geq j \geq 0}^{\text{NDim}} B_{ijk} \frac{\partial g(\mathbf{q}, \mathbf{q}_K, i, j, \alpha)}{\partial \mathbf{q}} \Big|_{\mathbf{q}=\mathbf{q}_L} \\ \frac{\partial^2 V_{12}^2(\mathbf{q})}{\partial \mathbf{q}^2} \Big|_{\mathbf{q}=\mathbf{q}_L} = \sum_K \sum_{i \geq j \geq 0}^{\text{NDim}} B_{ijk} \frac{\partial^2 g(\mathbf{q}, \mathbf{q}_K, i, j, \alpha)}{\partial \mathbf{q}^2} \Big|_{\mathbf{q}=\mathbf{q}_L} \quad (11)$$

If the number of Gaussian centers is equal to the number of points (i.e., if the number of coefficients is equal to the number of energy values, first derivatives, and second derivatives), this is simply the solution of a set of linear equations.

$$\mathbf{DB} = \mathbf{F} \quad (12)$$

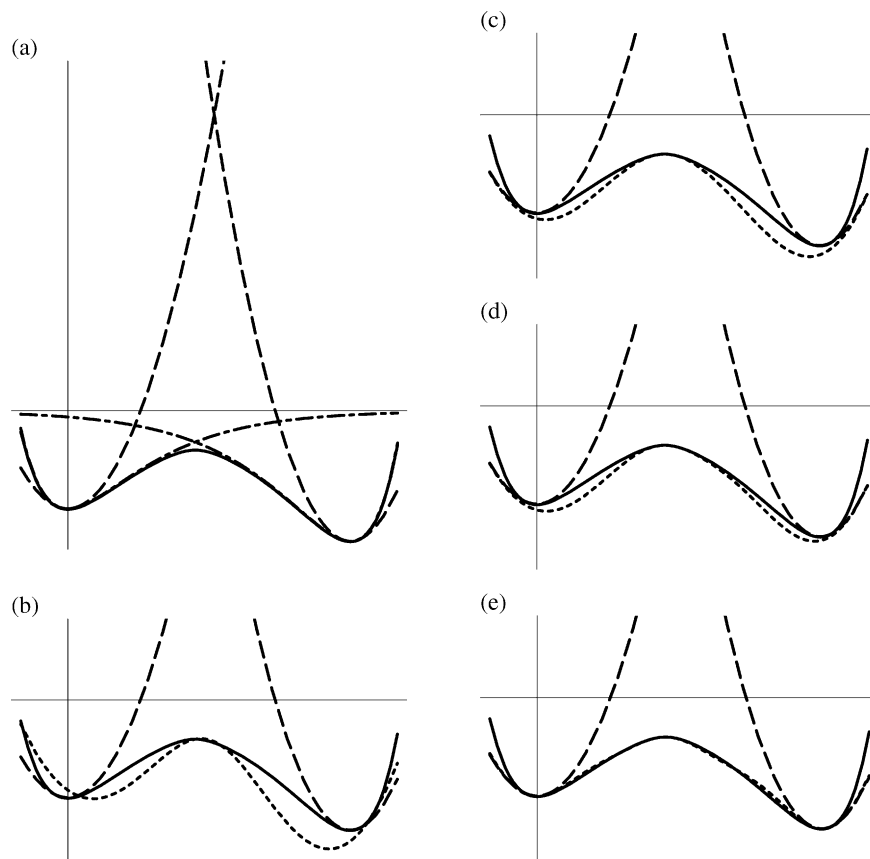


Figure 1. (a) One-dimensional potential energy curve (solid line) constructed from two interacting Morse curves (chain-dot). V_{11} and V_{22} (long dash) are quadratic functions fitted to the minima of the Morse curves, fitted using various EVB models (short dash). (b) EVB model with constant V_{12}^2 . (c) Chang–Miller EVB model with V_{12}^2 represented by a generalized Gaussian. (d) EVB model with V_{12}^2 represented by a quadratic polynomial times a Gaussian. (e) EVB model with a three-Gaussian fit.

where \mathbf{D} is a matrix containing the values of $g(\mathbf{q}_L, \mathbf{q}_K, i, j, \alpha)$, $\partial g(\mathbf{q}_L, \mathbf{q}_K, i, j, \alpha) / \partial \mathbf{q}|_{\mathbf{q}=\mathbf{q}_L}$, and $\partial^2 g(\mathbf{q}_L, \mathbf{q}_K, i, j, \alpha) / \partial \mathbf{q}^2|_{\mathbf{q}=\mathbf{q}_L}$ and \mathbf{F} is a column vector containing the values of $V_{12}^2(\mathbf{q}_L)$, $\partial V_{12}^2(\mathbf{q}) / \partial \mathbf{q}|_{\mathbf{q}=\mathbf{q}_L}$, and $\partial^2 V_{12}^2(\mathbf{q}) / \partial \mathbf{q}^2|_{\mathbf{q}=\mathbf{q}_L}$. Even with only a few expansion centers, the eigenvalues of \mathbf{D} become very small because of strong overlap between the Gaussians. In this case, the coefficients can be chosen in a least-squares manner. Similarly, if the number of Gaussian centers in the expansion is chosen to be smaller than the number of points where V_{12}^2 and its derivatives are evaluated, then the coefficients can also be obtained in a least-squares manner.

$$\begin{aligned} \text{minimize}(\mathbf{DB} - \mathbf{F})^T \mathbf{W} (\mathbf{DB} - \mathbf{F}) \\ \mathbf{D}^T \mathbf{W} \mathbf{D} \mathbf{B} = \mathbf{D}^T \mathbf{W} \mathbf{F} \end{aligned} \quad (13)$$

where \mathbf{W} is a diagonal weighting matrix. This can be solved easily using singular value decomposition.

Examples

One-Dimensional Test Case—Intersecting Morse Curves. A simple one-dimensional potential energy curve can be constructed from two intersecting Morse curves, as shown in Figure 1a. This resembles the potential energy along the reaction path for hydrogen abstraction reactions, $X-H + Y \rightarrow X + H-Y$, and similar atom-transfer processes involving the forming and breaking of single bonds. The parameters for the Morse curves are $D_e = 0.12$ and 0.16 au with force constants at the minima of 0.40 and 0.50 au, respectively; the curves are displaced by 3.00 au and interact by a small

matrix element, $V_{12}^2 = 0.010$ au. The empirical valence bond approximation to the surface is constructed from two quadratic potentials fitted to the individual Morse functions at their minima. As can be seen from Figure 1a, in the region of the transition state, V_{11} and V_{22} are much higher than the potential energy curve being modeled. Hence, V_{12}^2 will have to be quite large, providing a suitable challenge for the methodology.

Starting with Warshel and Weiss's method,² V_{12}^2 is set equal to a constant. In Figure 1b, the constant is chosen to reproduce the forward barrier height, and the curve is displaced to match the energies of the reactant and TS. Note that the resulting minima positions are shifted, the barrier width is too small, and the reaction exothermicity is too large. With only one parameter, fitting the potential energy curve well is difficult.

In the Chang–Miller approach, V_{12}^2 is represented by a generalized Gaussian, with the parameters fitted to the transition-state energy, gradient, and Hessian. For this example, the parameters for eq 5 are $A = 0.167$, $B = 0.385$, and $C = 1.988$. As shown in Figure 1c, this yields a significant improvement in fit to the potential energy curve. Because V_{12}^2 is not zero at the reactant and product geometries, the minima are slightly displaced, though not as much as in Figure 1b. When V_{12}^2 is represented by a single Gaussian times a quadratic polynomial, Figure 1d, the results are similar to the Chang–Miller approach. The Gaussian exponent α can be varied over the range 1.5 – 3.0 (bracketing the Chang–Miller

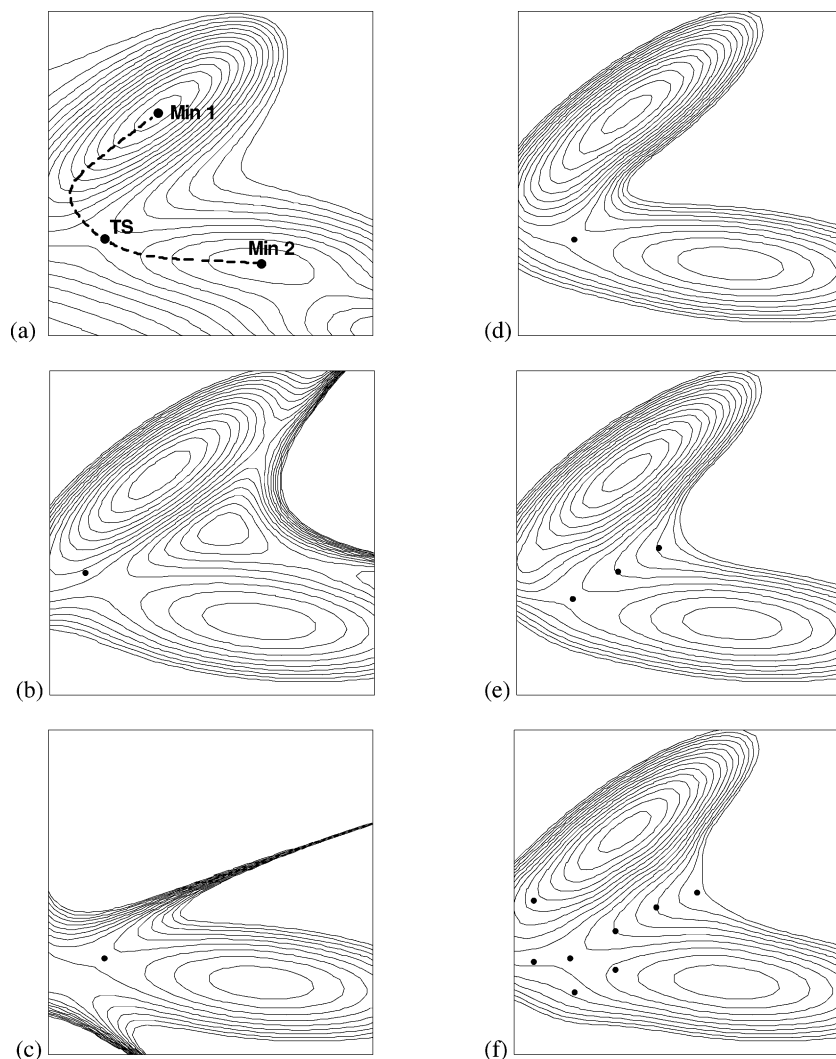


Figure 2. (a) Müller–Brown potential. (b) Chang–Miller EVB model with the V_{12}^2 Gaussian placed at the minimum on the intersection seam of V_{11} and V_{22} . (c) Chang–Miller EVB model with the V_{12}^2 Gaussian at the TS. (d) EVB model with V_{12}^2 represented by a quadratic polynomial times a Gaussian at the TS. (e) EVB model with a three-Gaussian fit. (f) EVB model with an eight-Gaussian fit. In parts b–f, the points indicate positions of the Gaussians used to construct V_{12}^2 .

exponent) and provides some additional flexibility in fitting the potential. If the exponent is chosen to be too large, V_{12}^2 is too narrow and the EVB curve no longer descends smoothly from the transition state.

A better fit is obtained by using three Gaussians times quadratic polynomials, for example, one at the transition state, another halfway between the TS and the reactant, and the third halfway between the TS and the product. Figure 1e shows that this approach produces a very good fit to the potential energy curve for suitably chosen exponents. The additional two Gaussians could also be placed at the minima, but this does not yield as smooth a curve. For more difficult cases, it could be beneficial to utilize five Gaussians: one at the TS, one at each minimum, and one halfway between the TS and each minimum.

Two-Dimensional Test Case—Müller–Brown Surface. The Müller–Brown surface²¹ is a convenient two-dimensional example frequently used as a test case for optimization algorithms and reaction-path-following methods:

$$V(x,y) = \sum A_i \exp[a_i(x - x_i^0)^2 + b_i(x - x_i^0)(y - y_i^0) + c_i(y - y_i^0)^2] \quad (14)$$

where $A = \{-200, -100, -170, 15\}$, $x^0 = \{1, 0, -0.5, -1\}$, $y^0 = \{0, 0.5, 1.5, 1\}$, $a = \{-1, -1, -6.5, 0.7\}$, $b = \{0, 0, 11, 0.6\}$, and $c = \{-10, -10, -6.5, 0.7\}$. As shown in Figure 2a, the surface has three minima. The upper two minima are connected by a rather curved reaction path and serve as a suitable test case for the EVB model. The V_{11} and V_{22} potentials are chosen as quadratic functions fitted to these two minima. Figure 2b demonstrates that the Chang–Miller method produces a good representation of the surface when the Gaussian for V_{12}^2 is placed at the lowest point on the intersection seam of V_{11} and V_{22} . Bofill et al. has used this approach in modeling potential energy surfaces for transition-state optimizations.¹⁵ However, placing a Gaussian for the Chang–Miller method at the TS yields a very poor approximation of the Müller–Brown surface, as seen in Figure 2c. This is because the matrix C has one negative eigenvalue, causing V_{12}^2 to diverge along the corresponding direction.

If V_{12}^2 is represented by a Gaussian times a quadratic polynomial placed at the transition state, then a good approximation to the Müller–Brown surface is obtained, as shown in Figure 2d. A better fit to the ridge separating the two minima may be constructed by placing two additional

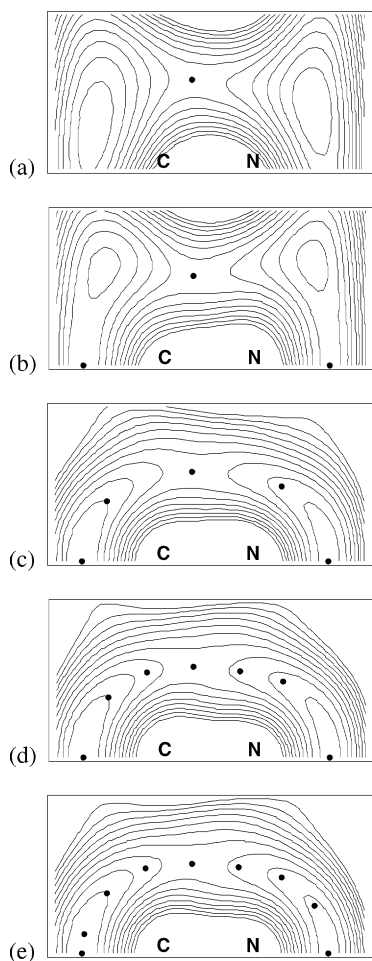


Figure 3. EVB fit to the potential energy surface for $\text{HCN} \rightarrow \text{HNC}$ using a Gaussian times a polynomial for V_{12}^2 in Cartesian coordinates. The carbon is at the origin; the nitrogen is at (1.116, 0.000), and the energy is plotted as a function of the Cartesian coordinates of the hydrogen. The points in a–e indicate the positions of the Gaussians used to construct V_{12}^2 .

Gaussians along the ridge, Figure 2e. The surface can be improved further by including more Gaussians, Figure 2f.

Molecular Case— $\text{HCN} \rightarrow \text{HNC}$. The isomerization of hydrogen cyanide is a simple unimolecular reaction often employed to test potential energy surface exploring algorithms. Because the C–N bond length changes little during this process, the key components of the potential energy surface can be easily visualized in two dimensions by plotting energy as a function of the hydrogen position. In internal coordinates involving bond lengths and angles, the reaction path is relatively linear. However, if Cartesian coordinates are used for the hydrogen, the reaction path is approximately a semicircle and fitting the surface should be more challenging. In particular, an EVB model with V_{11} , V_{22} , and V_{12}^2 in Cartesian coordinates is better suited for straight valleys rather than curved paths.

The transition state and reaction path for the $\text{HCN} \rightarrow \text{HNC}$ surface were calculated using the HF/3-21G level of theory.^{22–26} First and second derivatives were calculated at the transition state, the two minima, and selected points along the reaction path as input for the EVB model. The results are shown in Figure 3. Applying a Gaussian times a polynomial at the transition state yields a surface with some problems, Figure 3a. As a result of using Cartesian coordi-

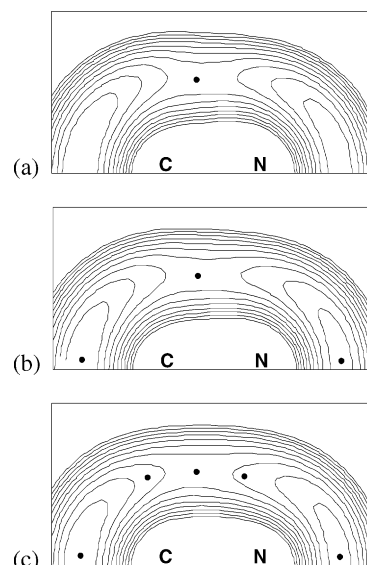


Figure 4. EVB fit to the potential energy surface for $\text{HCN} \rightarrow \text{HNC}$ using harmonic functions for V_{11} and V_{22} in redundant internal coordinates (C–N stretch, C–H stretch, N–H stretch, $\angle\text{H–C–N}$ bend, and $\angle\text{H–N–C}$ bend) using a Gaussian times a polynomial for V_{12}^2 in redundant internal coordinates. The carbon is at the origin; the nitrogen is at (1.116, 0.000), and the energy is plotted as a function of the Cartesian coordinates of the hydrogen. The points in a–c indicate the positions of Gaussians used to construct V_{12}^2 .

nates for the EVB surface, the minima valleys do not curve toward the transition state. The minima appear to have moved off the C–N axis as a consequence of fitting the V_{12}^2 only at the transition state. When two additional Gaussians ($\alpha = 0.5$) at the minima are included, Figure 3b, the energy, gradients, and Hessians at the minima are reproduced correctly by the EVB surface. However, the valleys still do not properly curve toward the TS, and there are spurious minima for bent structures. Adding two more points between the TS and the minima, Figure 3c, corrects the curvature of the valleys and eradicates the spurious minima. Two additional points near the transition state serve to improve the width of the potential energy surface through the transition state, Figure 3d. Extra points near the minima, Figure 3e, do not seem to provide any additional improvement.

As an alternative to Cartesian coordinates, internal coordinates can be used to construct V_{11} and V_{22} and to fit V_{12}^2 . Internal coordinates are more natural coordinates for this surface with a curved reaction path than Cartesian coordinates. To include the coordinates appropriate for both reactants and products, a redundant internal coordinate system consisting of $R(\text{CN})$, $R(\text{CH})$, $R(\text{NH})$, $\angle\text{HCN}$, and $\angle\text{HNC}$ was chosen. The simple Chang–Miller approach had difficulties because of negative eigenvalues in \mathbf{C} . A Gaussian times a quadratic polynomial provided a very reasonable fit to the surface, as shown in Figure 4a. Adding Gaussians near the reactant and product minima improves the surface somewhat, Figure 4b, primarily by providing a better fit around the minima. With an α value of 0.8 au for all Gaussians, the maximum error in the energy for points along the reaction path is 0.0025 au. Including two additional points along the reaction path on either side of the transition state reduced this error by a factor of 10 (Figure 4c, $\alpha = 1.5$ au).

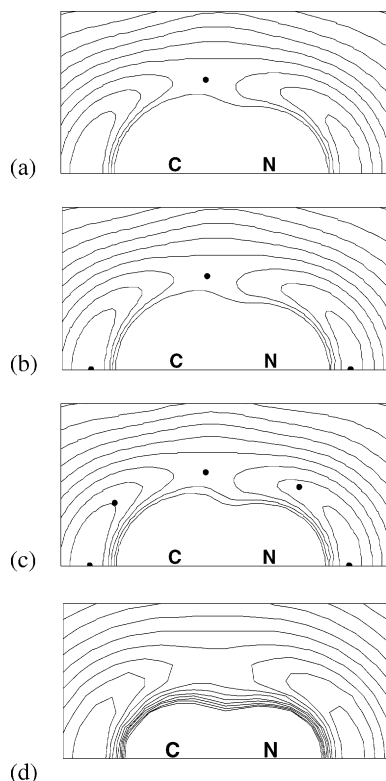


Figure 5. EVB fit using anharmonic functions for V_{11} and V_{22} in nonredundant (Z-matrix) internal coordinates (Morse for stretch, harmonic for bend, LJ for repulsion) and Gaussians times a polynomial in Cartesian coordinates for V_{12}^2 . The carbon is at the origin; the nitrogen is at (1.116, 0.000), and the energy is plotted as a function of the Cartesian coordinates of the hydrogen. The points in a–c indicate the positions of the Gaussians used to construct V_{12}^2 . (d) The potential energy surface for $\text{HCN} \rightarrow \text{HNC}$ calculated at RHF/3-21G.

Further reduction in the error can be achieved by adding more Gaussian centers at appropriate places on the surface.

As can be seen from Figures 3 and 4, the choice for V_{11} and V_{22} clearly has a profound effect on the shape of the potential energy surface in the regions away from the reaction path and fitting points. The simple harmonic functions used in Figures 3 and 4 were chosen to challenge the fitting procedure. More realistic potentials employed in molecular mechanics force fields include anharmonic stretching and bending potentials and nonbonded repulsions. Results employing such potentials are summarized in Figure 5 and compared to the actual $\text{HCN} \rightarrow \text{HNC}$ surface obtained by calculating the energy at the HF/3-21G level of theory on a suitable grid of points depicted in Figure 5d. To represent V_{11} in HCN, we employed Morse functions for the CN and CH bond stretches, harmonic potentials for the HCN bend and the CN–CH stretch–stretch interaction, and Lennard-Jones potentials for the nonbonded N–H interaction (an anti-Morse function works just as well; alternatively, a suitable anharmonic bend could have been used). Although the N–H nonbonded interaction would normally be covered by anharmonic bending terms in conventional force fields, a Lennard-Jones potential was employed to test the robustness of our fitting procedure. The corresponding coordinates were used in V_{22} for HNC. The interaction matrix element, V_{12}^2 , was represented by one or more Gaussians times polynomials in Cartesian coordinates and fit to energies, Cartesian

gradients, and Cartesian Hessians at selected points along the reaction path. A very good EVB surface is obtained with V_{12}^2 fit by only a single Gaussian times a quadratic polynomial at the transition state. The minima and shape of the reaction path are represented well. With suitably chosen dissociation energies for the Morse, the asymptotic form of the surface is also reproduced well. Including Gaussians at the minima does not change the surface, but adding two additional points between the minima and the TS improves the EVB surface. For the EVB surfaces shown Figure 5c, V_{12}^2 fit by five Gaussians with an exponent of 0.7 au yields a maximum error of 0.00013 au for the energy for points along the reaction path.

The logical extension of the tests cases illustrated in Figures 3–5 is the combination of anharmonic potentials for V_{11} and V_{22} in the natural internal coordinates for the reactants and products and V_{12}^2 represented by a series of Gaussians times quadratic polynomials in redundant internal coordinates. As in the quadratic synchronous transit transition-state optimization procedures,²³ these redundant internal coordinates are best chosen as the union of the reactant and product internal coordinates, augmented by any additional internal coordinates required to represent interactions found only in the reactive region of the potential energy surface. For an improved fit to V_{12}^2 , the Gaussians at the reactants, products, and transition states (and possible intermediates along the reaction path) should be augmented by additional Gaussians placed between those stationary points and the transition states along the reaction path. Extra fitting points can be added to represent special features such as the tunneling region near a saddle point or extended ridges separating reactant and product valleys. Molecular dynamics can locate additional areas of the potential energy surface where extra fitting points may be needed, in a manner akin to the “GROW” procedure of Collins.²⁷

Summary

The present work investigates some alternatives for representing V_{12}^2 employed in constructing EVB-type potential energy surfaces for later use in molecular dynamics calculations of chemical reactions. The use of a Gaussian times a quadratic polynomial for V_{12}^2 instead of the generalized Gaussian used in the Chang–Miller method has been proposed. This approach overcomes the divergence difficulties often encountered in practice when the generalized Gaussian is used to fit to the energy, gradient, and Hessian at a transition state. The approach is extended by representing V_{12}^2 as a linear combination of Gaussians times polynomials at selected points anywhere on the surface. The utility of the methodology is illustrated by applications to some simple one- and two-dimensional model surfaces along with the surface for the $\text{HCN} \rightarrow \text{HNC}$ isomerization reaction. A single Gaussian times a quadratic polynomial performs as well as the Chang–Miller approach where the latter succeeds and gives a good fit even when Chang–Miller has divergence difficulties. Better fits to potential energy surfaces are obtained with a distribution of Gaussians, particularly when the reaction path is curved or when the coordinates system makes the fit challenging. For $\text{HCN} \rightarrow \text{HNC}$, the effect of

the coordinate system on the quality of the EVB surface was explored. Internal coordinates performed better than Cartesian coordinates; however, both coordinate systems could be used to fit the potential energy along the reaction path to within chemical accuracy with as few as five fitting points. The quality of the surface away from the fitting points depends on the choice of V_{11} and V_{22} . Anharmonic, internal coordinate potentials with the proper asymptotic behavior produce a significantly improved global surface when compared to harmonic potentials in either Cartesian or internal coordinates. There is no restriction on the coordinate system or placement of the Gaussians representing V_{12}^2 in the current method, and extra points can be added to fine-tune special features on the surface. Practical methods for the automatic placement of the Gaussians will be explored in future work.

Acknowledgment. This research was supported by the Office of Naval Research (N00014-05-1-0457). The authors thank Drs. W. H. Miller, G. A. Voth, K. Wong, and F. Paesani for helpful discussions.

References

- (1) Villa, J.; Warshel, A. *J. Phys. Chem. B* **2001**, *105*, 7887–7907.
- (2) Warshel, A.; Weiss, R. M. *J. Am. Chem. Soc.* **1980**, *102*, 6218–6226.
- (3) Chang, Y.-T.; Miller, W. H. *J. Phys. Chem.* **1990**, *94*, 5884–5888.
- (4) Hansson, T.; Nordlund, P.; Aqvist, J. *J. Mol. Biol.* **1997**, *265*, 118–127.
- (5) Okuyama-Yoshida, N.; Nagaoka, M.; Yamabe, T. *J. Phys. Chem. A* **1998**, *102*, 285–292.
- (6) Nagaoka, M.; Okuyama-Yoshida, N.; Yamabe, T. *J. Phys. Chem. A* **1998**, *102*, 8202–8208.
- (7) Luzhkov, V.; Aqvist, J. *J. Am. Chem. Soc.* **1998**, *120*, 6131–6137.
- (8) Okuyama-Yoshida, N.; Kataoka, K.; Nagaoka, M.; Yamabe, T. *J. Chem. Phys.* **2000**, *113*, 3519–3524.
- (9) Karmacharya, R.; Antoniou, D.; Schwartz, S. D. *J. Phys. Chem. A* **2001**, *105*, 2563–2567.
- (10) Aqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523–2544.
- (11) Hwang, J. K.; Chu, Z. T.; Yadav, A.; Warshel, A. *J. Phys. Chem.* **1991**, *95*, 8445–8448.
- (12) Chang, Y. T.; Minichino, C.; Miller, W. H. *J. Chem. Phys.* **1992**, *96*, 4341–4355.
- (13) Jensen, F. *J. Comput. Chem.* **1994**, *15*, 1199.
- (14) Jensen, F. *J. Am. Chem. Soc.* **1992**, *114*, 1596.
- (15) Anglada, J. M.; Besalu, E.; Bofill, J. M.; Crehuet, R. *J. Comput. Chem.* **1999**, *20*, 1112–1129.
- (16) Bernardi, F.; Olivucci, M.; Robb, M. A. *J. Am. Chem. Soc.* **1992**, *114*, 1606–1616.
- (17) Minichino, C.; Voth, G. A. *J. Phys. Chem. B* **1997**, *101*, 4544–4552.
- (18) Kim, Y.; Cochado, J. C.; Villa, J.; Xing, J.; Truhlar, D. G. *J. Chem. Phys.* **2000**, *112*, 2718–2735.
- (19) Lin, H.; Pu, J.; Albu, T. V.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 4112–4124.
- (20) Albu, T. V.; Cochado, J. C.; Truhlar, D. G. *J. Phys. Chem. A* **2001**, *105*, 8465–8487.
- (21) Müller, K.; Brown, L. D. *Theor. Chim. Acta* **1979**, *53*, 75.
- (22) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.03 ed.; Gaussian, Inc.: Wallingford, CT, 2004.
- (23) Peng, C.; Schlegel, H. B. *Isr. J. Chem.* **1993**, *33*, 449–454.
- (24) Gonzalez, C.; Schlegel, H. B. *J. Phys. Chem.* **1990**, *94*, 5523.
- (25) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2145.
- (26) Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939–947.
- (27) Collins, M. A. *Theor. Chem. Acc.* **2002**, *108*, 313–324.

CT600084P

Acceleration of Classical Mechanics by Phase Space Constraints

Emilio Martínez-Núñez

*Departamento de Química Física, Universidad de Santiago de Compostela,
15782 Santiago de Compostela, Spain*

Dmitrii V. Shalashilin*

*Physical and Theoretical Chemistry Laboratory, Oxford University, South Parks Road,
Oxford OX1 3QZ, U.K.*

Received February 7, 2006

Abstract: In this article phase space constrained classical mechanics (PSCCM), a version of accelerated dynamics, is suggested to speed up classical trajectory simulations of slow chemical processes. The approach is based on introducing constraints which lock trajectories in the region of the phase space close to the dividing surface, which separates reactants and products. This results in substantial (up to more than 2 orders of magnitude) speeding up of the trajectory simulation. Actual microcanonical rates are calculated by introducing a correction factor equal to the fraction of the phase volume which is allowed by the constraints. The constraints can be more complex than previously used boosting potentials. The approach has its origin in Intramolecular Dynamics Diffusion Theory, which shows that the majority of nonstatistical effects are localized near the transition state. An excellent agreement with standard trajectory simulation at high energies and Monte Carlo Transition State Theory at low energies is demonstrated for the unimolecular dissociation of methyl nitrite, proving that PSCCM works both in statistical and nonstatistical regimes.

1. Introduction

Speeding up calculations of chemical reaction rates is an important goal of theoretical chemistry. For reactions with a high activation barrier, classical trajectory simulation, which is the most straightforward way to obtain reaction rates, can be very time-consuming even for reactions of moderate sized molecules in the gas phase. Current approaches to speeding up classical trajectory simulations include reduction of degrees of freedom (coarse graining),^{1,2} importance sampling of initial conditions near the transition state,³ and the hyperdynamics approach, which modifies interactions in the system by elevating potential energy wells in order to decrease the reaction barrier^{4–8} without changing the characteristics of the transition state. See also ref 9 for some preliminary ideas related to hyperdynamics. Often trajectory simulations^{10,11} are performed at temperatures well

in excess of those in experiment, and rates are then extrapolated.^{12,13}

In this article we introduce and test the method of Phase Space Constrained Classical Dynamics to speed up classical trajectory simulations of low rates of chemical processes. The main idea is to impose constraints which lock trajectories in the region of the phase space close to the dividing surface separating reactants and products (i.e. transition state). This pushes the trajectories toward the dividing surface, thereby decreasing the simulation time and substantially speeding up the trajectory simulation. Actual microcanonical rates are calculated as a product of the accelerated rate and a correction factor equal to the fraction of phase volume allowed by constraints. The justification is provided by Intramolecular Dynamics Diffusion Theory (IDDT),^{13–16} which shows that far from the dividing surface the dynamics do not disturb microcanonical distribution so that the majority of nonstatis-

tical effects are localized near the transition state. Therefore only the dynamics near the transition state need to be simulated. In the next section we give a brief summary of IDDT and describe the method of Phase Space Constrained Classical Mechanics (PSCCM). In section 3 various implementations of PSCCM are tested for the reaction of dissociation of methyl nitrite ($\text{CH}_3\text{ONO} \rightarrow \text{CH}_3\text{O} + \text{NO}$). We show that PSCCM reproduces the Monte Carlo Transition State Theory at low energies, where the trajectory rate is statistical, and trajectory simulations which account for nonstatistical effects at high energies. The last section provides a summary and a discussion.

2. Theory

2.1. Trajectory Calculations and Monte Carlo Transition State Theory. The calculation of the microcanonical rate constant $k(E)$ of a chemical reaction begins with choosing a dividing surface, S^* , which separates the reactants and the products in the phase space. Usually S^* is implicitly defined by a critical value of the reaction coordinate $q_r = q_r^*$. The dividing surface can also be defined in the phase space.¹⁷ After that the dynamics of reaction can then be treated either by classical trajectories (CT) or by transition state theory.

In trajectory simulations a microcanonical ensemble of initial phase space points in the region of the reactant is set by a Monte Carlo procedure, which are then propagated in time by numerical integration of Hamilton's equations of motion. The rate can be established by a fit to the first-order rate equation

$$\ln \frac{N(t)}{N(t=0)} = -k^{\text{traj}}(E)t \quad (1)$$

where N is the number of trajectories still in the region of the reactants.

These trajectory rates can be compared with those obtained by statistical Monte Carlo Transition State Theory (MCTST),¹⁸ which will have the form of a flux through the dividing surface as

$$k^{\text{stat}}(E) = \frac{1}{2} \frac{\int_{\Gamma} \delta[H(p,q) - E] |\dot{q}_r| \delta(q_r - q_r^*) d\Gamma}{\int_{\Gamma} \delta[H(p,q) - E] d\Gamma} \quad (2)$$

The standard numerical approach to MCTST is by the Metropolis random walk.¹⁹ In most cases the computed trajectory rates (1) are lower than those of the purely statistical theory (2), despite the fact that the initial ensemble in the trajectory simulation is microcanonical. This is known as intrinsic non-RRKM behavior.^{20,21}

2.2. Brief Summary of IDDT. The Intramolecular Dynamics Diffusional Theory¹³⁻¹⁶ was developed to explain and to quantify the intrinsic non-RRKM behavior. IDDT considers only the motion along the reaction coordinate and replaces the Liouville equation of Classical Mechanics by a diffusional equation along this coordinate. The nonstatistical trajectory rates after that can easily be extracted from the rate of diffusion along the reaction coordinate toward the dividing surface (transition state), which serves as an absorbing wall for the diffusion. Initial microcanonical distribution in the

reactant region results in the initial statistical rate. However, this initial uniform distribution evolves quickly so that it becomes depleted near the absorbing wall although it remains unchanged (i.e. still microcanonical and uniform) far from the dividing surface. Later the rate of reaction is determined by trajectory diffusion to the depleted region near the adsorbing wall—a process attributed to the intramolecular vibrational energy redistribution (IVR), which cannot be described by transition state theory.

Therefore IDDT predicts that the trajectory rate constant k^{traj} in (1) must be time dependent. By definition the initial ($t=0$) trajectory rate constant must be equal to the statistical k^{stat} given by eq 2. In practice however, this short time rate is never observed in trajectory simulations unless serious efforts are made to detect it.^{15,16} The rate observed in trajectory simulations is actually the rate of diffusion along the reaction coordinate toward the dividing surface due to intramolecular vibrational energy redistribution, which is smaller than the initial statistical rate.

$$k^{\text{traj}}(E) = k^{\text{IVR}} < k^{\text{stat}}(E) \quad (3)$$

Depletion of the distribution near the dividing surface and reduction of the trajectory below the Transition State Theory value k^{stat} is a nonstatistical effect, which becomes stronger as the initial energy increases. At lower energies the rate constant reaches its statistical limit.

$$k^{\text{traj}}(E) \approx k^{\text{stat}}(E) \quad (4)$$

The energies at which trajectory simulations start yielding statistical results are quite low, and straightforward trajectory simulation is expensive for those energies. Therefore methods of accelerated dynamics are required to reach the statistical limit of classical mechanics. Another conclusion of IDDT, which is important in the context of the present work, is that microcanonical distribution is disturbed only in the vicinity of the dividing surface, which means that all nonstatistical effects are well localized. The IDDT picture has been confirmed by a number of simulations.¹³⁻¹⁶

2.3. Phase Space Constrained Classical Mechanics. As mentioned above, straightforward application of trajectory simulations is extremely time-consuming for energies approaching the activation energy threshold. Calculation of the integrals in the MCTST eq 2 is also difficult. In practice, the δ -function, $\delta(q_r - q_r^*)$, is approximated by a narrow function, which will be nonzero only in the vicinity of the dividing surface. The Monte Carlo random walk only visits the region that contributes to the integral in eq 2 infrequently. This results in a very slow convergence of the MCTST.

An obvious method, which speeds up the convergence of MCTST, has been used in ref 22. It is based on the idea of importance sampling and is illustrated in Figure 1a. The random walk is restricted to the phase volume V_1 close to the dividing surface. The calculated auxiliary rate constant $k_1^{\text{stat}}(E)$ is given by eq 2 with the integral over Γ_1 only. The actual rate can then be calculated by

$$k^{\text{stat}}(E) = k_1^{\text{stat}}(E) \frac{\Gamma_1}{\Gamma} \quad (5)$$

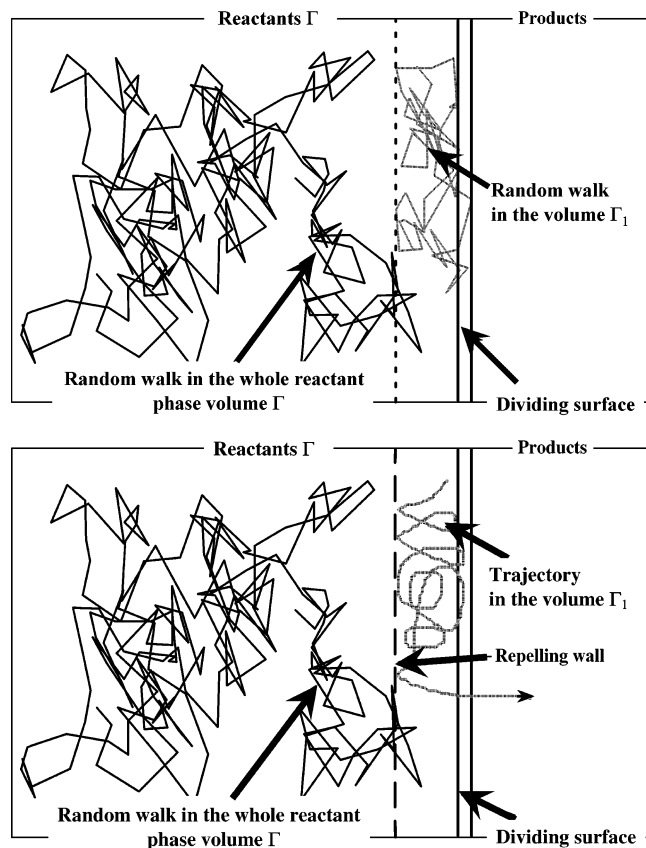


Figure 1. Sketch of the importance sampling approach to Monte Carlo Transition State Theory (frame a) and Phase Space Constrained Classical Mechanics (frame b). In the former the random walk, locked in the volume Γ_1 (dashed line in the frame a), visits the dividing surface more frequently yielding accelerated statistical rate $k_1^{\text{stat}}(E)$, while in the latter a trajectory locked in the volume Γ_1 (dashed line in the frame b) crosses the dividing surface and reaches products faster, producing an accelerated trajectory rate $k_1^{\text{traj}}(E)$. Both methods require additional random walk (solid line) to calculate the correction factor Γ_1/Γ .

A second random walk is then needed to estimate the volume ratio Γ_1/Γ . This method introduces no new approximations into MCTST but greatly reduces computational time because two short random walks converge much more quickly than a single random walk in the whole phase space.

In this article we propose a method of Phase Space Constrained Classical Mechanics, which expands the technique²² to trajectory simulations. An outline of the method is illustrated in Figure 1b. The motion of the molecule is restricted in the region close to the dividing surface by applying an auxiliary rigid wall, which repels trajectories in the direction of the products, thus increasing the probability that reactive conditions are reached and dramatically decreasing the cost of the calculation. An analogous expression to eq 5 for the true rate can be defined as

$$k_1^{\text{traj}}(E) = k_1^{\text{traj}}(E) \frac{\Gamma_1}{\Gamma} \quad (6)$$

where $k_1^{\text{traj}}(E)$ is the accelerated rate constant, and the correction factor Γ_1/Γ is the same as in (5). The advantage

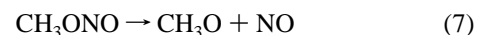
of PSCCM in comparison to accelerated MCTST is that PSCCM should be able to take into account nonstatistical intrinsic non-RRKM effects, described by the IDDT.

It should be noted that not only the microcanonical rate constant but also the canonical (i.e. thermal average) rate constant can be estimated using PSCCM. The generalization is not difficult.

The PSCCM method is an application of Importance Sampling, which is well-known within the domain of trajectory simulations. In traditional methods, initial conditions are often biased to the most important region.³ A new feature of the PSCCM method is the additional biasing of the dynamics itself. The only comparable approach to accelerating classical dynamics is the method suggested by Voter,^{4–8} which uses additional “boosting” potentials to push the dynamics toward important regions of the configuration space. There are two advantages of our version of the accelerated dynamics, based on constraints which lock the trajectory in a small portion of the total phase volume rather than on boosting potentials. First, PSCCM relies on a rigorous understanding of nonstatistical effects provided by IDDT, which therefore gives a good idea as to where to put constraints. Second, as will be shown in the next section, phase space constraints can be introduced in a way, which is hard to describe by a boosting potential. It should be also mentioned that phase space constraints were used previously in refs 23–26 for incorporating zero point energy effects into classical mechanics. Although the technique of PSCCM is somewhat similar to that of refs 23–26, our goal is to speed up simulations.

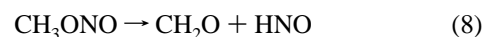
3. Implementation and Test of PSCCM

We have implemented the idea of PSCCM by calculation of the microcanonical rate of dissociation of methyl nitrite (CH_3ONO).



3.1. Potential Energy Surface for the Model System.

Previously, reaction 7 has been studied both by trajectory simulation and MCTST.²⁷ For the present study the potential energy surface has been modified in order to prevent another reaction channel considered before,²⁷ namely



Specifically, the potential energy surface of the present study reads

$$V = V(R_{\text{ON}}) + \sum_{i=1}^5 V(R_i) + \sum_{i=1}^8 V(\theta_i) + \sum_{i=1}^4 V(\tau_i) \quad (9)$$

where

$$V(R_{\text{ON}}) = D_e \{1 - \exp[-\beta \times (R_{\text{ON}} - R_{\text{ON}}^{\text{eq}})]\}^2 - D_e \quad (10)$$

is the same Morse function as used before²⁷ but without the switching functions.²⁷ For the remaining NO, CO, and the three CH bonds the harmonic functions

$$V(R_i) = 0.5K_i^s(R_i - R_i^{\text{eq}})^2 \quad (11)$$

Table 1. New Parameters of the Potential Energy Surface Used in the Present Study and Frequencies for *trans*-CH₃ONO

parameter	value (kcal/ mol/Å ²)	parameter	value (kcal/ mol/Å ²)	parameter	value (kcal/ mol/Å ²)
K_{NO}	250	K_{CO}	1250	K_{CH}	650
frequencies					
old PES (ref 27)			this work		
173	931	1459	173	953	1249
246	1049	1670	246	1049	1670
368	1099	2760	372	1103	2807
465	1430	2873	470	1430	2914
715	1430	2909	726	1430	2916

were used to prevent additional dissociation channels. The ONO, CON, HCO, and HCH bending potentials are the same as in ref 27 where they were modeled by the harmonic function

$$V(\theta_i) = 0.5K_i^b(\theta_i - \theta_i^{\text{eq}})^2 \quad (12)$$

Finally $V(\tau_i)$ are a cosine series to treat the dihedral interactions in the system (i.e. the CONO and the three HCON dihedrals)

$$V(\tau) = \sum_{i=0}^5 a_i \cos(i\tau) \quad (13)$$

The parameters of eqs 9–13 are reported in ref 27 except that the force constants of harmonic potentials (11) are used here instead of Morse functions. These K_i^s were adjusted to fit the vibrational frequencies of our previous PES (see Table 1), which was, in turn, developed to reproduce ab initio or experimental (when available) vibrational frequencies, reaction enthalpies, and geometrical parameters. The most crucial parameter is the dissociation energy of the ON bond $D_e = 42$ kcal/mol which is the same as in ref 27. The vibrational frequencies of *trans*-CH₃ONO (the most stable conformer) are collected in Table 1 together with those calculated with the previous PES.²⁷ As can be seen the removal of some flexibility in the PES has little effect on the vibrational frequencies of the reactant, and therefore the simplified PES is accurate.

3.2. Trajectory and MCTST Computational Details.

Before applying PSCCM we used the new PES to calculate the rate constant by means of standard trajectory simulation (1) and Monte Carlo Transition State Theory (2) in its accelerated version²² for the following vibrational energies $E = 70, 80, 100,$ and 150 kcal/mol with zero total angular momentum. An extensively modified version of the GEN-DYN code¹¹ has been used.

The dividing surface was positioned at the internuclear distance R_{ON} , which minimizes the statistical MCTST reaction rate, namely $R_{\text{ON}} = 4.3, 3.8, 3.5,$ and 3.3 Å for $E = 70, 80, 100,$ and 150 kcal/mol, respectively. The minimized MCTST rate constants are collected in Table 2. The accelerated MCTST rate constants were calculated by eq 5 with the restricted phase space volume (Γ_1/Γ) confined to the region between 2 Å and q_r^* . For 80 kcal/mol the

Table 2. Accelerated MCTST Rate Constants (in ps⁻¹) Obtained in This Study

energy/(kcal/mol)	q_r^*	accelerated MCTST	standard MCTST
70	4.3	0.000081 ± 0.00002	
80	3.8	0.0012 ± 0.0001	0.0011 ± 0.0001
100	3.5	0.025 ± 0.001	0.025 ± 0.001
150	3.3	0.56 ± 0.02	0.54 ± 0.02

calculation of accelerated rates is an order of magnitude faster than by the standard MCTST procedure. For 70 kcal/mol standard MCTST calculations are prohibitive, while accelerated MCTST takes less than 1 h of CPU time. As can be seen in Table 2, the accelerated MCTST rate constants are in very good agreement with the standard calculations, which is not surprising, because the method²² does not introduce any approximations to MCTST.

In trajectory simulations, ensembles of 1000 trajectories were employed in all cases (for standard and PSCCM calculations) except for the standard trajectory calculations at 80 kcal/mol, for which we needed 3000 trajectories to achieve the same statistics as in the accelerated computations. In trajectory simulation the dividing surface was chosen to be the same as in the above MCTST (i.e. $R_{\text{ON}} = 4.3, 3.8, 3.5,$ and 3.3 Å for $E = 70, 80, 100,$ and 150 kcal/mol, respectively). Due to a rapid decrease of the rate constant, straightforward trajectory simulation is very slow for $E = 80$ kcal/mol and is not feasible at all for $E = 70$ kcal/mol.

3.4. Implementations of the PSCCM Methods. The most straightforward implementation of PSCCM is to impose a constraint on the bond length R_{ON} . When the distance between the two atoms becomes smaller than a given value R_{min} , we invert the projection of the velocities of O and N on the bond in the system of their center of mass. This is equivalent to introducing a hard wall potential between N and O. Figure 2a shows the rate constant $k(E)$ (circles) obtained by PSCCM as a function of the repelling wall position together with the rate constant calculated by the straightforward trajectory simulation shown by the gray line, the width of which indicates the error in the trajectory calculation. The accelerated rate is within the error bar from the trajectory result up to $R_{\text{min}} = 1.8$ Å. The rate constants shown in Figure 2a are also summarized in Table 3a.

The above implementation referred to as implementation 1 introduces an extra potential, and although its nature is different from that of Voter^{4–7} (instead of lifting up the bottom of the potential energy well we modify the repulsive part of the potential), the approach is similar in spirit. In implementation 2 we impose a different condition and invert the velocities only when the energy of the ON bond, written as

$$E_{\text{ON}} = p_{\text{rel}}^2/2\mu_{\text{rel}} + V(R_{\text{ON}}) \quad (14)$$

with p_{rel} and μ_{rel} being the relative momentum and the reduced mass of the ON bond, respectively, and $V(R_{\text{ON}})$, the Morse potential of eq 9, becomes smaller than a given minimal energy E_{min} . The constraint

$$E_{\text{ON}} > E_{\text{min}} \quad (15)$$

includes momenta, whereas simple boosting potentials^{4–8}

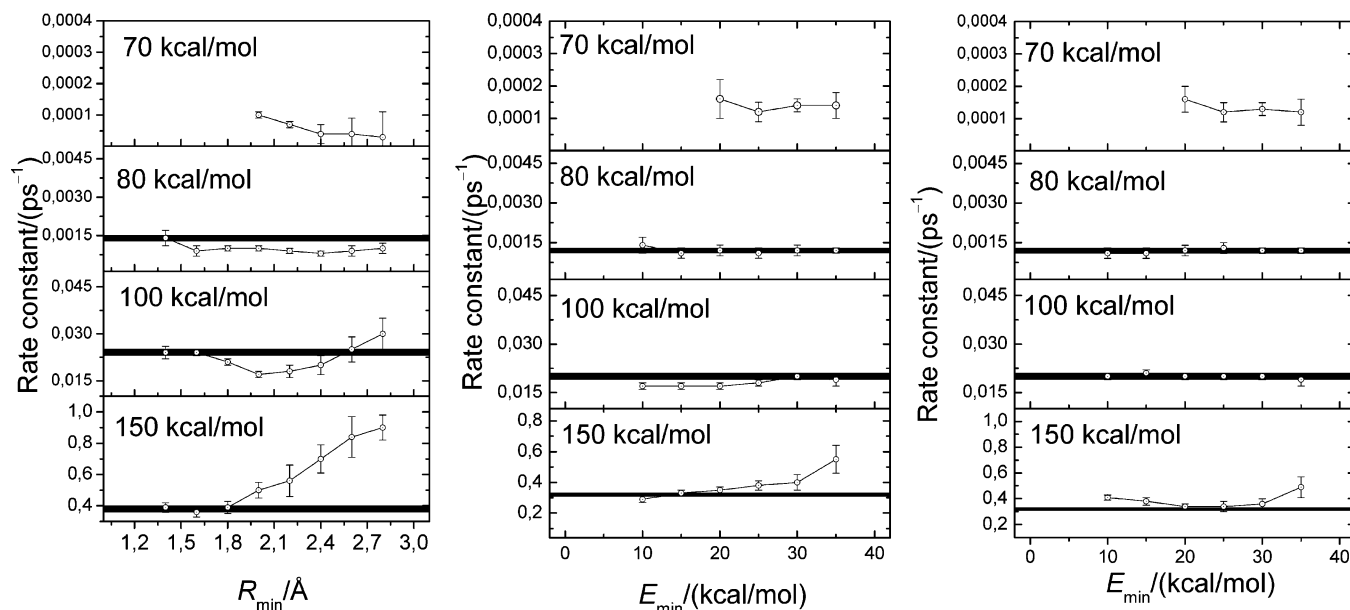


Figure 2. a. The microcanonical rate constant $k(E)$ obtained with the help of PSCM, compared with the result of straightforward trajectory simulation (gray line) for $E = 70, 80, 100,$ and 150 kcal/mol as a function of the constraint condition given by R_{\min} (the minimum ON bond length) for implementation 1. b. Microcanonical rate constant $k(E)$ obtained with the help of PSCCM compared with the result of straightforward trajectory simulation (gray line) for $E = 70, 80, 100,$ and 150 kcal/mol as a function of the constraint condition given by E_{\min} (the energy of the diatomic ON fragment) for implementation 2. c. Microcanonical rate constant $k(E)$ obtained with the help of PSCCM compared with the result of straightforward trajectory simulation (gray line) for $E = 70, 80, 100,$ and 150 kcal/mol as a function of the constraint condition given by E_{\min} (the energy of the diatomic ON fragment) for implementation 3.

depend only on coordinates. Figure 2b and Table 3b show the results of implementation 2 for the rate constants, which are again very close to those of straightforward trajectory simulations, even for the energies E_{\min} approaching the dissociation energy of the ON bond. Overall, the performance of implementation 2 is much better than that of implementation 1. However, at the highest energies of this study (100 and 150 kcal/mol) method 2 yields rates a bit smaller than the standard ones, particularly at low E_{\min} . We noticed that this behavior could be due to trapping of trajectories in the repulsive part of the ON potential well because on the repulsive part of the ON Morse potential, the energy can be higher than E_{\min} and therefore the inversion of O and N velocities can lead to shrinking rather than stretching of the ON bond distance, which may lower the value of the accelerated rate constant.

To prevent this trapping we devised implementation 3. In implementation 3, which is a combination of implementations 1 and 2, the velocity is inverted either when the energy becomes smaller than E_{\min} or when the internuclear distance becomes smaller than the equilibrium ON bond length. In this way, we prevent the system from moving to the repulsive part of the potential. Implementation 3 produced the best results, shown in Figure 2c (circles) and Table 3c. Particularly, for the highest energy employed in this study (150 kcal/mol) the method is accurate when $E_{\min} \leq 30$ kcal/mol (i.e. up to 70% of the dissociation energy of the molecule). For other energies the method is accurate even for the highest E_{\min} of 35 kcal/mol, which is up to 83% of the dissociation energy of the molecule.

Table 3 also shows the computational cost of calculations (CPU time). For example for $E = 80$ kcal/mol straightfor-

ward trajectory simulation (TS) required more than 3000 min of CPU time, while implementations 2 and 3 needed only about 100 minutes with approximately 60% of the effort going on running trajectories estimating $k_1^{\text{traj}}(E)$ (CPU time in parentheses). The rest was spent on calculating the (Γ_1/Γ) correction factor by random walk.

Figure 3 shows rate constants calculated by PSCCM, straightforward TS, and statistical MCTST. On the scale of the plot the result of PSCCM is indistinguishable from that of straightforward trajectory simulation. In agreement with the predictions of IDDT at high energies, the dynamical calculations produce rates higher than those of statistical MCTST, while at low energies nonstatistical effects are negligible. Accelerated PSCCM allows trajectory simulations for energies inaccessible for straightforward trajectory simulations and pushes the limit of dynamical calculations to low energies where the rate constant reaches its statistical limit.

Therefore PSCCM reproduces nonstatistical effects at high energies. On the other hand, it is most efficient at low energies where trajectory rates are close to statistical, which can be calculated very accurately and efficiently with the help of accelerated MCTST. To compare the PSCCM and MCTST methods at low energies methods Figure 4 shows variation of the rate constant with the position of the dividing surface. MCTST rate has a distinct minimum, which defines the transition state and the actual rate. On the other hand, the PSCCM are much less sensitive to the variations of the dividing surface so that the PSCCM rate constant is in good agreement with the minimized MCTST rate constant. This can be an advantage when variation of the transition state is difficult, for example, if the reaction coordinate cannot be

Table 3. Computational Details and Rate Constants (in ps⁻¹) Obtained in This Study with Implementations 1–3

a. Implementation 1																
traj	E = 70 kcal/mol				E = 80 kcal/mol				E = 100 kcal/mol			E = 150 kcal/mol				
	rate			time ^a	rate			time ^a	rate		time ^a	rate		time ^a		
	$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		0.0012± 0.0001			3348	0.020± 0.001		380	0.32± 0.01		36		
PSCCM	$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ	
R_{min}^b																
1.4					0.0014± 0.0003	0.0020	0.70	121 (119)	0.024± 0.002	0.033	0.72	114 (112)	0.39± 0.03	0.49	0.78	43 (42)
1.6					0.0009± 0.0002	0.0045	0.21	114 (108)	0.024± 0.001	0.085	0.29	110 (106)	0.36± 0.03	0.77	0.47	31 (29)
1.8					0.0010± 0.0001	0.023	0.041	107 (101)	0.021± 0.001	0.22	0.093	104 (98)	0.39± 0.04	1.46	0.27	23 (19)
2.0	0.00010± 0.00001	0.029	0.0032	1971 (1171)	0.0010± 0.0001	0.12	0.0087	100 (71)	0.017± 0.001	0.51	0.034	68 (53)	0.50± 0.05	2.95	0.17	20 (12)
2.2	0.00007± 0.00001	0.11	0.00058	1759 (1194)	0.0009± 0.0001	0.36	0.0025	105 (59)	0.018± 0.002	1.11	0.016	42 (27)	0.56± 0.10	4.97	0.11	16 (8)
2.4	0.00004± 0.00003	0.30	0.00015	1980 (1151)	0.0008± 0.0001	0.74	0.0011	85 (36)	0.020± 0.003	2.09	0.0094	40 (15)	0.70± 0.09	8.90	0.079	15 (5)
2.6	0.00004± 0.00005	0.64	0.000060	1575 (1104)	0.0009± 0.0002	1.43	0.00060	79 (22)	0.025± 0.004	4.17	0.0061	44 (10)	0.84± 0.13	15.4	0.054	13 (3)
2.8	0.00003± 0.00008	1.24	0.000028	1762 (1103)	0.0010± 0.0002	2.80	0.00035	105 (12)	0.030± 0.005	7.34	0.0041	95 (6)	0.90± 0.08	27.1	0.033	20 (3)
b. Implementation 2																
traj	E = 70 kcal/mol				E = 80 kcal/mol				E = 100 kcal/mol			E = 150 kcal/mol				
	rate			time ^a	rate			time ^a	rate		time ^a	rate		time ^a		
	$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		0.0012± 0.0001			3348	0.020± 0.001		380	0.32± 0.01		36		
PSCCM	$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ	
E_{min}^c																
10					0.0014± 0.0003	0.0072	0.19	143 (132)	0.017± 0.001	0.06	0.30	105 (92)	0.29± 0.02	0.56	0.52	59 (47)
15					0.0011± 0.0002	0.014	0.078	150 (130)	0.017± 0.001	0.10	0.16	106 (83)	0.33± 0.02	0.86	0.38	45 (33)
20	0.00016± 0.00006	0.010	0.016	175 (152)	0.0012± 0.0002	0.037	0.033	160 (125)	0.017± 0.001	0.19	0.088	93 (78)	0.35± 0.02	1.15	0.30	59 (31)
25	0.00012± 0.00003	0.022	0.0055	181 (135)	0.0011± 0.0002	0.081	0.013	144 (113)	0.018± 0.001	0.39	0.048	67 (47)	0.38± 0.03	1.82	0.21	31 (18)
30	0.00014± 0.00002	0.077	0.0018	142 (100)	0.0012± 0.0002	0.22	0.0053	126 (70)	0.020± 0.001	0.76	0.026	59 (31)	0.40± 0.05	2.67	0.15	26 (14)
35	0.00014± 0.00004	0.25	0.00055	195 (92)	0.0012± 0.0001	0.55	0.0022	132 (51)	0.019± 0.002	1.34	0.014	37 (16)	0.55± 0.09	5.04	0.11	38 (6)
c. Implementation 3																
traj	E = 70 kcal/mol				E = 80 kcal/mol				E = 100 kcal/mol			E = 150 kcal/mol				
	rate			time ^a	rate			time ^a	rate		time ^a	rate		time ^a		
	$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		0.0012± 0.0001			3348	0.020± 0.001		380	0.32± 0.01		36		
PSCCM	$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ		$k^{\text{traj}}(E)$	$k_1^{\text{traj}}(E)$	Γ_1/Γ	
E_{min}^b																
10					0.0011± 0.0002	0.0074	0.15	138 (128)	0.020± 0.001	0.078	0.25	100 (88)	0.41± 0.02	0.88	0.46	43 (32)
15					0.0011± 0.0002	0.016	0.068	145 (127)	0.021± 0.001	0.15	0.14	104 (81)	0.38± 0.03	1.08	0.35	42 (30)
20	0.00016± 0.00004	0.0098	0.016	170 (148)	0.0012± 0.0002	0.041	0.030	154 (120)	0.020± 0.001	0.24	0.080	92 (77)	0.34± 0.02	1.30	0.26	39 (25)
25	0.00012± 0.00003	0.022	0.0055	180 (132)	0.0013± 0.0002	0.11	0.012	136 (109)	0.020± 0.001	0.44	0.045	63 (45)	0.34± 0.04	1.70	0.20	25 (13)
30	0.00013± 0.00002	0.073	0.0018	138 (99)	0.0012± 0.0001	0.23	0.0051	99 (59)	0.020± 0.001	0.80	0.025	55 (28)	0.36± 0.04	2.50	0.14	24 (15)
35	0.00012± 0.00004	0.23	0.00053	167 (84)	0.0012± 0.0001	0.55	0.0021	124 (53)	0.019± 0.002	1.38	0.014	36 (16)	0.49± 0.08	5.04	0.098	17 (7)

^a Total CPU time (in min). In the PSCCM calculations the time is the sum of CPU times for the calculation of $k_1^{\text{traj}}(E)$ (in parentheses) and Γ_1/Γ . ^b E_{min} in kcal/mol. ^c R_{min} in Å

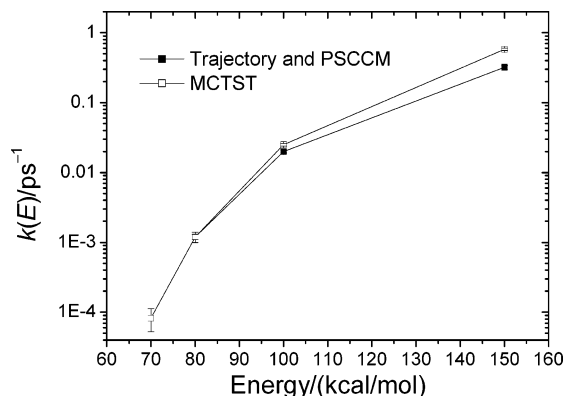


Figure 3. Rate constants $k(E)$ calculated by standard straightforward trajectory simulation (black square) [PSCCM (method 3 $E_{\min}=30$ kcal/mol) is indistinguishable from trajectory simulation] and the statistical Monte Carlo Transition State Theory (open square). Dynamical PSCCM calculation shows that nonstatistical effects are absent at low energies (70 and 80 kcal/mol), thereby confirming the predictions of the Intramolecular Dynamics Diffusional Theory.

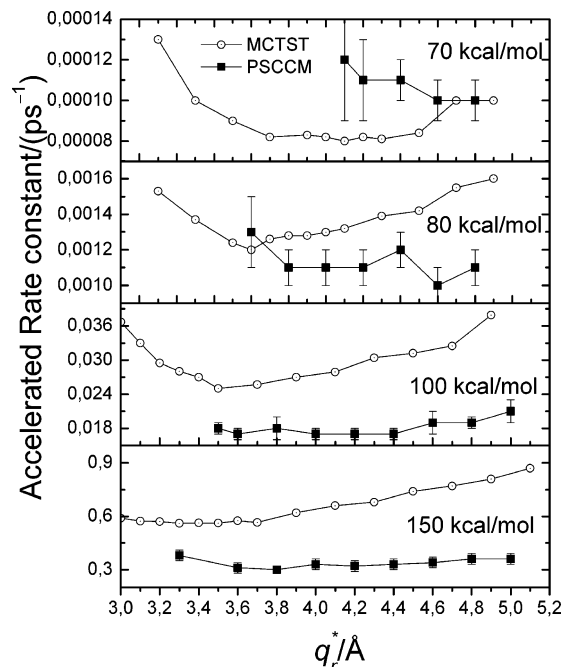


Figure 4. Variations of the rate constants obtained by MCTST (open circles) and PSCCM (black squares) with the position of the dividing surface. The actual position minimizes MCTST rate. The PSCCM rate is less sensitive to the position of the dividing surface.

defined as easily as in the current case of simple bond fission. PSCCM might be able to produce correct results even without a careful choice of the dividing surface. Of course, some general principles of defining constraints should be met. First, the allowed phase space volume should be far enough from the dividing surface to include all nonstatistical effects. On the other hand, the allowed phase space should be made as small as possible in order to make PSCCM numerically efficient.

4. Summary and Conclusions

In this article we propose a new version of accelerated dynamics based on phase space constraints rather than previously suggested boosting potentials.^{4–8} We demonstrated that the approach speeds up trajectory simulations with no loss of accuracy. For example as Table 3a–d shows, for $E = 80$ kcal/mol the CPU time required by PSCCM is more than an order of magnitude smaller than that of straightforward trajectory simulation. For $E = 70$ kcal/mol straightforward trajectory simulation was not feasible, but simple extrapolation of CPU time suggests acceleration by more than 2 orders of magnitude.

The method of PSCCM is based on the Intramolecular Dynamics Diffusional Theory. On the other hand, accelerated PSCCM helped to demonstrate that nonstatistical effects at lower energies become less important so that the dynamical simulation simply reproduces the statistical MCTST rate (see Figure 5). Previously this prediction of IDDT could not be directly verified without accelerating the dynamics. Therefore PSCCM fills the gap between trajectory simulations at high energies and MCTST at low energies. Like any other version of accelerated dynamics Phase Space Constrained Classical Mechanics should be used for energies low enough for straightforward trajectory simulations to be time-consuming but high enough for nonstatistical effects to be important for IVR limited intrinsic non-RRKM reactions. At low energies rates are statistical and therefore can (and should) be easily estimated by the accelerated Monte Carlo Transition State Theory.²² However, at low energies PSCCM is still useful since, as it is shown at Figure 4, PSCCM is less sensitive to variations of the dividing surface. Therefore, PSCCM, which could be even faster than accelerated MCTST at low energies, can be used when the dividing surface is difficult to define.

Acknowledgment. We thank “Centro de Supercomputación de Galicia” (CESGA) for the use of their facilities. E.M.-N. acknowledges Ministerio de Ciencia y Tecnología of Spain for financial support through “Ramón y Cajal” program. D.S. acknowledges the support from UK EPSRC. We also would like to thank Prof. M. S. Child for his useful comments.

Appendix

Some Details of the Calculations. The trajectories were integrated using the Runge–Kutta algorithm with a fixed step size of 0.05 fs, which is enough to ensure an excellent energy conservation: the maximum energy difference along the trajectories is $3 \times 10^{-5}\%$. For the highest energies considered in this study (100 and 150 kcal/mol), the trajectories were followed until R_{ON} reaches the dividing surface (see above) or 5 ps elapsed. For the lowest energies (70 and 80 kcal) the maximum time for a trajectory was 20 ps. Ensembles of 1000 trajectories were employed in all cases (for standard and PSCCM calculations) except for the standard trajectory calculations at 80 kcal/mol, for which we needed 3000 trajectories to get the same statistics as in the accelerated computations. Two different types of initial conditions were used in the present study for standard and

PSCCM calculations: efficient microcanonical sampling (EMS)¹¹ and Metropolis sampling.¹⁹ The microcanonical rate constants obtained from standard trajectory simulations with both initialization methods (EMS and Metropolis) are very similar. For both EMS and Metropolis we used warm-up random walks of 500 000 steps and walks of 10 000 steps between trajectories. Additionally, the maximum displacement for the atoms in each step of the random walk in the EMS was 0.07 Å. For the Metropolis sampling the maximum displacements for the Cartesian coordinates and momenta were both 0.1 (in Å and (kcal/mol×amu)^{1/2} respectively). In Metropolis sampling the probability for acceptance/rejection of a given point along the walk is given by

$$P = \begin{cases} \epsilon/\{\epsilon^2 + [E - H(q,p)]\} & \text{for } |E - H| < E_{\text{limit}} \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

In the present work we used $\epsilon = 10$ and $E_{\text{limit}} = 10$ kcal/mol. Acceptance/rejection ratios close to 0.5 were achieved in all cases. The same random walk was employed both for choosing the initial conditions of trajectory simulation and for calculating the statistical MCTST rates.

References

- (1) Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay S.; Klein M. L. *J. Phys. Chem B* **2001**, *105*, 4464.
- (2) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. *J. Chem. Phys.* **2003**, *119*, 7043.
- (3) Chandler, D. *J. Chem. Phys.* **1978**, *68*, 2959. Montgomery, J. A., Jr.; Chandler, D.; Berne, B. J. *J. Chem. Phys.* **1979**, *70*, 4056.
- (4) Voter, A. F. *Phys. Rev. B* **1998**, *57*, 13985.
- (5) Voter, A. F. *J. Chem. Phys.* **1997**, *106*, 4665.
- (6) Voter, A. F. *Phys. Rev. Lett.* **1997**, *78*, 3908.
- (7) Montalenti, F.; Sørensen, M. R.; Voter, A. F. *Phys. Rev. Lett.* **2001**, *87*, 126101.
- (8) Voter, A. F.; Montalenti, F.; Germann, T. G. *Annu. Rev. Mater. Res.* **2002**, *32*, 3219.
- (9) Grimmelmann, E. K.; Tully, J. C.; Helfand, E. *J. Chem. Phys.* **1981**, *74*, 5300.
- (10) (a) Brady, J. W.; Doll, J. D.; Thompson, D. L. *J. Chem. Phys.* **1980**, *73*, 2767. (b) Brady, J. W.; Doll, J. D.; Thompson, D. L. *J. Chem. Phys.* **1981**, *74*, 1026–1028. (c) Viswanathan, R.; Thompson, D. L.; Raff, L. M. *J. Chem. Phys.* **1984**, *80*, 4230. (d) NoorBatcha, I.; Raff, L. M.; Thompson, D. L. *J. Chem. Phys.* **1986**, *84*, 4341. (e) Rice, B. M.; Raff, L. M.; Thompson, D. L. *J. Chem. Phys.* **1986**, *85*, 4392. (f) Gai, H.; Thompson, D. L.; Raff, L. M. *J. Chem. Phys.* **1988**, *88*, 156. (g) Agrawal, P. M.; Thompson, D. L.; Raff, L. M. *J. Chem. Phys.* **1988**, *89*, 741. (h) Agrawal, P. M.; Thompson, D. L.; Raff, L. M. *J. Chem. Phys.* **1990**, *92*, 1069. (i) Sewell, T. D.; Thompson, D. L. *J. Chem. Phys.* **1990**, *93*, 4077. (j) Schranz, H. W.; Raff, L. M.; Thompson, D. L. *J. Chem. Phys.* **1991**, *94*, 4219. (k) Schranz, H. W.; Raff, L. M.; Thompson, D. L. *Chem. Phys. Lett.* **1991**, *182*, 455. (l) Sewell, T. D.; Schranz, H. W.; Thompson, D. L.; Raff, L. M. *J. Chem. Phys.* **1991**, *95*, 8089. (m) Sorescu, D. C.; Thompson, D. L.; Raff, L. M. *J. Chem. Phys.* **1994**, *101*, 3729. (n) Rice, B. M.; Adams, G. F.; Page, M.; Thompson, D. L. *J. Phys. Chem.* **1995**, *99*, 5016. (o) Chambers, C. C.; Thompson, D. L. *J. Phys. Chem.* **1995**, *99*, 15881.
- (11) Sewell, T. D.; Thompson, D. L. *Int. J. Mod. Phys. B* **1997**, *11*, 1067.
- (12) Sørensen, M. R.; Voter, A. F. *J. Chem. Phys.* **2000**, *112*, 9599.
- (13) Shalashilin, D. V.; Thompson, D. L. *J. Chem. Phys.* **1996**, *105*, 1833.
- (14) Guo, Y.; Shalashilin, D. V.; Krouse, J. A.; Thompson, D. L. *J. Chem. Phys.* **1999**, *110*, 5514; **1999**, *110*, 5521.
- (15) Shalashilin, D. V.; Thompson, D. L. *J. Chem. Phys.* **1997**, *107*, 6204.
- (16) Shalashilin, D. V.; Thompson, D. L. In *Highly Excited Molecules*; ACS Symposium Series; American Chemical Society: Washington, DC, 1997; Vol. 678, p 81.
- (17) Gray, S. K.; Rice S. A.; Davis, M. J. *J. Phys. Chem* **1986**, *90*, 3470.
- (18) Doll, J. D. *J. Chem. Phys.* **1981**, *74*, 1074.
- (19) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. J. *J. Chem. Phys.* **1953**, *21*, 1087.
- (20) Bunker, D. L.; Hase W. L. *J. Chem. Phys.* **1973**, *59*, 4621.
- (21) Hase, W. L. *J. Chem. Phys.* **1978**, *69*, 4711.
- (22) Shalashilin, D. V.; Thompson, D. L. *J. Phys. Chem. A* **1997**, *101*, 961.
- (23) Bowman, J. M.; Gazdy, B.; Sun, O. *J. Chem. Phys.* **1989**, *91*, 2859.
- (24) Miller, W. H.; Hase, W. L.; Darling, C. L. *J. Chem. Phys.* **1989**, *91*, 2863.
- (25) Sewell, T. D.; Thompson, D. L.; Gezelter, J. D.; Miller, W. H. *Chem. Phys. Lett.* **1992**, *193*, 512.
- (26) Peshherbe, G. H.; Hase, W. L. *J. Chem. Phys.* **1994**, *100*, 1179.
- (27) Martínez-Núñez, E.; Vázquez, S. A. *J. Chem. Phys.* **1998**, *109*, 8907.

CT060042Z

Molecular Dynamics Simulation of Argon, Krypton, and Xenon Using Two-Body and Three-Body Intermolecular Potentials

Elaheh K. Goharshadi* and Mohsen Abbaspour

Department of Chemistry, Ferdowsi University, Mashhad 91779, Iran

Received February 2, 2006

Abstract: We have performed the molecular dynamics simulation to obtain energy and pressure of argon, krypton, and xenon at different temperatures using a HFD-like potential which has been obtained with an inversion of viscosity data at zero pressure. The contribution of three-body dispersion resulting from third-order triple-dipole interactions has been computed using an accurate simple relation between two-body and three-body interactions developed by Marcelli and Sadus. Our results indicate that this simple three-body potential which was originally used in conjunction with the BFW potential is also valid when used with the HFD-like potential. This appears to support the conjecture that the relationship is independent of the two-body potential. The energy and pressure obtained are in good overall agreement with the experiment, especially for argon. A comparison of our simulated results with HMSA and ODS integral equations and a molecular simulation have been also included.

1. Introduction

It is well established that the physical properties of fluids are governed overwhelmingly by interactions involving pairs of molecules. However, the pair-potentials alone are insufficient for qualitatively accurate calculations. To obtain qualitative agreement with experiment, pair-potentials must be used in conjunction with three-body interactions.^{1–5}

The practical applications of the three-body interactions are well demonstrated in prediction of the phase-transition of not only pure substances^{2–5} but also the three-component mixtures.^{6,7} Three-body interactions joined with the mixing rules and an empirical equation of state leads to performance in phase behavior calculations of mixtures.⁷

The important three-body effects have previously remained undetected because earlier works were confined to effective potentials such as Lennard-Jones potential. Even, when regarded simply as effective potentials, the capacity of the pair-potential to reproduce known behavior has its limitation.

The knowledge of interactions in noble gases remains a fundamental question that is not completely solved. Despite the simplicity of their closed-shell electronic structure, it is well-known that a simple pair-potential, through giving the

essential features of the structural and thermodynamic properties, is not sufficient for a quantitative description, and many-body effects have to be taken into account.⁸

Molecular simulation is an ideal tool to investigate the role of intermolecular interactions, because, unlike conventional theoretical methods, the contributions from intermolecular potentials can be evaluated rigorously.⁹

Calculations of three-body interactions typically only consider contributions from the Axilrod–Teller¹⁰ (AT) term. The Axilrod–Teller term only accounts for triple-dipole interactions, whereas other three-body interactions arising from high multipoles are possible.

Marcelli and Sadus³ have shown that vapor–liquid equilibria of argon, krypton, and xenon are affected substantially by three-body interactions. They reported good results for the prediction of the vapor–liquid equilibria of argon, krypton, and xenon using two-body potentials such as the BFW potential¹¹ plus three-body contributions.

Recently, Jakse et al.⁸ performed molecular dynamics simulation to predict thermodynamic properties of liquid krypton using Aziz and Slaman^{12,13} plus the triple-dipole Axilrod–Teller potential. It has been shown that the AT potential gives an overall good description of liquid krypton, though other contributions such as higher order three-body dispersion and exchange terms cannot be ignored.

* Corresponding author phone: ++989155021947; e-mail: gohari@ferdowsi.um.ac.ir.

Table 1. Summary of the Intermolecular Potential Parameters Used in This Work

	argon	krypton	xenon
$(\epsilon/k)/K$	143.224	201.2	282.29
$\sigma/\text{\AA}$	3.3527	3.5709	3.8924
A	99744.4	543237.0	5.54437
α	11.9196	11.0068	-20.1659
β	-2.371328	-3.85189	-24.9602
C_6^*	0.651991	1.57171	7.76621
C_8^*	3.68594	0.580741	-33.5169
C_{10}^*	-2.99307	-0.786392	50.6382
D	1.36	1.37	2.78
$\nu/(\text{a.u.})$	518.3 ^a	1572.0 ^a	5573.0 ^a

^a Reference 19.

Bomont and Bretonent¹⁴ obtained the structural and thermodynamic properties of xenon at supercritical temperature and low densities with the use of an integral equation conjugated with an effective pair potential consisting of the Aziz–Slaman^{12,13} two-body potential plus the Axilrod–Teller three-body potential.

The aim of this work is to perform molecular dynamics simulation to obtain internal energy and pressure of argon, krypton, and xenon at different temperatures and densities using a HFD-like potential¹⁵ which has been obtained with an inversion of viscosity data and a simple and accurate expression for computing the three-body dispersion interactions.

2. Theory

2.1. Intermolecular Potential. The prediction of structural and thermodynamic properties of dense fluids requires an accurate knowledge of the intermolecular potential.¹⁴

In this work, a HFD-like potential¹⁵ which has been obtained from the inversion of the viscosity collision integrals at zero pressure has been used for the pair-interaction potential of argon, krypton, and xenon. It has the following functional form

$$V_2^*(x) = A \exp(-\alpha x + \beta x^2) - f(x) \left(\frac{C_6^*}{x^6} + \frac{C_8^*}{x^8} + \frac{C_{10}^*}{x^{10}} \right) \quad (1)$$

where $f(x)$ is in the form

$$f(x) = \exp\left[-\left(\frac{D}{x} - 1\right)^2\right] \quad x < D \quad (2)$$

$$f(x) = 1 \quad x \geq D \quad (3)$$

where $x = r/\sigma$ and $V_2^* = V_2/\epsilon$ (σ is the distance at which the intermolecular potential has zero value and ϵ is the well depth of potential). The values of parameters σ , ϵ , A , α , β , C_6^* , C_8^* , C_{10}^* , and D have been given in Table 1.

Marcelli and Sadus¹⁶ showed there is a simple and accurate relationship between the two-body (U_2) and three-body (U_3) energies of a fluid

$$U_3 = -\frac{2\nu\rho U_2}{3\epsilon\sigma^6} \quad (4)$$

where ν is the nonadditive coefficient¹⁷ (Table 1), ϵ is the characteristic depth of the pair-potential, σ is the characteristic depth of the pair-potential, and $\rho = N/V$ is the number density obtained by dividing the number of molecules (N) by the volume (V). The significance of this relationship is that it allows us to obtain an accurate overall intermolecular

Table 2. Results of the Reduced Two-Body and Total Pressure of Argon Obtained with the Different Methods

T^*	ρ^*	P_{exp}^* ²⁰	P_2^*				P_t^*			
			our work	HMSA ²¹	MC ²²	MC ³	our work	HMSA ²¹	MC ²²	MC ³
0.74409	0.73684	0.13062	0.48819			-0.90062	0.57430			-0.53111
0.81850	0.68025	0.04951	0.24982			-0.84150	0.31114			-0.57447
0.83645	0.03262	0.02226	0.02307		0.02800		0.02264		0.46000	
0.84330	0.66634	0.06308	0.23647			-0.81293	0.31923			-0.56658
0.86810	0.65344	0.08202	0.25363			-0.79716	0.27817			-0.56560
0.87827	0.03825	0.02703	0.02765		0.03400		0.02777		0.41100	
0.89291	0.63458	0.08052	0.21168			-0.77647	0.27613			-0.56166
0.91771	0.60873	0.06577	0.16400			-0.73213	0.21702			-0.54491
0.92010	0.59674	0.04373	0.12472		0.05500		0.16301		0.38100	
0.94251	0.59583	0.08933	0.17640			-0.70749	0.21288			-0.53505
0.96192	0.08430	0.05327	0.05607		0.06700		0.05561		0.33700	
0.96731	0.56010	0.07196	0.11881			-0.66315	0.14804			-0.51633
0.99212	0.50942	0.06592	0.07685			-0.58235	0.10680			-0.46706
1.00374	0.10229	0.06440	0.06655		0.08100		0.06599		0.30100	
1.04556	0.12014	0.07633	0.07952		0.09600		0.07807		0.24600	
1.33830	0.85389	0.36709	0.37090	4.36880			0.34751	5.28305		
	0.56926	0.24916	0.25110	0.58050			0.23510	0.80223		
	0.28463	0.13504	0.13660	0.21100			0.12927	0.23541		
1.67290	0.85389	0.56680	0.57440	5.95390			0.53351	6.91950		
	0.56926	0.35948	0.36330	1.18850	0.33541	1.39990				
	0.28463	0.18259	0.18340	0.37470	0.17478	0.39807				

Table 3. Results of the Reduced Two-Body and Total Pressure of Krypton Obtained with the Different Methods

T^*	ρ^*	P_{exp}^{*20}	P_2^*		P_t^*		
			our work	MC ³	our work	MD ⁸	MC ³
0.84000	0.66340	0.10119	-0.53140		-0.44205	0.04248	
	0.64750	0.03296	-0.53300		-0.47418	-0.03526	
0.98910	0.55100	0.11988	-0.23940		-0.19387	0.08134	
	0.53090	0.09135	-0.23290		-0.18868	0.06363	
	0.51500	0.07593	-0.21160		-0.17064	0.04248	
0.75261	0.71073	0.04991	-0.72324	-0.90058	-0.62544		-0.50989
0.82787	0.66980	0.08258	-0.57236	-0.84949	-0.46968		-0.54295
0.85296	0.64087	0.03838	-0.48708	-0.80841	-0.43467		-0.53494
0.87804	0.62988	0.06766	-0.45756	-0.78537	-0.38750		-0.52993
0.90313	0.61489	0.08398	-0.40180	-0.75933	-0.34952		-0.52592
0.92822	0.58397	0.06523	-0.37392	-0.70423	-0.29939		-0.50188
0.95331	0.52707	0.05404	-0.25748	-0.61708	-0.23984		-0.45980
0.97839	0.50809	0.06300	-0.21648	-0.57400	-0.14064		-0.43576

Table 4. Results of the Reduced Two-Body and Total Pressure of Xenon Obtained with the Different Methods

T^*	ρ^*	P_{exp}^{*20}	P_2^*			P_t^*			
			our work	ODS ¹⁴	MC ³	our work	ODS ¹⁴	MC ³	
0.74657	0.70725	0.02117	-1.16729		-0.94452	-0.25177		-0.50168	
0.82123	0.67220	0.08986	-0.94928		-0.87271	-0.17600		-0.50966	
0.84612	0.63514	0.02745	-0.87358		-0.82583	-0.22573		-0.51266	
0.87100	0.61810	0.03380	-0.78728		-0.77696	-0.21519		-0.48972	
0.89589	0.60006	0.04757	-0.76306		-0.74904	-0.15933		-0.48972	
0.92077	0.57904	0.05781	-0.66313		-0.69418	-0.18678		-0.45979	
0.94566	0.51792	0.05464	-0.49508		-0.60940	-0.15966		-0.43187	
0.97054	0.51191	0.06372	-0.46026		-0.59444	-0.16077		-0.42588	
1.05210	0.01000	0.00900	0.01010	0.00100		0.01004	0.01030		
	0.02000	0.01930	0.01910	0.01000		0.01904	0.01970		
	0.03000	0.02801	0.02710	0.02000		0.02687	0.02816		
	0.04000	0.03634	0.03410	0.03000		0.03363	0.03624		
	0.05000	0.04353	0.04010	0.03300		0.03978	0.04367		
	0.06000	0.05034	0.04530	0.04000		0.04517	0.05051		
	0.07000	0.05640	0.04970	0.05000		0.04858	0.05672		
	0.08000	0.06170	0.05190	0.05500		0.05171	0.06144		
	0.09000	0.06662	0.05560	0.05900		0.05467	0.06734		
	0.10000	0.07116	0.05770	0.06300		0.05629	0.07155		
	1.23990	0.10000	0.09538	0.07990	0.04000		0.07833	0.09558	
		0.20000	0.15352	0.08690	0.05000		0.08408	0.15676	
		0.30000	0.20098	0.03300	0.07000		0.03795	0.21107	
0.40000		0.26798	-0.04650	0.08000		-0.02441	0.28603		
0.50000		0.42014	-0.10280	0.10000		-0.06368	0.46271		
0.60000		0.79485	-0.03360	0.28000		0.36258	0.88335		
0.70000		1.65480	0.35880	0.50000		0.48781	1.91646		
1.48780	0.80000	3.47387	1.46250			1.64453			
	0.90000		3.73960			3.97214			
	1.00000		7.96360			8.30149			
	0.10000	0.12718	0.11020	0.13000		0.11015	0.13305		
	0.20000	0.22710	0.16050	0.14000		0.15539	0.25051		
	0.30000	0.33157	0.17870	0.15000		0.16600	0.35789		
	0.40000	0.48145	0.17870	0.24000		0.17266	0.54145		
	0.50000	0.75700	0.23920	0.37000		0.24185	0.87643		
	0.60000	1.31415	0.47390	1.00000		0.51115	1.54455		
	0.70000	2.43603	1.12640	2.00000		1.20065	2.76233		
0.80000		2.54350			2.64981				
0.90000		5.17790			5.43715				
1.00000		9.73500			10.10352				

Table 5. Results of the Reduced Two-Body and Total Energy of Argon Obtained with the Different Methods

T^*	ρ^*	U_{exp}^* ²⁰	U_2^*				U_t^*			
			our work	HMSA ²¹	MC ²²	MC ³	our work	HMSA ²¹	MC ²²	MC ³
0.74409	0.73684	-3.30527	-3.35367			-4.69269	-3.23940			-4.52502
0.81850	0.68025	-2.81756	-2.86545			-4.29584	-2.77531			-4.16488
0.83645	0.03262	0.95980	1.00584		-4.09025		1.00432		-3.91878	
0.84330	0.66634	-2.68473	-2.73035			-4.12718	-2.64622			-4.00317
0.86810	0.65344	-2.55862	-2.59525			-4.02797	-2.51683			-3.90991
0.87827	0.03825	0.98626	1.02802		-3.88950		1.02620		-3.73392	
0.89291	0.63458	-2.40051	-2.43705			-3.95852	-2.36554			-3.84642
0.91771	0.60873	-2.20584	-2.23684			-3.79979	-2.17387			-3.69760
0.92010	0.59674	-2.13602	-2.16291		-3.69712		-2.10322		-3.55576	
0.94251	0.59583	-2.08309	-2.10637			-3.68073	-2.04833			-3.58450
0.96192	0.08430	0.76634	0.83004		-3.52146		0.82680		-3.40018	
0.96731	0.56010	-1.84071	-1.84373			-3.46247	-1.79598			-3.37516
0.99212	0.50942	-1.56732	-1.52691			-3.16484	-1.49094			-3.08943
1.00374	0.10229	0.71840	0.76757		-3.21198		0.76394		-3.10407	
1.04556	0.12014	0.68109	0.74673		-2.91085		0.74258		-2.83223	
1.33830	0.85389	0.78712	0.80600	-2.71680			0.77417	-1.81477		
	0.56926	1.29048	1.31960	-1.26204			1.28486	-0.91675		
	0.28463	1.82797	1.84520	0.27971			1.82091	0.39347		
1.67290	0.85389	1.47807	1.49490	-1.90376			1.43587	-0.98199		
	0.56926	1.96802	1.98670	-0.61563			1.93440	-0.27268		
	0.28463	2.46931	2.48960	0.90337			2.45683	0.99705		

Table 6. Results of the Reduced Two-Body and Total Energy of Krypton Obtained with the Different Methods

T^*	ρ^*	U_{exp}^* ²⁰	U_2^*			U_t^*		
			our work	MD ⁸	MC ³	our work	MD ⁸	MC ³
0.84000	0.66340	1.26558	-3.18900	-4.12700		-3.06925	-3.93800	
	0.64750	1.35636	-3.10210	-4.03400		-2.98840	-3.85500	
0.98910	0.55100	1.24387	-2.17200	-3.35800		-2.10426	-3.23200	
	0.53090	1.27754	-2.05070	-3.24500		-1.98907	-3.12700	
	0.51500	1.29691	-1.96240	-3.16500		-1.90519	-3.05300	
0.75261	0.71073	0.81780	-3.70726		-4.50562	-3.55811		-4.32198
0.82787	0.66980	1.21665	-3.26815		-4.09419	-3.14424		-3.94167
0.85296	0.64087	1.43220	-3.03006		-3.98381	-2.92014		-3.84232
0.87804	0.62988	1.54769	-2.90028		-3.89350	-2.79687		-3.75903
0.90313	0.61489	1.68417	-2.74915		-3.76304	-2.65346		-3.63661
0.92822	0.58397	1.90491	-2.50161		-3.56235	-2.41892		-3.44695
0.95331	0.52707	2.15395	-2.16448		-3.24124	-2.09990		-3.14290
0.97839	0.50809	2.36644	-4.96536		-3.11079	-4.82255		-3.02047

potential (V) solely in terms of pair contributions (V_2) and well-known intermolecular parameters:

$$V = V_2 \left(1 - \frac{2\nu\rho}{3\epsilon\sigma^6} \right) \quad (5)$$

Therefore, the effect of three-body interactions can be incorporated into a simulation involving pair-interactions without any additional computational cost.⁴ Comparison of this approach with a full two-body plus three-body calculation indicates that there is no significant loss of accuracy.¹⁶ In this work, we have used this equation in our simulations.

2.2. Simulation Details. The molecular dynamics simulations for 1000 atoms of argon, krypton, and xenon have been performed. The simulations were performed in cubic boxes, and the conventional periodic boundary conditions were applied. The NVT ensemble was implemented using a Nose-

Hoover thermostat for the systems of argon, krypton, and xenon interacting via the two-body HFD-like (eqs 1–3) and then via the overall intermolecular potential (eq 5). Before including the three-body interactions, simulations were performed with the two-body part of the potential only for the same thermodynamic states as those intended for the overall intermolecular potential. The number of time steps, n_t , size of time steps, Δt^* , and the cutoff radius, r_c , have been chosen as 5000, 0.001, and 2.5σ , respectively. The long-range correction terms were evaluated to recover the contribution to the pressure and energy for the intermolecular potential.

3. Results and Discussion

We have used the HFD-like potential (eqs 1–3) in the MD simulations for two-body intermolecular potentials of argon,

Table 7. Results of the Reduced Two-Body and Total Energy of Xenon Obtained with the Different Methods

T^*	ρ^*	U_{exp}^{*20}	U_2^*			U_t^*		
			our work	ODS ¹⁴	MC ³	our work	ODS ¹⁴	MC ³
0.74657	0.70725	0.87782	-2.14716		-4.45953	-2.04718		-4.25148
0.82123	0.67220	1.24190	-1.92124		-4.08126	-1.83621		-3.90109
0.84612	0.63514	1.48500	-1.76169		-3.91204	-1.68802		-3.74779
0.87100	0.61810	1.64544	-1.67746		-3.78263	-1.60919		-3.62834
0.89589	0.60006	1.79626	-1.58138		-3.62337	-1.51890		-3.47903
0.92077	0.57904	1.96118	-1.48350		-3.51387	-1.42694		-3.37949
0.94566	0.51792	2.19060	-1.24671		-3.15551	-1.20420		-3.04203
0.97054	0.51191	2.40642	-1.21201		-3.11570	-1.17116		-3.00719
1.05210	0.01000	5.44816	1.50190	-0.36650		1.50091	-0.46651	
	0.02000	5.37399	1.42320	-0.42250		1.42133	-0.52249	
	0.03000	5.30024	1.33230	-0.05000		1.32967	-0.14720	
	0.04000	5.22437	1.22860	-0.55000		1.22536	-0.65047	
	0.05000	5.15275	1.14650	-0.61000		1.14273	-0.71844	
	0.06000	5.07943	1.06350	-0.68000		1.05930	-0.78641	
	0.07000	5.00782	0.99730	-0.74000		0.99270	-0.84641	
	0.08000	4.93962	0.88370	-0.81000		0.87905	-0.91039	
	0.09000	4.87013	0.82360	-0.87000		0.81872	-0.97038	
	0.10000	4.79980	0.74890	-0.90000		0.74397	-0.99950	
1.23990	0.10000	5.15275	1.05940	-1.07000		1.05242	-1.07338	
	0.20000	4.56279	0.32560	-1.53000		0.32131	-1.57375	
	0.30000	4.03085	-0.42490	-2.00000		-0.41651	-2.04602	
	0.40000	3.53723	-1.06000	-2.30000		-1.03208	-2.49943	
	0.50000	3.03176	-1.75150	-2.83000		-1.69384	-2.93400	
	0.60000	2.51563	-2.43680	-3.20000		-2.34054	-3.38742	
	0.70000	2.01664	-3.15350	-3.64000		-3.00816	-3.84084	
	0.80000	1.58713	-3.81620	-3.90000		-3.61519	-4.18155	
	0.90000		-4.36070	-4.00000		-4.10230	-4.40014	
	1.00000		-4.67150	-3.90000		-4.36393	-4.39332	
1.48780	0.10000	5.57050	1.48780	-1.11000		1.47800	-1.11870	
	0.20000	5.02231	0.78660	-1.60000		0.77624	-1.60641	
	0.30000	4.50653	0.09110	-1.92000		0.08930	-2.02687	
	0.40000	4.00890	-0.58080	-2.35000		-0.56550	-2.48081	
	0.50000	3.51097	-1.26400	-2.70000		-1.22239	-2.90083	
	0.60000	3.01390	-1.95530	-3.10000		-1.87806	-3.32099	
	0.70000	2.54590	-2.62940	-3.40000		-2.50822	-3.70738	
	0.80000		-3.24110	-3.50000		-3.07038	-4.03723	
	0.90000		-3.72600	-3.60000		-3.50521	-4.19822	
	1.00000		-3.98720	-3.20000		-3.72468	-4.14406	

krypton, and xenon. The total (two-body plus three-body) contributions have been considered in the simulation using eq 5.

Our results of reduced pressure and energy for argon, krypton, and xenon in the NVT ensemble have been compared at different temperatures and densities with experimental and previous theoretical works in Tables 2–7. We have also considered the corrections to calculation of pressure using the total intermolecular potential (eq 5) proposed by Smit et al.¹⁸ The normal conventions have been adopted for the reduced density ($\rho^* = \rho\sigma^3$), reduced temperature ($T^* = kT/\epsilon$), reduced energy ($U^* = U/\epsilon$), and reduced pressure ($P^* = P\sigma^3/\epsilon$). In Tables 2–7, the subscripts 2 and t denote two-body, three-body, and two-body plus three-body contributions, respectively.

Three-body interactions based on the triple-dipole dispersion term of Axilrod and Teller contribute commonly 5–10% to the overall energy of the liquid phase. The data in Tables

5–7 indicate that the three-body interactions via the expression of Marcelli and Sadus¹⁶ contribute to the total energy of argon, krypton, and xenon 0.15–4.11%, 2.96–4.19%, and 0.13–7.05%, respectively.

As Tables 2–7 show the contribution of the three-body interaction on pressure and energy based on the Marcelli and Sadus expression is almost the same contribution as the three-body interaction using the Axilrod–Teller expression. A situation such as this has been obtained by Marcelli and Sadus.^{16,19} They performed the nonequilibrium molecular dynamics (NEMD) for argon and found that the calculations of energy and pressure (by concerning the Smit et al.¹⁸ corrections) using eq 5 were in good agreement with the two-body (BFW potential) + three-body (Axilrod–Teller potential) energy and pressure. They have also used the Smit et al.¹⁸ correction to calculation of pressure using eq 5.

As Table 2 shows there is a better accordance between our simulated values of two-body and total pressure of argon

and the experimental values²⁰ than other simulation and theories.^{3,21,22} This agreement may be mainly due to the two-body potential of argon used in our calculations because the agreement with the experiment for the two-body pressure is better than total pressure.

Our results of two-body and total pressure of argon are better than that of an integral equation theory (HMSA)²¹ which has used the two-body potential of Aziz and Slaman^{12,13} and the three-body potential of Axilrod-Teller. Our simulation is also better than those calculated using Monte Carlo (MC) simulations of Sadus and Prausnitz²² and Marcelli and Sadus³ which have used the two-body Lennard-Jones and BFW¹¹ potentials, respectively. These two preceding works have used the Axilrod-Teller expression for three-body simulations.

Table 3 shows that our calculated two-body pressures are larger than those obtained using MC simulations³ but are smaller than the experimental values.²⁰ The same situation occurs for the total pressure of krypton. It is shown that our results of total pressure have more accordance with experimental values than two-body pressure values, and this is due to considering the three-body contribution of Marcelli and Sadus in our calculation. We have also compared our results with the molecular dynamics (MD) simulation of Jakse et al.⁸ which have used the HFD potential of Aziz and Slaman^{12,13} in conjunction with the three-body interactions relation of Axilrod-Teller. The MD results of Jakse et al.⁸ have good agreement with the experiment but our results underestimate the experimental values, and it can be referred to as the kind of two-body potential of krypton which has been used in the calculations.

We have compared our calculated reduced two-body and the two-body plus three-body pressure of xenon with the experiment²⁰ in Table 4. The results of the MC simulation of Marcelli and Sadus³ and an integral equation theory (ODS)¹⁴ which has used the HFD potential of Aziz and Slaman^{12,13} in conjunction with the three-body interactions of Axilrod-Teller have been also considered for this comparison. It is clear that our results are in a fairly good agreement with the experiment. It is shown that the three-body interactions have affected the total pressure of xenon especially at higher densities. Our results are also better than those obtained using MC simulations, but the results of the ODS theory are better than our calculations at some points.

It is evident from Table 5 that there is a very good accordance between our simulated values of two-body and the total energy of argon and the experimental values.¹⁶ Our results are also better than those calculated using the HMSA theory²¹ and the MC simulations.^{3,22} The reason for our good results for argon is due to our two-body potential with the three-body term of Marcelli and Sadus used in the calculations because the three-body term has improved our results in this table.

We have compared our reduced two-body and the total energy of krypton and xenon with the experiment²⁰ and the other theories and simulations^{3,10,14} in Tables 6 and 7. It is obvious that our results underestimate the experimental values but much better than other previous works, and this

may be attributed to the use of more accurate pair-potential for these compounds.

4. Concluding Remarks

We have performed the molecular dynamics simulation to obtain energy and pressure of argon, krypton, and xenon at different temperatures and densities using a two-body HFD-like potential which has been obtained with an inversion of viscosity data at zero pressure, and the three-body interactions have been calculated using the Marcelli and Sadus expression.

The energy and pressure obtained are in good overall agreement with the experiment, especially for argon, and this can be due to the two-body potential used in this work. A comparison of our simulated results with the corresponding values obtained from HMSA and ODS integral equations and molecular simulation is also included.

Our results indicate that the simple three-body potential of Marcelli and Sadus which was originally used in conjunction with the BFW potential is also valid when used with the HFD-like potential. This appears to support the conjecture that the relationship is independent of the two-body potential.

References

- (1) Maitland, G. C.; Rigby, M.; Smith, E. B.; Wakeham, W. A. *Introduction. In Intermolecular Forces, Their Origin and Determination*; Clarendon Press: Oxford, 1981; pp 1-4.
- (2) Sadus, R. J. Molecular dynamics simulation. In *Molecular Simulation of Fluids: Theory, Algorithms and Object-Oriented*; Elsevier: Amsterdam, 1999; pp 67-68.
- (3) Marcelli, G.; Sadus, R. J. *J. Chem. Phys.* **1999**, *111*, 1533-1540.
- (4) Marcelli, G.; Todd, B. D.; Sadus, R. J. *J. Chem. Phys.* **2001**, *115*, 9410-9413.
- (5) Sadus, R. J. *Fluid Phase Equilib.* **1998**, *144*, 351-360.
- (6) Benmekki, E. H.; Mansoori, G. A. *Proceedings of the 1986 Annual SPE Convention and Proceedings of the Eastern Regional Meeting of SPE, Society of Petroleum Engineers*; Richardson, TX, 1986.
- (7) Benmekki, E. H.; Mansoori, G. A. *Fluid Phase Equilib.* **1988**, *41*, 43-57.
- (8) Jakse, N.; Bomont, J. M.; Bretonnet, J. L. *J. Chem. Phys.* **2002**, *116*, 8504-8508.
- (9) Marcelli, G.; Sadus, R. J. *High Temp. - High Pressures* **2001**, *33*, 111-118.
- (10) Axilrod, B. M.; Teller, E. *J. Chem. Phys.* **1943**, *11*, 299-300.
- (11) Barker, J. A.; Fisher, R. A.; Watts, R. O. *Mol. Phys.* **1971**, *21*, 657.
- (12) Aziz, R. A.; Slaman, M. J. *Mol. Phys.* **1985**, *57*, 827.
- (13) Aziz, R. A.; Slaman, M. J. *Mol. Phys.* **1986**, *58*, 679.
- (14) Bomont, J. M.; Bretonnet, J. L. *Phys. Rev. B* **2002**, *65*, 224203-1-224203-5.
- (15) Goharshadi, E. K.; Jami-Alahmadi, M.; Najafi, B. *Can. J. Chem.* **2003**, *81*, 1-6.
- (16) Marcelli, G.; Sadus, R. J. *J. Chem. Phys.* **2000**, *112*, 6382-6385.

- (17) Marcelli, G.; Todd, B. D.; Sadus, R. J. *J. Chem. Phys.* **2004**, *120*, 3043.
- (18) Smit, B.; Hauschild, T.; Prausnitz, J. M. *Mol. Phys.* **1992**, *77*, 1021.
- (19) Leonard, P. J.; Barker, J. A. *In Theoretical Chemistry: Advances and Perspectives*; Eyring, H., Henderson, D., Eds.; Academic Press: London, 1975; Vol. 1.
- (20) National Institute of Standards and Technology. <http://webbook.nist.gov/chemistry/fluid>.
- (21) Bomont, J. M.; Bretonnet, J. L. *J. Chem. Phys.* **2001**, *114*, 5674–5681.
- (22) Sadus, R. J.; Prausnitz, J. M. *J. Chem. Phys.* **1996**, *104*, 4784–4787.

CT060039F

JCTC

Journal of Chemical Theory and Computation

Elucidating the Conformational Dependence of Calculated pK_a Values

Dennis R. Livesay,^{*,†,‡} Donald J. Jacobs,^{||} Julie Kanjanapangka,[§] Eric Chea,[§]
Hector Cortez,[†] Jorge Garcia,[†] Patrick Kidd,[§] Mario Pulido Marquez,[†]
Swati Pande,[§] and David Yang[§]

Department of Chemistry, Center for Macromolecular Modeling & Materials Design, and Department of Biological Sciences, California State Polytechnic University, Pomona, California, and Department of Physics and Optical Science, University of North Carolina, Charlotte, North Carolina

Received February 16, 2006

Abstract: The variability within calculated protein residue pK_a values calculated using Poisson–Boltzmann continuum theory with respect to small conformational fluctuations is investigated. As a general rule, sites buried in the protein core have the largest pK_a fluctuations but the least amount of conformational variability; conversely, sites on the protein surface generally have large conformational fluctuations but very small pK_a fluctuations. These results occur because of the heterogeneous or uniform nature of the electrostatic microenvironments at the protein core or surface, respectively. Atypical surface sites with large pK_a fluctuations occur at the interfaces between significant anionic and cationic potentials.

Introduction

Understanding amino acid pK_a fluctuations is key to a deeper understanding of enzyme catalysis.¹ This is especially important considering the dynamic nature of enzyme catalytic site pK_a values. For example, the catalytic Glu169 of the glycolytic enzyme triosephosphate isomerase changes its protonation state four times along its reaction pathway,² thus necessitating a dynamic pK_a value. At the beginning of the reaction cycle, Glu169 must be deprotonated (i.e., a low pK_a value) in order for it to act as a general base. Next, the Glu169 pK_a must shift upward such that it can give up the proton to form the enediol intermediate. This protonation/deprotonation cycle is repeated in the second half of the mechanism, finally resulting in the formation of glyceraldehyde-3-phosphate. Previously,³ we have attributed the first pK_a shift to changes in the local electrostatic environment

upon substrate binding. Using Poisson–Boltzmann (PB) continuum theory (described below), the pK_a of Glu169 in the apo structure is calculated to be 0.77, ensuring a deprotonated carboxylate. However, the pK_a is shifted to 8.00 upon substrate binding, making protonation energetically feasible. There are likely two primary factors mediating the remaining three protonation changes. The first, and likely most important, is that changes in the local electrostatics due to the various mechanistic intermediates substantially alter the pK_a of the catalytic site.⁴ The second is local conformational changes within the enzyme active site.^{5,6} Conformational changes represent a simple way to modulate pK_a values. As a first step toward a computational methodology to probe these complicated acid/base effects, we report the sensitivity of calculated pK_a values to local fluctuations about a native structure.

PB continuum electrostatic theory has become ubiquitous within the computational biology community, see Fogolari et al.⁷ for a recent review. One common application of PB theory is in the calculation of pK_a values. There are several similar, yet distinct, PB algorithms for calculating pK_a values using continuum theory, for example, see refs 8–15. However, all are based on the original method of Tanford and Roxby,¹⁶ which assumes that the equilibrium between the acid and base is governed by an *intrinsic* pK_a , where $pK_{a,int}$ is equal to the pK_a if every other titratable site is

* Corresponding author tel.: (909) 869-4409; fax: (909) 869-4434; e-mail: drlivesay@csupomona.edu.

† Department of Chemistry, California State Polytechnic University.

‡ Center for Macromolecular Modeling & Materials Design, California State Polytechnic University.

§ Department of Biological Sciences, California State Polytechnic University.

|| University of North Carolina.

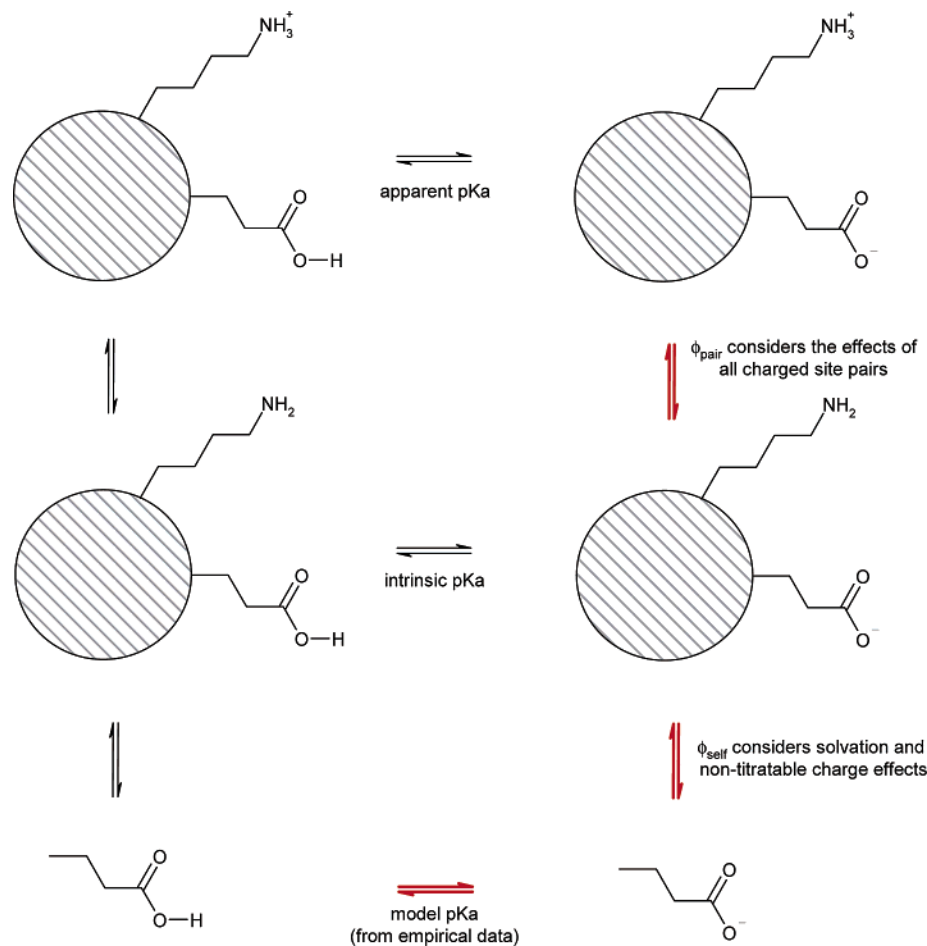


Figure 1. Schematic description of the pK_a calculation algorithm. The method is based on an energy cycle. An *intrinsic* pK_a , which is the hypothetical pK_a for a site if all other titratable sites are neutral, is calculated from the model pK_a by accounting for solvation effects and all nontitratable (partial) charge groups. The apparent pK_a , or real value, is calculated from the intrinsic value by accounting for all charge–charge pair interactions. The *ionic* pK_a is calculated from the model pK_a by ignoring the consequences of Φ_{self} . Despite the schematic shown above, the *apparent* pK_a is actually a mixture of the top two lines because a Boltzmann probability distribution is used to describe the ionization polynomial (see Methods section).

neutral. Common differences within the various pK_a calculation algorithms are related to how flexibility, H-bond networks, and dielectric constants are dealt with.¹⁷ The University of Houston Brownian Dynamics¹⁸ (UHBD) suite of programs calculates the $pK_{a,\text{int}}$ from the *model* pK_a , $pK_{a,\text{model}}$, which is the experimentally determined aqueous solution pK_a value of the amino acid side chain, by evaluating the effect of nontitratable partial charges and changes in solvation. Computation of the $pK_{a,\text{int}}$ requires calculation of the background potential, Φ_{self} , which models the effects of the above considerations. The *apparent* pK_a is calculated from the $pK_{a,\text{int}}$ after evaluating the effect of all charge–charge pairs. Each electrostatic potential between two charged sites is calculated by UHBD and is represented as Φ_{pair} . Figure 1 provides a schematic representation of the method; a more detailed description is provided in the Methods section. The approach implemented into UHBD uses a clustering algorithm to reduce the computational expense of evaluating all electrostatic pair potentials in order to compute the actual pK_a .^{10,12} The ionization polynomial is exactly solved within a titrating site cluster, whereas a mean-field approximation is used to treat intercluster interactions.

Calculated pK_a values are sensitive to a number of factors, including the chosen interior dielectric constant,¹¹ H-bond network,¹³ and the number of explicit water molecules included.¹⁹ Recently, several reports have focused on understanding the effects of slight conformational changes on calculated pK_a values. For example, a single torsion angle change in hen egg white lysozyme (HEWL) results in large pK_a differences of active site residues.²⁰ Nielsen and McCammon¹⁷ have investigated the conformational dependence of calculated pK_a values from 41 HEWL X-ray structures, focusing specifically on the ability to correctly identify proton donors and acceptors within two catalytic acids (Glu35 and Asp52). One intriguing conclusion from this work relates to the origins of the conformational dependence of the variability within these two positions. The variability within Glu35 is largely attributed to changes within the set of Φ_{pair} , whereas the variability within Asp52 is caused by changes within both Φ_{self} and Φ_{pair} . Similarly, Kumar and Nussinov have used continuum electrostatics to probe the stability of salt bridges from alternate NMR conformers.²¹ Their results indicate that stabilities of salt bridges vary considerably across the conformation ensemble. Moreover, most salt

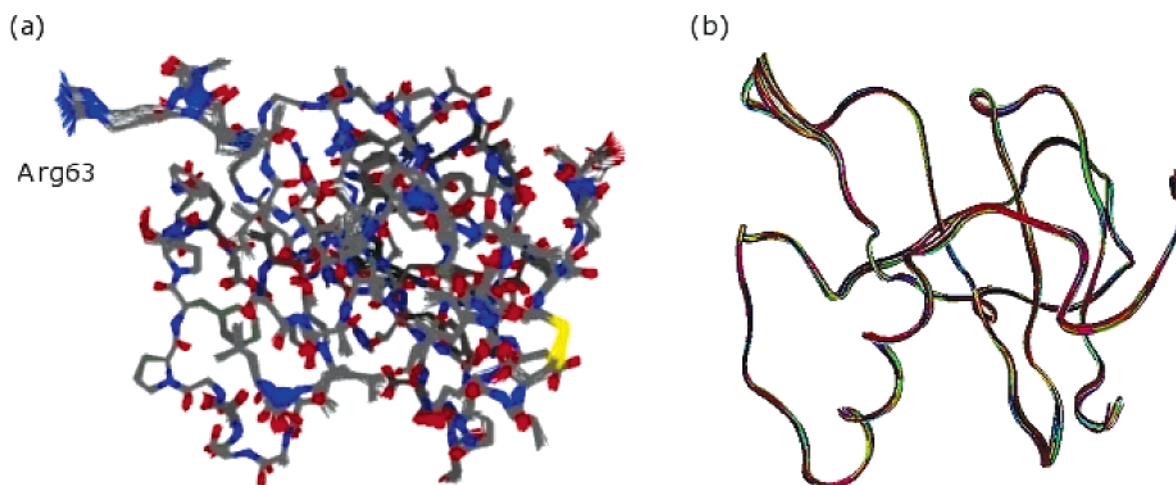


Figure 2. (a) Structural superposition of all RNase Sa conformers investigated. The all-atom RMSD is 0.42 Å. Arg63, which has the greatest structural fluctuations across the ensemble, is highlighted. (b) Backbone superposition, shown in the same orientation as that of part a, of all RNase Sa conformers investigated. The α -carbon RMSD is 0.31 Å.

bridge pairs vary between stabilizing and destabilizing at least once within the population. Changes in salt bridge stability arise because of changes in the location of the charged residues and their orientation (within the salt bridge pair and with respect to other charged sites).

In this report, molecular dynamic (MD) simulations are used to generate an ensemble of protein conformers for three test cases: ribonuclease Sa (RNase Sa) from *Streptomyces aureofaciens*; c-type lysozyme (LYS) from humans, and triosephosphate isomerase (TIM) from *Saccharomyces cerevisiae*. Subsequently, the pK_a values of all titratable sites are calculated and evaluated. Our results indicate that sites buried within the protein core are generally associated with increased pK_a variability. Moreover, we attempt to identify the exact molecular origins of the variability by scrutinizing electrostatic potentials, pK_a values using only Φ_{self} or Φ_{pair} , solvent accessibilities, and titratable site root-mean-square deviations (RMSDs). Finally, it is demonstrated that overall electrostatic free energies, G_{elec} , are generally insensitive to slight conformational changes, especially when compared against the variability within traditional force field potential energy calculations.

Results and Discussion

Variability within pK_a Values. RNase Sa is a small (96 residues) microbial enzyme whose residue pK_a values have been the focus of numerous experimental^{22–24} and combined (experimental and theoretical)²⁵ investigations. Interestingly, the enzyme has 12 acidic residues and only five basic residues. RNase Sa is an ideal starting point for this investigation because of its small size, the fact that several pK_a values of RNase Sa have been solved experimentally, and the fact that a crystal structure is available.²⁶

An all-atom structural superposition of the RNase Sa conformers is provided in Figure 2a. The fluctuations are small, as we are purposely investigating small-scale variations. The average pairwise all-atom RMSD is 0.42 Å. Figure 2b provides a backbone superposition of the RNase Sa conformers. The average α -carbon RMSD is 0.31 Å. Significant backbone variability is isolated within the loop region connecting

strands $\beta 3$ and $\beta 4$. RMSDs describing the structural variability within each titratable residue are provided in Table 1.

Table 1. Rank-Ordered List of All Titratable Averaged pK_a Values, Standard Deviations, Structural Variabilities, and Solvent Accessibilities of RNase Sa^a

rank order ^b	residue	average pK_a	std. dev.	RMSD ^c (Å)	RSA ^d (Å ²)
1	ASP33	1.47	0.49	0.15	0.5
2	ARG69	15.86	0.40	0.14	2.3
3	ARG65	15.94	0.38	0.14	9.9
4	TYR51	9.20	0.35	0.26	10.4
5	TYR86	8.10	0.32	0.20	11.2
6	TYR80	12.50	0.30	0.22	3.6
7	GLU54	2.49	0.27	0.14	4.2
8	TYR52	10.75	0.21	0.13	1.4
9	TYR55	9.00	0.19	0.28	6.8
10	HIS53	9.48	0.18	0.25	14.0
11	GLU78	4.63	0.18	0.15	6.0
12	ASP79	4.52	0.16	0.14	3.2
13	ASP84	2.84	0.15	0.20	13.7
14	TYR30	7.91	0.14	0.21	19.1
15	TYR81	8.22	0.14	0.20	7.2
16	ARG68	14.22	0.12	0.17	23.1
17	TYR49	6.89	0.12	1.12	43.4
18	GLU14	2.98	0.12	0.13	9.3
19	TRN1	9.49	0.12	0.36	33.0
20	ASP1	2.88	0.10	0.58	33.0
21	TRC96	3.73	0.10	0.27	11.0
22	ASP93	4.09	0.08	0.20	9.9
23	ARG40	12.98	0.07	0.36	47.8
24	GLU74	3.87	0.07	0.19	24.3
25	ASP17	3.98	0.07	0.18	22.1
26	GLU41	3.99	0.06	0.23	30.0
27	HIS85	6.00	0.06	0.37	26.5
28	ASP25	4.48	0.04	0.37	28.4
29	ARG63	12.31	0.03	1.19	58.6
average			0.17	0.29	17.7
std. dev.			0.12	0.26	15.0
correlation ^e				−0.36	−0.63

^a Average pK_a values and standard deviations are provided for $l = 150$ mM. Similar deviations are observed at $l = 100$ and 300 mM. The overall all-atom and α -carbon RMSDs for the structural ensemble are 0.42 and 0.31 Å, respectively. ^b The table is rank-ordered vis-à-vis (largest to smallest) pK_a standard deviation. ^c Titratable atom RMSD. ^d Side-chain solvent accessibility. ^e Linear correlation coefficient between the indicated column and pK_a standard deviation.

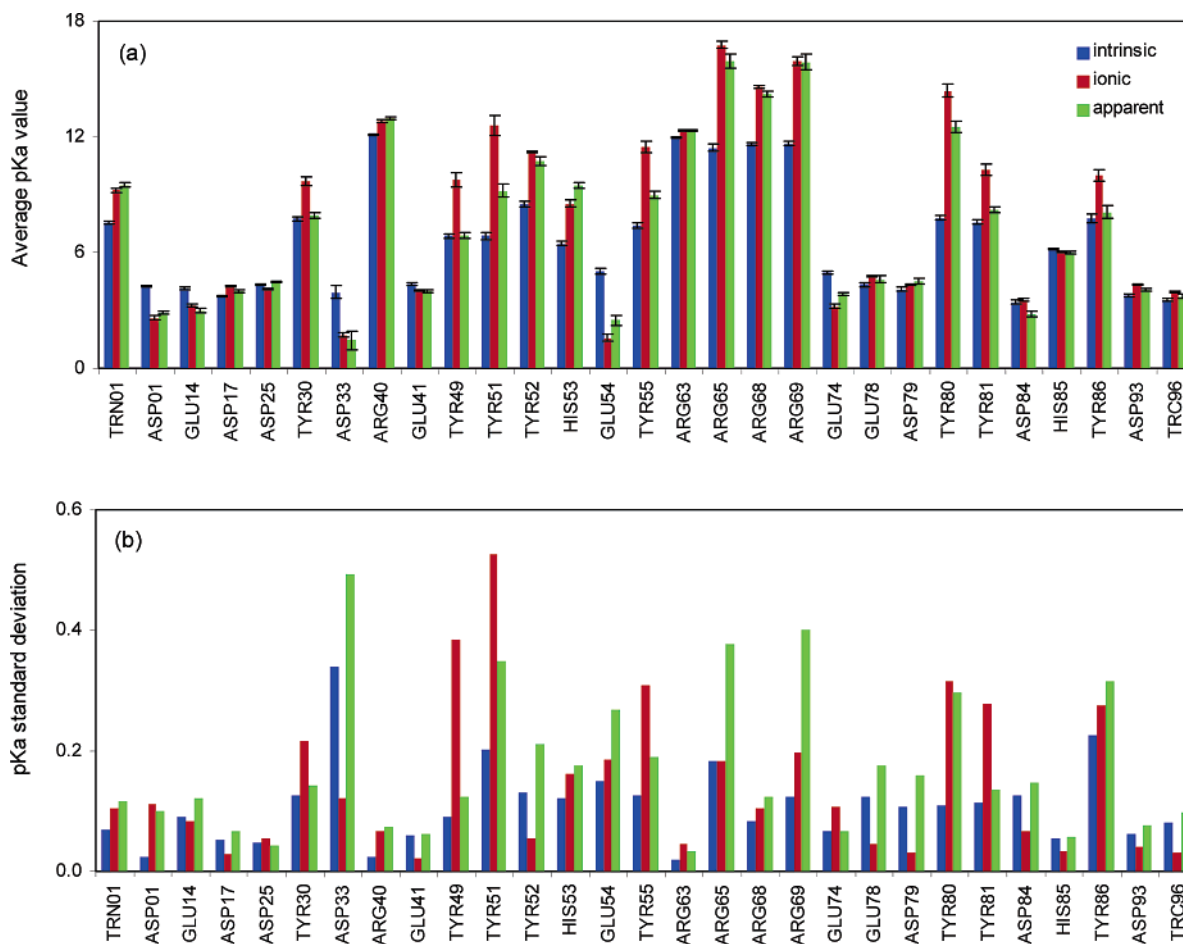


Figure 3. (a) Average intrinsic, ionic, and apparent pK_a values for all sites in RNase Sa ($I = 150$ mM). The intrinsic and ionic pK_a values are calculated by neglecting Φ_{pair} and Φ_{self} , respectively. Error bars represent \pm the standard deviations for each pK_a distribution. Standard deviations are expanded for clarity in part b. Qualitatively similar site-by-site distinctions are observed for $I = 100$ and 300 mM.

Figure 3 provides the average pK_a value and standard deviation for each titratable site. The values provided are for an ionic strength of 150 mM. Values have also been calculated at 100 and 300 mM; however, these results are not shown as their site-to-site distinctions are similar. Table 2 demonstrates that the calculated pK_a values compare favorably to the experimental values.

Table 1, which provides a rank-ordered list of the pK_a standard deviations, shows a wide spectrum of variability within the pK_a fluctuations. Naively, one might expect sites with large conformational fluctuations to also have large pK_a fluctuations. However, this is clearly not the case. In fact, the pK_a fluctuations are slightly anticorrelated with the structural fluctuations, meaning that sites with the smallest conformational fluctuations have the largest pK_a fluctuations. The overall RNase Sa correlation coefficient between the per residue structural variability (calculated as the RMSD for all side-chain target atoms) and the pK_a standard deviation is -0.36 (see Figure 4a). This initially counterintuitive result arises from simple protein structure considerations. Residues on the surface are free to orientate themselves in a variety of ways without drastically affecting their electrostatic surroundings, whereas this is not the case within the crowded protein core. Within the core, slight conformational rearrangements can lead to drastic changes in the electrostatic

microenvironments around the buried sites. Figure 4b plots side-chain atomic solvent accessibility against the pK_a variability. A similar negative correlation between the solvent accessibility and pK_a variability is observed in the LYS and TIM examples. Curiously, no significant correlation is observed between the pK_a fluctuations and fluctuations within the overall potential energy values (calculated using the CHARMM²⁷ force field). The range of pK_a /potential energy correlations for the 29 different RNase Sa titratable sites is $\{-0.22; 0.25\}$. Moreover, the correlation between the force field potential energy and the electrostatic free energy (G_{elec}) at pH 7.0 is also insignificant ($R = -0.20$). This result is discussed in more detail below.

The above points are exemplified by a single RNase Sa arginine pair. For example, the pK_a variability within Arg65 is quite high (standard deviation = 0.38), whereas the variability within Arg63 is the smallest (standard deviation = 0.03). Figure 5a compares the structural superposition of each residue's conformers. However, the structural variability within the guanidinium group of Arg65 (CZ RMSD = 0.14 Å) is much smaller than that of Arg63 (CZ RMSD = 1.18 Å). Arg65 is buried (side chain ASA = 9.92 Å²) within the core, which significantly reduces its conformational freedom. Nevertheless, because of the heterogeneous nature of the electrostatic environment within the core, the slight confor-

Table 2. Comparison of Calculated and Experimental pK_a Values for RNase Sa^a

	experiment $I = 100$ mM	calculated $I = 100$ mM	calculated $I = 150$ mM	calculated $I = 300$ mM
TRN1	9.14	9.83	9.49	9.33
ASP1	3.44	3.17	2.88	2.93
GLU14	5.02	3.94	2.98	3.06
ASP17	3.72	4.44	3.98	3.99
ASP25	4.87	4.83	4.48	4.43
TYR30	11.3	11.85	7.91	7.92
ASP33	2.39	2.47	1.47	1.60
GLU41	4.14	4.40	3.99	4.03
TYR49	10.6	10.58	6.89	6.88
HIS53	8.27	10.18	9.48	9.21
GLU54	3.42	5.14	2.49	2.61
GLU74	3.47	4.61	3.87	3.88
GLU78	3.13	7.65	4.63	4.60
ASP79	7.37	5.62	4.52	4.47
ASP84	3.01	3.49	2.84	2.94
HIS85	6.35	6.83	6.00	5.98
ASP93	3.09	4.50	4.09	4.05
TRC96	2.42	2.97	3.73	3.72

^a Calculated pK_a values are the average of all conformers using an interior (protein) dielectric of 20, and an exterior (solvent) dielectric of 80, at three different ionic strengths. Experimental values are taken from Laurents et al.²⁵ The correlation coefficient computed from the ensemble-averaged pK_a ($R = 0.90$) is similar to, albeit slightly less ($R = 0.93$) than, the correlation coefficient of the theoretical results reported in Laurents et al. Correlation coefficients are only computed for the 100 mM results, which is the same as the experimental conditions.

mational changes in Arg65 can have pronounced effects on its pK_a value. Conversely, Arg63 is completely solvent-exposed (side chain ASA = 58.57 Å²) and is, thus, able to explore a much larger conformational space. Because the electrostatic environment on the RNase Sa surface is more uniform, at least compared to the myriad electrostatic microenvironments within the protein core, the large conformational changes with Arg63 have little effect on its calculated pK_a value.

Arg65 is part of an electrostatic network that includes Asp33, Glu54, and Arg69 (Figure 5b). All three also have large fluctuations within their pK_a value distributions. In fact, the variability within Asp33 is the largest for RNase Sa. As one site is perturbed, there is a local change in the electrostatic microenvironment that affects all four pK_a values simultaneously. Comparable sensitivities are observed in other buried charged clusters. The variabilities within the LYS and TIM pK_a values are similar to the RNase Sa results. Moreover, the inverse correlation between ASA and RMSD is also qualitatively similar (see Tables 3 and 4). However, some interesting deviations to the overall trends, which are also discussed in the next section, do occur in TIM.

pK_a Variability within Solvent-Exposed Sites. It is demonstrated above that the extent of pK_a variability can generally be ascribed to solvent accessibility and structural variability, which are, of course, related. Like Arg63 of

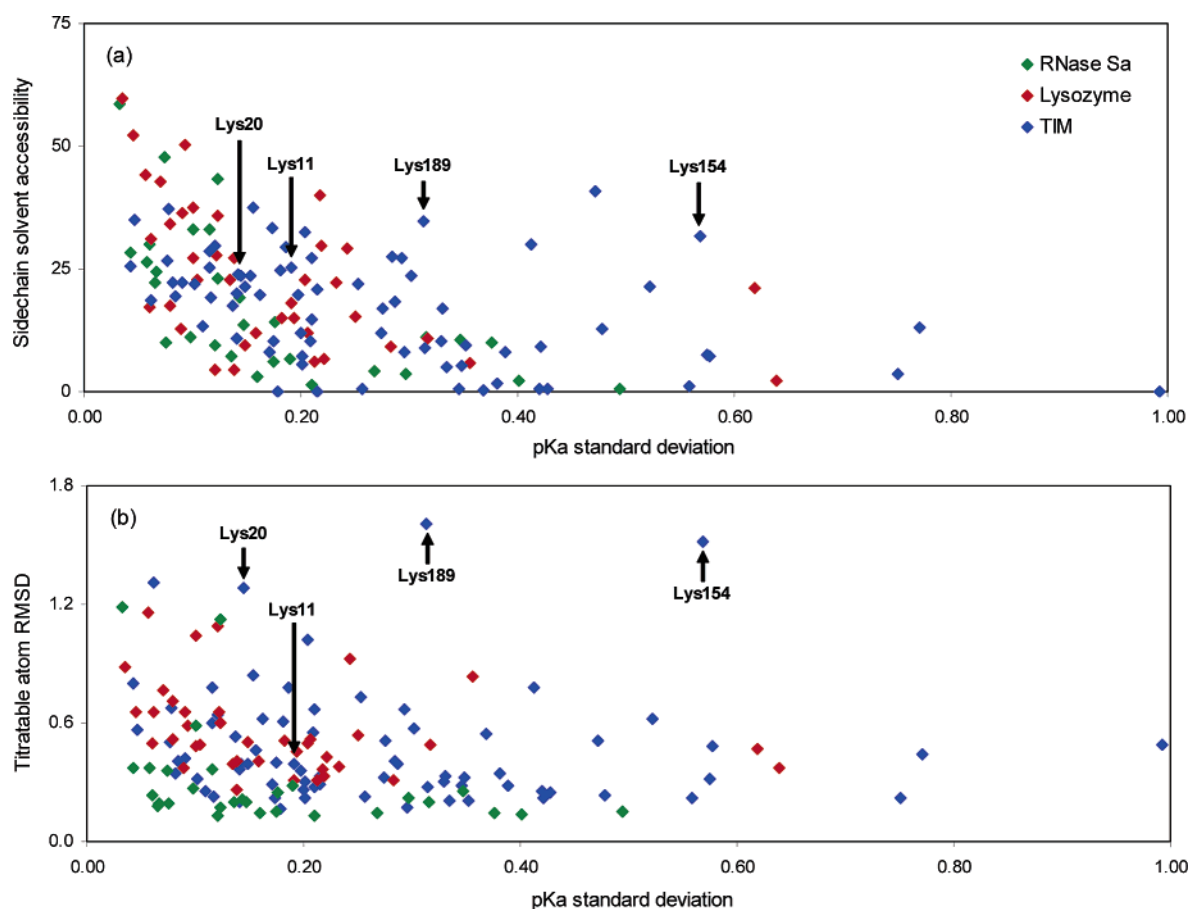


Figure 4. (a) Side-chain solvent accessibility vs the standard deviation for each pK_a distribution for the three investigated proteins. Note the nonlinear dependence of the solvent accessibility effects. (b) Structural RMSD for each titratable target atom vs the standard deviation for each pK_a distribution. A similar nonlinear correlation is observed. In both figures, the four TIM lysines discussed in the text are highlighted.

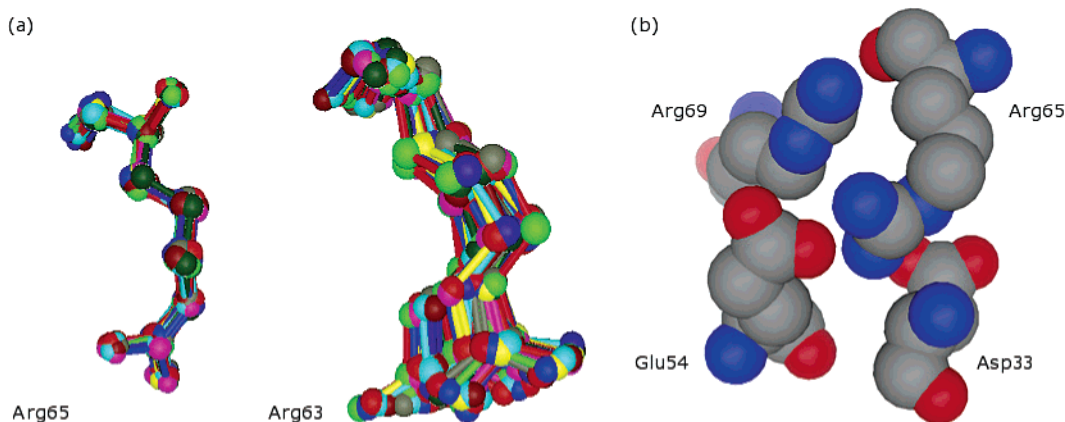


Figure 5. (a) Structural superposition of RNase Sa's Arg65 and Arg63. The structural variability within the solvent-exposed Arg63 is the largest of all RNase Sa residues. However, because the electrostatic microenvironment of solvent-exposed sites is generally uniform, its pK_a variability is quite small. On the other hand, the structural variability within the buried Arg65 is small, yet its pK_a variability is quite large—the third largest of all RNase Sa residues. (b) The buried Arg65 is sandwiched between Asp33, Glu54, and Arg69. As a consequence, small conformational fluctuations can significantly affect the electrostatic microenvironment of these residues. In fact, residues Asp33, Arg69, and Arg65 have the three largest pK_a value standard deviations (see Table 1). Glu54, ranked seventh, also displays significant pK_a variability.

RNase Sa, Lys20 of TIM typifies the normal situation of a solvent-accessible site (see Figure 6a). Despite the general consistency of these trends, there are some notable exceptions. It can be seen in Figure 4 that plotting RMSD or ASA versus pK_a variability loosely approximates a quadratic curve in all three examples, meaning sites with large pK_a fluctuations are almost always buried (or conformationally constrained). The most egregious exceptions to this trend occur in TIM. For example, Lys154 and Lys189, which are both solvent-accessible, have atypically large pK_a fluctuations. The basis of the observed pK_a variability is revealed in Figure 6. The electrostatic potential maps in Figure 6 clearly indicate that the positions of Lys154 and Lys189 are located at interfaces between significant anionic and cationic regions. As the position of the titratable atom fluctuates, its electrostatic microenvironment also changes. Consequently, a large range of pK_a values is observed for these sites. Normally, the distinctions in the electrostatic microenvironments on the protein surface are not very significant; however, in these two cases, they clearly are. The pK_a ranges for Lys154 and Lys189 are {10.68; 12.45} and {10.74; 12.29}, respectively.

Across all three protein examples, solvent-exposed sites with constrained conformational variability can occur. In all cases analyzed, the dynamics within these sites are constrained by some sort of noncovalent interaction. For example, Lys11, which has a solvent accessibility quantitatively similar to that of Lys20, is highly constrained because of an extended electrostatic network within the active-site region.³ Lys11 is most strongly interacting with Glu96 via a strong ionic bond. This salt bridge constricts the dynamics of Lys11. Lys11 is also electrostatically interacting with the catalytic Glu164, His94, Tyr100, and Cys125 (Figure 6d). The pK_a variability within Lys11 is ranked in the middle third of all TIM sites. The bulk of the sites with pK_a fluctuations of similar scale are inaccessible to solvents.

Elucidating the Origins of pK_a Variability within the Core. As discussed above (see Figure 1), the standard procedure of computing pK_a values uses a two-step process. The first

step calculates the *intrinsic* pK_a from the *model* pK_a by accounting (via Φ_{self}) for the solvation and background charge changes that occur when going from a fully solvated side chain to the hypothetical neutral protein environment. The *apparent* pK_a , which is the final calculated value, is calculated from the intrinsic pK_a by accounting (via Φ_{pair}) for a more realistic charged-protein environment. In this step, the electrostatic potential between all titratable charge pairs is evaluated. In RNase Sa, the largest pK_a variability occurs in Asp33. Figure 7 plots the difference between all electrostatic potentials between the two RNase Sa conformers with the most extreme Asp33 pK_a values. Curiously, there is significant and consistent variability within Φ_{self} (on diagonal), whereas the variability within Φ_{pair} is more intermittent. Moreover, the difference within the self-potential of Asp33 (1.3 kcal/mol/e) dwarfs all other differences (the second largest difference is 0.3 kcal/mol/e).

To better understand the effects of potential variability on calculated pK_a values, we compute pK_a values using only one of the two steps from the normal procedure. Values calculated using only Φ_{self} are deemed *intrinsic* pK_a 's, whereas values calculated using only Φ_{pair} are called *ionic* pK_a 's. These values are also presented in Figure 3a alongside the *apparent* pK_a values. The difference between the three different " pK_a values" is small for most solvent-exposed sites (e.g., Arg40, Arg63, and His85). However, large differences are common within buried sites. There is a slight negative correlation ($R = -0.34$) between site accessibility and the ΔpK_a (defined as $|pK_{a\text{intrinsic}} - pK_{a\text{ionic}}|$). The ionic pK_a values are generally closer to the apparent values than the intrinsic values, which highlights the increased importance of the various formal charges within the protein. This result is especially true for sites that are largely inaccessible to solvents.

In all but three sites, the intrinsic pK_a is calculated to be less than the ionic pK_a . Two of the exceptions correspond to Asp33 and Glu54, both of which are discussed above. Figure 3b expands the standard deviations observed within

Table 3. Rank-Ordered List of All Titratable Averaged pK_a Values, Standard Deviations, Structural Variabilities, and Solvent Accessibilities of LYS^a

rank order ^b	residue	average pK_a	std. dev.	RMSD ^c (Å)	RSA ^d (Å ²)
1	ASP67	3.43	0.64	0.37	2.3
2	LYS69	15.16	0.62	0.47	21.1
3	TYR54	8.65	0.36	0.83	5.9
4	ARG62	15.01	0.32	0.49	10.9
5	TYR38	7.78	0.28	0.31	9.1
6	GLU7	2.53	0.25	0.53	15.3
7	TYR45	7.51	0.24	0.92	29.0
8	LYS13	12.60	0.23	0.38	22.2
9	TYR124	7.39	0.22	0.43	6.7
10	ARG101	13.34	0.22	0.33	29.7
11	ARG10	13.18	0.22	0.37	39.9
12	ASP53	3.64	0.21	0.31	6.1
13	ASP91	3.67	0.21	0.52	12.1
14	LYS1	12.88	0.20	0.50	22.7
15	ASP49	2.29	0.19	0.46	15.0
16	ARG98	14.07	0.19	0.31	18.0
17	TYR20	6.43	0.18	0.51	14.9
18	LYS97	11.38	0.16	0.41	12.0
19	ASP18	2.38	0.15	0.51	9.4
20	GLU35	5.17	0.14	0.26	4.4
21	ARG21	13.06	0.14	0.41	27.1
22	TRN1	8.28	0.13	0.39	22.7
23	TYR63	7.14	0.12	0.60	35.8
24	ARG5	13.73	0.12	0.65	27.7
25	TRC130	2.43	0.12	1.09	4.4
26	ASP87	2.65	0.10	0.49	22.8
27	ARG119	13.24	0.10	0.48	27.3
28	HIS78	6.09	0.10	1.04	37.6
29	ARG122	13.49	0.09	0.59	50.4
30	ARG115	12.68	0.09	0.65	36.3
31	ASP102	2.35	0.09	0.37	12.9
32	LYS33	11.49	0.08	0.71	17.4
33	ARG113	13.11	0.08	0.51	34.2
34	ARG107	12.04	0.07	0.77	42.9
35	GLU4	3.93	0.06	0.66	31.1
36	ASP120	2.81	0.06	0.50	17.1
37	ARG50	12.71	0.06	1.16	44.2
38	ARG41	12.79	0.05	0.65	52.3
39	ARG14	12.38	0.03	0.88	59.8
average			0.18	0.56	23.4
std. dev.			0.13	0.22	14.6
correlation ^e				-0.30	-0.48

^a Average pK_a values and standard deviations are provided for $I = 150$ mM. Similar deviations are observed at $I = 300$ mM. The overall all-atom and α -carbon RMSDs for the structural ensemble are 0.73 and 0.58 Å, respectively. ^b The table is rank-ordered vis-à-vis (largest to smallest) pK_a standard deviation. ^c Titratable atom RMSD. ^d Side-chain solvent accessibility. ^e Linear correlation coefficient between the indicated column and pK_a standard deviation.

all calculated pK_a values in order to facilitate comparisons. As suggested by Figure 7, Asp33 is unique because of its large variability within its intrinsic pK_a distribution. This result indicates that Asp33 is very sensitive to local fluctuations within the background electrostatics. To explore this result more closely, correlations between the intrinsic pK_a values for all titratable site pairs are computed (data not shown). The site most strongly correlated with Asp33 is

Arg65, which suggests that the relative location of these two sites has a pronounced effect on the background electrostatics. This initially counterintuitive result (one might expect variability between two interacting *charged* residues to affect the *ionic* pK_a more than the intrinsic pK_a) is explained by the fact that the carboxylate of Asp33 is doubly hydrogen-bonded to the nontitrating NE and NH1 atoms of Arg65. As a consequence, a slight structural rearrangement between the two significantly affects the local background electrostatics (as exemplified in Figure 7). Changes in the protonation or deprotonation state of Asp33 or Arg65 has no effect on the presence of the two hydrogen bonds, which explains the reduced ionic pK_a correlation for this pair. Because of their structural proximity, significant fluctuations are observed within $\Phi_{\text{Asp33-Arg65}}$ and $\Phi_{\text{Asp33-Arg69}}$. However, large fluctuations are not observed in any other Asp33 site pairs, which keeps its $pK_{a,\text{ionic}}$ from varying significantly. While the correlation between the contacting Asp33-Arg65 pair is the strongest observed for Asp33, it should be noted that several structurally remote sites are also strongly correlated with it. The origin of these correlations is unclear. Future work will attempt to identify their origin.

Sites with the largest variability within their ionic pK_a values are generally tyrosines. As can be seen in Figure 3b, Tyr51 is identified as the RNase Sa site with the most significant ionic pK_a variability. Figure 7a reveals that significant changes within the pairwise potentials occur within $\Phi_{\text{Tyr51-Glu74}}$, $\Phi_{\text{Tyr51-Glu78}}$, and $\Phi_{\text{Tyr51-Tyr80}}$, which result in the ionic pK_a fluctuations. Figure 7b demonstrates that these four sites constitute a second electrostatic tetrad (distinct from the Asp33-Glu54-Arg65-Arg69 tetrad discussed above). In this charge cluster, slight conformational changes significantly affect ionic pK_a values. Nevertheless, a significant fraction of the apparent pK_a variability within Glu74 and Glu78 is also attributed to fluctuations within the intrinsic pK_a (Figure 3b). This point illustrates one of the main results of this investigation, that being apparent pK_a fluctuations within the protein core can arise from changes in both the background and pairwise electrostatic interactions. Apparent pK_a fluctuations that arise from convolutions of Φ_{self} and Φ_{pair} are also frequently observed in LYS and TIM. As mentioned previously, Nielsen and McCammon¹⁷ report an identical conclusion regarding the origins of the variability within Asp52 from their comparison of 41 HEWL X-ray structures.

Variability within Overall G_{elec} Values. G_{elec} , which is also calculated by UHBD as part of the pK_a calculation, is the purely electrostatic portion of the overall protein free energy. (A brief description of how G_{elec} is determined is provided in Livesay et al.²⁸) Because G_{elec} is frequently used to assess the electrostatic portions of molecular recognition events²⁸⁻³¹ and overall protein stability,^{32,33} understanding the conformational sensitivity of this quantity is also paramount. For all three proteins, it is found that the average snapshot-to-snapshot $\Delta G_{\text{elec}} \approx 0$, meaning that the stabilizing and destabilizing changes tend to cancel each other out. Table 5 lists the average snapshot-to-snapshot absolute value of ΔG_{elec} for the three protein examples; the standard deviation of $|\Delta G_{\text{elec}}|$ and its overall range is also provided. As with

Table 4. Rank-Ordered List of All Titratable Averaged pK_a Values, Standard Deviations, Structural Variabilities, and Solvent Accessibilities of TIM^a

rank order ^b	residue	average pK_a	std. dev.	RMSD ^c (Å)	RSA ^d (Å ²)	rank order ^b	residue	average pK_a	std. dev.	RMSD ^c (Å)	RSA ^d (Å ²)
1	TYR207	16.77	0.99	0.49	0.0	38	LYS113	12.07	0.21	0.67	27.3
2	TYR48	11.17	0.77	0.44	13.1	39	ASP80	2.53	0.21	0.55	10.2
3	GLU103	1.63	0.75	0.22	3.7	40	LYS134	11.57	0.20	1.02	32.5
4	TYR45	10.77	0.58	0.49	7.2	41	ARG204	15.88	0.20	0.31	7.3
5	TYR66	11.20	0.58	0.32	7.5	42	ASP179	2.42	0.20	0.22	5.7
6	LYS154	11.70	0.57	1.52	31.7	43	GLU132	2.79	0.20	0.26	11.9
7	TYR163	19.49	0.56	0.22	1.1	44	GLU202	2.70	0.20	0.36	19.7
8	ASP197	2.36	0.52	0.62	21.3	45	LYS11	12.42	0.19	0.39	25.2
9	GLU76	1.86	0.48	0.24	12.9	46	GLU131	2.80	0.19	0.78	29.5
10	TYR100	10.09	0.47	0.51	40.9	47	LYS220	11.84	0.18	0.60	24.8
11	LYS111	14.94	0.43	0.25	0.5	48	GLU128	5.73	0.18	0.17	0.1
12	ARG97	14.67	0.42	0.22	9.3	49	GLU238	2.65	0.18	0.40	10.3
13	ASP226	-0.41	0.42	0.25	0.6	50	ARG144	14.03	0.17	0.22	33.4
14	LYS16	11.05	0.41	0.78	29.9	51	GLU96	1.32	0.17	0.29	8.0
15	ASP110	2.83	0.39	0.29	8.0	52	LYS194	11.14	0.16	0.62	19.8
16	HIS94	6.15	0.38	0.35	1.7	53	LYS198	11.96	0.16	0.46	37.6
17	CYS40	11.08	0.37	0.55	0.2	54	LYS68	10.53	0.15	0.84	23.5
18	ARG188	17.67	0.35	0.21	9.4	55	ASP221	2.05	0.15	0.40	21.5
19	ASP224	2.39	0.35	0.32	5.3	56	LYS20	11.49	0.14	1.28	23.6
20	CYS125	15.06	0.35	0.28	0.5	57	GLU21	3.16	0.14	0.39	23.8
21	ASP105	0.42	0.33	0.21	5.0	58	GLU143	4.57	0.14	0.20	10.8
22	ASP140	2.57	0.33	0.33	17.1	59	LYS55	11.71	0.14	0.36	20.1
23	ARG2	14.46	0.33	0.30	10.2	60	LYS222	11.84	0.14	0.53	17.5
24	GLU36	3.26	0.32	0.28	8.8	61	LYS88	11.21	0.12	0.64	29.8
25	LYS189	11.14	0.31	1.61	34.7	62	ASP104	3.09	0.12	0.23	19.2
26	ASP47	3.11	0.30	0.57	23.5	63	LYS137	12.01	0.12	0.60	25.2
27	ARG98	17.62	0.30	0.18	8.0	64	LYS236	11.50	0.12	0.78	28.7
28	LYS54	12.00	0.29	0.67	27.1	65	ARG25	13.96	0.11	0.26	13.4
29	LYS133	12.90	0.29	0.39	18.4	66	ASP241	3.86	0.10	0.32	22.1
30	LYS106	13.37	0.28	0.41	27.6	67	ASP182	3.94	0.09	0.42	22.2
31	GLU24	3.18	0.28	0.51	16.8	68	TRC247	3.48	0.08	0.40	19.5
32	GLU152	2.32	0.27	0.32	11.8	69	ASP155	2.89	0.08	0.35	22.3
33	GLU164	-0.41	0.26	0.23	0.7	70	HIS102	6.78	0.08	0.67	37.1
34	LYS83	12.07	0.25	0.73	22.0	71	GLU151	4.28	0.08	0.50	26.7
35	HIS184	7.31	0.21	0.29	0.0	72	TRN1	7.79	0.06	1.31	18.7
36	GLU178	2.79	0.21	0.33	20.7	73	GLU33	4.24	0.05	0.56	34.9
37	ARG246	14.63	0.21	0.28	14.6	74	ASP84	4.14	0.04	0.80	25.4
average			0.27	0.48	17.0						
std. dev.			0.18	0.30	10.9						
correlation ^e				-0.09	-0.40						

^a Average pK_a values and standard deviations are provided for $I = 150$ mM. Similar deviations are observed at $I = 300$ mM. The overall all-atom and α -carbon RMSDs for the structural ensemble are 0.59 and 0.46 Å, respectively. ^b The table is rank-ordered vis-à-vis (largest to smallest) pK_a standard deviation. ^c Titratable atom RMSD. ^d Side-chain solvent accessibility. ^e Linear correlation coefficient (for all 74 titratable sites) between the indicated column and pK_a standard deviation.

the pK_a variations, the G_{elec} fluctuations within RNase Sa are the smallest ($\langle |\Delta G_{\text{elec}}| \rangle = 0.35$ kcal/mol) of the three examples investigated. Somewhat surprisingly, the variation with the UHBD G_{elec} values is uncorrelated with the CHARMM potential energy values (see Figure 8). The lack of correlation arises from the reduced variability within the G_{elec} values. For example, the standard deviation within the UHBD G_{elec} values is 10% of the average value, whereas it is 74% of the average CHARMM potential energy value. From this result, it can be inferred that G_{elec} is fairly insensitive to slight conformational changes, especially when compared to traditional force field methods. Moreover, the observed robustness within G_{elec} strengthens the conclusions

of studies that use G_{elec} to probe single protein conformations, such as those referenced above.

Conclusions

Our ultimate goal is to develop a robust computational framework to understand pK_a changes along a reaction coordinate. In this report, we use MD simulations to generate a conformational ensemble within three protein examples (RNase Sa, TIM, and LYS) to determine the conformational sensitivity of calculated pK_a values. The conformational variability is explicitly designed to be small in order to focus this investigation on the effects of *slight conformational changes*. These results provide a baseline of pK_a fluctuations

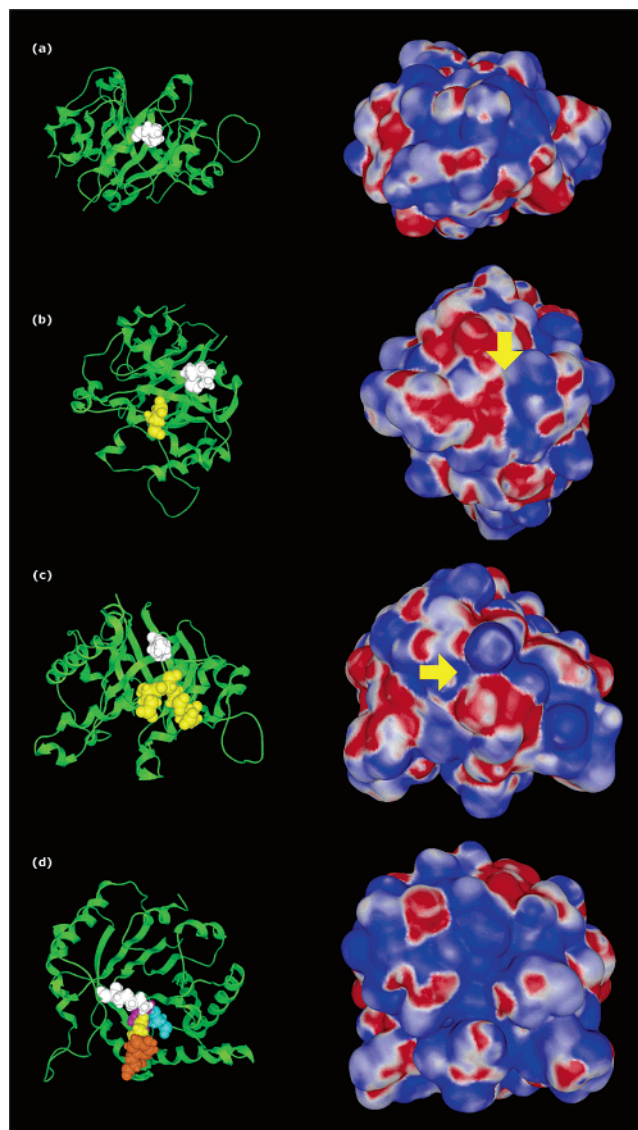


Figure 6. Comparison of the electrostatic environment around four TIM lysine residues. Structural cartoons, oriented the same as the electrostatic potential maps, are provided to facilitate comparisons. The target lysine residues are colored white, and are (from top to bottom) Lys20, Lys189, Lys154, and Lys11. (a) Lys20 typifies the normal case where solvent-exposed residues exhibit large structural fluctuations, yet their pK_a variability is small. This result occurs because the electrostatic microenvironment of the surface is more uniform than in the core; the electrostatic environment surrounding Lys20 is predominantly cationic, with a few small anionic regions. The solvent-exposed (b) Lys189 and (c) Lys154 are atypical because these sites have significant pK_a variability (see Figure 4). This result occurs because there are stark anionic/cationic electrostatic potential regions near these sites. The culprit anionic/cationic potential interfaces are highlighted by the yellow arrow. Acidic residues that predominantly define the anionic regions near the two lysine residues are colored yellow. (d) The solvent-exposed Lys11 is also atypical because its structural variability is significantly constrained. The constrained structural variability within Lys11 is due to a strong salt bridge between it and Glu96 (colored yellow). Also displayed are Glu164 (cyan), His94 (violet), and Tyr100 (orange), which make up an extended electrostatic network at the active site of the enzyme.

that can be used in subsequent investigations. Future work will attempt to incorporate protein flexibility and changes in electrostatic environment due to substrates and reaction intermediates, similar to our previous work,³ to better model these effects.

Our results indicate that sites buried in the protein core are very sensitive to slight structural fluctuations, whereas sites on the surface are generally robust. A few exceptions to the latter trend are observed in TIM, which can be explained by their proximity to drastic changes in the anionic/cationic character of the electrostatic potential surfaces. In summary, the results presented herein (for both buried and exposed sites) highlight the structural sensitivity of calculated pK_a values within heterogeneous electrostatic environments. Heterogeneity within the local electrostatic environment is usually associated with the crowded protein core; however, as demonstrated by TIM, it can also be significant on the protein surface. Finally, overall G_{elec} values are generally robust to slight conformational changes. This final result is especially apparent when compared against the increased variability within traditional force field techniques.

Separating the apparent pK_a calculation into its intrinsic pK_a and ionic pK_a constituent parts indicates that the observed pK_a variability arises from effects associated with both nontitratable and titratable charges. For example, in the case of RNase Sa, Asp33 is hydrogen-bonded to the nontitrating NE and NH1 atoms of Arg65. As a consequence, much of the variability within Asp33, which has the large apparent pK_a variability of all RNase Sa sites, is due to nontitratable (background) charges. Conversely, slight conformational fluctuations have a more significant effect on the pairwise electrostatic potentials of Tyr51 than its self-potential. Similar results are observed in LYS and TIM.

Methods

Calculation of pK_a Values. Titratable residue pK_a values are calculated using the UHBD suite of programs.³⁴ All calculations employ the same approach that we have reported previously.^{3,28,31,32,35} In the approach, Φ_{self} is used to calculate the *intrinsic* pK_a from the model values. When calculating Φ_{pair} , all background charges are set to zero, because they are already included in the *intrinsic* pK_a . With the Φ_{self} and Φ_{pair} potentials in hand, the pK_a is determined after considering all possible ionization states, meaning that, despite the schematic shown in Figure 1, the *apparent* pK_a is actually a mixture of the top two lines. For example, Figure 1 encapsulates four different ionization states: $\text{Lys}^{+1}/\text{Glu}^{1-}$, $\text{Lys}^{+1}/\text{Glu}^0$, $\text{Lys}^0/\text{Glu}^{1-}$, and $\text{Lys}^0/\text{Glu}^0$. A Boltzmann probability distribution is used to describe each possible ionization state. Over a series of pH values, the fractional charge of each site is calculated as the sum of the probabilities when ionized. From the Henderson–Hasselbalch equation, the pK_a is simply defined as the pH at which the fractional charge is ± 0.5 , for acids and bases, respectively. Because sites can be either neutral or ionized, it also follows that both sides of the horizontal equilibria are evaluated when the pK_a values are determined. For large numbers of titratable sites, the computational cost of considering all 2^N possible ionization states is prohibitive. To make the problem computationally

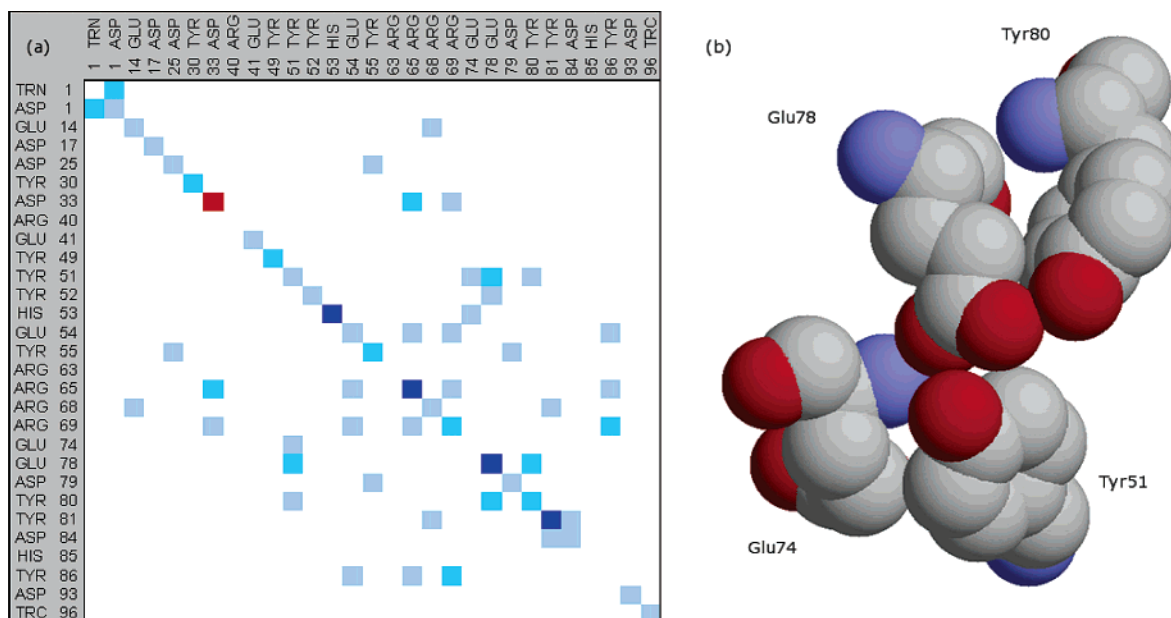


Figure 7. (a) Differences in electrostatic potentials between the RNase Sa conformer pair with the most extreme Asp33 pK_a values. Off-diagonal values correspond to Φ_{pair} , whereas on-diagonal values correspond to Φ_{self} . The three shades of blue (light to dark) correspond to differences of 0.1 kcal/mol/e, 0.2 kcal/mol/e, and 0.3 kcal/mol/e; red corresponds to a difference of 1.3 kcal/mol/e. (b) Tyr51 has the most significant ionic pK_a variability (see Figure 3b). For this site, the variability arises from changes within $\Phi_{\text{Tyr51-Glu74}}$, $\Phi_{\text{Tyr51-Glu78}}$, and $\Phi_{\text{Tyr51-Tyr80}}$, which constitute a tight, solvent-exposed cluster of four titratable sites.

Table 5. $|\Delta G_{\text{elec}}|$ Variability Statistics^a

protein	$\langle \Delta G_{\text{elec}} \rangle$ (kcal/mol)	std. dev. (kcal/mol)	minimum (kcal/mol)	maximum (kcal/mol)
RNase Sa	0.35	0.25	0.00	1.59
LYS	0.90	0.67	0.01	2.56
TIM	1.08	0.92	0.06	4.72

^a Statistics describing the distribution of contiguous snapshot-to-snapshot $|\Delta G_{\text{elec}}|$ values.

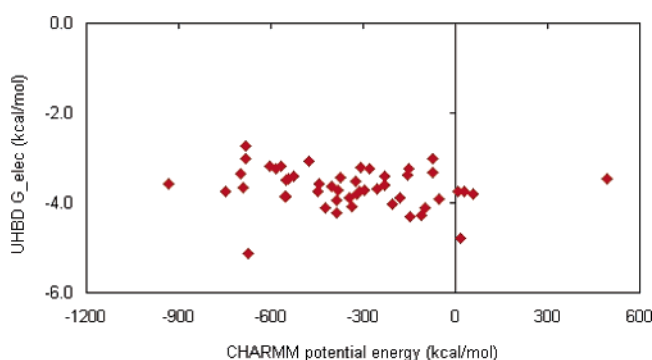


Figure 8. No obvious correlation exists between the UHBD-calculated G_{elec} values of the RNase Sa conformers and the corresponding potential energies computed from the CHARMM force field ($R = -0.20$). This result arises from the general lack of conformational sensitivity within the G_{elec} values. As a consequence, G_{elec} is determined to be rather insensitive to small structural fluctuations.

tractable, UHBD calculates pK_a values using the cluster method described in Gilson¹² and Antosiewicz et al.¹⁰ The ionic pK_a is simply calculated from the model value by setting all Φ_{self} values to zero.

The linear Poisson–Boltzmann equation (LPBE) is solved using the Choleski preconditioned conjugate gradient method. The LPBE is used, versus the computationally more expen-

sive nonlinear Poisson–Boltzmann equation (NLPBE), because of the large number of electrostatic potential calculations required to calculate all pK_a values within a given protein structure. The protein is centered on a $65 \times 65 \times 65$ grid with each grid unit equaling 1.5 Å. Focusing is used around each titrating site with the grid spacing becoming 1.2, 0.75, and 0.25 Å. In all calculations, a solvent dielectric constant of 80 and a protein dielectric constant of 20 are used. Protein partial charges are taken from the CHARMM parameter set²⁷ and radii from the Optimized Potentials for Liquid Systems.³⁶ The ionic strength varies between 100 and 300 mM, and the temperature is 298 K. Intrinsic and ionic pK_a values are calculated using the standard procedure, but without including the background and charge pair effects, respectively.

Electrostatic Potential Maps. Electrostatic potential maps are calculated using the NLPBE solver within the Molecular Operating Environment (MOE) software package. The proteins are centered on a $33 \times 33 \times 33$ grid. A solvent dielectric constant of 80 is used, with a protein dielectric constant of 4, which are standard values in electrostatic potential map calculations.³⁷ Electrostatic potential maps calculated using an interior dielectric constant of 20 are qualitatively similar (results not shown). Protein partial charges are taken from the CHARMM parameter set.²⁷ The temperature is 298 K; the ionic strength is 150 mM, and the protein concentration is 0.001 M. Electrostatic potentials are rendered in blue and red onto the protein solvent-accessible surface at ± 1.0 kcal/mol/e, respectively. Electrostatic potential maps are provided for only one exemplar conformer; however, all qualitative conclusions based on that exemplar are robust to structural variations.

Protein Structures and Molecular Dynamics. Protein structures are modified versions of the coordinates retrieved

from the Protein Databank (PDB). The continuum electrostatics method implemented in the UHBD suite of programs requires explicit polar hydrogen atoms, which are added using the MOE software package. Proteins (and PDB identification codes) for the protein structures used are RNase Sa from *S. aureofaciens* (1RGG),²⁶ triosephosphate isomerase from *S. cerevisiae* (7TIM),³⁸ and human c-type lysozyme (1JSF).³⁹ Canonical ensemble (fixed *NVT*) in vacuo molecular dynamics simulations, as implemented in MOE, are used to generate the ensemble of conformers. In each example, the time scale of the simulations is 1 ns, and the time step is 0.001 ps. A steepest-descent minimization (till convergence) and an equilibration phase (1000 iterations) precede the sampling phase of the simulation. In the cases of RNase Sa, conformers are sampled uniformly over a narrow 100 ps range (sampled every 1.5 ps) to specifically focus on slight fluctuations. The structural variability within the nonsampled phase of the simulation suggests that its pK_a variability should be consistent with the sampled conformations. In the case of TIM and LYS, conformers are uniformly sampled every 40 and 50 ps, respectively, over the entire MD simulation. Despite the reduced sampling time of the RNase Sa simulation, Figure 4a clearly indicates that the scale of the RNase Sa structural variations is in line with the other two examples. It is obvious that this in vacuo simulation protocol is unacceptable to determine realistic aqueous-phase dynamics. However, it is acceptable for the aims of this work because the simulation is simply used to generate a conformational distribution. The reduced computational complexity of the in vacuo simulation freed up computer time to perform the computationally intensive pK_a calculations.

It should be noted that MD simulations fail to accurately represent true conformational variability. For example, MD simulations tend to have difficulty sufficiently sampling rotamer space,⁴⁰ which is why MD simulations sometimes fail to reproduce all NMR side-chain order parameters.⁴¹ Nevertheless, the results presented here reveal clear trends within the conformational dependence of the method and represent a first step toward a better understanding of how conformational variability affects calculated pK_a values. Work is currently underway in extending this investigation to a broader range of protein flexibility as probed by NMR conformational ensembles and various crystallographic isoforms.

Side-chain accessible surface areas (ASAs) are calculated using WhatIf.⁴² All ASA values are calculated using a contact surface procedure, meaning WhatIf identifies the molecular surface that a spherical probe (representing a water molecule) can come into contact with. Reported ASA values are for a single exemplar conformer; however, all qualitative conclusions based on that exemplar are robust to the slight structural perturbations. The side-chain structural variability of the titratable residues is calculated as the RMSD of a *target* atom representative of the charge location. RMSDs for each conformer are calculated relative to the average position within the ensemble. Charged residue target atoms are defined by Livesay et al.³⁵ In the cases of lysine, tyrosine, cysteine, and the N terminus, the target atoms are simply the charged atoms NZ, OH, SG, and N, respectively. In the

cases of aspartic acid, glutamic acid, C-terminus histidine, and arginine, the target atoms are CG, CD, C, CE2, and CZ, respectively, which are all central to the multiple partially charged atoms.

Acknowledgment. This research project began as a class exercise in CHM 416 (Macromolecular Modeling) at California State Polytechnic University, Pomona. The class was taught by D.R.L., and the students were J.K., H.C., J.G., P.K., M.P.M., S.P., and D.Y. E.C. is supported by a Howard Hughes Medical Institute undergraduate fellowship. D.R.L. thanks Marty Scholtz and C. Nick Pace for insightful discussions concerning RNase Sa. The reviewers are also thanked for helpful comments.

References

- (1) Harris, T. K.; Turner, G. J. Structural Basis of Perturbed pK_a Values of Catalytic Groups in Enzyme Active Sites. *IUBMB Life* **2002**, *53*, 85–98.
- (2) Kursula, I.; Partanen, S.; Lambeir, A. M.; Antonov, D. M.; Augustyns, K.; Wierenga, R. K. Structural Determinants for Ligand Binding and Catalysis of Triosephosphate Isomerase. *Eur. J. Biochem.* **2001**, *268*, 5189–5196.
- (3) Livesay, D. R.; La, D. The Evolutionary Origins and Catalytic Importance of Conserved Electrostatic Networks Within TIM-Barrel Proteins. *Protein Sci.* **2005**, *14*, 1158–1170.
- (4) Ha, N. C.; Kim, M. S.; Lee, W.; Choi, K. Y.; Oh, B. H. Detection of Large pK_a Perturbations of an Inhibitor and a Catalytic Group at an Enzyme Active Site, a Mechanistic Basis for Catalytic Power of Many Enzymes. *J. Biol. Chem.* **2000**, *275*, 41100–41106.
- (5) Rozovsky, S.; Jogl, G.; Tong, L.; McDermott, A. E. Solution-State NMR Investigations of Triosephosphate Isomerase Active Site Loop Motion: Ligand Release in Relation to Active Site Loop Dynamics. *J. Mol. Biol.* **2001**, *310*, 271–280.
- (6) Rozovsky, S.; McDermott, A. E. The Time Scale of the Catalytic Loop Motion in Triosephosphate Isomerase. *J. Mol. Biol.* **2001**, *310*, 259–270.
- (7) Fogolari, F.; Brigo, A.; Molinari, H. The Poisson–Boltzmann Equation for Biomolecular Electrostatics: A Tool for Structural Biology. *J. Mol. Recognit.* **2002**, *15*, 377–392.
- (8) Alexov, E. G.; Gunner, M. R. Incorporating Protein Conformational Flexibility into the Calculation of pH-Dependent Protein Properties. *Biophys. J.* **1997**, *72*, 2075–2093.
- (9) Alexov, E. G.; Gunner, M. R. Calculated Protein and Proton Motions Coupled to Electron Transfer: Electron-Transfer From QA- to QB in Bacterial Photosynthetic Reaction Centers. *Biochemistry* **1999**, *38*, 8253–8270.
- (10) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. Prediction of pH-Dependent Properties of Proteins. *J. Mol. Biol.* **1994**, *238*, 415–436.
- (11) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. The Determinants of pK_a s in Proteins. *Biochemistry* **1996**, *35*, 7819–7833.
- (12) Gilson, M. K. Multiple-Site Titration and Molecular Modeling: Two Rapid Methods for Computing Energies and Forces for Ionizable Groups in Proteins. *Proteins* **1993**, *15*, 266–282.

- (13) Nielsen, J. E.; Vriend, G. Optimizing the Hydrogen-Bond Network in Poisson–Boltzmann Equation-Based PK(a) Calculations. *Proteins* **2001**, *43*, 403–412.
- (14) van Vlijmen, H. W.; Schaefer, M.; Karplus, M. Improving the Accuracy of Protein PKa Calculations: Conformational Averaging Versus the Average Structure. *Proteins* **1998**, *33*, 145–158.
- (15) Zhou, H. X.; Vijayakumar, M. Modeling of Protein Conformational Fluctuations in PKa Predictions. *J. Mol. Biol.* **1997**, *267*, 1002–1011.
- (16) Tanford, C.; Roxby, R. Interpretation of Protein Titration Curves. Application to Lysozyme. *Biochemistry* **1972**, *11*, 2192–2198.
- (17) Nielsen, J. E.; McCammon, J. A. On the Evaluation and Optimization of Protein X-Ray Structures for PKa Calculations. *Protein Sci.* **2003**, *12*, 313–326.
- (18) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Lutty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Gagheri, B.; Scott, L. R.; McCammon, J. A. Electrostatics and Diffusion of Molecules in Solution, Simulations with the University of Houston Brownian Dynamics Program. *Comput. Phys. Commun.* **1995**, *91*, 57–95.
- (19) Gibas, C. J.; Subramaniam, S. Explicit Solvent Models in Protein PKa Calculations. *Biophys. J.* **1996**, *71*, 138–147.
- (20) Nielsen, J. E.; Andersen, K. V.; Honig, B.; Hooft, R. W.; Klebe, G.; Vriend, G.; Wade, R. C. Improving Macromolecular Electrostatics Calculations. *Protein Eng.* **1999**, *12*, 657–662.
- (21) Kumar, S.; Nussinov, R. Fluctuations in Ion Pairs and Their Stabilities in Proteins. *Proteins* **2001**, *43*, 433–454.
- (22) Laurents, D.; Perez-Canadillas, J. M.; Santoro, J.; Rico, M.; Schell, D.; Pace, C. N.; Bruix, M. Solution Structure and Dynamics of Ribonuclease Sa. *Proteins* **2001**, *44*, 200–211.
- (23) Laurents, D. V.; Perez-Canadillas, J. M.; Santoro, J.; Rico, M.; Schell, D.; Hebert, E. J.; Pace, C. N.; Bruix, M. Sequential Assignment and Solution Secondary Structure of Doubly Labelled Ribonuclease Sa. *J. Biomol. NMR* **1999**, *14*, 89–90.
- (24) Laurents, D. V.; Scholtz, J. M.; Rico, M.; Pace, C. N.; Bruix, M. Ribonuclease Sa Conformational Stability Studied by NMR-Monitored Hydrogen Exchange. *Biochemistry* **2005**, *44*, 7644–7655.
- (25) Laurents, D. V.; Huyghues-Despointes, B. M.; Bruix, M.; Thurlkill, R. L.; Schell, D.; Newsom, S.; Grimsley, G. R.; Shaw, K. L.; Trevino, S.; Rico, M.; Briggs, J. M.; Antosiewicz, J. M.; Scholtz, J. M.; Pace, C. N. Charge–Charge Interactions Are Key Determinants of the PK Values of Ionizable Groups in Ribonuclease Sa (PI=3.5) and a Basic Variant (PI=10.2). *J. Mol. Biol.* **2003**, *325*, 1077–1092.
- (26) Sevcik, J.; Dauter, Z.; Lamzin, V. S.; Wilson, K. S. Ribonuclease From *Streptomyces Aureofaciens* at Atomic Resolution. *Acta Crystallogr., Sect. D* **1996**, *52*, 327–344.
- (27) Brooks, R. B.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM, A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (28) Livesay, D. R.; Linthicum, S.; Subramaniam, S. PH Dependence of Antibody–Hapten Association. *Mol. Immunol.* **1999**, *36*, 397–410.
- (29) Gibas, C. J.; Subramaniam, S.; McCammon, J. A.; Braden, B. C.; Poljak, R. J. PH Dependence of Antibody/Lysozyme Complexation. *Biochemistry* **1997**, *36*, 15599–15614.
- (30) Gibas, C. J.; Jambeck, P.; Subramaniam, S. Continuum Electrostatic Methods Applied to PH-Dependent Properties of Antibody–Antigen Association. *Methods* **2000**, *20*, 292–309.
- (31) Livesay, D. R.; Subramaniam, S. Conserved Sequence and Structure Association Motifs in Antibody–Protein and Antibody–Hapten Complexes. *Protein Eng., Des. Sel.* **2004**, *17*, 463–472.
- (32) Torrez, M.; Schultehenrich, M.; Livesay, D. R. Conferring Thermostability to Mesophilic Proteins Through Optimized Electrostatic Surfaces. *Biophys. J.* **2003**, *85*, 2845–2853.
- (33) Zhou, H. X.; Dong, F. Electrostatic Contributions to the Stability of a Thermophilic Cold Shock Protein. *Biophys. J.* **2003**, *84*, 2216–2222.
- (34) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Lutty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Gagheri, B.; Scott, L. R.; McCammon, J. A. Electrostatics and Diffusion of Molecules in Solution, Simulations With the University of Houston Brownian Dynamics Program. *Comput. Phys. Commun.* **1995**, *91*, 57–95.
- (35) Livesay, D. R.; Jambeck, P.; Rojnuckarin, A.; Subramaniam, S. Conservation of Electrostatic Properties within Enzyme Families and Superfamilies. *Biochemistry* **2003**, *42*, 3464–3473.
- (36) Jorgensen, W. L.; Tirado-Rives, J. The OPLS Potential Function for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.
- (37) Sharp, K. A.; Honig, B. Electrostatic Interactions in Macromolecules: Theory and Applications. *Annu. Rev. Biophys. Chem.* **1990**, *19*, 301–332.
- (38) Davenport, R. C.; Bash, P. A.; Seaton, B. A.; Karplus, M.; Petsko, G. A.; Ringe, D. Structure of the Triosephosphate Isomerase–Phosphoglycolohydroxamate Complex: An Analogue of the Intermediate on the Reaction Pathway. *Biochemistry* **1991**, *30*, 5821–5826.
- (39) Harata, K.; Abe, Y.; Muraki, M. Full-Matrix Least-Squares Refinement of Lysozymes and Analysis of Anisotropic Thermal Motion. *Proteins* **1998**, *30*, 232–243.
- (40) Berndt, K. D.; Guntert, P.; Wuthrich, K. Conformational Sampling by NMR Solution Structures Calculated with the Program DIANA Evaluated by Comparison with Long-Time Molecular Dynamics Calculations in Explicit Water. *Proteins* **1996**, *24*, 304–313.
- (41) Philippopoulos, M.; Lim, C. Exploring the Dynamic Information Content of a Protein NMR Structure: Comparison of a Molecular Dynamics Simulation With the NMR and X-Ray Structures of *Escherichia coli* Ribonuclease HI. *Proteins* **1999**, *36*, 87–110.
- (42) Vriend, G. WHAT IF: A Molecular Modeling and Drug Design Program. *J. Mol. Graphics* **1990**, *8*, 52–6, 29.

Increasing the Efficiency of Free Energy Calculations Using Parallel Tempering and Histogram Reweighting

Steven W. Rick*

Department of Chemistry, University of New Orleans, New Orleans, Louisiana 70148,
and Chemistry Department, Southern University of New Orleans,
New Orleans, Louisiana 70126

Received August 19, 2005

Abstract: Free energy calculations from molecular simulations using thermodynamic integration or free energy perturbation require long simulation times to achieve sufficient precision. If entropic and enthalpic components of the free energy are desired, then the computational requirements are larger still. Here we present how parallel tempering (PT) Monte Carlo and weighted histogram analysis method (WHAM) can be used to improve the efficiency of free energy calculations. For both methods, which can be used separately or together, simulations at more than one temperature are performed. The same additional temperatures are often used to determine entropy changes. The results, for the aqueous solvation of *n*-butane and methane, show noticeable improvement in the precision of the free energy and entropy changes. The PT and WHAM methods can give similar error bars as conventional molecular dynamics in half the simulation time. The methods offer an efficient procedure for calculating free energy, entropy, and enthalpy changes in which free energy calculations are performed in parallel for a small number of closely spaced temperatures (for example, as here, at three temperatures: 298 K and 298 ± 15 K), and WHAM is used to enhance the data at each temperature.

I. Introduction

Free energy differences for processes involving changes in noncovalent interactions can be calculated through free energy perturbation (FEP) or thermodynamic integration (TI).^{1,2} These methods give an exact free energy change, ΔG , and are limited only by the accuracy of the potential models and large computational requirements. The calculation of fully converged ΔG values can involve extensive sampling of phase space as demonstrated by one recent study which used over 300 ns of simulation time to calculate a single ΔG value.³ More thermodynamic information can be found by calculating ΔG over a range of temperatures, from which changes in entropy, enthalpy, and heat capacity can be found.^{4–10} The calculations at different temperatures are from independent simulations. By combining the simulations at different temperatures, the efficiency of the free energy calculations may be improved in two ways. First, the sampling over phase space at one temperature can be

enhanced from phase space sampling at different temperatures using parallel tempering (PT).^{11–14} Second, the data from one temperature can be used to determine ensemble averages at another temperature using WHAM.^{15,16}

The most common free energy methods are the potential of mean force (PMF), in which free energies as a function of a physical coordinate is determined, and FEP and TI, in which a free energy for adding particles to a system is determined. In these methods, the potential energy of the system, $E(\mathbf{r})$, is scaled by a parameter λ which can couple to a biasing potential for PMF or the interaction of the added particles in TI or FEP, as

$$E(\mathbf{r}) = E_0(\mathbf{r}) + \lambda V(\mathbf{r}) \quad (1)$$

A variety of methods use replica exchange and WHAM in combination with PMF, FEP, and TI (see Table 1). Parallel tempering can improve sampling efficiency by running several identical replicas of the system at different temperatures. In replica exchange, the replicas are simulated not

* Corresponding author e-mail: sricks@uno.edu.

Table 1. Free Energy Methods, Free Energy Perturbation (FEP), Potential of Mean Force (PMF), and Thermodynamic Integration (TI), Which Use the Weighted Histogram Analysis Method (WHAM) and Replica Exchange Swaps in Both Temperature, T , and Hamiltonian, λ , Variables

free energy method	swaps	WHAM	ref
PMF		λ, T	16
FEP		λ	17
TI, FEP	λ or T		18
PMF	λ, T	λ, T	19
TI	T	T	present work

only at different temperatures but also other thermodynamic conditions or Hamiltonians.^{13,14} Each replica is simulated with conventional Monte Carlo (MC) or molecular dynamics (MD), and, in addition to the local sampling of phase space, global moves are attempted which involve exchanges between replicas at, for example, different temperatures or values of λ . Swaps in either T and λ (but not both) were used by Woods, Essex, and King in combination with FEP and TI,¹⁸ and swaps in both T and λ were used by Sugita, Kitao, and Okamoto with PMF.¹⁹ Reference 19 also described a FEP method using replica exchange and WHAM, but this method was not applied. Histogram reweighting provides a method to reweight data generated with a different Hamiltonian or temperature for the desired Hamiltonian or temperature. The use of WHAM is extremely common for PMF calculations, and, due to the similarities in the energies given by eq 1, what might be termed FEP or PMF is somewhat arbitrary in certain cases. One method combining FEP and WHAM was described by Nina, Beglov, and Roux.¹⁷ Histogram reweighting can be used to find a system's free energy as a function of temperature.^{15,20–24} The method of expanded ensembles provides another method for finding free energy as a function of temperature.²⁵ Some of the WHAM studies have used parallel tempering to aid in the simulations at different temperatures.^{20,22–24}

This paper examines how PT and WHAM can improve the convergence of the ensemble average quantities that are needed by TI to find ΔG and entropy changes. Calculations are done for the hydration free energy of methane and of butane. For methane, a united atom, single interaction site model is used, so this calculation is predominantly dependent on the solvent degrees-of-freedom. For butane, which has important intramolecular degrees-of-freedom, the calculation is dependent on both solute and solvent coordinates.

II. Methods

Thermodynamic Integration. The free energy difference between two systems can be obtained by thermodynamic integration from

$$\Delta G = \int_0^1 \left\langle \frac{dE_\lambda}{d\lambda} \right\rangle_\lambda d\lambda \quad (2)$$

where λ is a parameter that connects the two systems, E_λ is the λ dependent Hamiltonian, and $\langle \dots \rangle_\lambda$ corresponds to an isothermal, isobaric ensemble average with potential energy,

E_λ . In the examples studied here, the $\lambda = 0$ state corresponds to pure water, and $\lambda = 1$ corresponds to water with the addition of a single solute molecule, butane or methane. The entropy change, ΔS , can be found by taking the temperature derivative of eq 2 giving^{26–30}

$$\Delta S = \frac{-1}{kT^2} \int_0^1 \left(\left\langle (PV + E_\lambda) \frac{\partial E_\lambda}{\partial \lambda} \right\rangle_\lambda - \langle PV + E_\lambda \rangle_\lambda \left\langle \frac{\partial E_\lambda}{\partial \lambda} \right\rangle_\lambda \right) d\lambda \quad (3)$$

Heat capacity changes can be found from the second temperature derivative of eq 2.⁹ Interactions between the solvent and the solute are scaled by the parameter λ using the separation-shifted scaling method of Zacharias et al., which eliminate the singularities in the potential energy terms as λ approaches zero.³¹ The λ dependent energy is

$$E_\lambda = E_{\text{water, water}} + \lambda \sum_i \sum_j 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}^2}{r_{ij}^2 + \delta(1-\lambda)} \right)^6 - \left(\frac{\sigma_{ij}^2}{r_{ij}^2 + \delta(1-\lambda)} \right)^3 \right] + q_i q_j / (r_{ij}^2 + \delta(1-\lambda))^{1/2} \quad (4)$$

where the sum over i is for solvent atoms and over j is for solute atoms, ϵ_{ij} and σ_{ij} are the Lennard-Jones parameters, q_i is the charge of atom i , r_{ij} is the distance between i and j , and δ is the shifting parameter that avoids the singularities at $r_{ij} = 0$. The value of δ is chosen so that the integrand is as linear as possible, giving the most direct path from $\lambda = 0$ to $\lambda = 1$.^{31,32} In this study, δ is set equal to 7 \AA^2 . The entropy change can also be found using a finite difference expression for the temperature derivative

$$\Delta S = -[\Delta G(T + \Delta T) - \Delta G(T - \Delta T)]/2\Delta T \quad (5)$$

This requires calculating ΔG at two additional temperatures ($T \pm \Delta T$). The finite difference expression is strictly valid only for small ΔT and assumes that higher temperature derivatives of the free energy are not large. For hydration free energy calculations, ΔT around 15 Kelvin is a good approximation, as indicated by agreement between calculated with the two different methods.^{5,6,9,33}

Parallel Tempering. In a parallel tempering simulation, the system is replicated N times, and each replica is simulated at a different temperature. Each replica is simulated with conventional Monte Carlo (MC) or constant temperature molecular dynamics (MD), and, in addition to the local sampling of phase space, global moves are attempted which involve exchanges between replicas. The swapping moves introduce configurations from higher temperature simulations into the ensemble averages of lower temperature simulations (and vice versa). The swapping then provides a way for the lower temperature simulations to escape local minima of phase space. In the isothermal, isobaric ensemble attempted swaps of replicas i and j are accepted with a probability

$$\text{acc}(i \leftrightarrow j) = \min[1, \exp(\beta_i - \beta_j)(E_i + PV_i - E_j - PV_j)] \quad (6)$$

where $\beta = 1/kT$, P is pressure, and V_i is the volume of replica i .²⁴ This method would swap both the coordinates and the

volume of the two replicas. Alternatively, the volume could not be swapped, and only the coordinates are exchanged.^{24,34} In this method, the coordinates from each replica would have to be rescaled to be contained within the volume of the other replica. The acceptance probability would require a recalculation of the energy of each configuration after volume rescaling. One advantage of exchanging both the coordinates and the volume is that the energies E_i and E_j are already known, and no new energy calculations are required for the replica exchange moves. This is the method used in this study. Each replica is simulated using constant temperature, constant pressure molecular dynamics. Exchanges between neighboring replicas are attempted every 0.1 ps. After each successful exchange of coordinates at a temperature T_i to a temperature of T_j , the velocities need to be rescaled by a factor $(T_j/T_i)^{1/2}$.³⁴ Although not done here, replica exchange moves can also be made between replicas with different Hamiltonians, as for instance in refs 18 and 19, and in other applications.^{35,36} In these cases, the acceptance probabilities are slightly different than eq 6, with terms involving the energy function of one replica with the coordinates of the other replica and vice versa.

The Weighted Histogram Analysis Method. The configurational partition function for a system at a temperature T_k is

$$Z_k(N, T, P) = e^{-\beta_k G_k} = \int \mathbf{dr} dV e^{-\beta_k H(\mathbf{r}, V)} = \sum_H e^{-\beta_k H} \int \mathbf{dr} dV \delta(H(\mathbf{r}, V) - H) = \sum_H e^{-\beta_k H} \Omega_k(H) \quad (7)$$

where G_k is the Gibbs free energy, \mathbf{r} are the system coordinates, and H is the enthalpy ($E(\mathbf{r}) + PV$). The function $\Omega_k(H)$ is the temperature independent enthalpy density of states and can be found from a single simulation at a temperature T_k by

$$\Omega_k(H) = \int \mathbf{dr} dV \delta(H(\mathbf{r}, V) - H) = \frac{\int \mathbf{dr} dV e^{-\beta_k H(\mathbf{r}, V)} \delta(H(\mathbf{r}, V) - H) e^{\beta_k H(\mathbf{r}, V)}}{(Z_k/Z_k)} = \frac{e^{\beta_k H} \int \mathbf{dr} dV e^{-\beta_k H(\mathbf{r}, V)} \delta(H(\mathbf{r}, V) - H)}{1/Z_k} = \frac{e^{\beta_k H} Z_k \langle \delta(H(\mathbf{r}, V) - H) \rangle_k}{Z_k} = e^{\beta_k H} e^{-\beta_k G_k} N_k(H) \quad (8)$$

where $N_k(H)$ is the histogram of enthalpies from the simulation at T_k .

The density of states can be constructed from M different simulations at M different temperatures using a weighted sum

$$\Omega(H) = \sum_{k=1}^M w_k(H) \Omega_k(H) = \sum_{k=1}^M w_k(H) N_k(H) e^{\beta_k H} e^{-\beta_k G_k} \quad (9)$$

The Ferrenberg and Swendsen optimized weights are given by¹⁵

$$w_k(H) = e^{-\beta_k H} e^{\beta_k G_k} / \sum_{j=1}^M e^{-\beta_j H} e^{\beta_j G_j} \quad (10)$$

and eq 9 becomes

$$\Omega(H) = \sum_{k=1}^M N_k(H) / \sum_{k=1}^M e^{-\beta_k H} e^{\beta_k G_k} \quad (11)$$

The Gibbs free energy at T_k can be found from

$$e^{-\beta_k G_k} = Z_k = \sum_H e^{-\beta_k H} \Omega(H) = \sum_H e^{-\beta_k H} \left(\sum_{j=1}^M N_j(H) / \sum_{j=1}^M e^{-\beta_j H} e^{\beta_j G_j} \right) \quad (12)$$

Equation 12 can be solved iteratively to find the set of G_j 's, provided the histogram $N_k(H)$ has some region of overlap with $N_{k+1}(H)$.

Average quantities of a property A , given by

$$\langle A \rangle_k = \int \mathbf{dr} dV A(\mathbf{r}, V) e^{-\beta_k H} / Z_k \quad (13)$$

can be expressed as

$$\langle A \rangle_k = \sum_H e^{-\beta_k H} A_k(H) / \sum_H e^{-\beta_k H} \Omega_k(H) \quad (14)$$

The function $A_k(H)$ can be found from simulation data using

$$A_k(H) \equiv \int \mathbf{dr} dV A(\mathbf{r}, V) \delta(H(\mathbf{r}, V) - H) = e^{\beta_k H} Z_k \int \mathbf{dr} dV A(\mathbf{r}, V) e^{-\beta_k H} \delta(H(\mathbf{r}, V) - H) / Z_k = e^{\beta_k H} e^{-\beta_k G_k} \langle A(H) \rangle_k \quad (15)$$

where $\langle A(H) \rangle_k$ is the average value of A for a particular value of H . Data from simulations at different temperatures can be combined using

$$A(H) = \sum_{j=1}^M w_j(H) A_j(H) = \sum_{j=1}^M w_j(H) \langle A(H) \rangle_j e^{\beta_j H} e^{-\beta_j G_j} \quad (16)$$

Putting this expression for $A(H)$ and the weights from eq 10 into eq 14 gives

$$\langle A \rangle_k = \left[\sum_H \left(\sum_j \langle A(H) \rangle_j \right) e^{-\beta_k H} / \sum_j e^{-\beta_j H} e^{\beta_j G_j} \right] / Z_k \quad (17)$$

where Z_k can be found from eq 12. Equation 17, after eq 12, is used to find the G_j 's and provides a method to use data from other temperatures to calculate averages at a given temperature.

For thermodynamic integration, averages of $\langle \partial E_\lambda / \partial \lambda \rangle_\lambda$ are needed. To calculate this using eq 17 requires calculating the function $\langle \partial E_\lambda / \partial \lambda(H) \rangle_\lambda$. For free energy perturbation, the function $\langle \exp[-\beta(E_{\lambda+\Delta\lambda} - E_\lambda)](H) \rangle_\lambda$ would be needed. Other useful averages can be calculated from

$$\langle H \rangle_k = \sum_H [H \left(\sum_j N_j(H) \right) e^{-\beta_k H} / \sum_j e^{-\beta_j H} e^{\beta_j G_j}] / Z_k \quad (18)$$

and

$$\left\langle \frac{\partial E_\lambda}{\partial \lambda} \right\rangle_k = \sum_H \left[H \sum_j \left\langle \frac{\partial E_\lambda}{\partial \lambda} \right\rangle_{j|H} \right] e^{-\beta_k H} / \sum_j e^{-\beta_j H} e^{\beta_j G_j} / Z_k \quad (19)$$

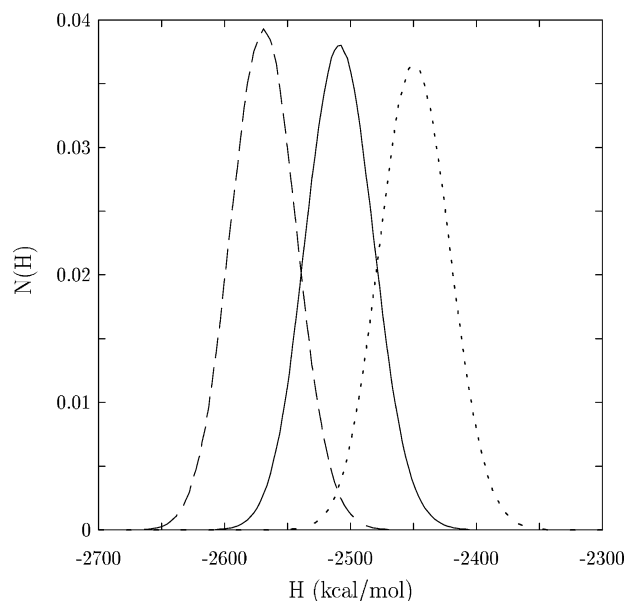


Figure 1. Histograms of the enthalpy at 283 K (dashed line), 298 K (solid line), and 313 K (dotted line), for butane in water at $\lambda = 1$.

All the averages needed to calculate ΔS from eq 3 can be found from the two functions $N_j(H)$ and $\langle \partial E_j / \partial \lambda(H) \rangle_j$.

Histogram reweighting can be used to combine data not only at different temperatures but also for different Hamiltonians.^{15,16} The Hamiltonian corresponding to eq 4 is not linear in λ , due to the separated-shifted scaling terms, so WHAM cannot be used to combine data at different values of λ . If the Hamiltonian were linear or another power of λ , then WHAM could be used.

Simulation Details. The simulations used 256 water molecules, treated using the TIP4P model,³⁷ and one solute molecule, methane or butane. The united atom, one-site OPLS model is used for methane,³⁸ and the all atom, OPLS-AA model is used for butane.³⁹ The Lorentz–Berthelot combining rules were used for the Lennard-Jones interactions, and all bonds were treated as rigid, using SHAKE.⁴⁰ The bond angles and torsional angles in butane are treated as flexible. The simulations were done in the isothermal–isobaric (constant T,P,N) ensemble, by coupling to a pressure bath (at 1 atm) and a Nosé–Hoover temperature bath for three temperatures (283, 298, and 313 K).^{41–45} For this system size, the distributions of enthalpies at the neighboring temperatures have enough overlap of the enthalpy histograms for histogram reweighting and are close enough for good acceptance ratios for the parallel tempering moves (Figure 1). For butane at $\lambda = 1$, the acceptance ratio between the replicas at 283 and 298 K is 0.10 and between the 298 and 313 K replicas is 0.12. The acceptance ratios at other values of λ and for methane are similar. Long-ranged electrostatic interactions were treated using Ewald sums,⁴⁰ and no tail corrections^{3,46} were made for the Lennard-Jones interactions, which were cut off at half the box length. The thermodynamic integration calculations used 15 different λ values (11 at equally spaced intervals of 0.10 from 0 to 1.0 plus 4 additional points at 0.025, 0.05, 0.15, and 0.25). Data for each λ value were generated from six 1 ns simulations, using a 1 fs time step.

Table 2. Solvation Free Energies and Error Estimates (in kcal/mol) for the Four Different Methods at Three Different Temperatures^a

method	283 K		298 K		313 K	
	ΔG	$\delta \Delta G$	ΔG	$\delta \Delta G$	ΔG	$\delta \Delta G$
Butane						
CMD	2.862	0.034(2)	3.180	0.034(6)	3.418	0.030(2)
CMD-WHAM	2.862	0.027(2)	3.175	0.024(5)	3.421	0.027(1)
PT	2.891	0.032(1)	3.196	0.024(2)	3.449	0.027(3)
PT-WHAM	2.886	0.029(1)	3.189	0.021(4)	3.459	0.025(3)
Methane						
CMD	2.081	0.018(1)	2.265	0.022(3)	2.406	0.020(1)
CMD-WHAM	2.083	0.016(1)	2.268	0.015(2)	2.401	0.017(1)
PT	2.102	0.017(1)	2.252	0.015(2)	2.414	0.016(1)
PT-WHAM	2.102	0.016(1)	2.253	0.012(3)	2.414	0.014(1)

^a Numbers in parentheses give error estimates for $\delta \Delta G$.

III. Results

Four methods to calculate the aqueous solvation free energies for the two molecules, butane and methane, are compared. The methods are as follows.

1. CMD: conventional molecular dynamics. Independent simulations at three temperatures, and for each value of λ , are used to calculate ΔG and ΔS using the thermodynamic integration eqs 2 and 3.

2. CMD-WHAM: conventional molecular dynamics plus histogram reweighting. From the independent simulations at different temperatures, the histograms $N_k(H)$ and $\langle \partial E_k / \partial \lambda(H) \rangle_k$ are found. From these histograms and eqs 12 and 17–19, ΔG and ΔS can be calculated.

3. PT: parallel tempering. Parallel tempering is used to generate the ensemble averages at the three temperatures, for each value of λ .

4. PT-WHAM: parallel tempering plus histogram reweighting. From the parallel tempering simulations the histograms are found and used to calculate ΔG and ΔS .

The ΔG values are in fair agreement with the experimental values (1.932 kcal/mol for methane and 2.148 kcal/mol for butane)⁴⁷ and in good agreement with other calculated values for methane^{3,4,9,48,49} and butane³ (Table 2). Error estimates represent two standard deviations of the data calculated from

$$\delta x = \frac{2}{\sqrt{N-1}} \sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} / N \quad (20)$$

where x_i is the value from the i th 1 ns simulation, and \bar{x} is the average of all N x_i values. Estimates of the errors of the error bars can be made by splitting the data into halves and calculating the standard deviation of the δx values from each half. Equation 20 only gives valid error estimates if the data points are uncorrelated. The correlation time was found by calculating the time correlation function of $\langle dE_i/d\lambda \rangle_\lambda$.⁵⁰ For $\langle dE_i/d\lambda \rangle_\lambda$ we find that the correlation time is less than 5 ps, consistent with the results for similar models,³ so each 1 ns simulation is uncorrelated with the others.

Differences in the error bars are in some cases not much bigger than the errors estimates in the error bars, but by looking at all six calculated free energies (3 temperatures

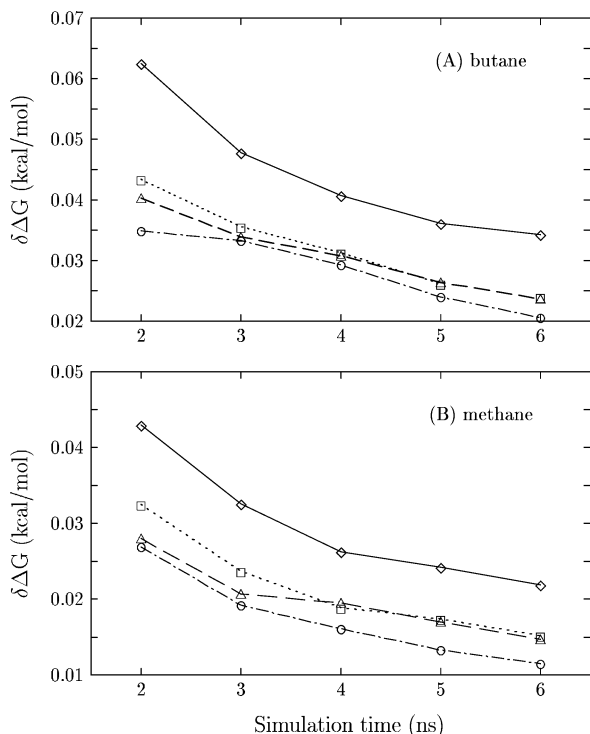


Figure 2. Error estimates of ΔG for (A) butane and (B) methane, comparing the various methods: CMD (solid line, \diamond), CMD-WHAM (dotted line, \square), PT (dashed line, \triangle), and PT-WHAM (dot-dashed line, \circ).

and 2 solute molecules) better assessments about the effectiveness of the different models can be made. (The error bars of the error bars are themselves difficult to determine with much precision, but they are about 5–15% of the magnitude of the $\delta\Delta G$ values, see Table 2.) When either CMD and CMD-WHAM or PT and PT-WHAM are compared, using WHAM lowers the error bars. The PT results are lower than the CMD results for all six ΔG calculations, although the CMD-WHAM results are comparable and in some cases lower than the PT or PT-WHAM results. The improvement of one method over another can be judged by looking at the ratio of the error bars. For example, for butane at 298 K, the ratio of the error bars of CMD to those of PT is $(0.034 \pm 0.006)/(0.024 \pm 0.002)$ or 1.42 ± 0.27 . For methane at 298, the same ratio is $(0.022 \pm 0.003)/(0.015 \pm 0.002)$ or 1.47 ± 0.28 . Since the error bars will decrease as the square root of the simulation time, improving the error bars by a factor of 1.4 (or $\sqrt{2}$) means error bars comparable to CMD can be achieved in half the simulation time. Figure 2 shows the error estimates for the four methods as a function of simulation length, for butane and methane at 298 K. This shows again the improvement of the PT and WHAM methods over CMD. The PT and WHAM methods give error bars after 3 ns which are lower than CMD gives after 6 ns of simulation time. It should be kept in mind for these comparisons, that the PT and WHAM results, involving three separate simulations, use three times the computer time. As long as only ΔG values are desired, gains of $\sqrt{2}$ are not enough to overcome the need for the additional simulations. If ΔS is also desired, then simulations at other temperatures would be typically performed anyway and are not additional.

Table 3. Solvation Entropy Changes, $-T\Delta S$ (in kcal/mol) for the Four Different Methods at $T = 298$ K, Using the Two Different Entropy Expressions, Eqs 3 and 5^a

method	eq 3		eq 5	
	$-T\Delta S$	$T\delta\Delta S$	$-T\Delta S$	$T\delta\Delta S$
Butane				
CMD	6.14	0.69(7)	5.52	0.45(3)
CMD-WHAM	5.94	0.51(4)	5.56	0.38(2)
PT	6.45	0.75(9)	5.55	0.41(3)
PT-WHAM	5.75	0.53(3)	5.69	0.38(3)
Methane				
CMD	2.83	0.55(8)	3.24	0.27(1)
CMD-WHAM	3.25	0.33(3)	3.17	0.23(1)
PT	3.33	0.47(7)	3.10	0.23(1)
PT-WHAM	3.13	0.42(3)	3.10	0.21(1)

^a Numbers in parentheses give error estimates of $T\delta S$.

The entropy changes, as found from both eqs 3 and 5, are given in Table 3. The values are not too far off from the experimental values ($-T\Delta S$ is 4.8 kcal/mol for methane and 7.7 kcal/mol for butane).⁴⁷ Entropy changes have been calculated for the solvation of methane,^{4,9,49} and they are close to the present values, especially for the study that used the same potential.⁹ The improvement in the error bars using WHAM is about the same or slightly better for the entropy calculations, through eq 3, than it is for ΔG . Using PT does not appear to improve the error bars in the entropy calculations. The ratio of the CMD error bars over the CMD-WHAM error bars are 1.4 for butane and 1.7 for methane. This may reflect the fact that the integrand of eq 3 requires more sampling than that of eq 2. The error bars decrease more using WHAM for ΔS calculated through eq 3 than for eq 5, but still the error bars are less using eq 5. The finite difference expression, eq 5, has been previously shown to have smaller error bars.²⁶ The agreement in ΔS using the two different equations is very good, particularly when WHAM is used.

Histogram reweighting improves the error bars by supplementing the data at one temperature with data at other temperatures, as given by eq 17. The CMD-WHAM averages at a temperature T_k can be found from $\langle \partial E / \partial \lambda \rangle = \sum_H \langle \partial E / \partial \lambda (H) \rangle_{k, \text{WHAM}}$. The function $\langle \partial E / \partial \lambda (H) \rangle_{k, \text{WHAM}}$ is the sum of the appropriately weighted histograms from the CMD at three different temperatures

$$\left\langle \frac{\partial E}{\partial \lambda} (H) \right\rangle_{k, \text{WHAM}} = \sum_{j=1}^M \left\langle \frac{\partial E}{\partial \lambda} (H) \right\rangle_j e^{-\beta_k H} / \sum_{l=1}^M e^{-\beta_l H} e^{\beta_l G_l} / Z_k \quad (21)$$

In Figure 3, the reweighted histograms at each of the three temperatures is shown as well as $\langle \partial E / \partial \lambda (H) \rangle_{k, \text{WHAM}}$ for butane at 298 K and $\lambda = 1$. It is clear that data from all three temperatures contribute to the average. The areas under each curve are 0.925 kcal/mol (283 K), 3.46 kcal/mol (298 K), and 1.06 kcal/mol (313 K), giving a total of 5.45 kcal/mol. So 0.64 comes from the CMD simulation at 298 K (3.46/5.45), and the rest, about 1/3, comes from CMD simulations at different temperatures. The infusion of data from other temperatures helps reduce the error bars. If all WHAM did was add 1/3 more information, then using WHAM would

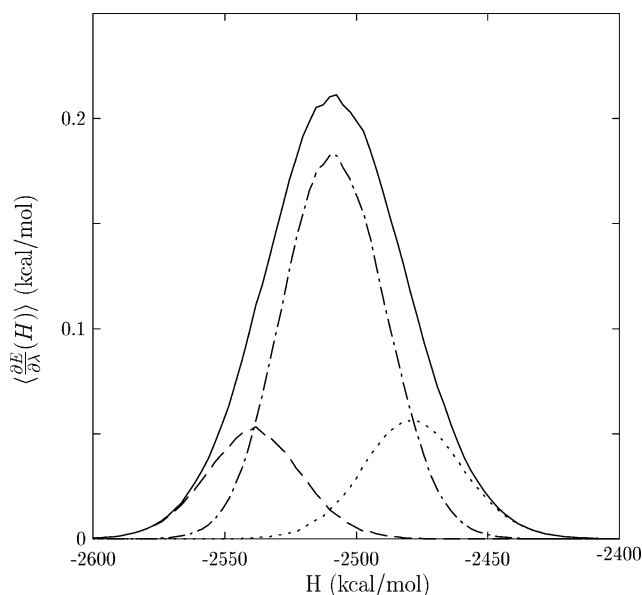


Figure 3. The function $\langle \partial E/\partial \lambda(H) \rangle$ from histogram reweighting (solid line) for butane at 298 K and $\lambda = 1$. Also shown are the reweighted histograms from CMD at 283 K (dashed line), 298 K (dot-dashed line), and 313 K (dotted line). The three reweighted histograms sum to give the solid line.

be like running CMD for 1/3 longer, and the error bars would only be smaller by a factor of $\sqrt{4/3}$ or 1.15 instead of $\sqrt{2}$. Additional improvements in precision are found because the histograms at other temperatures preferentially sample different regions of enthalpy than the 298 K simulation and therefore increase the precision of $\langle \partial E/\partial \lambda(H) \rangle_{k, \text{WHAM}}$ away from the peak. The error estimates of the histograms from CMD and from CMD-WHAM for butane at 298 K and $\lambda = 1$ shows that the errors are only slightly smaller for CMD-WHAM at the peak (around -2500 kcal/mol), because at the peak the CMD-WHAM histogram is primarily made up of the CMD histogram at 298 K. Away from the peak, the error bars are significantly smaller due to the contributions from the other temperatures. Note that the largest decrease in error bars with WHAM is for the 298 K data. For the other two temperatures the decrease is not as great because data at neighboring temperatures contributes the most to the reweighted histograms and these temperatures only have one neighboring temperature, whereas 298 K combines data from two nearby temperatures.

Parallel tempering improves the error bars of the calculations but not as significantly as other applications, in which PT improves sampling efficiency by an order of magnitude.^{20,51} In the study of Woods, Essex, and King which used combined PT and TI to calculate the free energy of converting a water to a methane molecule, PT (using 16 replicas, a larger number was needed in this study because it had about 6 times more molecules) reduced the error by a factor of 1.4, similar to what is found here.¹⁸ The sampling of the aqueous methane and butane systems does not appear to involve motion over large energy barriers. If it did, then parallel tempering would improve the error bars more significantly. Also, the improvement in using PT rather than CMD would be better at low temperatures, which does not appear to be the case in this study (see Table 2). The largest

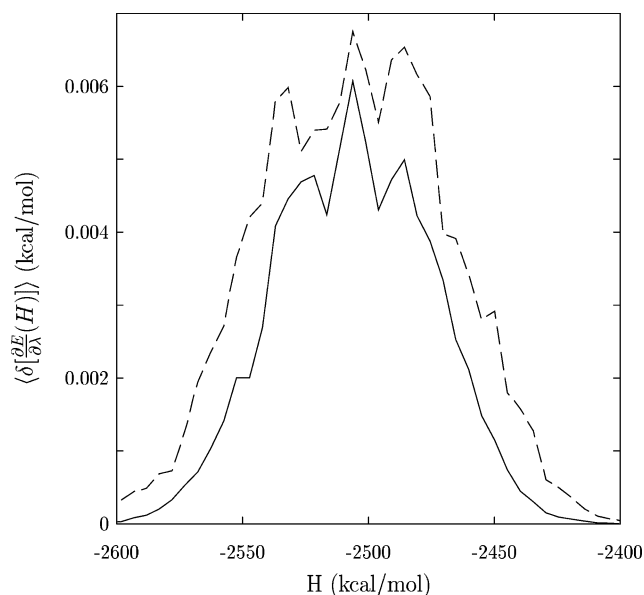


Figure 4. Error estimates of the $\langle \partial E/\partial \lambda(H) \rangle$ from histogram reweighting (solid line) and from conventional molecular dynamics (dashed line), for butane at 298 K and $\lambda = 1$.

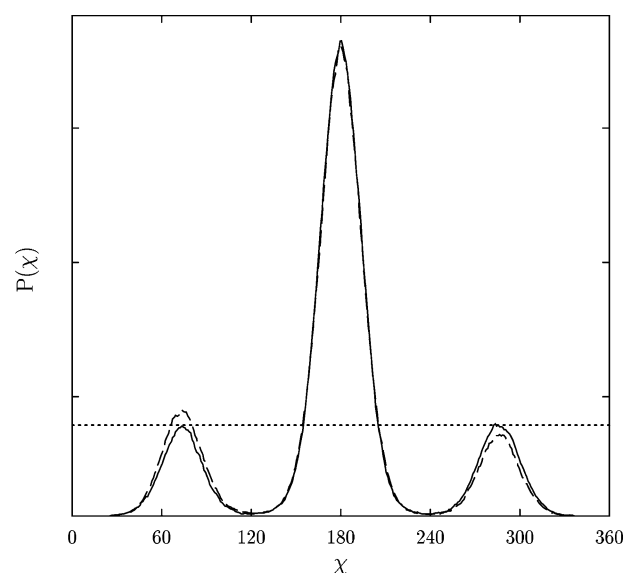


Figure 5. Distribution of the dihedral angle, χ , for butane at 283 K from 5 ns of simulations using parallel tempering (solid line) and conventional molecular dynamics (dashed line). The dotted line shows the value of the gauche peak from parallel tempering.

molecular barrier of these systems is the torsional angle of butane, involving the four carbon atoms, which a previous study has indicated may not be properly sampled over nanosecond simulations.³ The barrier for rotation in the OPLS-AA model is 3.68 kcal/mol.³⁹ In our simulations, even at the lowest temperature, the torsion angle, χ , is sampled fairly adequately (Figure 5). If the sampling were completely converged, the heights of both gauche peaks (around 70 and 390 degrees) would be the same, as they are for the PT simulations, but not quite for the CMD simulations at 283 K. At 298 K and 313 K, the distribution of χ appear to be completely converged. Even though there are differences between the distributions of χ at 283 K between CMD and

PT, they are not enough to indicate large sampling problems. In addition, Shirts et al. showed that the value of ΔG was not sensitive to the value of χ .³ These results all indicate that large barriers are not present in either of these free energy calculations. The improvements in sampling efficiency using PT is most likely not as much due to help in crossing over barriers as in providing independent trajectories.

Conclusion

The results show that notable improvements in the efficiency of calculating solvation free energies and entropies can be achieved for solvation free energies when using parallel tempering (PT) or the weighted histogram analysis method (WHAM). Error estimates when using PT and/or WHAM can be a factor or $\sqrt{2}$ less than those using conventional molecular dynamics (CMD), which means similar error bars can be achieved with simulation times half as long. The PT/WHAM error bars are lower than those of the extremely precise values of Shirts et al. which reports a ΔG of OPLS-AA butane of 3.10 ± 0.06 kcal/mol at 298 K (with 2σ error bars).³ The calculations did not reveal that solvating a methane or a butane molecule involved sampling over large barriers. For other systems which do involve crossing barriers, use of PT would result in larger increases in the efficiency of the ΔG calculations. The WHAM method uses data generated either from CMD or PT at other temperatures to generate the quantities $\langle \partial E_i / \partial \lambda \rangle_\lambda$, $\langle H \rangle_\lambda$, and $\langle H \partial E_i / \partial \lambda \rangle_\lambda$ needed to calculate ΔG and ΔS . For the temperatures and system sizes used here, there is enough overlap of the enthalpy histograms (Figure 1) to give good improvement in the precision of the free energy calculations. Both PT and WHAM are computationally inexpensive, and if simulations at additional temperatures were required to find entropy and enthalpy changes,^{4–10} then using PT or WHAM does not add additional simulation time. The WHAM method is particularly easy to implement, simply requiring the calculation of two one-dimensional histograms, and can be easily combined with standard simulation programs.

Acknowledgment. This material is based upon work supported by the American Chemical Society Petroleum Research Fund under grant number 40212-GB6. A donation of hurricane relief computer time from Dr. Rigoberto Hernandez (through the National Science Foundation CRIF grant CHE 04-43564) is gratefully acknowledged.

References

- Beveridge, D. L.; DiCapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431.
- Kollman, P. A. *Chem. Rev.* **1993**, *93*, 2395.
- Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740.
- Guillot, B.; Guissani, Y.; Bratos, S. *J. Chem. Phys.* **1991**, *95*, 3643.
- Smith, D. E.; Haymet, A. D. J. *J. Chem. Phys.* **1993**, *98*, 6445.
- Rick, S. W. *J. Phys. Chem. B* **2000**, *104*, 6884.
- Lüdemann, S.; Schreiber, H.; Abseher, R.; Steinhauser, O. *J. Chem. Phys.* **1996**, *104*, 286.
- Shimizu, S.; Chan, H. S. *J. Chem. Phys.* **2000**, *113*, 4683.
- Rick, S. W. *J. Phys. Chem. B* **2003**, *107*, 9853.
- Olano, L. R.; Rick, S. W. *J. Am. Chem. Soc.* **2004**, *126*, 7991.
- Geyer, C. J. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*; Interface Foundation: Fairfax Station, 1991; p 156.
- Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19*, 451.
- Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604.
- Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.
- Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195.
- Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011.
- Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239.
- Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13703.
- Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042.
- Yamamoto, R.; Kob, W. *Phys. Rev. E* **2000**, *61*, 5473.
- Hernández-Cobos, J.; Mackie, A. D.; Vega, L. F. *J. Chem. Phys.* **2001**, *114*, 7527.
- Chang, J.; Sandler, S. I. *J. Chem. Phys.* **2003**, *118*, 8390.
- Mitsutake, A.; Okamoto, Y. *J. Chem. Phys.* **2004**, *121*, 2491.
- Doxastakis, M.; Mavrantzas, V. G.; Theodorou, D. N. *J. Chem. Phys.* **2001**, *115*, 11352.
- Lyubatshev, A. P.; Martinsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776.
- Smith, D. E.; Zhang, L.; Haymet, A. D. J. *J. Am. Chem. Soc.* **1992**, *114*, 5875.
- Guillot, B. *J. Chem. Phys.* **1991**, *95*, 1543.
- Yu, H.; Karplus, M. *J. Chem. Phys.* **1988**, *89*, 2366.
- Mezei, M.; Beveridge, D. L. *Ann. N. Y. Acad. Sci.* **1986**, *482*, 1.
- Chialvo, A. A. *J. Chem. Phys.* **1990**, *92*, 673.
- Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1994**, *100*, 9025.
- Rick, S. W.; Berne, B. J. *J. Am. Chem. Soc.* **1996**, *118*, 672.
- Lüdemann, S.; Abseher, R.; Schreiber, H.; Steinhauser, O. *J. Am. Chem. Soc.* **1997**, *119*, 4206.
- Bedrov, D.; Smith, G. D. *J. Chem. Phys.* **2001**, *115*, 1121.
- Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116*, 9058.
- Jang, S.; Shin, S.; Pak, Y. *Phys. Rev. Lett.* **2003**, *91*, 058305.
- Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106*, 6638.
- Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.
- Allen, L. C. *Acc. Chem. Res.* **1990**, *23*, 175.

- (41) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384.
- (42) Ciccotti, G.; Ryckaert, J. P. *Comput. Phys. Rep.* **1986**, *4*, 345.
- (43) Martyna, G. J.; Tobias, D. J.; Klein, M. L. *J. Chem. Phys.* **1994**, *101*, 4177.
- (44) Nosé, S. *Mol. Phys.* **1984**, *52*, 255.
- (45) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.
- (46) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: from Algorithms to Applications*; Academic Press: San Diego, CA, 1996.
- (47) Ben-Naim, A.; Marcus, Y. *J. Chem. Phys.* **1984**, *81*, 2016.
- (48) Jorgensen, W. L.; Blake, J. F.; Buckner, J. K. *Chem. Phys.* **1989**, *129*, 193.
- (49) Guillot, B.; Guissani, Y. *J. Chem. Phys.* **1993**, *99*, 8075.
- (50) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637.
- (51) Yan, Q.; de Pablo, J. J. *J. Chem. Phys.* **2000**, *113*, 1276.

CT0502070

JCTC

Journal of Chemical Theory and Computation

The Structure of Liquid Benzene

Christopher M. Baker^{†,§} and Guy H. Grant^{*,‡}

Department of Chemistry, Physical and Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford, U.K. OX1 3QZ, and Unilever Centre for Molecular Informatics, The University Chemical Laboratory, Lensfield Road, Cambridge, U.K. CB2 1EW

Received January 17, 2006

Abstract: The interactions of aromatic groups have been identified as playing a crucial role in many systems of interest. Unfortunately, conventional atom-centered force fields provide only an approximate representation of these molecules owing to their failure to consider the quadrupole moment arising from the π electrons. In this paper the structure of liquid benzene, the prototypical aromatic system, is investigated using a novel approach to Monte Carlo simulation, parametrized against experimental thermodynamic data, which incorporates an explicit representation of the aromatic π electrons. In contrast to previous simulations of liquid benzene it is found that a perpendicular arrangement of benzene molecules is preferred to a parallel arrangement. This result is in good agreement with experimental data.

Introduction

In recent years interactions involving aromatic residues have been shown to be of crucial significance in a number of important problems including protein–ligand binding,^{1,2} the determination of protein structure,³ and DNA base stacking.⁴ As the realization of the importance of aromatic interactions grows, so too does the requirement for accurate model potentials which can reproduce these experimental observations. The charge separation model of Hunter and Sanders⁵ is one such model and has received much attention as a simple and physically reasonable method for modeling these interactions. It represents the aromatic π electrons as a series of explicit points lying in two planes above and below the aromatic C atoms. While it has been successfully applied in a variety of situations, varying from porphyrin rings⁵ to aromatic amino acids⁶ and molecular clips,⁷ it has not been used to investigate the nature of molecular liquids such as benzene. As the prototypical case of the aromatic interaction,

a thorough understanding of the characteristics of the benzene molecule is essential if we are to progress to modeling more complicated systems involving aromatic interactions. Here we present such an application, giving an improved parametrization of the model which reproduces the experimental properties of liquid benzene as well as providing insights into the intermolecular geometries that give rise to these properties at a molecular level.

Background

Although the isolated benzene dimer has received much attention, both theoretically and experimentally, there is still much debate as to the true structure of its global energy minimum. The two candidates are those structures that would be anticipated given the quadrupolar nature of the benzene molecule: a parallel displaced (PD) structure and a T-shaped (TS) structure in which one molecule lies perpendicular to the second, forming a hydrogen bond to the π system. In reality this is not a hydrogen bond in the conventional sense; when the dimer is formed, the C–H bond length shortens to allow for the maximization of the favorable quadrupole–quadrupole interactions. As such, it has been termed an ‘anti-hydrogen bond’.⁸

In the early days of theoretical calculations on the benzene dimer, it was generally believed that the structure of the dimer was T-shaped^{9,10} or a slightly distorted T-shaped

* Corresponding author e-mail: ghg24@cam.ac.uk.

[†] Department of Chemistry, Physical and Theoretical Chemistry Laboratory, University of Oxford.

[‡] The University Chemical Laboratory.

[§] Current address: Unilever Centre for Molecular Informatics, The University Chemical Laboratory, Lensfield Road, Cambridge, U.K. CB2 1EW.

structure.¹¹ As new experimental evidence came to light,¹² the structure of the benzene dimer received new focus, and, performing calculations at the MP2/6-31+G* level of theory, Hobza et al.¹³ identified that the parallel displaced structure was actually an energetic minimum and that it even lay lower in energy than the T-shaped structure. Jaffe and Smith,¹⁴ again working with the MP2 theory, also favored the PD structure, concluding that the TS structure was not in fact a minimum at all but rather a saddle point on the transition between two parallel displaced structures. The complexity of the problem was well illustrated when, within a few months, Hobza et al.¹⁵ presented a study using the CCSD(T) theory, which concluded that both the PD and TS structures were true minima, and almost isoenergetic, but with the TS structure lying marginally lower in energy. Since this time, the debate had bounced back and forth between the two competing structures, Gonzalez and Lim¹⁶ concluded that the TS dimer is marginally lower in energy than the PD dimer, although their work was limited by the small size of their basis sets. Hobza et al. used CCSD(T) calculations to parametrize the NEMO model,¹⁷ suggesting that only one minimum, a TS structure, is present and that the PD structure is actually a transition state.¹⁸ Tsuzuki et al.^{19,20} have concluded that the two dimer structures are approximately isoenergetic,¹⁹ though CCSD(T) calculations reveal that the PD structure is slightly lower in energy.²⁰ The authors conclude, however, that these calculations are likely to overestimate the attraction in the PD case and that, in reality, the TS structure may be lower in energy. The most recent and rigorous calculations on this system by Sinnokrot et al.,^{21,22} performed using CCSD(T), also conclude that the two dimer structures are isoenergetic.

The results from experimental studies on the benzene dimer are just as inconclusive as those from theoretical calculations. The earliest experimental studies were performed using molecular beams and concluded that the benzene dimer is polar.^{23,24} The authors interpreted this to mean that the molecules adopt a T-shaped arrangement, as is found in the solid.²⁵ This view was also backed up by experiments performed using resonant two photon ionization (R2PI) techniques,²⁶ ionization-detected stimulated Raman spectroscopy (IDSRS),^{27,28} and rotational spectroscopy.²⁹

The experimental evidence favoring the TS dimer, however, is far from conclusive. Various vibronic spectra of isotopically substituted benzenes have been measured;^{30–33} all of these studies conclude that the structure of the dimer is symmetric, precluding the TS structure, but not the PD structure, which has been suggested by Bernstein et al.^{30,31} Additionally, Schlag et al. have proposed from these results a 'V-shaped' dimer structure.^{32,33} In addition to the IDSRS experiments performed by Henson et al.,^{27,28} the same technique has been used by Ebata et al.,³⁴ who came to the conclusion that two isomers exist, having center of mass separations of 3.6 Å and 5.0 Å, which would correspond to the PD and TS dimers, respectively. The conclusion that more than one dimer structure is present was also reached by Scherzer et al.,¹² who found that at least two dimers exist.

It is clear that the theoretical and experimental study of the benzene dimer has given results for its structure which

are, thus far, inconclusive. However, there are some broad conclusions that can be drawn: 1. The PD and TS structures lie very close in energy. 2. The potential energy surface for the benzene dimer is very flat in the region around the minima.^{19,35} 3. In reality the benzene dimer is likely to be highly fluxional, constantly moving between the two structures.²¹

Although the structure of the benzene dimer has received much attention, and the structure of solid benzene is well defined,²⁵ what is less well understood is the structure of the liquid phase of benzene. Atom-centered force field simulations have suggested that the liquid is comprised of well-defined solvation spheres around each molecule but that within each sphere there is either no orientational preference for the individual molecules³⁶ or a very slight preference for the orthogonal arrangement.³⁷ Evidence from X-ray diffraction³⁸ and neutron scattering³⁹ experiments, however, as well as recent experimental results⁴⁰ from optical Kerr effect spectroscopy⁴¹ all conclude that the local ordering in liquid benzene is perpendicular.

Conventional all atom force fields perform well in many situations, for example in the reproduction of the dipole moment of molecules, but actually provide a poor description of the electronic distribution in the benzene molecule. In benzene, the delocalized π orbitals above and below the plane of the ring contain substantial amounts of electron density that give rise to a quadrupole moment, the first nonzero multipole moment present in the benzene molecule. This quadrupole moment is completely neglected by atom-centered approaches. Hunter and Sanders⁵ proposed that this charge distribution could be accounted for by placing two points above and below each C atom in the ring, each having a negative charge but no volume, to represent the π electrons. This approach, termed charge separation, has been applied widely for the inclusion of lone pairs in, for example, water⁴² and sulfur in proteins⁴³ and has been found to perform well.

In this work we will begin by presenting a brief study of the benzene dimer, which will illustrate the effect of the charge separation model on a simple aromatic system. We will then move on to consider the case of liquid benzene and examine how the new model affects previous ideas on the structure of the liquid at a molecular level.

Methods

Ab Initio Calculations. In parametrizing their model Hunter and Sanders⁵ proposed that each C atom would contribute one electron to the π system, meaning that each π point would have a charge, q_π , of -0.50 e. The value of the separation between the nuclear site and the π point, δ , was determined by fitting to the gas-phase quadrupole moment of the benzene molecule, to give a value of 0.47 Å.

To consider the effect of charge separation on the benzene dimer we adopted a different approach, parametrizing the model via a comparison with ab initio data. Tran et al.⁴⁴ identified 10 minimum energy conformations of the benzene dimer, labeled a–j. The energies of these structures were recalculated at the MP2/6-311+G** level of theory. Although CCSD(T) methods have become the de facto method of choice for benzene dimer calculations, it has been shown⁴⁵

that MP2 methods employing medium sized basis sets give very good results, due to a fortuitous cancellation of errors, at much reduced computational cost. All ab initio calculations were performed using the Gaussian98⁴⁶ program, and all calculated energies were corrected for basis set superposition error (BSSE)⁴⁷ using the counterpoise method.⁴⁸

With the ab initio energies calculated, the equivalent energies were calculated for the same 10 structures using a charge separation force field in which the values of q_π and δ were varied from 0 to $-2e$ and 0 to 1 Å, respectively. The ability of the force field to reproduce the ab initio calculated energies at each set of parameters was measured via eq 1.

$$\Delta E = \sum_{n=a}^j (E_n^{\text{MP2}} - E_n^{\text{CS}})^2 \quad (1)$$

Monte Carlo Calculations. Previous studies have found that parametrizing force fields by fitting to ab initio data in vacuo is inappropriate for modeling condensed phases in general⁴⁹ and aromatic interactions in solution in particular,⁵⁰ and this observation was found to hold true in this case, with the ab initio derived parameters performing poorly for the case of the liquid simulations (results not shown). To surmount this problem when parametrizing the original OPLS all atom model for liquid simulations of benzene Jorgensen and Severance⁵⁴ fitted their parameters to experimental thermodynamic and structural data. In this instance, a similar approach has been adopted, with the additional constraint that the model must reproduce the experimental value of the benzene quadrupole moment.

Because OPLS is an effective potential, any properties not explicitly accounted for in the model will have been ‘mixed into’ the model during the parametrization process. This means that rather than just adding extra points so that they reproduce the correct quadrupole moment, it is necessary to reexamine all of the parameters within the system. To do this we have followed a methodology similar to that used in the parametrization of the TIP5P⁴² water model. The bond lengths and angles used are the experimentally derived values for the isolated benzene molecule.⁵⁴ To these we then add a series of charge and van der Waals parameters. The charge parameters are subject to the constraints that the individual molecules must be charge neutral and that $q_\pi = -q_H$ where q_π and q_H are the charges on the π electron points and H atoms, respectively. q_π and δ were then varied systematically along with the van der Waals parameters, σ and ϵ , until the models give the minimum deviation from experimental thermodynamic results. For the purpose of parametrization a series of Monte Carlo simulations including 267 benzene molecules in the NPT ensemble (with $P=1$ atm and $T=298$ K) was performed. Each simulation consisted of 6.0×10^7 steps of equilibration followed by 6.0×10^7 steps of averaging.

With the necessary parameters in place three Monte Carlo simulations of liquid benzene were performed. The first treated the benzene molecules using a 12 site model, the OPLS all atom potential⁵¹ (denoted OPLS), the second used a 24 site model consisting of the OPLS all atom potential

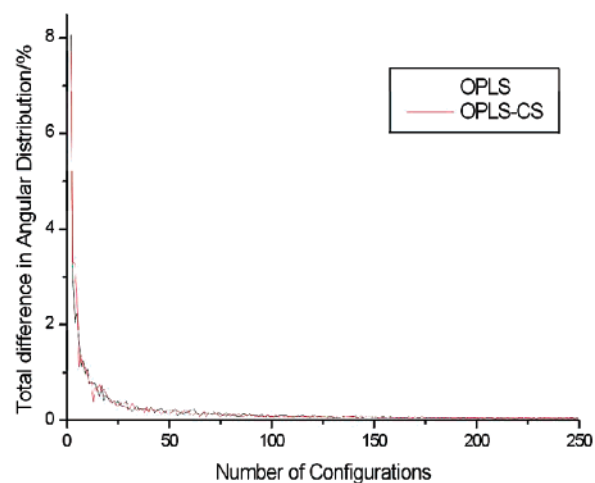


Figure 1. Convergence of angular distributions.

modified in such a way as to incorporate the π electron points of the charge separation model (denoted OPLS-CS), and the third used the original charge separation model of Hunter and Sanders (HS). All simulations were performed using BOSS version 4.2;⁵² in the OPLS simulation the standard OPLSAA parameters were used, in the OPLS-CS simulation the OPLSAA parameters were modified so as to incorporate the π parameters described above, and in the HS simulation the parameters used were those obtained in the original work by Hunter and Sanders.⁵ In all cases a system consisting of 267 benzene molecules was used in simulations that were run in the NPT ensemble with $T = 298$ K and $P = 1.0$ atm. The simulations were begun from a configuration in which all of the benzene molecules were arranged in a parallel fashion, and in all simulations 4×10^6 equilibration steps were performed followed by 2.5×10^8 steps of averaging. The orientational distributions shown in Figure 7 were calculated as the average of configurations extracted from the simulation every 1×10^6 steps, which was found to be sufficient for the angular distributions to have converged to their limiting values (Figure 1).

Results and Discussion

Benzene Dimer: Ab Initio Calculations. The results of the parametrization against ab initio data can be seen in Figure 2.

From this we can see that the best parameter values lie some way from those of an all atom force field at $q_\pi = 0e$, $\delta = 0$ Å and also from those used by Hunter and Sanders at $q_\pi = -0.5e$ and $\delta = 0.47$ Å. The values obtained from this parametrization are $q_\pi = -0.30e$ and $\delta = 0.30$ Å.

It has previously been shown that atom-centered force fields perform badly when modeling T-shaped structures of the benzene dimer.¹⁶ As an illustration of the improvement that can be brought about via the use of the charge separation model Figure 3 shows potential energy surfaces calculated for the region around the T-shaped minimum. In all cases, the surfaces have been calculated by keeping the z separation of the molecules fixed at their equilibrium separation of 4.9 Å²² and scanning over the x and y directions, calculating the energy every 0.2 Å.

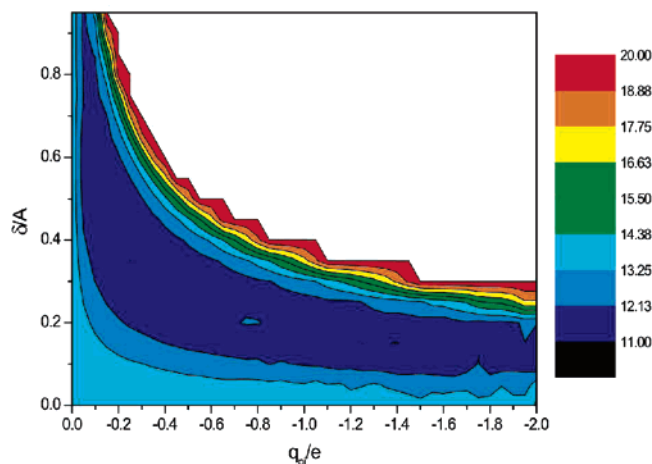


Figure 2. Parametrizing the CS model against ab initio data.

The general shape of the surface produced using the charge separation model is much closer to the ab initio surface than is the surface calculated using an all atom approach. Although the agreement between the charge separation and ab initio surfaces is not quantitative, the charge separation model does give qualitatively correct results. Since the objective of this work is not the accurate reproduction of benzene dimer energies, we have not chosen to refine this model further but have rather considered these data to be evidence that the use of explicit points to represent aromatic π electrons can improve the representation of the benzene molecule. As such, rather than providing a definitive solution to the problem, these calculations demonstrate the potential of the charge separation approach, and recommend it for further study.

Liquid Benzene: Monte Carlo Simulations

From the initial simulations performed, a set of parameters was determined as the best OPLS-CS model. These values are listed in Table 1, and the comparison with the thermodynamic data from experiment can be seen in Table 2.

The agreement between the OPLS-CS calculated and experimental values is generally at least as good as that of the OPLS model and offers a large improvement in terms of the reproduction of the quadrupole moment of the molecule. Of all the models employed, the Hunter and Sanders model performs worst in reproducing the experimental thermodynamic properties of the liquid, providing a good representation of the quadrupole moment but overestimating the attraction between the molecules. That this model performs badly is perhaps no great surprise. It was never developed with the simulation of liquids in mind and indeed was not intended to treat benzene at all, actually being developed for the treatment of porphyrins.⁵

As a first measure of the structure of the liquid we can consider the center of mass radial distribution functions, $g_{\text{CMCM}}(r)$ (Figure 4). These distributions provide little information about the detailed structure within the liquid but do provide information on the size and number of molecules within the first solvation shell. Furthermore, by comparing the calculated values to experimental results obtained from neutron diffraction,³⁹ we can begin to judge the quality of

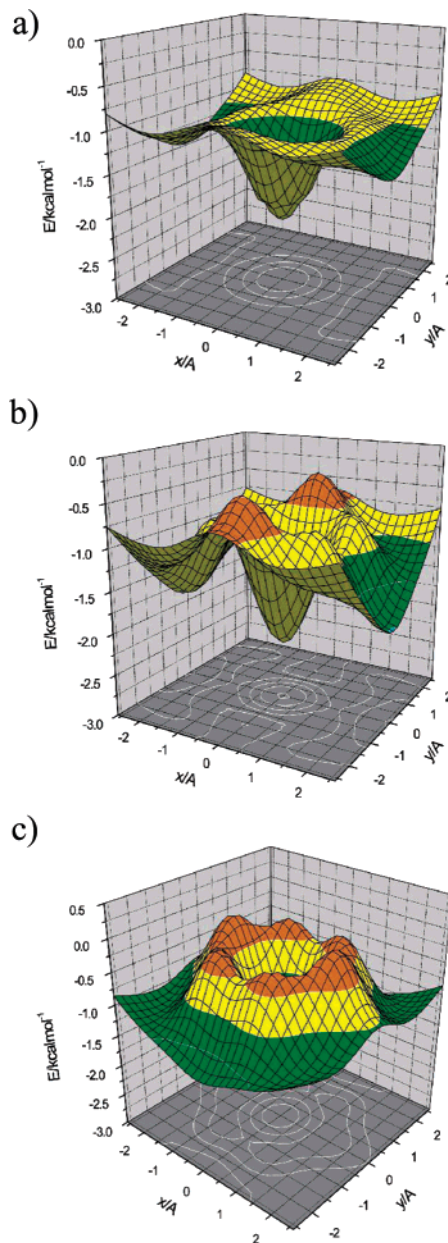


Figure 3. Potential energy surface around the T-shaped minimum, calculated using (a) MP2/6-311+G**, (b) the CS model, and (c) an all atom potential.

our potentials. In this case, both the OPLS and OPLS-CS models perform reasonably well in terms of reproducing the general shape and position of the distribution, with the OPLS-CS model more accurately predicting the height of the first peak and the OPLS model its slope.

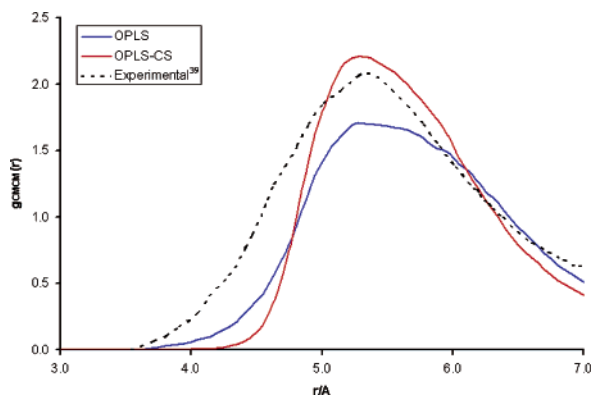
The relative orientation of molecules in liquid benzene can be analyzed more closely via consideration of the constituent radial distribution functions, $g(r)$. The experimental $g(r)$ have been determined by X-ray diffraction,³⁸ and the same functions have also been calculated as a result of several simulations using Monte Carlo⁵⁴ and molecular dynamics^{35,37,55,56} techniques. In this study we have calculated the radial distribution functions using both the OPLS and OPLS-CS potentials, with $g_{\text{CC}}(r)$ also evaluated for the HS model. The calculated $g(r)$ are shown in Figure 5.

Table 1. Parameters Used in OPLS-CS and OPLS Benzene Models

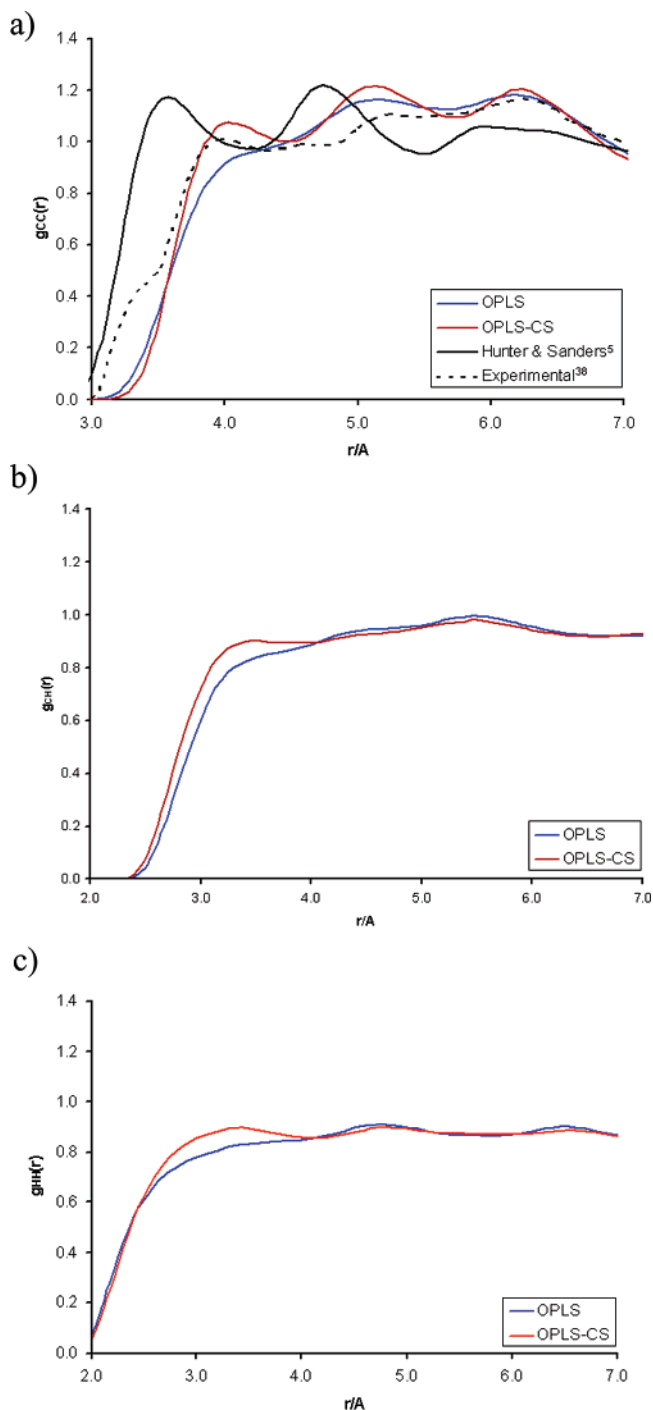
parameter	OPLS-CS	OPLS
$R_{CC}/\text{\AA}$	1.40	1.40
$R_{CH}/\text{\AA}$	1.08	1.08
$\delta/\text{\AA}$	0.90	n/a
$\theta_{CCC}/^\circ$	120.0	120.0
$\theta_{CCH}/^\circ$	120.0	120.0
$\theta_{CC\pi}/^\circ$	90.0	n/a
$\Psi_{CCCC}/^\circ$	0.0	0.0
$\Psi_{CCCH}/^\circ$	180.0	180.0
$\Psi_{CC\pi}/^\circ$	90.0/−90.0	n/a
q_C/e	0.1435	−0.115
q_H/e	0.1435	0.115
q_π/e	−0.1435	n/a
$\sigma_C/\text{\AA}$	3.69	3.55
$\sigma_H/\text{\AA}$	2.52	2.42
$\epsilon_C/\text{kcalmol}^{-1}$	0.07	0.07
$\epsilon_H/\text{kcalmol}^{-1}$	0.03	0.03

Table 2. Thermodynamic Properties of Liquid Benzene

	OPLS	OPLS-CS	HS	experiment ⁵⁴
quadrupole/ ea_0^2	0.00	−6.7	−6.4	−6.7 ⁵³
dipole/ ea_0	0.00	0.00	0.00	0.00
density/ g cm^{-3}	0.865	0.872	0.997	0.874
$\Delta H_{\text{vap}}/\text{kcalmol}^{-1}$	7.89	7.58	25.50	8.09
C_p/k_B	15.0	15.9	10.91	15.5
molecular volume/ \AA^3	149.8	148.7	130.1	148.4

**Figure 4.** $g_{\text{CCM}}(r)$ for liquid benzene.

Of these radial distribution functions, the most useful in terms of elucidating structural information is $g_{\text{CC}}(r)$. The HS model performs poorly in reproducing the experimental $g_{\text{CC}}(r)$, the first peak is found at too small a distance, and the height of the peaks is far larger than that observed experimentally. The model seems to be predicting an excessively solidlike structure. Although the OPLS potential provides a reasonable reproduction of the experimental data, the first peak in the experimental $g_{\text{CC}}(r)$ is not well reproduced, instead being merged into the second peak. With the OPLS-CS model, however, we see a better reproduction of this first peak within the experimental $g_{\text{CC}}(r)$. It follows that this difference in the $g_{\text{CC}}(r)$ values must be related to a structural difference within the liquid. To investigate this difference, we have performed a geometrical analysis on the results of

**Figure 5.** Radial distribution functions for liquid benzene: (a) $g_{\text{CC}}(r)$ (b) $g_{\text{CH}}(r)$, and (c) $g_{\text{HH}}(r)$.

these two simulations. For each molecule within the system, we have extracted the coordinates of every molecule that lies within the first solvation shell, defined by analysis of $g_{\text{CCM}}(r)$ as having an intermolecular centroid distance less than 7.7 \AA , and then for each pair of molecules calculated the angle between the vectors normal to the planes of the two rings (Figure 6). The resulting orientational distributions can be seen in Figure 7.

While the OPLS simulation gives a sinusoidal distribution, indicating that there is an isotropic arrangement of molecules and hence no preference for either of the two energetic minima, the OPLS-CS distribution deviates significantly from

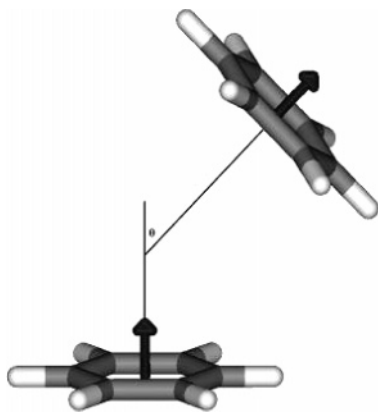


Figure 6. Calculation of θ , the angle between the normals to two benzene molecules.

the sinusoidal shape, revealing a preference for the orthogonal arrangement over the parallel arrangement, in effect the OPLS-CS model predicts a much more solidlike structure than does the OPLS (Figure 7). This idea, that the orthogonal arrangement is preferred, can also be seen in the first solvation shell of a molecule taken from the simulation (Figure 8).

It is informative to consider both the radial and orientational distributions as sources of structural information, and the process can be refined further via the consideration of angular distribution functions (Figure 9), which consider simultaneously both the radial and orientational dependence of the molecular structure.

The first peak in $g(r, \theta)$ represents the first solvation shell of the molecule, and it is clear to see that in the OPLS simulations, there is a small preference for a perpendicular arrangement of the molecules. This result is in good agreement with angular distribution functions calculated from previous simulations using atom-centered potentials.^{36,57,58,59} In contrast, $g(r, \theta)$ obtained from the OPLS-CS simulations shows a clear preference for the perpendicular arrangement of molecules within the first solvation shell.

Such a result has been predicted theoretically for quadrupolar fluids. Streett and Tildesley⁶⁰ performed molecular dynamics simulations on an idealized diatomic liquid, both with and without the inclusion of quadrupole–quadrupole interactions. When these interactions were omitted from their simulation, it was found that the molecules exhibited no orientational preference. Once quadrupolar interactions were switched on, however, the structure of the liquid showed a clear preference for a T-shaped orientation of the molecules. In the case of benzene, conventional all atom force fields do not account for the quadrupole moment that arises from the π electron clouds above and below the plane of the ring. By incorporating the charge separation model we have reproduced this quadrupole moment within the benzene molecule, and the theoretically predicted behavior has been recovered. Such a result has also been seen in the case of liquid bromine, where Monte Carlo simulations including quadrupolar interactions were found to predict more accurately experimental results than those without, and also favored a T-shaped arrangement of the molecules.⁶¹ Furthermore, Brown and Swinton⁶² found that in the prediction

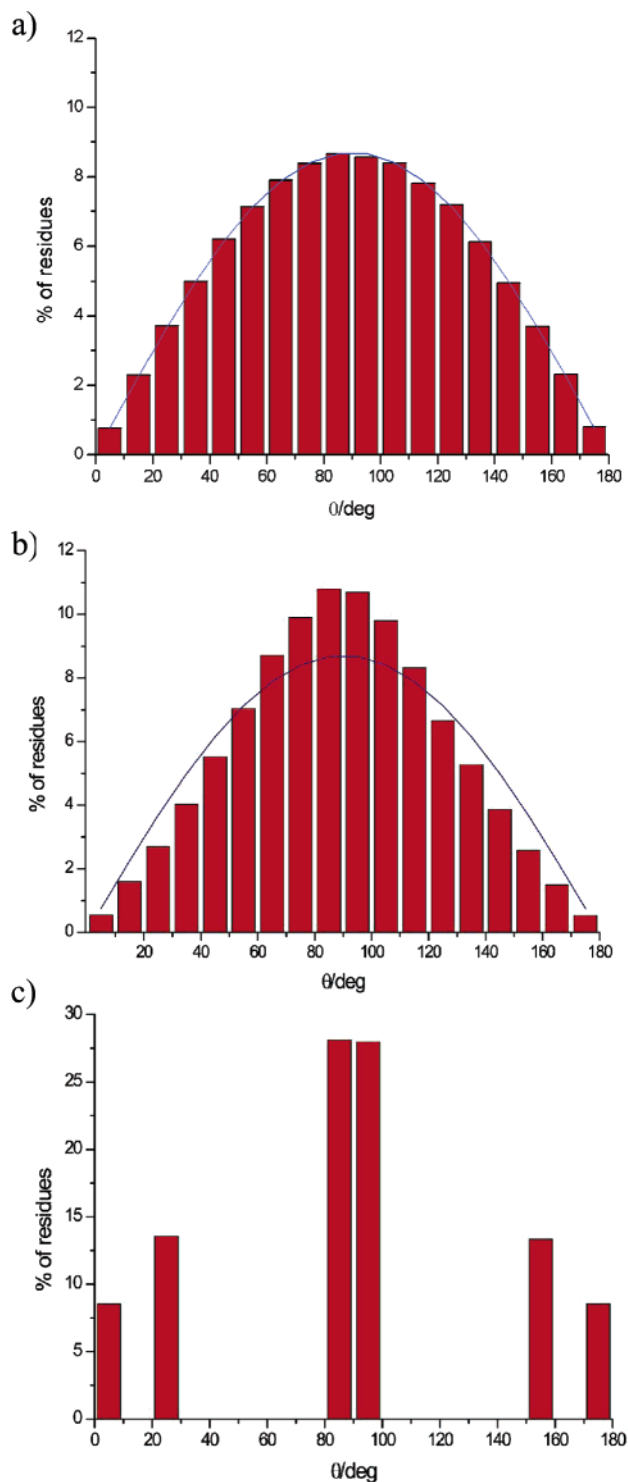


Figure 7. Angular distributions of benzene molecules within the first solvation shell of benzene, obtained from (a) the OPLS and (b) OPLS-CS simulations as well as (c) solid benzene.²⁵ 0° corresponds to a parallel structure and 90° to a perpendicular structure.

of the structures of solid benzene and hexafluorobenzene, the quadrupole moment was the most important factor.

These results, when combined with the experimental data available, support the view that the structure of liquid benzene is well ordered with an orthogonal arrangement of benzene molecules existing wherever possible. The parallel

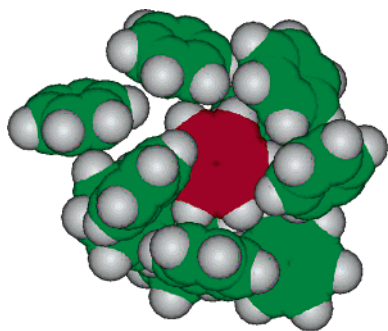


Figure 8. First solvation shell of a single benzene molecule (in red), taken from the OPLS-CS simulation.

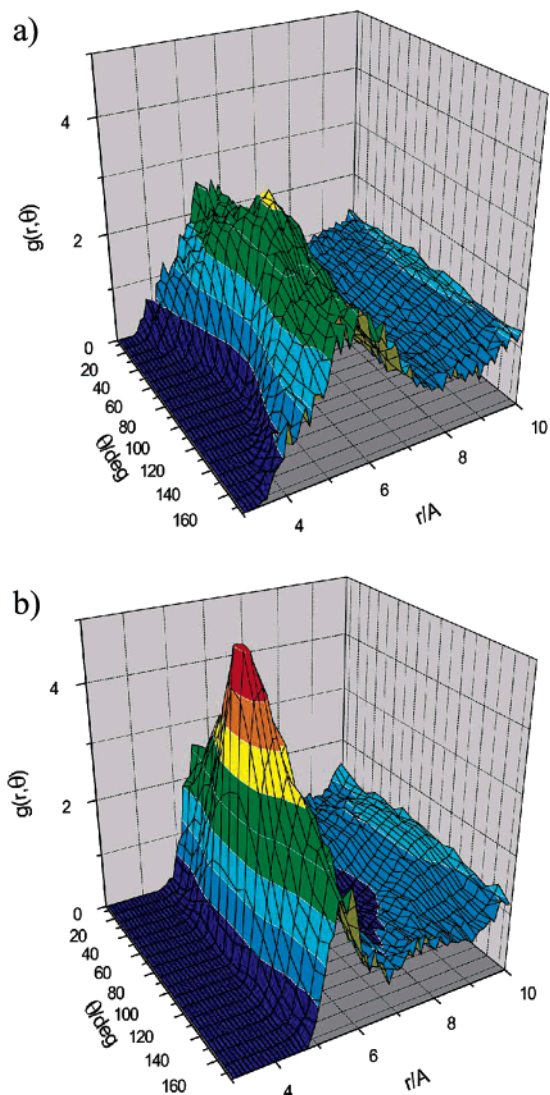


Figure 9. Angular distribution functions: (a) OPLS and (b) OPLS-CS.

displaced structure still features in the liquid phase but is less common than previously believed.

These results are also in good agreement with a variety of other condensed phase results, in which it has been found that aromatic–aromatic interactions tend to favor the T-shaped structure over the parallel displaced structure. Examples of such cases include solid benzene, in which 8 out of the 12 nearest neighbors of any molecule are found to be

orthogonal,²⁵ and proteins,⁶³ in which aromatic–aromatic interactions are believed to be an important factor in the determination of the structure a protein adopts.⁶⁴

Although this study suggests that the effect of including the quadrupole moment of the benzene molecule is significant, we must be aware of the fact that OPLS-CS (like OPLS) is an effective potential. While the basis of the OPLS model is physical, for example using experimental geometries, the simplicity of the model means that fitting to experimental data is necessary if we are to achieve an accurate reproduction of experimental data, and the model becomes an ‘effective’ potential. The result of this is that the original physical characteristics lose their precise meanings, and any properties not explicitly accounted for are ‘mixed into’ the model. In the case of the OPLS-CS model, one of the physical properties that we originally considered was the quadrupole moment of benzene, and it follows that the parametrization process will have resulted in the ‘mixing in’ of other properties into this term. Thus, although the only physical addition we have made to the model is that of the quadrupole moment, the effects that we are seeing may also be arising from properties other than the quadrupole. This, however, is a problem inherent to any effective potential and the ability of the OPLS-CS potential to reproduce the thermodynamic data gives us confidence that it is a reasonable potential, but it is important to be aware of the possibility that some of the effects that we observe may not be due entirely to the inclusion of the quadrupole.

When considering potentials for simulation of a molecular liquid, it is wise to be aware of the development of potential functions for liquid water, by far the most intensively studied of all molecular liquids. For 20 years from the early 1980s simple three site models of water, such as TIP3P,⁶⁵ were the methods of choice for molecular mechanics simulation. In 2000 Mahoney and Jorgensen⁴² demonstrated that a physically intuitive 5 site model (with the extra sites located at the O lone pair sites) offered a considerable improvement in terms of the reproduction of both thermodynamic and structural data. Water models have also increased significantly in both sophistication and accuracy via the inclusion of, for example, polarizability⁶⁶ or diffuse charges⁶⁷ into the potential. Over the same period the models of aromatic groups available within the commonly used force fields has remained at the level of an all atom potential. We would acknowledge that there is still work to be done before the available models of benzene reach the same level of sophistication as those of water but view this work as a necessary step toward that goal and a step that is readily compatible with existing molecular mechanics methodologies.

Conclusions

The charge separation model of Hunter and Sanders⁵ has been reparametrized to model the liquid phase of benzene by fitting to experimental thermodynamic data. This model has then been applied to the study of the structure of liquid benzene via Monte Carlo simulation and has been shown to offer a better reproduction of experimental results than a conventional all atom force field. The charge separation

model indicates that the structure within the first solvation shell of liquid benzene is largely perpendicular, in agreement with several experimental studies but in contrast to previous molecular mechanics based calculations.

The agreement between the experimental and calculated results, though improved, is still not perfect, and, if we have learned from the case of water, models of increasing sophistication will be required before we can truly hope to model the full range of aromatic interactions with complete confidence. This work might be considered to be only a first step toward that goal, but the development of a new force field that demonstrates an improved ability to treat the interactions of aromatic molecules bodes well for the study of many important biological systems.

Acknowledgment. C.M.B. thanks the National Foundation for Cancer Research for funding and Prof. W. L. Jorgensen for his generous provision of the BOSS program.

References

- (1) Zacharias, N.; Dougherty, D. A. *Trends Pharmacol. Sci.* **2002**, *23*, 281.
- (2) Sussman, J. L.; Harel, M.; Frolow, F.; Oefner, C.; Goldman, A.; Toker, L.; Silman, I. *Science* **1991**, *253*, 872.
- (3) Vondráček, J.; Bendová, L.; Klusák, V.; Hobza, P. *J. Am. Chem. Soc.* **2005**, *127*, 2615.
- (4) Hunter, C. A. *Philos. Trans. R. Soc. London, Ser. A* **1993**, *345*, 77.
- (5) Hunter, C. A.; Sanders, J. K. M. *J. Am. Chem. Soc.* **1990**, *112*, 5525.
- (6) Hunter, C. A.; Singh, J.; Thornton, J. M. *J. Mol. Biol.* **1991**, *218*, 837.
- (7) Reek, J. N. H.; Priem, A. H.; Engelkamp, H.; Rowan, A. E.; Elemans, J. A. A. W.; Nolte, R. J. M. *J. Am. Chem. Soc.* **1997**, *119*, 9956.
- (8) Hobza, P.; Špirko, V.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem. A* **1998**, *102*, 2501.
- (9) Karlström, G.; Linse, P.; Wallqvist, A.; Jönsson, B. *J. Am. Chem. Soc.* **1983**, *105*, 3777.
- (10) Čárský, P.; Selzle, H. L.; Schlag, E. W. *Chem. Phys.* **1988**, *125*, 165.
- (11) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Chem. Phys.* **1990**, *93*, 5893.
- (12) Scherzer, W.; Kratzschmar, O.; Selzle, H. L.; Schlag, E. W. *Z. Naturforsch. A* **1992**, *47*, 1248.
- (13) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1993**, *97*, 3937.
- (14) Jaffe, R. L.; Smith, G. D. *J. Chem. Phys.* **1996**, *105*, 2780.
- (15) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1996**, *100*, 18790.
- (16) Gonzalez, C.; Lim, E. C. *J. Phys. Chem. A* **2000**, *104*, 2953.
- (17) Engkvist, O.; Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Chem. Phys.* **1999**, *110*, 5758.
- (18) Špirko, V.; Engkvist, O.; Soldán, P.; Selzle, H. L.; Schlag, E. W.; Hobza, P. *J. Chem. Phys.* **1999**, *111*, 572.
- (19) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104.
- (20) Tsuzuki, S.; Uchimaru, T.; Sugawara, K.; Mikami, M. *J. Chem. Phys.* **2002**, *117*, 11216.
- (21) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887.
- (22) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200.
- (23) Janda, K. C.; Hemminger, J. C.; Winn, J. S.; Novick, S. E.; Harris, S. J.; Klemperer, W. *J. Chem. Phys.* **1975**, *63*, 1419.
- (24) Steed, J. M.; Dixon, T. A.; Klemperer, W. *J. Chem. Phys.* **1979**, *70*, 4940.
- (25) Jeffrey, G. A.; Ruble, J. R.; McMullan, R. K.; Pople, J. A. *Proc. R. Soc. London, Ser. A* **1987**, *414*, 47.
- (26) Hopkins, J. B.; Powers, D. E.; Smalley, R. E. *J. Phys. Chem.* **1981**, *85*, 3739.
- (27) Henson, B. F.; Hartland, G. V.; Venturo, V. A.; Hertz, R. A.; Felker, P. M. *Chem. Phys. Lett.* **1991**, *176*, 91.
- (28) Henson, B. F.; Hartland, G. V.; Venturo, V. A.; Felker, P. M. *J. Chem. Phys.* **1992**, *97*, 2189.
- (29) Arunan, E.; Gutowsky, H. S. *J. Chem. Phys.* **1993**, *98*, 4294.
- (30) Law, K. S.; Schauer, M.; Bernstein, E. R. *J. Chem. Phys.* **1984**, *81*, 4871.
- (31) Schauer, M.; Bernstein, E. R. *J. Chem. Phys.* **1985**, *82*, 3722.
- (32) Fung, K. H.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1983**, *87*, 5113.
- (33) Börnsen, K. O.; Selzle, H. L.; Schlag, E. W. *Z. Naturforsch. A* **1984**, *39*, 1255.
- (34) Ebata, T.; Hamakado, M.; Moriyama, S.; Morioka, Y.; Ito, M. *Chem. Phys. Lett.* **1992**, *199*, 33.
- (35) Cacelli, I.; Cinacchi, G.; Prampolini, G.; Tani, A. *J. Am. Chem. Soc.* **2004**, *126*, 14278.
- (36) Cabaço, M. I.; Danten, Y.; Besnard, M.; Guissani, Y.; Guillot, B. *J. Phys. Chem. B* **1997**, *101*, 6977.
- (37) Chelli, R.; Cardini, G.; Ricci, M.; Righini, R.; Califano, S. *Phys. Chem. Chem. Phys.* **2001**, *3*, 2803.
- (38) Narten, A. H. *J. Chem. Phys.* **1977**, *67*, 2102.
- (39) Misawa, M.; Fukunaga, T. *J. Chem. Phys.* **1990**, *93*, 3495.
- (40) Fourkas, J. T. In *Ultrafast Infrared and Raman Spectroscopy*; Fayer, M. D., Ed.; Marcel Dekker: New York, 2001; pp 473–512.
- (41) Righini, R. *Science* **1993**, *262*, 1386.
- (42) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910.
- (43) Wennmohs, F.; Schindler, M. *J. Comput. Chem.* **2005**, *26*, 283.
- (44) Tran, F.; Weber, J.; Wesolowski, T. A. *Helv. Chim. Acta* **2001**, *84*, 1489.
- (45) Šponer, J.; Hobza, P. *Chem. Phys. Lett.* **1997**, *267*, 263.
- (46) 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.; Baboul, A.

- G.; 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.
- (47) Ransil, B. J. *J. Chem. Phys.* **1961**, *34*, 2109.
- (48) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
- (49) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665.
- (50) Aschi, M.; Mazza, F.; Di Nola, A. *J. Mol. Struct. (THEOCHEM)* **2002**, *587*, 177.
- (51) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.
- (52) Jorgensen, W. L. *BOSS 4.2*; Yale University: New Haven, CT, 2001.
- (53) Stone, A. J. *The Theory of Intermolecular Forces*; Oxford University Press: Oxford, 1996.
- (54) Jorgensen, W. L.; Severance, D. L. *J. Am. Chem. Soc.* **1990**, *112*, 4768.
- (55) Laaksonen, A.; Stilbs, P.; Waysylishen, R. E. *J. Chem. Phys.* **1998**, *108*, 455.
- (56) Dang, L. X. *J. Chem. Phys.* **2000**, *113*, 266.
- (57) Tassaing, T.; Cabaço, M. I.; Danten, Y.; Besnard, M. *J. Chem. Phys.* **2000**, *113*, 3757.
- (58) Chelli, R.; Cardini, G.; Procacci, P.; Righini, R.; Califano, S.; Albrecht, A. *J. Chem. Phys.* **2000**, *113*, 6851.
- (59) Lorenz, S.; Walsh, T. R.; Sutton, A. P. *J. Chem. Phys.* **2003**, *119*, 2903.
- (60) Streett, W. B.; Tildesley, D. J. *Proc. R. Soc. London, Ser. A* **1977**, *355*, 239.
- (61) Agrawal, R.; Sandler, S. I.; Narten, A. H. *Mol. Phys.* **1978**, *35*, 1087.
- (62) Brown, N. M. D.; Swinton, F. L. *J. Chem. Soc., Chem. Commun.* **1974**, *19*, 770.
- (63) Burley, S. K.; Petsko, G. A. *Science* **1985**, *229*, 23.
- (64) Chelli, R.; Gervasio, F. L.; Procacci, P.; Schettino, V. *J. Am. Chem. Soc.* **2002**, *124*, 6133.
- (65) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (66) Mackerell, A. D., Jr. *J. Comput. Chem.* **2004**, *25*, 1584.
- (67) Guillot, B.; Guissani, Y. *J. Chem. Phys.* **2001**, *114*, 6720.

CT060024H

Trisilaallene and the Relative Stability of Si₃H₄ Isomers

Monica Kosa, Miriam Karni,* and Yitzhak Apeloig*

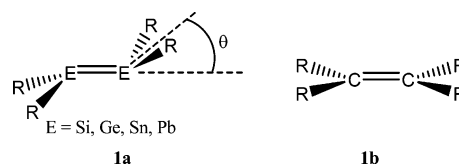
Department of Chemistry and the Lise Meitner, Minerva Center for Computational Quantum Chemistry, Technion-Israel Institute of Technology, Haifa 32000, Israel

Received June 17, 2005

Abstract: A theoretical quantum-mechanical study of trisilaallene, H₂Si=Si=SiH₂, and of 15 other Si₃H₄ isomers was carried out using ab initio and DFT methods with a variety of basis sets. Values given below are at B3LYP/6-31G(d,p). Unlike H₂C=C=CH₂ which is linear, H₂-Si=Si=SiH₂ is highly bent at the central silicon atom, with a SiSiSi bending angle of 69.4°. The Si=Si bond length is 2.269 Å, longer than a regular Si=Si double bond (2.179 Å) but shorter than a Si–Si single bond (2.351 Å). The distance between the terminal silicon atoms is 2.583 Å, significantly longer than a Si–Si single bond. The geometry and electronic properties of H₂-Si=Si=SiH₂ are similar to those of the corresponding trisilacyclopropylidene, which is only 2.7 kcal/mol higher in energy. A barrier of only 0.1 kcal/mol separates trisilacyclopropylidene and trisilaallene which can be described as bond-stretch isomers. Sixteen minima were located on the Si₃H₄ PES, most of them within a narrow energy range of ca. 10 kcal/mol. Six of the Si₃H₄ isomers are analogous to the classic C₃H₄ minima structures; however, the other Si₃H₄ isomers do not have carbon analogues, and they are characterized by hydrogen-bridged structures.

Introduction

The chemistry of compounds containing multiple bonds to silicon developed rapidly since the isolation of the first stable silene and disilene in 1981.¹ A variety of compounds with C=E and E=E (E = Si, Ge, Sn, Pb) bonds were isolated and characterized, and these developments were accompanied by numerous theoretical studies.^{1f} These studies revealed that silicon compounds as well as other heavier group 14 analogues can form stable multiply bonded compounds provided that the double bonds are protected by bulky substituents. One of the most interesting conclusions which developed from this new chemistry is the realization that multiply bonded silicon compounds usually adopt structures that are very different from those of the analogous carbon species.^{1,2} For example, heavier group 14 doubly bonded compounds usually have a trans-bent geometry as shown in **1a**, in contrast to olefins which are generally planar (**1b**). For H₂E=EH₂, E = Si, Ge, Sn, and Pb the calculated bending angle θ is 36.1°, 47.3°, 51.0°, and 53.6°, respectively.^{1f}

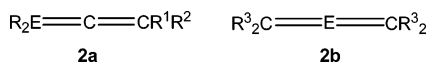


The origin of the trans-bent geometry of heavier group 14 doubly bonded compounds was discussed extensively by us³ and by others.⁴ It was suggested that the degree of trans-bending of R₂E=E'R'₂ is a function of the sum of the singlet–triplet energy separation ($\Sigma\Delta E_{st}$) of its constituent divalent species, R₂E and R'₂E', and the double bond energy, $E_{\sigma+\pi}$. According to this model, the double bond adopts a trans-bent structure when $\Sigma\Delta E_{st}$ is larger than half of $E_{\sigma+\pi}$.^{4c} A complementary explanation suggests that trans-bending results from effective $\pi-\sigma^*$ mixing for the heavier group 14 elements.^{4c,d} It was also demonstrated that the degree of trans-bending is strongly dependent on the substituents, R.³

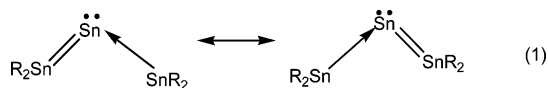
The experimental and theoretical knowledge on compounds containing an extended skeleton of heavier group 14 multiple bonds, e.g., E=E=C, E=C=E, or E=E=E is quite limited.^{5–10} The first such compounds, R₂E=C=CR¹R², E = Si, Ge, R = 2,4,6-triisopropylphenyl, R¹ = *t*-Bu, R² =

* Corresponding author e-mail: chrapel@tx.technion.ac.il (Y.A.) and chrmiri@tx.technion.ac.il (M.K.).

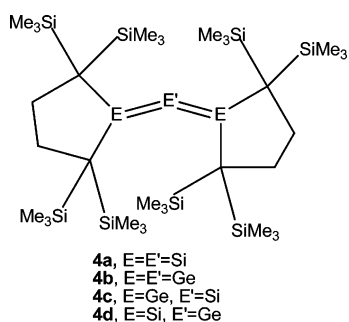
phenyl (**2a**), were synthesized and characterized by X-ray crystallography only recently,⁶ revealing that the heavy atom skeletons of **2a** are slightly bent (173.5^{06a} for E = Si and 159.2^{06c} for E = Ge). An additional 1-germaallene with R = Tbt, Mes, CR¹R² = fluorenyl, was reported by Tokitoh et al.⁷ A theoretical study has shown that 1- and 2-silaallenes, **2a** and **2b**, E = Si, with R = H, CH₃, SiH₃, R¹ = R² = R³ = H, all have a linear central skeleton and in **2a**, E = Si the terminal R₂Si and R₂C fragments are planar and perpendicular to each other, similarly to allene. With R = F, the central skeleton of **2a** is bent with a SiCC bond angle of 148.7° . For E = Ge, R = H, CH₃, SiH₃, R¹ = R² = R³ = H, both **2a** and **2b** are bent.⁸



The isolation of the first heavier group 14 allenic compound, (*t*-Bu₃Si)₂Sn=Sn=Sn(*t*-Bu₃)₂ (**3**), was reported by Wiberg in 1999.^{9a} The X-ray structure of **3** showed significant bending at the central Sn atom with a SnSnSn bond angle of 155.9° and an average Sn=Sn bond length of 2.683 \AA —which is shorter than other reported Sn=Sn double bond lengths ($2.77\text{--}2.91^{9b}$). However, the authors argued that **3** is not a real analogue of allene and that it is better described by the donor–acceptor resonance structures shown in eq 1.^{9a}



A recent spectacular achievement by Kira et al. is the isolation and characterization by X-ray crystallography of the first trisilaallene **4a**.^{10a} The X-ray structure of **4a** showed that the central SiSiSi skeleton is strongly bent with a bond angle of 136.5° . The Si=Si bond lengths of 2.177 \AA and 2.187 \AA are in the range of other known Si=Si double bond lengths. Most recently, Kira has synthesized the analogous trigermaallene, **4b**, and 1,3-digermasilaallene, **4c**,^{10b} and 2-germadisilaallene, **4d**,^{10c} and they are all strongly bent at E' (EE'E bond angle of 122.6° , 125.7° , and 132.4° , respectively).



Kira's impressive achievements demonstrate that these interesting compounds are experimentally accessible, and this prompted us to try to understand their basic properties, their bonding characteristics, and their relationship to other isomers. In this study we report a detailed computational quantum-mechanical study, using both traditional *ab initio*¹¹ and density functional (DFT) methods,¹² of the molecular

structure and the electronic properties of the parent H₂Si=Si=SiH₂ as well as of its relationship to other Si₃H₄ isomers. This study reveals an unexpected complex Si₃H₄ potential energy surface, much more complex than that of C₃H₄, with many interesting novel structures, including a bond-stretch isomer¹³ of trisilaallene.

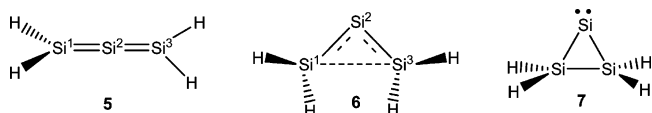
Computational Methods

Calculations were performed using both *ab initio*¹¹ and density functional theory (DFT)¹² techniques, as implemented in the Gaussian 98 series of programs.¹⁴ The geometries of all molecules were fully optimized, and vibrational frequencies were computed at the same level of theory in order to characterize the stationary points as minima (no imaginary eigenvalues), transition states (one imaginary eigenvalue), or saddle points of second order (two imaginary eigenvalues). For the DFT calculations we have used mostly the hybrid B3LYP density functional¹² with the doubly polarized 6-31G-(d,p) basis set. The *ab initio* calculations were performed mostly at the MP2/6-31G(d,p)//MP2/6-31G(d,p)¹¹ level of theory. The geometries of trisilaallene and of several of its isomers were also optimized at the correlated CCSD/6-311+G(2df,p) and CAS(6,6)/6-31G(d,p) levels of theory.¹¹

The discussion below is based mainly on the B3LYP/6-31G(d,p)//B3LYP/6-31G(d,p) results (unless otherwise specified), and the values given in parentheses are at MP2/6-31G(d,p)//MP2/6-31G(d,p). The energies reported include zero-point energy (ZPE) corrections at either the B3LYP or MP2 level (unless otherwise specified). The calculated geometries, total energies, and ZPEs of all calculated species are given in the Supporting Information.

Results and Discussion

1. Trisilaallene. The linear (*D*_{2d} symmetry) trisilaallene **5** is *not* a minimum on the Si₃H₄ potential energy surface (PES). Rather, **5** is a second-order saddle point with two degenerate imaginary frequencies. Full geometry optimization of **5** leads to **6** having an unusual highly bent structure of *C*_s symmetry, which is a minimum on the Si₃H₄ PES. The linear *D*_{2d} structure **5** lies 20.6 (22.7) kcal/mol above **6**. Another minimum which has quite a similar geometry to that of **6** is the *C*_{2v} trisilacyclopropylidene **7**. Other Si₃H₄ isomers are discussed below.



a. Geometry. The optimized structures of trisilaallene **6** as well as those of the hypothetical linear **5** and of cyclic **7** calculated using several theoretical methods are given in Table 1. The notations for the geometrical parameters are shown in Figure 1.

The structure of trisilaallene **6**¹⁵ is dramatically different from those of the carbon analogue, H₂C=C=CH₂, and from 1-silaallenes.⁸ The most unusual geometrical feature of trisilaallene is its very acute SiSiSi bond angle of only $67.1^\circ\text{--}70.4^\circ$ (depending on the computation level). In contrast, 1-silaallene, H₂Si=C=CH₂ is linear⁸ and in 1,2-disilaallene,

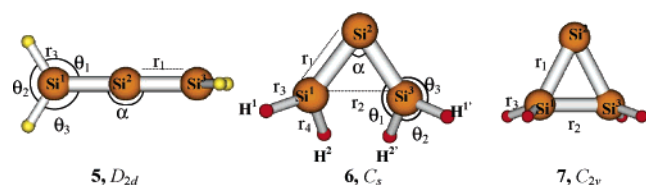


Figure 1. Geometry parameters of linear (**5**) and bent (**6**) trisilaallene and of cyclic silylene **7**. The notation of the geometrical parameters in **6** apply also to **7**.

$\text{H}_2\text{Si}=\text{Si}=\text{CH}_2$, the SiSiC bending angle is 140.9° (B3LYP/6-31G(d,p)). The structural contrast between trisilaallene and allene demonstrates that carbon chemistry is a poor guide for predicting the geometry of low-coordination silicon compounds.^{2a} The calculated bending angle in **6** is much smaller than the angle α determined experimentally for **4a** (136.5°), but it is similar to α of 74.2° calculated for $\text{Me}_2\text{-Si}=\text{Si}=\text{SiMe}_2$.¹⁶

The two H_2Si fragments in **6** are essentially planar ($\Sigma\theta = 359.95^\circ$ (359.85°)), but the hydrogens adopt unusual orientations. Thus, the planes defined by the H_2Si atoms are not mutually perpendicular, as in allene (or in **5**). Instead, the $\text{H}^1\text{Si}^1\text{Si}^3\text{H}^1$ and $\text{H}^1\text{Si}^1\text{Si}^3\text{H}^2$ dihedral angles are 0° and 115.5° , respectively (Table 1, these angles are 90° and -90° in allene).

The distance between the central silicon atom (Si^2) and the two terminal silicon atoms (Si^1 , Si^3) is 2.269 \AA (2.246 \AA), by 0.090 \AA (0.078 \AA) longer than the $\text{Si}=\text{Si}$ double bond in $\text{H}_2\text{Si}=\text{SiH}_2$ of 2.179 \AA (2.168 \AA), but it is much shorter than a typical Si–Si single bond, e.g., 2.351 \AA (2.338 \AA) in $\text{H}_3\text{Si}-\text{SiH}_3$. This indicates that these bonds have only a partial $\text{Si}=\text{Si}$ double bond character. The distance between the terminal Si^1 and Si^3 atoms is 2.583 \AA (2.583 \AA), 0.232 \AA longer than a typical Si–Si single bond. However, this distance is short enough to allow significant bonding interaction between these atoms as demonstrated by the existence of stable molecules with even longer Si–Si bonds, e.g., 2.697 \AA in $(t\text{-Bu})_3\text{Si}-\text{Si}(t\text{-Bu})_3$.¹⁷ The nature of the bonding interactions between the terminal silicon atoms in **6** is further discussed below.

b. Ring Opening of Trisilacyclopropylidene **7 to Trisilaallene **6**.** The geometry of trisilacyclopropylidene, **7**, is quite similar to that of trisilaallene **6**. The SiSiSi bending angle of 55.8° in **7** is smaller than that in **6** (of 69.4°), but both are in the range of that of a trisilacyclopropyl ring (60°). The Si^1-Si^3 bond distance of 2.291 \AA in **7** is shorter than in **6** (2.583 \AA), and the Si^1-Si^2 and Si^2-Si^3 bonds of 2.446 \AA in **7** are longer than in **6** (2.291 \AA), and they are also longer than a regular Si–Si single bond of 2.345 in a trisilacyclopropyl ring.¹⁸

Trisilacyclosilylidene **7** is by only 2.7 kcal/mol (3.66 kcal/mol without ZPE correction) higher in energy than **6**, which is expected as their geometries are quite similar (Figure 1, Table 1). The transition state connecting **7** and **6**, TS_{6-7} , lies only 0.1 kcal/mol above **7** and 2.8 kcal/mol above **6** (0.02 and 3.7 kcal/mol without ZPE correction respectively, 0.1 and 6.2 kcal/mol , respectively, at MP2/6-31G(d,p)//MP2/6-31G(d,p)+ZPE). At CCSD/6-311+G(2df,p)//CCSD/6-311+G(2df,p) the relative stability of **6** and **7** is reversed with trisilaallene **6** lying by 2.5 kcal/mol above **7** and the $\text{6} \rightarrow \text{7}$ and $\text{7} \rightarrow \text{6}$ barriers being of 3.0 and 5.5 kcal/mol , respectively.¹⁹ The very small energy barriers calculated at several theoretical levels imply that in practice silylidene (**7**) and trisilaallene (**6**), which are very close in energy, undergo rapid rearrangement even at low temperatures, with **6** being the dominant molecule. The energy profile at three theoretical levels for the ring opening of **7** to **6** is shown in Figure 2.

A comparison of the ring opening of **7** to **6** with that of the all-carbon analogue, cyclopropylidene to allene is of interest. The ring opening of **7** to **6** follows a simple disrotatory motion of the H_2Si groups. The ring opening of cyclopropylidene to allene also starts with a disrotatory motion of the methylene groups, but additional geometry changes are required to reach the final linear geometry of allene.²⁰ The overall barrier for ring opening of cyclopropylidene to allene is also low, 4.8 kcal/mol , but in this case the reaction is highly exothermic, by 69.3 kcal/mol (B3LYP/TZP//B3LYP/TZP).²⁰

The similar geometries and electronic structures (see below) of **6** and **7** indicate that they can be regarded as one

Table 1. Calculated Bond Lengths (\AA) and Bond Angles (deg) of **5–7** and of TS_{6-7} at Several Theoretical Levels^a

level of theory	species	α	r_1	r_2	r_3	r_4	$\Sigma\theta^b$	$\angle(\text{H}^1\text{Si}^1\text{Si}^3\text{H}^1)^c$	$\angle(\text{H}^1\text{Si}^1\text{Si}^3\text{H}^2)^c$
B3LYP/6-31G(d,p),	5	180.0	2.125		1.477		360.0	90.0	-90.0
MP2/6-31G(d,p),		180.0	2.127		1.467		360.0	90.0	-90.0
CCSD/6-311+G(2df,p)		180.0	2.126		1.472		360.0	90.0	-90.0
B3LYP/6-31G(d,p),	6	69.4	2.269	2.583	1.491	1.489	359.9	0.0	115.5
B3LYP/6-311G(2d,p)		70.4	2.262	2.607	1.487	1.485	359.9	0.0	115.0
MP2/6-31G(d,p)		70.2	2.246	2.583	1.479	1.479	359.8	0.0	114.4
MP2/6-311G(2d,p)		71.2	2.260	2.631	1.478	1.478	359.9	0.0	113.9
CCSD/6-311+G(2df,p)		68.6	2.260	2.548	1.484	1.483	360.0	0.0	117.5
CAS(6,6)/6-31G(d,p)		67.1	2.283	2.523	1.477	1.476	359.8	0.0	119.7
B3LYP/6-31G(d,p)	7	55.8	2.446	2.291	1.486	-	351.8	0.0	146.3
MP2/6-31G(d,p)		56.0	2.424	2.277	1.475	-	352.1	0.0	147.3
CCSD/6-311+G(2df,p)		56.0	2.442	2.295	1.481	-	352.2	0.0	147.2
CAS(6,6)/6-31G(d,p)		57.0	2.432	2.322	1.475	-	346.5	0.0	147.8
B3LYP/6-31G(d,p)	TS_{6-7}	56.5	2.431	2.301	1.487	1.486	350.7	0.0	143.8
MP2/6-31G(d,p)		56.8	2.405	2.290	1.476	1.474	350.8	0.0	144.1
CCSD/6-311+G(2df,p) ^d		55.9	2.440	2.290	1.480	1.480	352.3	0.0	147.4

^a Notation of the geometrical parameters and atom numbering is given in Figure 1. ^b $\Sigma\theta_i = \theta_1 + \theta_2 + \theta_3$. ^c Dihedral angle. ^d Reference 19.

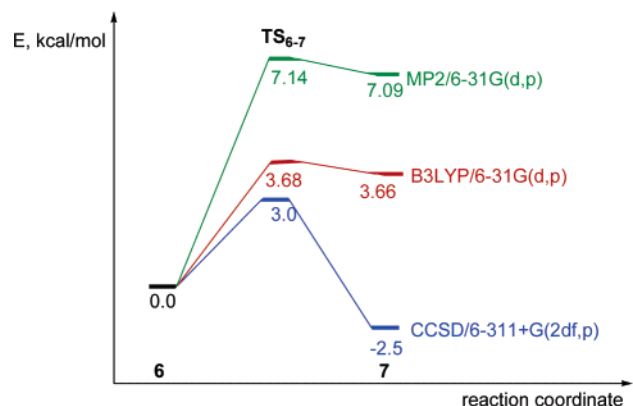


Figure 2. Reaction profile for ring opening of **7** to **6** (the energies are without ZPE corrections).

of a few known examples of “bond-stretch” isomers.¹² Bond-stretch isomerism is defined as the phenomenon whereby molecules of the same spin state, on the same potential energy surface, differ only in the length of one or several

bonds. This is indeed the case for **6** and **7**. However, unlike in ideal “bond-stretch” isomers, the terminal H_2Si groups rotate and the HSiSiH dihedral angle is changed upon stretching the $\text{Si}^1\text{—Si}^3$ bond and converting **7** to **6** (Figure 1, Table 1).

c. Electronic Structure. To describe the electronic structure and the chemical bonding in trisilaallene **6**, we find it convenient to compare it with those of the hypothetical linear trisilaallene **5** on one hand and with the cyclic trisilacyclopropylidene **7** on the other.

i. Frontier Molecular Orbitals of 5–7. The Frontier Molecular Orbitals (FMOs) of **5–7** calculated at the HF/6-31G(d,p)//B3LYP/6-31G(d,p) level are shown in Figure 3.²¹

The shapes of the FMOs of the hypothetical linear trisilaallene **5** are similar to those of allene. The HOMO and LUMO are both degenerate (as in allene), and they have the classic shape of π and π^* orbitals. The HOMO-1 of **5**, which is 4.7 eV lower in energy than the HOMO, is a σ -type orbital with a node at the central silicon atom, again, similar in

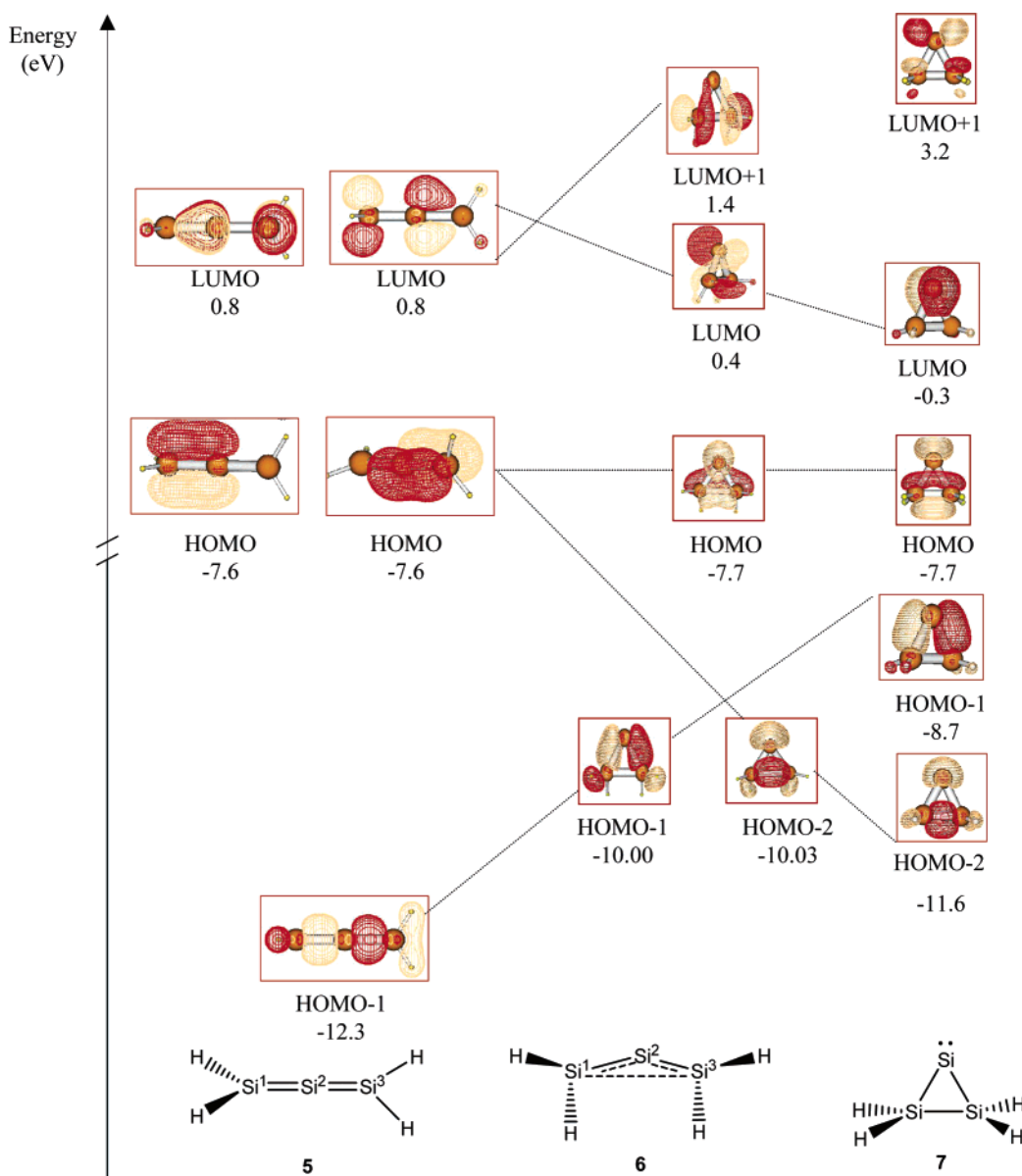


Figure 3. Frontier molecular orbitals (FMOs) of **5–7**. Orbital energies (HF/6-31G(d,p)//B3LYP/6-31G(d,p)) are given in eV.

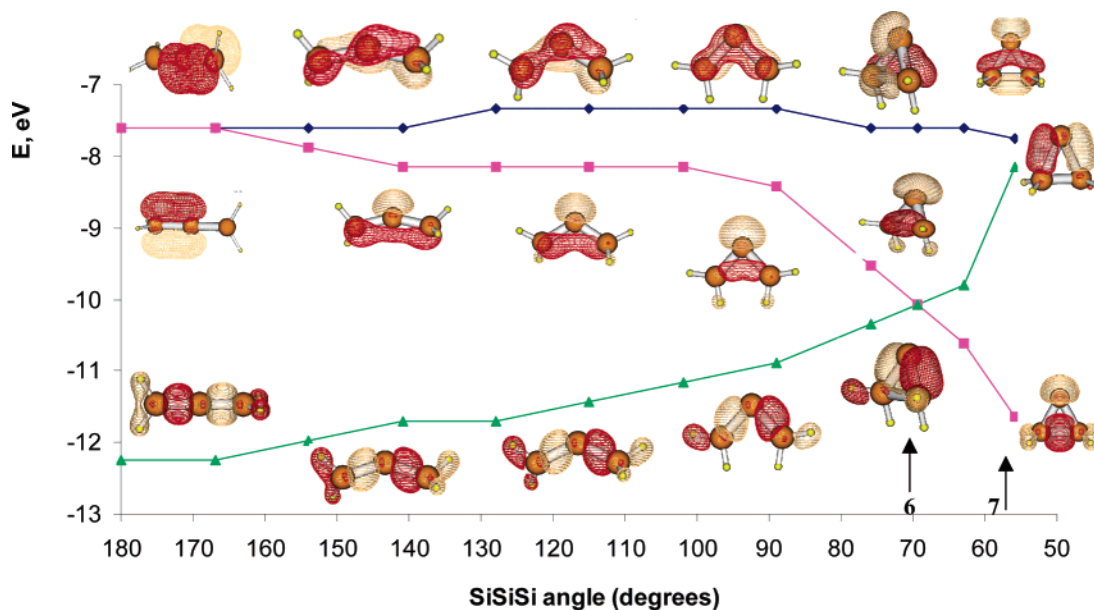


Figure 4. Walsh diagram (calculated at HF/6-31G(d,p)//B3LYP/6-31G(d,p)) for the $5 \rightarrow 6 \rightarrow 7$ transformation. The SiSiSi angle was fixed at the indicated values, while all other geometry parameters were optimized.

character to HOMO-1 of allene. The HOMO and the HOMO-1 of **5** are however much higher in energy than those of allene which are at -10.1 eV and -16.9 eV, respectively (at the same level of theory).

Upon bending of **5** to trisilaallene **6** the degenerate HOMO orbitals split: the HOMO of **6** has essentially the same energy as the HOMO of **5**. However, the second orbital (HOMO-2) drops strongly by 2.4 eV becoming almost degenerate in energy with the HOMO-1 orbital (σ -type) at -10.00 eV. The HOMO-2 of **6** is further stabilized to -11.6 eV upon ring closure to **7**. On the other hand, the energy of the HOMO-1 σ -orbital of **5** rises in energy by 2.3 eV upon bending to **6** and by an additional 1.3 eV upon ring closure to **7**. The LUMO of **6** is not degenerate (unlike in **5**) with LUMO+1 lying 1.0 eV above the LUMO. The HOMO–LUMO gap in **6** is 8.1 eV (8.4 eV in **5**), much smaller than in allene (15.0 eV).

The shapes of the FMOs of **6** and of **7** are similar. The HOMO of **6** (and of **7**) have a pronounced lone pair character at the central silicon atom, and they have the same energy. The shapes of the HOMO-1 and HOMO-2 of **6** and of **7** are also very similar, but in **6** these orbitals are almost degenerate, while in **7** the HOMO-1 lies 2.9 eV above HOMO-2. The LUMO of **6** and of **7** are similar. In **7** the LUMO is the empty $3p$ orbital on Si^2 , while in **6** the LUMO is a mixture of the $3p$ orbital of Si^2 and $\sigma(\text{Si}^1\text{–Si}^3)$.

A Walsh-type diagram²² showing the transformation of the degenerate HOMO π -orbitals and the HOMO-1 σ -orbital of linear **5** upon bending to **6** and to **7** is shown in Figure 4. Upon bending of **5** to **7** through **6**, the degeneracy of the HOMO orbital is lifted. The energy of one of the HOMO orbitals remains essentially unchanged along the bending process. The energy of the second HOMO is lowered from -7.6 eV in **5** to -11.0 eV in **7**, reflecting the build-up of the $\text{Si}^1\text{–Si}^3$ σ -bond which is evident in the orbital shape. On the other hand, the energy of the HOMO-1 (σ -orbital) is raised upon bending due to increased antibonding interactions

between the molecule's ends. The Walsh curves of the descending HOMO-1 and ascending HOMO-2 cross at a SiSiSi bond angle of 70° , i.e., practically at the bond angle of **6** (69.4°), where the HOMO-1 and HOMO-2 orbitals become degenerate. An additional small bending of the SiSiSi angle to 58° (reaching **7**) causes a significant decrease in the energy of the original (i.e. in **5**) π -orbital and a considerable increase in the energy of the original σ -orbital, resulting in a 2.9 eV energy difference between the HOMO-1 and HOMO-2 in **7**.

We note that the Walsh diagram in Figure 4 does not explain quantitatively the significantly lower energy of **6** and **7** relative to **5** since the sum of the FMOs energies of **5** (-27.5 eV) is almost identical to that of **6** (-27.7 eV) and **7** (-28.0 eV).

In summary, the FMOs of bent trisilaallene **6** are very similar in shape but are significantly different in energy compared to those of cyclic silylene **7**. However, both sets of FMOs are very different from those of hypothetical linear trisilaallene **5** or of allene.

ii. Charge Distribution. The atomic charges, bond orders, and orbital occupancies were calculated at the MP2/6-31G(d,p)//B3LYP/6-31G(d,p) level using Natural Bond Orbital (NBO) analysis.^{23a} The main results are given in Table 2.

The charge on the central silicon atom (Si^2) changes gradually upon bending the $\text{Si}^1\text{Si}^2\text{Si}^3$ bond angle, from a negative charge of -0.23 electrons in the linear trisilaallene **5** to neutral in **6** and to a positive charge of $+0.27$ electrons in trisilacyclopropylidene **7**. So, Si^2 is nucleophilic in linear **5** and electrophilic in silylene **7**. The charge on Si^2 in **7** is very similar to that in the disilylsilylene (H_3Si)₂ Si : (**8**), in line with its silylenic character. The positive charge on the terminal silicon atoms Si^1 and Si^3 decreases gradually from $+0.38$ el. in **5** to $+0.32$ el. in **6** to $+0.11$ el. in **7**. The hydrogens are negatively charged in all molecules, -0.13 el. in **5** and **7** and -0.16 el. in **6**.

d. The Nature of the Bonding in Trisilaallene 6. What is the nature of the bonding in trisilaallene and how is it

Table 2. Calculated (MP2/6-31G(d,p)//B3LYP/6-31G(d,p)) Charge Distributions, Orbital Occupancies, and Bond Orders in **5–7** and in (H₃Si)₂Si (**8**)^a

property		5	6	7	8
NPA charge	Si ²	-0.23	0.00	0.27	0.25
	Si ¹ , Si ³	0.38	0.32	0.11	0.37
	H	-0.13	-0.16	-0.13	-0.16
Wiberg bond index	Si ¹ –Si ²	1.81	1.25	0.90	0.93
	Si ² –Si ³	1.81	1.25	0.90	0.93
	Si ¹ –Si ³		0.58	0.98	
NBO occupancy ^b	Si ² –Si ³	3.85 ^c	1.89	1.92	1.93
	Si ² –Si ¹	3.85 ^c	1.89	1.92	1.93
	Si ¹ –Si ³		1.47	1.94	
	Si ² (LP ¹) ^d		1.87	1.96	1.94
	Si ² (LP ²) ^e		0.52	0.03	0.05

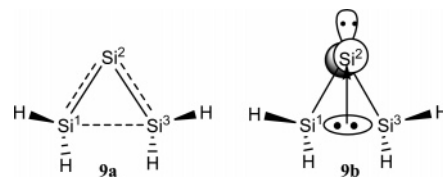
^a Atom numbering is given in Figure 2. ^b Occupancy in the indicated natural bond orbital. ^c In the σ and π bonds. ^d LP = lone pair; occupancy in the (SiSiSi) plane. ^e LP = lone pair; occupancy in the p orbital perpendicular to the (SiSiSi) plane.

different from the classic familiar bonding in allene? In particular, the strongly bent structure of trisilaallene **6** raises the question if there is a chemical bond between its terminal silicon atoms. To answer this question we used several criteria: The Wiberg Bond Index (WBI),^{23b} the electron occupancy of the Si¹–Si³ orbital space, and an analysis of the Si¹–Si³ orbital interactions. The calculated WBI of the Si¹–Si³ bond in **6** is 0.58. This WBI value indicates significant bonding, although weaker than in silylene **7** where the WBI is 0.98. For comparison, in cyclic Si₃H₆ and in Si₂H₆ the WBI of the Si–Si bond is 0.94 and 0.95, respectively (MP2/6-31G(d,p)//B3LYP/6-31G(d,p)). According to NBO analysis, 1.47 electrons occupy the Si¹–Si³ bond space in **6** compared with 1.94 in **7** and 1.93 and 1.95 el. in trisilacyclopropane and Si₂H₆, respectively (MP2/6-31G(d,p)//B3LYP/6-31G(d,p)). This analysis strongly supports the existence of a fairly strong partial Si–Si bond between the terminal silicon atoms of trisilaallene.²⁴

The WBI of the allenic Si¹–Si² (or Si²–Si³) bonds of only 1.25 in **6** as well as their calculated electron occupancy of only 1.89 el. indicate a significant reduction in the double bond character in **6** compared to linear **5** where the occupancy of each of the Si¹–Si² (Si³–Si²) bonds is 3.85 and the WBI of 1.81 is close to the classic value of 2. The reduced bond order of the Si¹–Si² (Si²–Si³) bonds in **6** is consistent with the fact that these bonds are longer in **6** (2.269 Å) relative to those in **5** (2.125 Å) or in H₂Si=SiH₂ (2.179 Å). It is interesting to note how the eight valence electrons connecting the three silicon atoms of **6** are distributed; 1.89 el. are assigned to each of the Si¹–Si² and Si²–Si³ bonds, 1.47 el. to the Si¹–Si³ bond, 1.87 el. to the in-plane lone pair at Si², and 0.52 el. to the formally empty out-of-plane orbital at Si².

In conclusion, according to the calculated geometry parameters and the above analysis of the electronic structure and charge distribution, the bonding in **6** is best described as consisting of two partial double bonds between Si² atom and the Si¹ and Si³ atoms and a partial bond between the terminal silicon atoms, as shown schematically in **9a**.

NBO analysis reveals that in **6** there is a substantial stabilizing interaction²⁵ between the Si¹–Si³ bonding elec-



trons and the formally empty 3p orbital on Si² which is manifested by charge transfer between these orbitals,²⁶ as schematically shown in **9b**. In contrast, in **7** this stabilizing interaction cannot occur, because the empty 3p orbital is strictly perpendicular to the Si¹–Si³ bond.

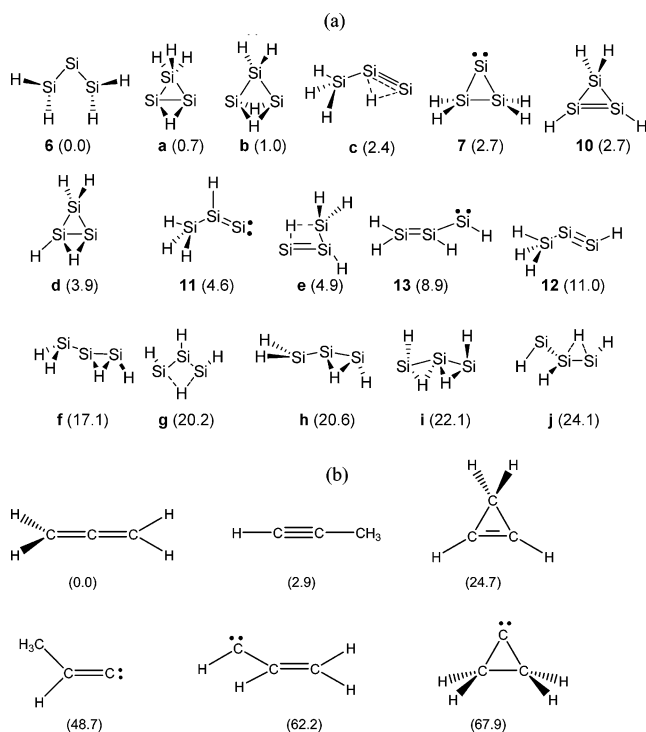
We conclude, based on the molecular geometry as well as on the WBI and NPA analysis, that in trisilaallene **6** a partial bond exists between the terminal Si¹ and Si³ atoms. Although formally being an allene, trisilaallene **6** has a very different electronic structure from that of allene, including a strong silylenic character at the central silicon atom. Trisilaallene is yet an additional example of low-valent silicon compounds, where traditional valence-bond Lewis structures cannot describe properly the bonding situation.^{2a}

2. Relative Energies of Si₃H₄ Isomers. The singlet potential energy surface (PES) is the lowest PES of Si₃H₄. The triplets of **6** and **7** lie 22.4 and 6.3 kcal/mol above the corresponding singlets, respectively.^{27,28} The quintet state of **6**, which involves unpairing of all four π electrons of trisilaallene, lies 53.7 kcal/mol above the singlet state. We therefore discuss below only the singlet Si₃H₄ PES.

The Si₃H₄ PES is much more complex than that of C₃H₄. As many as 16 minimum structures, i.e., Si₃H₄ isomers, were located on the Si₃H₄ singlet PES, and their relative energies (at B3LYP/6-31G(d,p)) are shown in Chart 1a.²⁹ Furthermore, all these Si₃H₄ isomers are within a relatively narrow energy range of only ~25 kcal/mol, and the energies of the nine lowest energy isomers are clustered within a range of only 5 kcal/mol. It is interesting to contrast the complex Si₃H₄ PES with the PES of C₃H₄^{30a} where only 6 minima exist and where the energy differences between the isomers are much larger, reaching 68 kcal/mol (Chart 1b).^{30a} The PESs of C₂SiH₄,^{30b} Si₄H₆, and Si₄R₆³¹ are also simpler than that of Si₃H₄.

The Si₃H₄ isomers include six structures which are analogous to the six C₃H₄ isomers (Chart 1b). The global minimum on the C₃H₄ PES is allene with propyne and cyclopropene lying 2.9 and 24.7 kcal/mol higher in energy, respectively. The other minima are the three carbenes, propenylidene, vinylmethylene, and cyclopropenylidene, lying by 48.7, 62.2, and 67.9 kcal/mol above allene, respectively.^{30a} The relative energies of the analogous silicon isomers are very different. The global Si₃H₄ minimum is the bent trisilaallene, **6**, with trisilacyclopropylidene, **7**, lying very close in energy. As the energy barrier separating **6** and **7** is very small (Figure 2), **7** will collapse to **6** even at very low temperatures. Trisilacyclopropene, **10**, lies only 2.7 kcal/mol above **6**. Silyldisilyne (**12**), the silicon analogue of propyne, is only 11–13 kcal/mol higher in energy than either **6** or **7**. **12** is trans-bent at the SiSi triple bond, as expected from previous theoretical^{1c} and consistent with recent experimental³² studies. **11** and **13**, two other silylene-type species, are by 4.6 and 8.9 kcal/mol higher in energy than **6**, respectively.

Chart 1. Relative Energies (kcal/mol) of Singlet M_3H_4 Isomers: (a) $M = Si$ (at B3LYP/6-31G(d,p)//B3LYP/6-31G(d,p)), All Structures Are Minima and (b) $M = C$ (at B3LYP/6-31G(d)//B3LYP/6-31G(d))^{30a}



An unusual feature of the Si_3H_4 PES is the existence of several hydrogen-bridged Si_3H_4 isomers (structures a–j in Chart 1a). This contrasts the C_3H_4 or the C_2SiH_4 PESs where hydrogen-bridged minima structures were not located. For example, **e** is a minimum on the Si_3H_4 PES, but the analogous C_3H_4 structure is a transition state that connects the carbon analogues of **10** and **11**.^{30a} The existence of hydrogen-bridged structures for heavier group 14 elements has been noted in other systems and was attributed to the larger size and higher polarizability of these atoms compared to those of carbon.^{1f} Interestingly, there are two hydrogen bridged structures, **b** (having two bridging hydrogens) and **d** (having one bridging hydrogen), which are very close in energy to the classic trisilacyclopene (**10**). Another interesting hydrogen bridged structure is **i**, which can be thought of as originating from a linear trisilaallene and in which a hydrogen bridges each of the two allenic double bonds.³³

Conclusions

Trisilaallene, the silicon analogue of allene has an unusual geometry, electronic structure, and bonding. It is strongly bent at the central silicon atom with a SiSiSi bond angle of only 69.4° (B3LYP/6-31G(d,p)) and has planar terminal H_2 -Si groups which adopt an unusual mutual orientation. A partial bond exists between the terminal silicon atoms and the two formal π -bonds have only partial occupancy. The formal trisilaallene is close in its geometry and energy to trisilacyclopene, and these two molecules which are connected by a very low barrier can be regarded as bond-stretch isomers.³⁴

The singlet PES surface of the Si_3H_4 isomers is very complex and includes at least 16 isomers, many having

nonclassical hydrogen-bridged structures. These isomers lie in a narrow energy range of less than 25 kcal/mol (11 isomers are in the range of 11 kcal/mol) suggesting the possible existence of a very complex mixture of isomers even at moderate temperatures. Many interesting questions are open for future studies, such as the effect of substituents on the structure and energetics of Si_3R_4 isomers. For example, in a recent paper¹⁶ we have demonstrated computationally that boryl-substituted trisilaallenes (and trigermaallenes) have linear classical allenic-type structures. We are continuing our studies of this intriguing group of compounds.

Acknowledgment. We thank the referees for their helpful comments. This research was supported by the Minerva Foundation in Munich and by the U.S.–Israel Binational Science Foundation (BSF).

Supporting Information Available: Cartesian coordinates, total energies, and ZPE of all calculated species. This material is available free of charge via Internet at <http://pubs.acs.org>.

References

- (1) For reviews see: (a) Müller, T.; Ziche, W.; Auner, N. In *The Chemistry of Organosilicon Compounds*; Rappoport, Z., Apeloig, Y., Eds.; John Wiley & Sons: Chichester, 1998; Vol. 2, Chapter 16, pp 857–1062. (b) Raabe, G.; Michl, J. In *The Chemistry of Organosilicon Compounds*; Patai, S., Rappoport, Z., Eds.; John Wiley & Sons: Chichester, 1989; Chapter 17, pp 1015–1142. (c) Brook, A. G.; Brook, M. A. *Advances in Organometallic Chemistry*; 1996; Vol. 39, p 71. (d) Brook, M. A. *Silicon in Organic, Organometallic, and Polymer Chemistry*; John Wiley & Sons: New York, 2000; Chapter 3, pp 39–96. (e) Eichler, B.; West, R. *Adv. Organomet. Chem.* 2000; Vol. 46, pp 1–46. (f) Karni, M.; Apeloig, Y.; Kapp, J.; Schleyer, P. von R. In *The Chemistry of Organic Silicon Compounds*; Rappoport, Z., Apeloig, Y., Eds.; John Wiley & Sons: Chichester, 2001; Vol. 3, Chapter 1, pp 1–163. (g) Tokitoh, N.; Okazaki, R. In *The Chemistry of Organic Germanium, Tin and Lead Compounds*; Rappoport, Z., Ed.; John Wiley & Sons: Chichester, 2002; Vol. 2, Chapter 13, pp 843–901.
- (2) (a) Karni, M.; Apeloig, Y. *Chem. Isr.* **2005**, *19*, 22. (b) Ichinohe, M.; Tanaka, T.; Sekiguchi, A. *Chem. Lett.* **2001**, *11*, 1074. (c) Power, P. P. *Chem. Rev.* **1999**, *99*, 3463. (d) Kutzelnigg, W. *Angew. Chem. Int. Ed. Engl.* **1984**, *23*, 272.
- (3) Karni, M.; Apeloig, Y. *J. Am. Chem. Soc.* **1990**, *112*, 8589.
- (4) (a) Carter, E. A.; Goddard, W. A., III. *J. Phys. Chem.* **1986**, *90*, 998. (b) Goldberg, D. E.; Hitchcock, P. B.; Lappert, M. F.; Thomas, K. M.; Thorne, A. J.; Fjeldberg, T.; Haaland, A.; Schilling, B. E. R. *J. Chem. Soc., Dalton Trans.* **1986**, 2387. (c) Trinquier, G.; Marlieu, J. P. *J. Am. Chem. Soc.* **1987**, *109*, 5303. (d) Marlieu, J. P.; Trinquier, G. *J. Am. Chem. Soc.* **1989**, *111*, 5916. (e) Trinquier, G.; Marlieu, J. P. *J. Phys. Chem.* **1990**, *94*, 6184.
- (5) Escudie, J.; Ranaivonjatovo, H.; Rigon, L. *Chem. Rev.* **2000**, *100*, 3639–3696.
- (6) (a) Miracle, G. E.; Ball, J. L.; Powell, D. R.; West, R. *J. Am. Chem. Soc.* **1993**, *115*, 11598. (b) Trommer, M.; Miracle, G. E.; Eichler, B. E.; Powell, D. R.; West, R. *Organometallics* **1997**, *16*, 5737. (c) Eichler, B. E.; Powell, D. R.; West, R. *Organometallics* **1998**, *17*, 2147. (d) Eichler, B. E.; Powell, D. R.; West, R. *Organometallics* **1999**, *18*, 540.

- (7) Tokitoh, N.; Kishikawa, K.; Okazaki, R. *Chem. Lett.* **1998**, 811.
- (8) Sigal, N.; Apeloig, Y. *Organometallics* **2002**, *21*, 5486.
- (9) (a) Wiberg, N.; Lerner, H. W.; Vasisht, S. K.; Wagner, S.; Karaghiosoff, K.; Nöth, H.; Ponikwar, W. *Eur. J. Inorg. Chem.* **1999**, 1211. (b) The only reported double bond length that is shorter than that of **3** is that of cytotristannene of 2.601 Å.^{9a}
- (10) (a) Ishida, S.; Iwamoto, T.; Kabuto, C.; Kira, M. *Nature* **2003**, *421*, 725–727. (b) Iwamoto, T.; Masuda, H.; Kabuto, C.; Kira, M. *Organometallics* **2005**, *24*, 197. (c) Iwamoto, T.; Abe, T.; Kabutu, C.; Kira, M. *Chem. Commun.* **2005**, *41*, 5190.
- (11) (a) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab initio Molecular Orbital Theory*; John Wiley & Sons: New York, 1986. (b) Young, D. C. *Computational Chemistry*; John Wiley & Sons: New York, 2001. (c) Jensen, F. *Introduction to Computational Chemistry*; John Wiley & Sons: Chichester, 1999.
- (12) (a) Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford University Press: Oxford, 1989. (b) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: Weinheim, 2000.
- (13) For a definition of bond-stretch isomerism, see: (a) Stohrer, W. D.; Hoffmann, R. *J. Am. Chem. Soc.* **1972**, *94*, 1661. (b) Stohrer, W. D.; Hoffmann, R. *J. Am. Chem. Soc.* **1972**, *94*, 779. For examples of bond-stretch isomers in heavier group 14 systems, see: (c) Nagase, S.; Nakano, M. *J. Chem. Soc., Chem. Commun.* **1988**, 1077. (d) Schleyer, P. v. R.; Sax, A. F.; Kalcher, J.; Janoschek, R. *Angew. Chem., Int. Ed. Engl.* **1987**, *26*, 364. (e) Nagase, S.; Kudo, T. *J. Chem. Soc., Chem. Commun.* **1988**, *1*, 54. (f) Kudo, T.; Nagase, S. *J. Phys. Chem.* **1992**, *96*, 9189.
- (14) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A.; 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.
- (15) A similar structure for trisilaallene was reported also by: Xu, W.; Yang, J.; Xiao, W. *J. Phys. Chem. A* **2004**, *108*, 11345–11353.
- (16) Kosa, M.; Karni, M.; Apeloig, Y. *J. Am. Chem. Soc.* **2004**, *126*, 10544. A very similar geometry was reported also in ref 10a.
- (17) Wiberg, N.; Schuster, H.; Simon, A.; Peters, K. *Angew. Chem.* **1986**, *98*, 100.
- (18) Kudo, T.; Akiba, S.; Kondo, Y.; Watanabe, H.; Morokuma, K.; Vreven, T. *Organometallics* **2003**, *22*, 4721.
- (19) At CCSD/6-311+G(2df,p), **TS₆₋₇** is not a real transition state but a third-order saddle point with the largest negative eigenvalue corresponding to a disrotatory motion of the SiH₂ fragments.
- (20) Bettinger, H. B.; Schreiner, P. R.; Schleyer, P. v. R.; Schaefer, H. F., III. *J. Phys. Chem.* **1996**, *100*, 16147.
- (21) The CAS(6,6)/6-31G(d,p) natural orbitals are almost identical to the HF molecular orbitals.
- (22) Albright, T. A.; Burdett, J. K.; Whangbo, M.-H. In *Orbital Interactions In Chemistry*; John Wiley & Sons: 1985; Chapter 7.
- (23) (a) As implemented in the NBO 5.0 version. Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Weinhold, F. Theoretical Chemistry Institute, University of Wisconsin, Madison, WI, 2001; <http://www.chem.wisc.edu/~nbo5>. (b) Wiberg, K. B. *Tetrahedron* **1968**, *24*, 1083.
- (24) (a) Structure **6** and **7** were also analyzed using the AIM^{24b} theory. No bond critical point was found between the terminal silicons in **6**. (b) Bader, R. F. W. In *Atoms in Molecules a Quantum Theory*; Clarendon Press: Oxford, 1994.
- (25) The strong interaction between the Si¹–Si³ bond orbital and the empty Si²(3p) orbital in **6** is evident in the magnitude of the second-order perturbation stabilization energy (ΔE) resulting from this interaction. Thus, while ΔE is zero in **7** it increases to 177 kcal/mol in **6** (calculated using NBO 5.0). These interaction energies change significantly with the method of calculation and basis set used, but the same qualitative picture emerges with several used methods.
- (26) The sum of the electron density residing between Si¹ and Si³ and in the formally empty Si² 3p orbital (which is not strictly perpendicular to the Si¹Si²Si³ plane and which becomes partially occupied due to this interaction) is close to 2 (1.47 (Si¹–Si³) + 0.52 (3p(Si²)) = 1.99 el., Table 2).
- (27) Fully optimized at UB3LYP/6-31G(d,p). The geometries are given in the Supporting Information.
- (28) The lowest triplet state of **6** (calculated by TDDFT/6-31G-(d,p)) corresponds to the biradical structure of **6** where a radical center is located on each of the terminal silicon atoms and the partial Si¹–Si³ bond is broken. The lowest triplet state of silylene **7** (vertical transition) has one electron in the in-plane orbital and one electron in the out-of-plane Si²-(3p) orbital. The different electronic structures of the triplet states of **6** and **7** point to the different electronic structures of the singlet ground states of **6** and **7**.
- (29) The most important Si₃H₄ isomers were optimized also using the larger B3LYP/6-311G(2d,p) basis set. It was found that this does not change significantly their relative energies. There is a relatively good agreement between the relative energy of **12** with respect to **6**, calculated at B3LYP/6-31G(d,p) (11.0 kcal/mol), MP2/6-31G(d,p) (8.5 kcal/mol), and CCSD/6-311+G(2df,p) (12.8 kcal/mol).
- (30) (a) Kakkar, R. *Int. J. Quantum Chem.* **2003**, *94*, 93–104, and references therein. (b) (i) Barthelat, J. C.; Trinquier, G.; Bertrand, G. *J. Am. Chem. Soc.* **1979**, *101*, 3785. (ii) Gordon, M. S.; Koob, R. D. *J. Am. Chem. Soc.* **1981**, *103*, 2939. (iii) Lien, M. H.; Hopkinsom, A. C. *Chem. Phys. Lett.* **1981**, *80*, 114.
- (31) (a) Most Si₃H₄ isomers lie in a relatively narrow energy range of ca. 10 kcal/mol, while the Si₄H₆^{31b} isomers lie in an energy range of ca. 35 kcal/mol. The Si₄Me₆ isomers (example of Si₄R₆^{31c}) lie in an energy range of ca. 145 kcal/mol. No hydrogen bridged structures were located on the Si₄H₆ PES. (b) Müller, T. In *Organosilicon Chemistry IV: From Molecules to Materials* **2000**, 110. (c) Koch, R.; Bruhn, T.; Weidenbruch, M. *Theochem.* **2004**, *680*, 91.

- (32) Sekiguchi, A.; Kinjo, R.; Ichinohe, M. *Science* **2004**, 305, 1755.
- (33) We suspect that additional stable structures may exist on the Si_3H_4 singlet surface.
- (34) While our paper was in press another paper that discusses the bonding in trisilaallene was published: Veszprémi, T.; Petrov, K.; Nguyen, C. T. *Organometallics* **2006**, 25, 1480.
CT050154A

Covalency in Highly Polar Bonds. Structure and Bonding of Methylalkalimetal Oligomers (CH₃M)_n (M = Li–Rb; n = 1, 4)

F. Matthias Bickelhaupt,^{*,†} Miquel Solà,^{*,‡} and Célia Fonseca Guerra[†]

Afdeling Theoretische Chemie, Scheikundig Laboratorium der Vrije Universiteit, De Boelelaan 1083, NL-1081 HV Amsterdam, The Netherlands, and Institut de Química Computacional, Universitat de Girona, Campus Montilivi, E-17071 Girona, Catalonia, Spain

Received December 28, 2005

Abstract: We have carried out a theoretical investigation of the methylalkalimetal monomers CH₃M and tetramers (CH₃M)₄ with M = Li, Na, K, and Rb and, for comparison, the methyl halides CH₃X with X = F, Cl, Br, and I, using density functional theory (DFT) at BP86/TZ2P. Our purpose is to determine how the structure and thermochemistry (e.g., C–M bond lengths and strengths, oligomerization energies) of organoalkalimetal compounds depend on the metal atom and to understand the emerging trends in terms of quantitative Kohn–Sham molecular orbital (KS-MO) theory. The C–M bond becomes longer and weaker, both in the monomers and tetramers, if one descends the periodic table from Li to Rb. Quantitative bonding analysis shows that this trend is not only determined by decreasing electrostatic attraction but also, even to a larger extent, by the weakening in orbital interactions. The latter become less stabilizing along Li–Rb because the bond overlap between the singly occupied molecular orbitals (SOMOs) of CH₃[•] and M[•] radicals decreases as the metal ns atomic orbital (AO) becomes larger and more diffuse. Thus, the C–M bond behaves as a typical electron-pair bond between the methyl radical and alkalimetal atom, and, in that respect, it is covalent. It is also shown that such an electron-pair bond can still be highly polar, in agreement with the large dipole moment. Interestingly, the C–M bond becomes less polar in the methylalkalimetal tetramers because metal–metal interactions stabilize the alkalimetal orbitals and, in that way, make the alkalimetal effectively less electropositive.

1. Introduction

Organoalkalimetal compounds, in particular organolithium reagents, are widely used in synthetic organic and organometallic chemistry.¹ Their methyl derivatives constitute the simplest organometallic compounds and contain the archetypal carbon–metal bond. Numerous theoretical^{2–5} and experimental^{6–8} studies have been undertaken to obtain

information about structure, stability, and bonding of this class of systems. Recently, Grotjahn and co-workers^{6a–c} published the first highly accurate gas-phase experimental structures for monomeric methyllithium, -sodium, and -potassium. These species have C_{3v} symmetry and consist of a pyramidal methyl group bound to the metal atom (Chart 1, left). The C–M bond distance increases along this series, as one might expect, from 1.961 to 2.299 to 2.633 Å (Table 1).

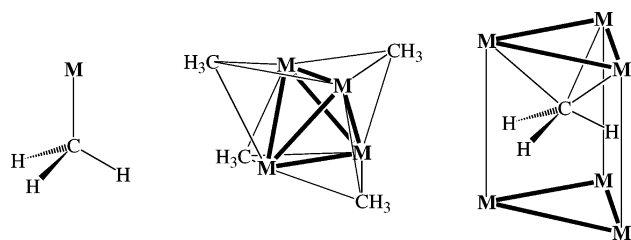
In the condensed phase, organoalkalimetal systems tend to oligomerize. Weiss and co-workers^{7a–d} have determined the crystal structure of deuterated methyllithium, -sodium,

* Corresponding author fax: +31 - 20 - 59 87 629; e-mail: FM.Bickelhaupt@few.vu.nl (F.M.B.) and fax: +34 - 972 - 41 83 56; e-mail: miquel.sola@udg.es (M.S.).

† Scheikundig Laboratorium der Vrije Universiteit.

‡ Universitat de Girona.

Chart 1



and -potassium oligomers. Descending the periodic table, the C–M bond again elongates, but it is systematically longer, by up to ca. 0.4 Å, than in the corresponding monomers (Table 1). The crystal structure of methyl lithium is composed of tetramers with T_d symmetry in which a central, tetrahedral lithium cluster is surrounded by four pyramidal methyl groups, one on each face of the metal tetrahedron, in a staggered orientation with respect to the adjacent Li_3 group (Chart 1, center). The methylsodium crystal has a somewhat more involved structure with a $(\text{CD}_3\text{Na})_{16}$ unit cell that, however, still consists 50% of tetramers similar to those of methyl lithium but slightly distorted. The methylpotassium crystal, on the other hand, has a $(\text{CD}_3\text{K})_6$ unit cell in which pyramidal methyl groups are located within a trigonal prism of potassium atoms and point with the vacant site of the sp^3 -carbon atom toward one of the K_3 faces (Chart 1, right).

Many studies have been directed toward unraveling the nature of the bonding in organoalkalimetal oligomers. The current picture^{2–4f} of the carbon–lithium bond is that of an ionic bond which can best be understood in terms of a CH_3^- anion and an Li^+ cation interacting predominantly electrostatically with only marginal covalent character. Streitwieser and co-workers² were the first to emphasize the highly polar character of this bond, based on atomic charges computed with the integrated projected populations (IPP) scheme. This approach yields an atomic charge of Li in methyl lithium of +0.8 au. Also other studies have been in support of a lithium atomic charge close to +1 au, for example, natural population analysis (NPA)^{3d,5} and atoms in molecules (AIM),^{4f,i} which yield charges close to +0.9 au. In addition, Streitwieser, Bushby, and Steel have shown that a simple electrostatic model is able to reproduce the ratio of carbon–carbon and lithium–lithium distances in the methyl lithium tetramer.^{3j,k} These results have led to the current idea that the C–Li bond is 80–90% ionic.

There are also data supporting a more prominent role of covalency in the C–Li bond. Early pioneering studies by the groups of Schleyer and Pople,^{4h} Lipscomb,⁴ⁱ and Ahlrichs^{4g} have highlighted these covalent aspects, especially in organolithium aggregates. This view is experimentally supported by the large carbon–lithium NMR coupling constants of up to 17 Hz that have been measured for organolithium aggregates^{8a–d} and by the solubility of simple organolithium compounds in nonpolar solvents.^{8e,f} Also, Streitwieser's aforementioned ideal distance ratio is not found for $(\text{LiH})_4$, $(\text{LiOH})_4$, and $(\text{LiF})_4$, and for the two latter, a simple electrostatic model erroneously predicts a planar eight-membered ring to be more stable than a tetrahedral structure.^{4j} Moreover, it has been pointed out that atomic charges are no absolute quantities and can, therefore, not serve as

absolute bond-polarity indicators.⁹ Different atomic-charge schemes yield different absolute values for one and the same atom in exactly the same chemical environment. For example, while the AIM atomic charge of +0.9 au for Li in methyl lithium suggests a nearly complete transfer of one electron, GAPT, Hirshfeld, and VDD sketch a far more moderate picture with lithium atomic charges of only +0.4, +0.5, and +0.4 au, respectively.^{4k,5,9} This undermines the main argument in support of an ionic C–Li bonding mechanism.

In the present study, we have undertaken a detailed investigation of methylalkalimetal monomers CH_3M and tetramers $(\text{CH}_3\text{M})_4$, with $\text{M} = \text{Li}, \text{Na}, \text{K},$ and Rb , using the generalized gradient approximation (GGA) of density functional theory (DFT) at the BP86/TZ2P level.¹⁰ We aim at three objectives. First, we wish to obtain a set of consistent structural and thermochemical data for methylalkalimetal monomers and tetramers (geometries, C–M bond strengths, tetramerization energies); all obtained at the same level of theory. This complements the available experimental and theoretical data, which are scarce for the monomers and missing for the oligomers of the heavier methylalkalimetal systems (beyond lithium), and it enables a systematic analysis of trends.

Second, our main purpose is to better understand the physics and the nature of the carbon–alkalimetal bond based on quantitative molecular orbital (MO) theory as contained in Kohn–Sham density functional theory.¹⁰ In particular, we wish to obtain a bonding *mechanism*, that is, an understanding of how the MO electronic structures of the methyl radical and metal atom interfere, how this provides C–M bonding, and how this makes the bond polar. Through a quantitative bond energy decomposition, we assess the importance of electrostatic attraction and orbital interactions for providing the C–M bond, and we reveal their role in determining trends therein along Li, Na, K, and Rb.¹⁰ Here, we anticipate that all C–M bonds have a strong intrinsic preference for homolytic over ionic dissociation. Interestingly, the weakening of the C–M bond that we find along $\text{M} = \text{Li}, \text{Na}, \text{K},$ and Rb is largely determined by the decreasing bond overlap between the SOMOs of the $\text{CH}_3^\bullet + \text{M}^\bullet$ radicals. Thus, the C–M bond behaves as a typical electron-pair bond between the methyl radical and alkalimetal atom, and, in that respect, it is covalent. We also show that such an electron-pair bond can still be highly polar, in agreement with the large dipole moment. Virtually covalent C–M electron-pair bonding occurs if the metal forms clusters, in our case, tetramers.

Third, we discuss various descriptors (atomic charge, wave function, energy decomposition) of covalency and ionicity, how these concepts and their descriptors can be interpreted, and to what extent they really provide information about the *mechanism* of highly polar (or less polar or nonpolar) bonds. Furthermore, we discuss homolytic ($\text{CH}_3^\bullet + \text{M}^\bullet$) versus ionic ($\text{CH}_3^- + \text{M}^+$) dissociation, and we compare carbon–metal with carbon–halogen bonds.

2. Theoretical Methods

2.1. General Procedure. All calculations were performed using the Amsterdam Density Functional (ADF) program.¹¹

Table 1. Structures (in Å, deg) of Methyl Alkalimetal Monomers and Tetramers^k

system	method	C–M	C–H	∠HCH	M–M	C–C	ref
CH ₃ Li	BP86/TZ2P	2.010	1.105	106.48			this work
	exp: mm-wave, gas phase ^a	1.961(5)	1.122(5)	107.2(1)			6a
	exp: mm-wave, gas phase ^b	1.959	1.111	106.2			6c
	exp: IR, argon matrix	2.10	1.12	107.3–109.5 ^c			6d
	HF/6-31G*	2.0013	1.0934	106.2			4f
	MP2/6-31+G*	2.005	1.099	107.3			5
	MP2/6-311G*	1.983	1.098	106.2			3e
	MP2/6-31++G**	2.004	<i>d</i>	<i>d</i>			3a
	MP2/TZ+spd	1.984	1.092	106.4			4e
	MP2/6-311+G(3df,2pd)	1.971	1.094	106.27			4c
	CASSCF(10/13)/cc-pVTZ	1.998	1.111	105.7			3h
	CCSD(T)/6-311+G(3df,2pd)	1.969	1.099	106.01			4c
	CCSD(T)/cc-pV(5,Q)Z	1.9799	1.0987	105.88			4b
CCSD(T)/MT(ae)	1.9619	1.0960	105.86			4b	
B3LYP/6-31+G*	1.986	<i>d</i>	106.4			4a	
(CH ₃ Li) ₄ ecl	BP86/TZ2P	2.199	1.111	102.23	2.418	3.579	this work
	MP2/6-31+G*	2.188	1.107	102.9	2.363	3.582	5
	HF/3-21G	2.236	1.102	103.9	2.420	3.657	3e
	B3LYP/6-31+G*	2.195	<i>d</i>	102.3	2.400	3.579	4a
(CH ₃ Li) ₄ stag	BP86/TZ2P	2.213	1.110	103.02	2.408	3.614	this work
	exp: neutron diffraction, 1.5 K ^e	2.256(6)	1.072(2)	108.2(2)	2.591(9)	3.621(6)	7a
	exp: neutron diffraction, 290 K ^e	2.209(14)	0.993(7)	110.7(6)	2.605(19)	3.511(10)	7a
	exp: X-ray, powder crystal, 290 K	2.31(5)	0.96(5)	111(8)	2.68(5)	3.68(5)	7e
	HF/3-21G	2.240	<i>d</i>	<i>d</i>	2.415	3.668	3e
CH ₃ Na	BP86/TZ2P	2.376	1.098	109.84			this work
	exp: mm-wave, gas phase ^b	2.299	1.091 ^f	107.3			6c
	HF/6-31G*	2.3236	1.0910	107.2			4f
	MP2/6-31++G**	2.342	<i>d</i>	<i>d</i>			3a
	MP2/TZ+spd	2.315	1.089	108.5			4e
	MP2/6-311+G(3df,2pd)	2.306	1.091	108.43			4c
	CCSD(T)/6-311+G(3df,2pd)	2.310	1.095	110.73			4c
(CH ₃ Na) ₄ ecl	BP86/TZ2P	2.684	1.110	102.20	3.070	4.315	this work
	BP86/TZ2P	2.675	1.109	103.09	3.059	4.300	this work
(CH ₃ Na) ₄ stag	exp: neutron+synchrotron diff., 1.5 K ^g	2.57–2.68	1.094 ^h	106.2 ^h	2.97–3.17	<i>d</i>	7b,c
CH ₃ K	BP86/TZ2P	2.747	1.100	109.14			this work
	exp: mm-wave, gas phase ^a	2.633(5)	1.135(5)	107.0(1)			6a
	HF/DZP+	2.754	1.094	106.2			3b
	MP2/DZP+	2.743	1.097	107.7			3b
	MP2/[6-31++G**] ^j	2.694	<i>d</i>	<i>d</i>			3a
	MP2/DZ+spd	2.675	1.102	106.7			4e
	CCSD(T)/(C,H)VTZ	2.661	1.097	106.6			3i
	CCSD(T)/(C,H)VTZ	2.661	1.097	106.6			3i
(CH ₃ K) ₄ ecl	BP86/TZ2P	3.000	1.112	101.47	3.724	4.658	this work
(CH ₃ K) ₄ stag	BP86/TZ2P	2.962	1.111	102.39	3.714	4.575	this work
	exp: neutron diffraction, 1.35 K ⁱ	2.947(2), 3.017(4)	1.082(4), 1.103(2)	104.8(2), 105.8(2)			7d
CH ₃ Rb	BP86/TZ2P	2.821	1.096	109.63			this work
	HF/DZP+	2.906	1.095	106.2			3b
	MP2/DZP+	2.897	1.097	107.9			3b
	MP2/[6-31++G**] ^j	2.855	<i>d</i>	<i>d</i>			3a
(CH ₃ Rb) ₄ ecl	BP86/TZ2P	3.118	1.112	101.41	3.906	4.817	this work
(CH ₃ Rb) ₄ stag	BP86/TZ2P	3.068	1.110	102.35	3.893	4.707	this work

^a Partial r_s structure. ^b r_0 structure. ^c Estimated value range. ^d Value not specified. ^e (CD₃Li)₄ staggered. ^f Estimated value from ref 4f. ^g Staggered (CD₃Na)₄ unit in more complex (CD₃Na)₁₆ unit cell. ^h Average of six similar values. ⁱ Basis for K and Rb: 9VE-ECP MWB 6s6p2d/5s5p2d. ^j Staggered [(CD₃)K₃] unit in more complex (CD₃K)₆ unit cell which contains no (CH₃K)₄ units! Instead, each methyl group is located in the center of a trigonal prism of six K atoms. C–K distances in this table correspond to close contacts, i.e., distances between C and the three K atoms of the trigonal prism toward which the lone pair of the methyl group is oriented. ^k See also Chart 1 and Figure S1.

The numerical integration was performed using the procedure developed by Boerrigter, te Velde, and Baerends.^{11e,f} The MOs were expanded in a large uncontracted set of Slater type orbitals (STOs) containing diffuse functions, which is of triple- ζ quality for all atoms and has been augmented with two sets of polarization functions: 3d and 4f on C, Li, Na;

4d and 4f on K, Rb; and 2p and 3d on H.^{11g} In addition, an extra set of p functions was added to the basis sets of Li (2p), Na (3p), K (4p), and Rb (5p). The 1s core shell of carbon and lithium, the 1s 2s 2p core shells of sodium and potassium, and the 1s 2s 3s 2p 3p 3d core shells of rubidium were treated by the frozen-core (FC) approximation.^{11d} 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-correlation potentials accurately in each SCF cycle.^{11h}

Energies and geometries were calculated using the generalized gradient approximation (GGA) of DFT at the BP86 level. GGA proceeds from the local density approximation (LDA) where exchange is described by Slater's $X\alpha$ potential¹¹ⁱ and correlation is treated in the Vosko-Wilk-Nusair (VWN) parametrization^{11j} which is augmented with nonlocal corrections to exchange due to Becke^{11k,l} and correlation due to Perdew^{11m} added self-consistently.¹¹ⁿ All open-shell systems were treated with the spin-unrestricted formalism. Bond enthalpies at 298.15 K and 1 atm (ΔH_{298}) were calculated from electronic bond energies (ΔE) and our frequency computations using standard statistical-mechanics relationships for an ideal gas.¹²

2.2. Bond Energy Decomposition. The overall bond energy ΔE is made up of two major components (eq 1):

$$\Delta E = \Delta E_{\text{prep}} + \Delta E_{\text{int}} \quad (1)$$

In this formula, the preparation energy ΔE_{prep} is the amount of energy required to deform the separate molecular fragments that are connected by the chemical bond from their equilibrium structure to the geometry that they acquire in the overall molecular system. The interaction energy ΔE_{int} corresponds to the actual energy change when the prepared fragments are combined to form the overall molecule. It is analyzed for our model systems in the framework of the Kohn–Sham MO model using a Morokuma-type decomposition of the bond into electrostatic interaction, exchange repulsion (or Pauli repulsion), and (attractive) orbital interactions (eq 2).^{10,13}

$$\Delta E_{\text{int}} = \Delta V_{\text{elstat}} + \Delta E_{\text{Pauli}} + \Delta E_{\text{oi}} \quad (2)$$

The term ΔV_{elstat} corresponds to the classical electrostatic interaction between the unperturbed charge distributions of the prepared (i.e. deformed) fragments and is usually attractive. The Pauli repulsion ΔE_{Pauli} comprises the destabilizing interactions between occupied orbitals. It arises as the energy change associated with going from the superposition of the unperturbed electron densities of two fragments, say CH_3^* and M^* , i.e., $\rho_{\text{CH}_3(\alpha)} + \rho_{\text{M}(\beta)}$, to the wave function $\Psi^0 = N A [\Psi_{\text{CH}_3(\alpha)} \Psi_{\text{M}(\beta)}]$, that properly obeys the Pauli principle through explicit antisymmetrization (A operator) and renormalization (N constant) of the product of fragment wave functions.¹⁰ It comprises the four-electron destabilizing interactions between occupied orbitals and is responsible for any steric repulsion. The orbital interaction ΔE_{oi} in any MO model, and therefore also in Kohn–Sham theory, accounts for electron-pair bonding, charge transfer (i.e., donor–acceptor interactions between occupied orbitals on one fragment with unoccupied orbitals of the other, including the HOMO–LUMO interactions), and polarization (empty-occupied orbital mixing on one fragment due to the presence of another fragment). In case of open-shell fragments, the bond energy analysis yields, for technical reasons, interaction energies that differ consistently in the order of a kcal/mol (too much stabilizing) from the exact BP86 result (because,

only in the bond energy analysis, the spin-polarization in the fragments is not accounted for). To facilitate a straightforward comparison, the results of the bond energy analysis were scaled to match exactly the regular BP86 bond energies.

The orbital interaction energy can be decomposed into the contributions from each irreducible representation Γ of the interacting system (eq 3) using the extended transition state (ETS) scheme developed by Ziegler and Rauk.^{13d,e}

$$\Delta E_{\text{oi}} = \sum_{\Gamma} \Delta E_{\Gamma} \quad (3)$$

The electron density distribution is analyzed using the Voronoi deformation density (VDD) method¹⁴ and the Hirshfeld scheme (see ref 15) for computing atomic charges, which are discussed in detail in ref 9.

3. Results and Discussion

3.1. Structures. Monomers. The structural results of our BP86/TZ2P computations are summarized in Table 1, Chart 1, and Figure S1 in the Supporting Information. The C–M bond distance in the C_{3v} symmetric methylalkalimetal monomers CH_3M increases descending the periodic table from 2.010 (Li) to 2.376 (Na) to 2.747 (K) to 2.821 Å (Rb). The C–H bond distance does not change much and amounts to ca. 1.10 Å throughout the series. The methyl group however becomes significantly less pyramidal if one goes from CH_3Li with an HCH angle of 106° to the heavier congeners which all have HCH angles of 109–110°.

This agrees satisfactorily with previous theoretical work^{3a,b,e,h,i,4a–c,e,f,5} and mm-wave experiments,^{6a,c} which also yield a monotonic increase of the C–M bond along CH_3Li , CH_3Na , CH_3K , and CH_3Rb (note that no experimental data are available for CH_3Rb and that the MP2 study by Schleyer and co-workers^{3a} is the only other theoretical study that treats the whole series consistently at the same level of theory). The experimental C–M bond lengths are however systematically shorter, by 2–4%, than our BP86/TZ2P and other theoretical values. A striking discrepancy between theory and experiment is that all theoretical approaches, up to the CCSD-(T) level,^{4c} show that from Li to the heavier alkalimetals the methyl group in CH_3M becomes significantly less pyramidal, whereas the mm-wave experiments yield little change in the HCH angle, which is always close to 107°.

Tetramers. All methylalkalimetal tetramers $(\text{CH}_3\text{M})_4$ have T_d symmetry and consist of a tetrahedral cluster of alkalimetal atoms surrounded by four methyl groups, one on each M_3 face, oriented with respect to the latter either eclipsed, for Li, or staggered, for Na, K, and Rb (see Figure S1). Tetramerization, i.e., going from CH_3M to $(\text{CH}_3\text{M})_4$, causes the C–M bond to elongate substantially by 0.2–0.3 Å, whereas the C–H bond distance increases only slightly by 0.01 Å (see Table 1). There is a remarkable increase in pyramidalization of the methyl groups as follows from the HCH angle, which decreases by 4° for lithium and up to 8° for the heavier alkalimetals. The C–M bond distance in the methylalkalimetal tetramers with the methyl groups eclipsed, $(\text{CH}_3\text{M})_4 \text{ecl}$, increases again monotonically from 2.199 (Li) to 2.684 (Na) to 3.000 (K) to 3.118 Å (Rb), as does the M–M bond distance in the central metal cluster. However,

Table 2. Homolytic and Heterolytic C–M Bond Strength (in kcal/mol) of Methyl Alkalimetal Monomers

monomer	method ^a	bond energies ^b			bond enthalpies ^d		ref
		ΔE_{homo}	ΔE_{hetero}	NIMAG ^c	ΔH_{homo}	ΔH_{hetero}	
CH ₃ Li	BP86/TZ2P	−44.8	−174.3	0	−44.0	−172.4	this work
	BP86/TZ2P//MP2/6-31+G*	−45.5	−174.2	0 ^e			5
	MP2/TZ+spd	−46.1		0			4e
	MP4(SDTQ)/6-311+G**	−44.18		0 ^f			3e
	scaled CPF/C	−46.4 ± 1.2					4 g
CH ₃ Na	B3LYP/6-311++G(2d,2p)//MP2/6-31++G**	−43.7 ^g					3a
	BP86/TZ2P	−31.0	−155.8	0	−30.3	−154.0	this work
	MP2/TZ+spd	−31.6		0			4e
CH ₃ K	B3LYP/6-311++G(2d,2p)//MP2/6-31++G**	−29.4 ^g					3a
	BP86/TZ2P	−26.6	−131.0	0	−26.2	−129.5	this work
	MP2/DZ+spd	−26.1		0			4e
CH ₃ Rb	B3LYP/basis C//MP2//basis B ^h	−26.0 ^g					3a
	BP86/TZ2P	−25.0	−124.6	0	−24.6	−123.1	this work
	B3LYP/basis C//MP2//basis B ^h	−23.4 ^g					3a

^a Energy and structure obtained at the same level of theory (unless stated otherwise). ^b Zero K electronic energies (unless stated otherwise). ^c Number of imaginary frequencies. ^d 298.15 K enthalpies. ^e Vibrational analysis at HF/6-31+G*. ^f Vibrational analysis at MP2/6-311G*. ^g Zero K electronic energy + Δ ZPE correction at HF (with 6-31+G* for C, H, Li, Na, and 9VE-ECP MWB 6s6p2d/5s5p2d for K, Rb). ^h Basis B: 6-31++G** for C, H; 9VE-ECP MWB 6s6p2d/5s5p2d for K, Rb. Basis C: 6-311++G(2d,2p) for C, H; 9VE-ECP MWB 6s6p2d/5s5p2d for K, Rb.

at variance with the monomers, not only the C–H bond distance of 1.11 Å but also the extent of pyramidalization of the methyl groups is practically constant with $\angle\text{HCH} = \text{ca. } 102^\circ$. These trends are similar for the methylalkalimetal tetramers with the methyl groups staggered, (CH₃M)₄ stag, in which the C–M bond is slightly longer for Li (by 0.01 Å) and somewhat shorter for Na–Rb (by 0.01–0.05 Å) than in the corresponding (CH₃M)₄ ecl. The C–H bonds are only marginally shorter (by 0.001–0.002 Å), and the methyl groups are only slightly less pyramidal ($\angle\text{HCH}$ increases by less than 1°) in (CH₃M)₄ stag compared to (CH₃M)₄ ecl.

There is, to the best of our knowledge, no other theoretical work on methylalkalimetal tetramers except for tetramethyl lithium. This prevents a comparison of structural trends found by us with other computations. However, our BP86/TZ2P geometries for (CH₃Li)₄ ecl are in excellent agreement with MP2/6-31+G*⁵ and B3LYP/6-31+G*^{4a} geometries (Table 1). The C–Li bond, for example, is 2.199, 2.188, and 2.195 Å at BP86, MP2, and B3LYP, respectively. Crystal structures of methyl lithium, -sodium, and -potassium always yield the staggered conformation of the methyl groups with respect to the M₃ face to which they are coordinated, whereas we find the eclipsed orientation (CH₃M)₄ ecl to be the lower-energy structure for M = Li (vide infra). The preference for (CH₃Li)₄ ecl over (CH₃Li)₄ stag that we find is confirmed by other theoretical studies^{3e,4a} (at HF/3-21G and B3LYP/6-31+G*), as is the elongation of the C–Li bond going from eclipsed to staggered (see HF/3-21G in Table 1). This is further evidence for intermolecular interactions and crystal packing effects being responsible for the experimentally observed staggered conformation (CH₃Li)₄ stag in methyl lithium crystals.^{7a,c} The importance of crystal environment and packing effects is also suggested by the increasing extent to which the methylalkalimetal aggregates in the crystal deviate from the tetrahedral tetramer structure along Li, Na, and K, as pointed out in the Introduction (see also Chart 1).⁷ The trend of increasing C–M bond distances that we find along M = Li–Rb agrees again well with the trend

emerging from crystal structures, which are available for Li–K (Table 1). No experimental data are available for methylrubidium aggregates. Our C–Li bond distance of 2.213 Å for (CH₃Li)₄ stag is between the 1.5 and 290 K neutron diffraction values of 2.256(6) and 2.204(14) Å found in the tetramers that constitute the methyl lithium crystal. Our C–Na distance of 2.675 Å for (CH₃Na)₄ stag is also between the range of 2.57–2.68 Å found for the slightly distorted tetramer units in the more complex (CD₃Na)₁₆ unit cell of the methyl sodium crystal using neutron and synchrotron diffraction techniques at 1.5 K.^{7c} And, finally, also our C–K distance of 2.962 Å for (CH₃K)₄ stag is nicely between the 1.35 K neutron diffraction values of 2.947 and 3.017 Å found for the CD₃K₃ entities in the (CD₃K)₆ unit cell of the methyl potassium crystal.

3.2. Thermochemistry. Monomers. The thermochemical results of our BP86/TZ2P calculations are collected in Tables 2 (monomers) and 3 (tetramers). Homolytic dissociation of the C–M bond in methylalkalimetal monomers (eq 4) is strongly favored over heterolytic or ionic dissociation (eq 5) for all methylalkalimetal monomers with heterolytic bond dissociation enthalpies (BDE = $-\Delta H$ in Table 2) being up to 5 times higher than the homolytic ones. This is because of the charge separation in the latter (eq 5), which is energetically highly unfavorable in the gas phase (vide infra).



Note that the C–M bond strength in methylalkalimetal monomers decreases if one descends group 1 in the periodic table. The bond enthalpies for both homolytic ($\Delta H_{\text{homo}} = -44.0, -30.3, -26.2$ and -24.6 kcal/mol) and heterolytic dissociation ($\Delta H_{\text{hetero}} = -172.4, -154.0, -129.5,$ and -123.1 kcal/mol) become less bonding along M = Li, Na, K, and Rb (Table 2). This finding is remarkable because it is *not* in line with the current idea of an “ionic” carbon–alkalimetal bond. We will come back to this in section 3.3.

Table 3. Tetramerization Energies and Enthalpies (in kcal/mol) of Methyl Alkalimetal Monomers

monomer	method ^a	tetramerization energies ^b		NIMAG ^c		tetramerization enthalpies ^d		ref
		$\Delta E_{\text{tetra}} \text{ ecl}^e$	$\Delta E_{\text{tetra}} \text{ stag}^f$	ecl ^e	stag ^f	$\Delta H_{\text{tetra}} \text{ ecl}^e$	$\Delta H_{\text{tetra}} \text{ stag}^f$	
CH ₃ Li	BP86/TZ2P	-125.3	-120.8	0	4	-120.3	-118.7	this work
	MP4(SDQ)/6-31+G* ^g	-131.5 ^g		0 ^g				5
	ab initio estimate ^h	-122.9 ^h	-116.0 ^h	0 ^h				3e
CH ₃ Na	BP86/TZ2P	-73.5	-73.6	i	i	-73.1	-73.5	this work
CH ₃ K	BP86/TZ2P	-82.1	-85.2	4	0	-81.4	-82.5	this work
CH ₃ Rb	BP86/TZ2P	-80.4	-85.2	i	i	-79.8	-87.1	this work

^a Energy and structure obtained at the same level of theory, unless stated otherwise. ^b Zero K electronic energies, unless stated otherwise. ^c Number of imaginary frequencies. ^d 298.15 K enthalpies. ^e Tetramer with methyl C–H bonds and metal atoms eclipsed. ^f Tetramer with methyl C–H bonds and metal atoms staggered. ^g Single-point calculation using MP2/6-31+G* geometry, with HF/6-31+G* ZPE correction. ^h Energies computed at HF/3-21G geometry. Difference between MP2/6-31G and HF/6-31G added to the HF/6-31G+6d(C) energy, with MNDO ZPE correction. ⁱ Numerical instabilities prevent accurate determination of NIMAG with ADF. Thermal energies and, thus, enthalpies are less sensitive.

There are no experimental data to which our bond enthalpies can be compared. However, the agreement with other theoretical studies is excellent (Table 2). These studies do not report bond enthalpies but zero K bond energies, and they do not cover heterolytic dissociation (eq 5). The trend of the decreasing homolytic C–M bond energy $\Delta E_{\text{hom}}^{\text{h}}$ is confirmed both at MP2 (computed for Li–K)^{4c} and UB3LYP (computed for Li–Rb),^{3a} and deviations with respect to our values are 1 kcal/mol or less.¹⁶ For the homolytic C–Li bond energy, there is also an MP4(SDTQ)/6-311+G** value (-44.18 kcal/mol) that differs only 0.6 kcal/mol from our BP86/TZ2P result (-44.8 kcal/mol).

Tetramers. Tetramerization is considerably more exothermic for methyllithium than for the heavier methylalkalimetals (see Table 3, Chart 1, and Figure S1). The tetramerization enthalpy of CH₃M is -120.3, -73.5, -82.5, and -87.1 kcal/mol along M = Li, Na, K, and Rb. The equilibrium structure of tetramethylithium has the methyl groups oriented eclipsed with respect to the Li₃ face to which they are coordinated. Enthalpically, the eclipsed structure ($\Delta H_{\text{tetra}} = -120.3$ kcal/mol) is however only slightly preferred over the staggered one ($\Delta H_{\text{tetra}} = -118.7$ kcal/mol), i.e., by 1.6 kcal/mol. Going from Li to Na, the staggered structure (CH₃Na)₄ stag ($\Delta H_{\text{tetra}} = -73.5$ kcal/mol) becomes slightly more stable than the eclipsed structure (CH₃Na)₄ ecl ($\Delta H_{\text{tetra}} = -73.1$ kcal/mol). Note, that the difference in tetramerization energy ΔE_{tetra} is practically zero. In other words, the methyl groups have hardly any barrier for rotation. Proceeding from Na via K to Rb, the enthalpic preference for the staggered structure increases from 0.4 to 1.1 to 7.3 kcal/mol.

No experimental tetramerization enthalpies are available. However, the fact that the methyl groups in crystal structures of methylalkalimetal compounds are found always (i.e., for Li, Na, and K) staggered with respect to the M₃ face to which they are coordinated (Table 1) agrees well with our finding that for Li this orientation is disfavored only very slightly (and can thus easily become the preferred structure through crystal packing effects) and favored for Na–Rb. Other theoretical studies on methylalkalimetal tetramers are only available for lithium (Table 1).^{3e,5} Our tetramerization energy ΔE_{tetra} of -125.3 kcal/mol is between the ab initio estimate of -122.9 kcal/mol by Kaufmann et al.^{3e} and the value of -131.5 obtained by Bickelhaupt et al.⁵ at MP4(SDQ)/6-

Table 4. Analysis of the Carbon–Metal Bond between CH₃* and M* in Methylalkalimetal Monomers^a

	CH ₃ –Li	CH ₃ –Na	CH ₃ –K	CH ₃ –Rb
Bond Energy Decomposition (in kcal/mol)				
ΔE_{A1}	-62.1	-41.7	-37.9	-41.5
ΔE_{E1}	-1.0	-0.5	-0.5	-0.7
ΔE_{O1}	-63.1	-42.2	-38.4	-42.2
ΔE_{Pauli}	38.4	27.8	23.4	30.7
ΔV_{elstat}	-30.3	-23.3	-19.0	-20.4
ΔE_{int}	-55.0	-37.7	-34.0	-31.9
ΔE_{prep}	10.2	6.7	7.4	6.9
ΔE_{hom}	-44.8	-31.0	-26.6	-25.0
Fragment Orbital Overlaps				
$\langle 1a_1 ns \rangle^b$	0.32	0.27	0.24	0.23
$\langle 2a_1 ns \rangle^b$	0.31	0.28	0.21	0.19
$\langle 1e_1 np_{\pi} \rangle^b$	0.22	0.18	0.14	0.14
Fragment Orbital Interaction Matrix Elements (in kcal/mol)				
$\langle 2a_1 F ns \rangle^b, c$	-42.2	-39.1	-25.3	c
Fragment Orbital Populations (in electrons)				
	CH ₃			
1a ₁	1.95	1.99	2.00	2.01
2a ₁	1.40	1.42	1.48	1.45
1e ₁	1.96	1.96	1.97	1.96
	M			
ns ^b	0.50	0.56	0.46	0.49
np _o ^b	0.18	0.03	0.02	0.03
np _π ^b	0.03	0.01	0.01	0.01

^a At BP86/TZ2P. See section 2.2 for explanation of energy terms. ^b $n = 2, 3, 4$ and 5 for $M = \text{Li, Na, K, and Rb}$, respectively. ^c Computed with the fully converged SCF density of CH₃M. Cannot yet be computed for Rb, for technical reasons.

31+G*. The former study^{3e} also confirms the energetic preference for the structure with eclipsed over that with staggered methyl groups (in the latter, this has not been investigated).

3.3. Analysis of the C–M Bond in CH₃M Monomers.

The analyses of the electronic structure and bonding mechanism in methylalkalimetal monomers CH₃M reveal both substantial covalent character for the C–M bond (see Table 4 and Figures 1–3) and a high polarity (see also the section on “Heterolytic Dissociation”, below). In the first place, for all four alkalimetals, the C–M bond is characterized by substantial mixing between the methyl 2a₁ SOMO and the alkalimetal ns AO in the 2a₁ + ns electron-pair bonding

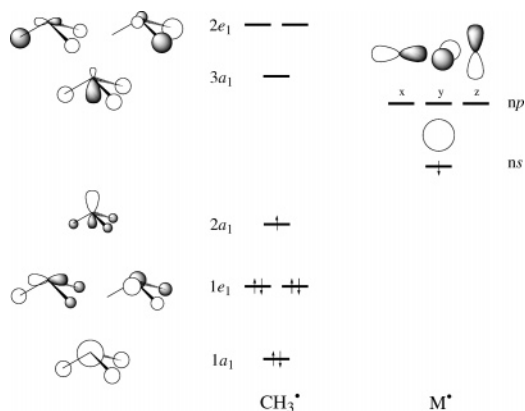


Figure 1. Schematic representation of the valence orbitals of CH_3^* and M^* ($n = 2, 3, 4,$ and 5 for $\text{M} = \text{Li}, \text{Na}, \text{K},$ and Rb).

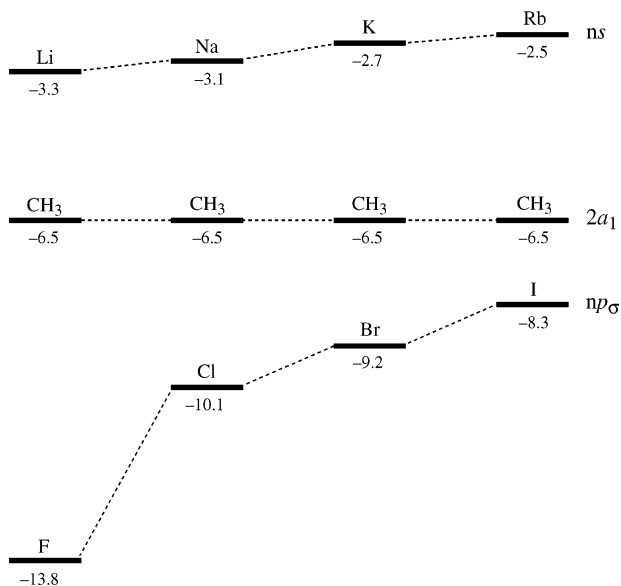


Figure 2. Energies (in eV) of the SOMOs of CH_3^* (in the geometry it adapts in CH_3Li), alkalimetal atoms M^* , and halogen atoms X^* at BP86/TZ2P.

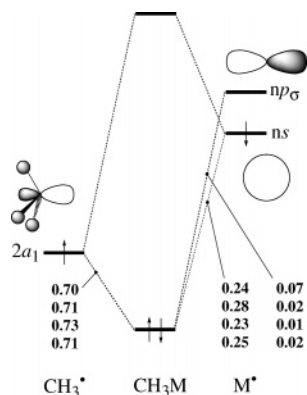


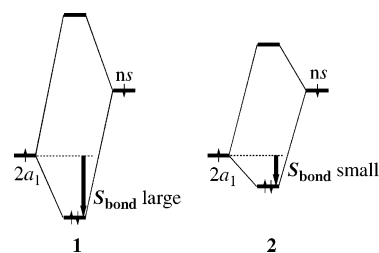
Figure 3. Orbital interaction diagram for CH_3M with Gross Mulliken contributions at BP86/TZ2P of CH_3^* and M^* fragment orbitals to the C–M electron-pair bonding MO for $\text{M} = \text{Li}, \text{Na}, \text{K},$ and Rb .

combination (see Figures 1 and 3). While it is true that the electron-pair bonding $2a_1 + ns$ combination is polarized toward methyl, the alkalimetal ns contribution is significant and not at all marginal: in terms of Gross Mulliken

contributions¹⁷ the composition is approximately 70% $2a_1 + 25\%$ ns (see Figure 3). In case of methyl lithium, the situation is 70% $2a_1 + 24\%$ $2s$ with, in addition, a sizable contribution of 7% from the lithium $2p_\sigma$ AO. In terms of mixing coefficients, this is $0.72 2a_1 + 0.53 2s (+ 0.32 2p_\sigma)$.

The above mixing is indicative for substantial $2a_1 + ns$ orbital interaction, which is confirmed by further analyses. Indeed, the bond interaction-matrix elements $F_{\text{bond}} = \langle 2a_1 | F | ns \rangle$ between the two SOMOs are strongly stabilizing with values ranging from -42.2 (Li) via -39.1 (Na) to -25.3 kcal/mol (K) (see Table 4; F is the effective one-electron Hamiltonian or Fock operator evaluated with the fully converged SCF density of the molecule). We recall that the stabilization $\Delta\epsilon$ of our electron-pair bonding $2a_1 + ns$ combination with respect to $\epsilon(2a_1)$ is, in second order (and neglecting the effect of other occupied and virtual orbitals!), given by $\langle 2a_1 | F | ns \rangle^2 / \epsilon(2a_1) - \epsilon(ns)$, that is, the interaction-matrix element squared divided by the difference in orbital energies.¹⁸ Thus, according to this approximate relationship, the stabilization $\Delta\epsilon$ is a sizable 24 kcal/mol for the C–Li bond, 19 kcal/mol for the C–Na bond, and 7 kcal/mol for the C–K bond (see $\epsilon(2a_1)$, $\epsilon(ns)$, and $\langle 2a_1 | F | ns \rangle$ values in Figure 2 and Table 4). This is a weakening along the C–Li, C–Na, and C–K bonds.

This trend can be straightforwardly understood in terms of the corresponding bond overlap $S_{\text{bond}} = \langle 2a_1 | ns \rangle$, which is sizable and decreases from 0.31 to 0.28 to 0.21 to 0.19 along $\text{M} = \text{Li}, \text{Na}, \text{K},$ and Rb (Table 4). This is caused by the metal ns AOs becoming more diffuse and extended along this series, leading to smaller optimum overlap at longer bond distance.¹⁹ This mechanism, which causes the C–M bond to weaken along Li, Na, K as observed, is illustrated by 1 and 2, below:



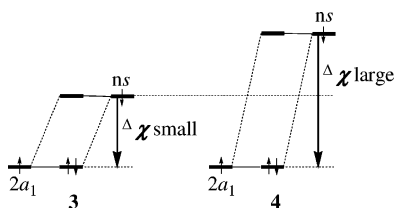
These illustrations show how the stabilization of the electrons in the bonding $2a_1 + ns$ combination is reduced if one goes from a situation with stronger (1) to a situation with weaker (2) $\langle 2a_1 | ns \rangle$ overlap and orbital interaction.

The above picture is confirmed by our quantitative bond energy decomposition. The exact (within our Kohn–Sham MO approach) values of the orbital interactions ΔE_{oi} are clearly larger than the above estimates. Importantly, they are of decisive importance and they show again the same trend. Along the C–Li, C–Na, and C–K bonds, the orbital interactions ΔE_{oi} weaken from -63.1 to -42.2 to -38.4 kcal/mol. The trend resulting from the orbital interactions is further enhanced by the electrostatic attraction ΔV_{elstat} . The latter also weakens along this series owing to the decreasing overlap between occupied orbitals of CH_3^* and M^* because the metal AOs become more extended and diffuse and the C–M bond length increases. For the same reason, the Pauli

repulsion becomes less repulsive along Li, Na, and K. Note that the dominant contributor to the trend in the overall C–M bond energy ΔE is the orbital interactions ΔE_{oi} . Thus, the trend in the thermodynamic stability ΔE (or $\Delta H^{298} = -\text{BDE}$) of the C–M bond, i.e., the weakening along Li, Na, and K, can be related directly to covalent features in the bonding mechanism: the bond overlap between and mixing of the SOMOs that yield the electron-pair bond.

From K to Rb the trend is determined by a more involved and subtle interplay of factors, and we restrict ourselves to the main effect. The step from K to Rb involves the introduction of the first subvalence d shell, i.e., $3d$. This has relatively little effect on the spatial extent of the ns AO, which expands slightly. The bond overlap $\langle 2a_1 | ns \rangle$ further decreases from K to Rb but more slightly so than before (from Na to K). Note however that the $2a_1 + ns$ mixing in CH_3Rb remains substantial (see Figure 3). In the end, the effect of the slight reduction in bond overlap is delicately overruled by that of the increase in stabilization of the electron stemming from the metal ns AO as the orbital energy $\epsilon(ns)$ rises from -2.7 (K) to -2.5 eV (Rb): the orbital interaction ΔE_{oi} becomes somewhat more stabilizing (Table 4 and Figure 2). The presence of the 10 electrons in the subvalence $3d$ shell has a more pronounced effect on the Pauli repulsion: going from K to Rb it becomes 7 kcal/mol more repulsive. This is the reason why overall the C–M bond strength continues to decrease.

Finally, it is interesting to note that if the ionic picture^{2–4f} were correct, one would obtain a trend in orbital interactions that is opposite to the actually observed one. If the C–M bond were predominantly ionic with marginal covalent contributions, the MO carrying the bonding electron pair would have only a slight contribution of the metal ns AO. In other words, this MO would resemble the methyl anion $2a_1$ lone-pair orbital rather than a bonding $2a_1 + ns$ combination. Consequently, it would be hardly stabilized with respect to the methyl $2a_1$ fragment MO. This ionic bonding mechanism is schematically shown in **3** ($\Delta\chi$ refers to the electronegativity difference defined in terms of the orbital-energy difference $\epsilon(2a_1) - \epsilon(ns)$, see ref 20):



In this (fictitious) ionic picture, the electron simply drops from the metal ns into the methyl $2a_1$ giving rise to a stabilization that equals the orbital energy difference $\epsilon(2a_1) - \epsilon(ns)$, indicated in **3** by a bold arrow. Thus, one would expect that the C–M orbital interaction ΔE_{oi} increases if the metal AO energy $\epsilon(ns)$ rises, that is, if the alkali metal becomes more electropositive, because, as shown in **4**, the electron originating from the metal would experience a larger stabilization energy $\epsilon(2a_1) - \epsilon(ns)$. But, above, we have already seen that the opposite happens: the C–M orbital interaction ΔE_{oi} decreases (Table 4: $\Delta E_{oi} = -63.1, -42.2,$

Table 5. Analysis of the Carbon–Metal Bond between $(\text{CH}_3)_4$ and $(\text{M})_4$ in Methyl Alkali Metal Tetramers^a

	$(\text{CH}_3\text{-Li})_4^b$	$(\text{CH}_3\text{-Na})_4^b$	$(\text{CH}_3\text{-K})_4^c$	$(\text{CH}_3\text{-Rb})_4^c$
Bond Energy Decomposition (in kcal/mol)				
ΔE_{A1}	-84.0	-58.5	-63.3	-62.4
ΔE_{T2}	-388.1	-269.0	-265.5	-271.6
ΔE_{rest}	-18.0	-4.5	-4.7	-5.0
ΔE_{oi}	-490.1	-332.0	-333.5	-339.0
ΔE_{Pauli}	502.2	243.4	254.7	268.1
ΔV_{elstat}	-377.8	-190.1	-187.0	-194.6
ΔE_{int}	-365.7	-278.7	-265.8	-265.5
$\Delta E_{\text{prep}}[\text{M}_4]^d$	-6.7	16.4	10.7	16.9
$\Delta E_{\text{prep}}[(\text{CH}_3)_4]^d$	67.9	64.8	63.5	63.4
ΔE_{homo}^d	-304.5	-197.5	-191.6	-185.2
Fragment Orbital Overlaps				
$\langle (\text{CH}_3)_4 (\text{M})_4 \rangle$				
$\langle 2a_1 qa_1 \rangle^e$	0.55	0.49	0.39	0.35
$\langle 3t_2 rt_2 \rangle^f$	0.28	0.24	0.18	0.17
$\langle \text{CH}_3 \text{CH}_3 \rangle$				
$\langle 2a_1 2a_1 \rangle$	0.08	0.04	0.03	0.02
$\langle \text{M} \text{M} \rangle$				
$\langle ns ns \rangle^g$	0.63	0.52	0.49	0.49
$\langle ns np_o \rangle^g$	0.41	0.40	0.40	0.33
Fragment Orbital Populations (in electrons)				
$(\text{CH}_3)_4$				
$2a_1$	1.07	1.29	1.55	1.58
$3t_2$	1.48	1.66	1.65	1.55
$(\text{M})_4$				
qa_1^e	0.84	0.67	0.26	0.26
rt_2^f	0.57	0.21	0.12	0.15

^a At BP86/TZ2P. ^b Tetramer with methyl C–H bonds and metal atoms eclipsed. ^c Tetramer with methyl C–H bonds and metal atoms staggered. ^d $\Delta E_{\text{prep}}[(\text{CH}_3)_4] = \Delta E_{\text{prep}}[4\text{CH}_3^* \rightarrow (\text{CH}_3)_4]$, $\Delta E_{\text{prep}}[\text{M}_4] = \Delta E_{\text{prep}}[4\text{M}^* \rightarrow \text{M}_4]$ and $\Delta E = \Delta E[4\text{CH}_3^* + 4\text{M}^* \rightarrow (\text{CH}_3\text{M})_4]$. ^e $q = 1$ (Li, Na) or 3 (K, Rb). ^f $r = 1$ (Li, Na) or 4 (K, Rb). ^g $n = 2, 3, 4, 5$ for $\text{M} = \text{Li, Na, K, and Rb}$, respectively.

-38.4 kcal/mol) as the metal becomes more electropositive (Figure 2: $\epsilon(ns) = -3.3, -3.1, -2.7$) along $\text{M} = \text{Li, Na, and K}$ (see also ref 20a).

In conclusion, the C–M bond has substantial covalent character stemming from bond overlap that determines largely the trend in bond strength descending the periodic table in group 1. The fact that part of the stabilization stems from bond overlap is not in contradiction with this bond being highly polar.

3.4. Analysis of the C–M Bond in CH_3M Tetramers.

Tetramerization further enhances the covalent character of the C–M bond, as follows from our computations (see Table 5 and Figures 4–6). The C–M bond in the methylalkali metal tetramers has been analyzed in terms of the interaction between the outer tetrahedron of methyl groups $(\text{CH}_3)_4$ and the inner tetrahedral metal cluster M_4 . The frontier orbitals of both fragments $(\text{CH}_3)_4$ and M_4 are energetically arranged in the three-over-one pattern characteristic for tetrahedral species: the bonding combination of four methyl $2a_1$ or alkali metal ns AOs at low orbital energy in A_1 symmetry and the corresponding three antibonding combinations at high orbital energy in T_2 symmetry (see Figure 4). In the valence state of $(\text{CH}_3)_4$ and M_4 , each of these orbitals is singly occupied, and the lowest energy for this configuration is achieved if the four unpaired electrons on either fragment

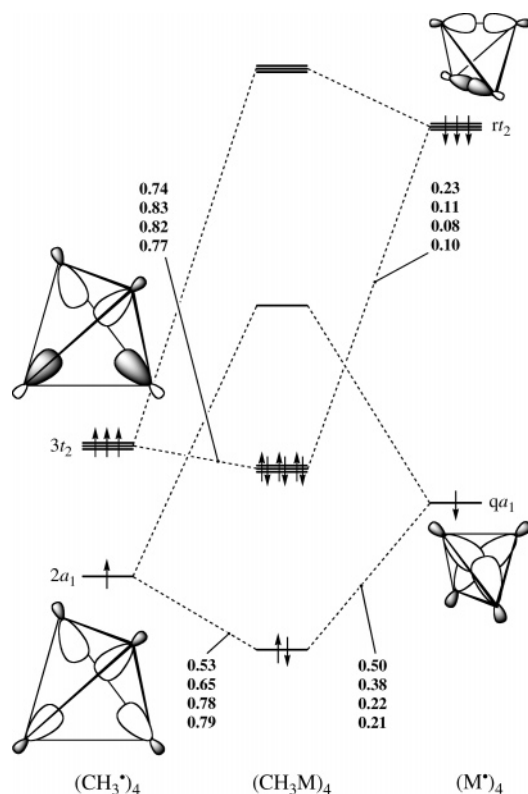


Figure 4. Orbital interaction diagram for $(\text{CH}_3\text{M})_4$ with Gross Mulliken contributions of $(\text{CH}_3^*)_4$ and $(\text{M}^*)_4$ fragment orbitals to the C–M electron-pair bonding MOs in A_1 and T_2 symmetry for $M = \text{Li, Na, K, and Rb}$. For clarity, only one of the 3-fold degenerate $3t_2$ and rt_2 orbitals of $(\text{CH}_3^*)_4$ and $(\text{M}^*)_4$, respectively, is visualized.

have equal spin. Note that this configuration leads in principle to $\text{CH}_3\text{--CH}_3$ and M--M repulsion. In the methylalkalimetal tetramer, four C–M electron-pair bonds are formed between $(\text{CH}_3)_4$ and M_4 : the $2a_1 + qa_1$ combination and the three degenerate $3t_2 + rt_2$ combinations.

Before further examining these bonds, it is important to take a closer look at the formation of the $(\text{CH}_3)_4$ and M_4 fragments from individual methyl radicals and alkalimetal atoms, respectively. The formation of the $(\text{CH}_3)_4$ tetrahedron from four methyl radicals is relatively endothermic with preparation energies $\Delta E_{\text{prep}}[(\text{CH}_3)_4]$ of 63–68 kcal/mol (see Table 5). The major part of this preparation energy, i.e., 62–63 kcal/mol (not shown in Table 5), is associated with methyl pyramidalization caused by the eventual interaction with the alkalimetal cluster in $(\text{CH}_3\text{M})_4$. The remaining part of $\Delta E_{\text{prep}}[(\text{CH}_3)_4]$, that is, the repulsion between the methyl radicals in the $(\text{CH}_3)_4$ tetrahedron is relatively small and decreases from 5.1 to 1.6 to 1.5 to 1.4 kcal/mol along $M = \text{Li, Na, K, and Rb}$ (not shown in Table 5). The reason is simply that the methyl radicals in $(\text{CH}_3)_4$ are far away from each other ($\text{C--C} = 3.597\text{--}4.707 \text{ \AA}$ along Li--Rb , see Table 1) and the methyl $2a_1$ SOMOs cannot build up much overlap ($\langle 2a_1|2a_1 \rangle = 0.08\text{--}0.02$, see Table 5). Therefore, they enter into an only weakly repulsive orbital interaction. This is also reflected by the small energy splitting between the bonding $2a_1$ and antibonding $3t_2$ orbitals of $(\text{CH}_3)_4$ shown in Figure 5.

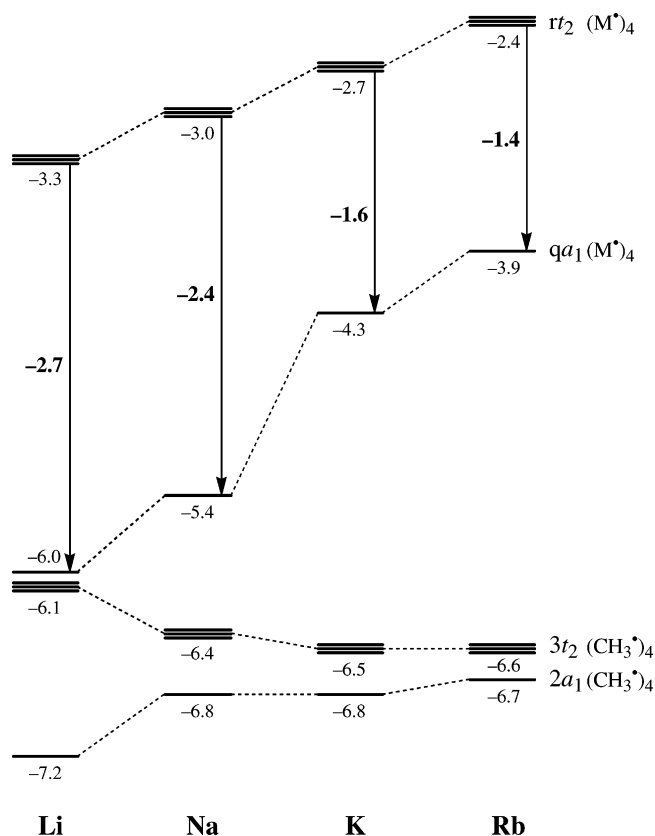


Figure 5. Energies (in eV) of the SOMOs of $(\text{CH}_3^*)_4$ and $(\text{M}^*)_4$ fragments in their valence state in the corresponding methyl alkalimetal tetramers $(\text{CH}_3\text{M})_4$ (eclipsed for Li, Na , staggered for K, Rb , see Figure S1; $q = r = 1$ for $M = \text{Li and Na}$, or $q = 3$ and $r = 4$ for $M = \text{K and Rb}$).

An interesting phenomenon occurs in the alkalimetal clusters M_4 . The metal atoms are in close contact ($\text{M--M} = 2.418\text{--}3.893 \text{ \AA}$ along Li--Rb , see Table 1) leading to remarkably large overlaps between the metal ns AOs ($\langle ns|ns \rangle = 0.63\text{--}0.49$, see Table 5). On the basis of this, one would expect a large energy splitting between the bonding qa_1 and antibonding rt_2 orbitals of the M_4 cluster and, accordingly, a strong M--M repulsion for all alkalimetals. The energy splitting between the bonding qa_1 and antibonding rt_2 combinations is indeed large, especially for Li_4 (see Figure 5). Yet, the preparation energy $\Delta E_{\text{prep}}[\text{M}_4]$ for the lithium cluster is not repulsive but stabilizing by -6.7 kcal/mol. The origin of this effect is stabilizing $ns\text{--}np$ mixing, as illustrated in Figure 6. This occurs in all four metal clusters, but the effect is particularly strong in case of lithium whose $2p$ AOs are at rather low energy. This makes the alkalimetal cluster effectively more electronegative than the isolated alkalimetal atom: the antibonding rt_2 orbitals of the M_4 cluster end up approximately at the same energy as the corresponding ns AOs (instead of at higher orbital energy), and the bonding qa_1 combinations drop enormously in energy, by 2.7, 2.4, 1.6, and 1.4 eV along Li--Rb (Figure 5).

The above has important consequences for the four C–M electron-pair bonds between $(\text{CH}_3)_4$ and M_4 in the methylalkalimetal tetramers. The qa_1 SOMO of Li_4 (at -6.0 eV) is stabilized so much that it begins to approach the energy of the $2a_1$ SOMO of $(\text{CH}_3)_4$ (at -7.2 eV) with which it forms an electron-pair bond (see Figures 4 and 5). This results in

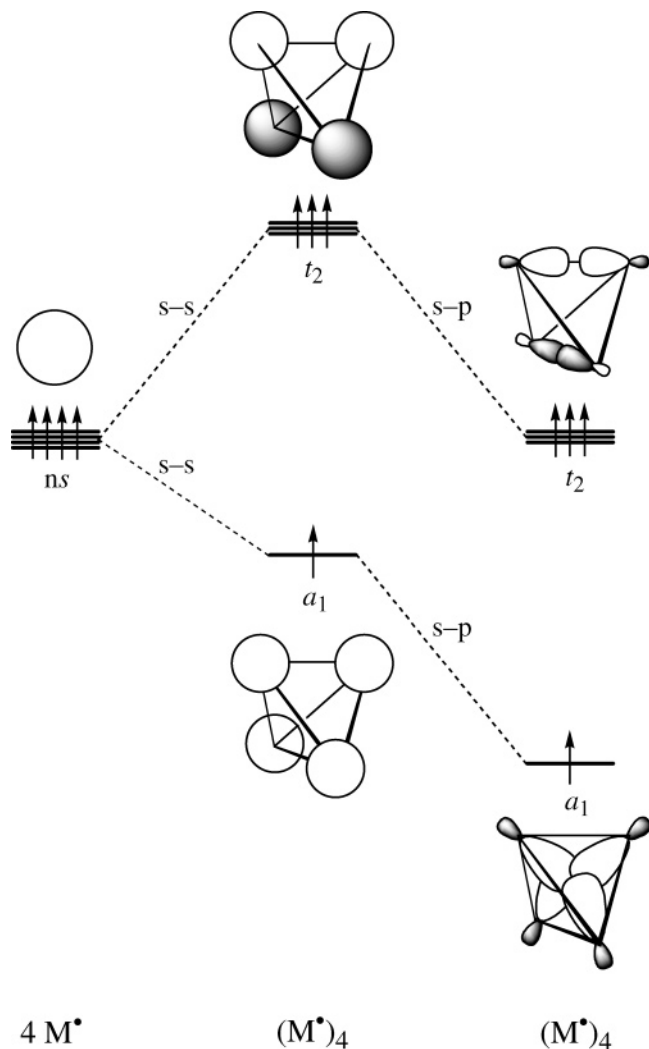


Figure 6. Formation of the SOMOs of the tetrahedral $(M^*)_4$ fragment in its valence state involves two principal interactions: (i) repulsive 4-center-4-electron interaction between the ns AOs of the four atoms (s - s mixing) and (ii) stabilizing admixture of np -derived orbitals (s - p mixing).

a virtually covalent $2a_1 + qa_1$ electron-pair bond with Gross Mulliken contributions¹⁷ of $(CH_3)_4$ (53%) and M_4 (50%) nearly in perfect balance. Thus, in this bond, there is essentially no transfer of charge from lithium to carbon. Going to the methylsodium tetramer, the $2a_1 + qa_1$ electron-pair bond becomes more polarized, but polarity is still reduced if compared to the situation in the monomer (compare Figures 3 and 4). Thereafter, if one goes to potassium and rubidium, the energy of the qa_1 SOMO of M_4 increases steeply, and the $2a_1 + qa_1$ electron-pair bond is no longer less polar in the tetramer than in the monomer. Note that the trend in the corresponding orbital interaction term ΔE_{A1} is dominated by the bond overlap $\langle 2a_1 | qa_1 \rangle$, which decreases from 0.55 to 0.35 along Li–Rb, except for the step from Na to K for which the qa_1 orbital energy leaps from -5.4 to -4.3 eV causing a bond polarization-driven strengthening of ΔE_{A1} as illustrated by **3** and **4** (see Table 5 and Figures 4 and 5). On the other hand, the three degenerate $3t_2 + rt_2$ combinations are, for all four alkali metals, polarized

74–83% toward the methyl tetrahedron, similar to but somewhat more polar than the $2a_1 + ns$ electron-pair bonding combination in the corresponding methylalkalimetal monomers (70–73%, Figure 3). Accordingly, the corresponding C–M orbital interaction energy ΔE_{T2} in the tetramer (see Table 5) is larger than but shows the same trend as ΔE_{oi} in the monomer (see Table 4): there is a pronounced weakening from Li to Na followed by a more subtle decrease from Na to K and increase from K to Rb. The combined orbital interactions ΔE_{oi} between $(CH_3)_4$ and M_4 drop markedly from Li to Na, following the trend in bond overlaps, and they increase marginally along Na–Rb, as a result of a more subtle interplay between the trend in bond overlap and orbital-energy (or electronegativity) difference. Basically, the same picture emerges for the net interaction ΔE_{int} : a strong weakening from Li to Na and marginal changes along Na, K, and Rb. The increased covalency is also found for the C–Li electron-pair bond between one single CH_3 and the Li_4 cluster in CH_3 – Li_4 (see Supporting Information).

In conclusion, the covalent character of the C–M bond increases substantially on tetramerization, especially for Li and to a lesser extent Na, because metal–metal interactions in the central M_4 cluster stabilize the alkali metal orbitals and, in that way, make the alkali metal effectively less electro-positive.

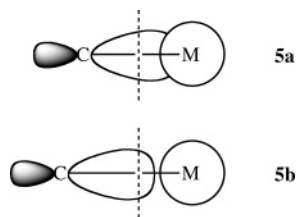
3.5. Analysis of Monomer–Monomer Interactions in CH_3M Tetramers. The most straightforward approach to understanding the stability of the methylalkalimetal tetramers toward dissociation into the four monomers is directly analyzing the interaction between these monomers in the tetramer. The decomposition of the tetramerization energy, shown in Table S1 in the Supporting Information, reveals that the electrostatic interaction ΔV_{elstat} is the dominant bonding force. This term first decreases from -292.7 (CH_3 –Li) to -205.7 (CH_3Na) and increases thereafter to -209.5 (CH_3K) and further to -240.1 kcal/mol (CH_3Rb). The sudden decrease of ΔV_{elstat} from the methyl lithium to the methylsodium tetramer is partially caused by the relatively large increase in C–M distance if one goes from Li to Na. Furthermore, the trend in ΔV_{elstat} parallels the trend in charge separation as reflected by the alkali metal atomic charges collected in Table 6. According to both the VDD and Hirshfeld method, the metal atomic charge in CH_3M decreases from Li to Na and then increases along Na, K, and Rb. This agrees perfectly with the trend in dipole moment: $\mu = 5.6, 5.2, 6.9,$ and 7.7 D along Li, Na, K, and Rb (see Table 6).

These trends can be understood as resulting from the interplay of two effects. The first one is the increasing extent of charge separation that results as the methyl group moves farther away from the metal atom as the latter becomes larger going from Li to Rb. Thus, the negative charge gained by the methyl group due to the formation of the polar $2a_1 + ns$ electron-pair bond penetrates less into the region of the metal atom and is increasingly associated with the carbon atom. This is schematically depicted by **5a** and **5b**, which represent the situation in a shorter and a longer C–M bond, respectively (the dashed lines represent the bond midplanes):

Table 6. Metal Atomic Charge $Q(M)$ of Methylalkalimetal Oligomers, Carbon Atomic Charge $Q(C)$ in Methyl Fluoride, and Dipole Moment μ^a

	Q(M) in CH ₃ M				Q(M) in (CH ₃ M) ₄				Q(F) in CH ₃ F	
	Li	Na	K	Rb	Li ^b	Na ^b	K ^c	Rb ^c	CH ₃ F	CH ₃ F ^d
VDD (au)	0.386	0.351	0.428	0.466	0.143	0.311	0.343	0.333	-0.142	-0.312
Hirshfeld (au)	0.495	0.417	0.493	0.534	0.306	0.428	0.509	0.527	-0.137	-0.295
μ (D)	5.629 ^e	5.212	6.855	7.723	0	0	0	0	1.808	3.989

^a At BP86/TZ2P. ^b Tetramer with methyl C–H bonds and metal atoms eclipsed. ^c Tetramer with methyl C–H bonds and metal atoms staggered. ^d CH₃F with C–F bond elongated to C–Li distance in CH₃Li (2.010 Å). ^e Agrees well with CCSD(T)/MT(ae) value of 5.643 D, see ref 4b.



The above provides an important insight. It shows that atomic charge values not only depend on the extent of interaction and mixing between fragment orbitals or wave functions but also on the bond distance: the larger the bond distance, the larger the charge separation. This is nicely illustrated by a numerical experiment with methyl fluoride: the fluoride atomic charge in CH₃F amounts to -0.142 and -0.137 au using VDD and Hirshfeld, respectively (see Table 6). This is, in absolute terms, much less than the corresponding lithium atomic charges of +0.386 and +0.495 au in CH₃-Li (see Table 6). However, if we elongate the C–F bond in methyl fluoride from its equilibrium value of 1.395 to 2.010 Å (the length of the C–Li bond in methyllithium), the negative fluoride atomic charge increases significantly, although the SOMO–SOMO mixing across the C–F bond slightly decreases: $Q(F)$ in this deformed CH₃F amounts to -0.312 and -0.295 au using VDD and Hirshfeld (see Table 6). Likewise, the weight of the methyl $2a_1$ SOMO in the C–M bonding $2a_1 + ns$ combination increases only marginally along Li–Rb and cannot be held responsible for the significant changes of the atomic charges along this series.

Superimposed on the above mechanism, which on its own would cause a steady increase of the charge separation (see **5a** and **5b**), there is a second effect, namely the loss (or strong reduction) of the participation of the alkalimetal np_σ AO if one goes from Li to Na because the lithium $2p_\sigma$ AO (-1.3 eV) is at lower energy and therefore a better acceptor orbital than, e.g., the sodium $3p_\sigma$ AO (-0.5 eV) (see Figure 3; orbital energies not shown in figure). This counteracts the former mechanism and causes the alkalimetal atomic charge to decrease from Li to Na. This is because the admixture of the lithium $2p_\sigma$ AO to the electron-pair bonding $2a_1 + 2s$ combination enhances polarization of the charge distribution away from the metal and toward carbon thus increasing the charge separation in CH₃Li, while the effect is absent (or negligible) in CH₃Na. This is schematically illustrated by **6** and **7**, respectively.



The orbital interactions ΔE_{oi} between the CH₃M monomers, although much smaller than ΔV_{elstat} , are still important for the cohesion between the monomers, with values ranging from -82.8 kcal/mol for the methyllithium tetramer to -53.1 kcal/mol for the methylrubidium tetramer (Table S1 in the Supporting Information). Note that these orbital interactions do not involve the formation of an electron-pair bond. They are mainly provided by donor–acceptor interactions between occupied σ_{C-M} and unoccupied σ^*_{C-M} orbitals of the monomers. The net interaction energy ΔE_{int} between CH₃M monomers decreases along M = Li–Rb, steeply at first, from -162.9 to -124.5, and then more gradually to -118.5 and further to -112.8 kcal/mol (see Table S1). The main feature of this trend, that is, the steep decrease in monomer–monomer interaction from methyllithium to the heavier methylalkalimetal systems, is preserved in the overall tetramerization ΔE_{tetra} , which varies from -125.3 for CH₃Li to -73.5, -85.2, and -85.2 for CH₃Na, CH₃K, and CH₃-Rb, respectively.

The analyses of the C–M electron-pair bond in the preceding section also provide insight, in a complementary and maybe a somewhat more indirect fashion, into the stabilizing effect of tetramerization, in particular, the cohesion within the inner alkalimetal cluster in the methylalkalimetal tetramers. In the first place, we have seen that considerable $ns-np$ hybridization (Figure 6) of the alkalimetal relieves the M–M repulsion in the valence state of M_4 , and, in case of Li, it even leads to an overall stabilizing interaction of -6.7 kcal/mol (see $\Delta E_{prep}[M_4]$ in Table 5). The cohesion within M_4 is further enhanced by the interaction with the outer tetrahedron of methyl radicals, especially for Li and Na, because for these metals the M–M bonding qa_1 SOMO of M_4 keeps much of its population, whereas the three M–M antibonding rt_2 SOMOs of M_4 are always more strongly depopulated (see Table 5). This is naturally reflected by the overall energy change $\Delta E_{homo} = \Delta E_{int} + \Delta E_{prep}[(CH_3)_4] + \Delta E_{prep}[M_4]$ for the formation of $(CH_3M)_4$ from $4CH_3 + 4M$ (Table 5), the value of which exceeds (i.e., is more stabilizing than) four times the value of ΔE_{homo} for one monomer (Table 4). This excess stabilization is by definition the tetramerization energy ΔE_{tetra} . As we have seen, it decreases indeed steeply if one goes from methyllithium ($\Delta E_{tetra} = -125.3$ kcal/mol) to the heavier congeners ($\Delta E_{tetra} = -73.5$ to -85.2 kcal/mol, see Table S1).

3.6. Polar Bonds and the Concepts of Covalency and Ionicity. Covalency and Ionicity. In the preceding sections, we have established that the C–M bond in methylalkalimetal oligomers has substantial covalent character, especially the C–Li bond in methyllithium, if one considers the sizable

orbital mixing and the fact that the trend in bond strength is dominated by the bond overlap. This is quite at variance with the current picture of this bond being predominantly “ionic”.^{2–4f} Note however that this current view is not based on bond energies and mechanisms but instead on analyses of the charge distribution, using methods such as Streitwieser’s integrated projected population (IPP), Weinhold’s natural population analysis (NPA), and Bader’s atoms in molecules (AIM) approach.²¹ These analyses yield Li atomic charges in methyllithium of +0.8 au with IPP at HF/SS+d,^{2a} +0.85 au with NPA at MP4(SDQ)/6-31+G*,⁵ and +0.90 au with AIM at HF/6-31G**^{4f,1,22} (this agrees well with our AIM value of +0.89 au at BP86/TZ2P).⁹ This led to the idea that the C–Li bond is 80–90% ionic. The problem with quantifying the extent of ionicity on the basis of atomic charges is that different approaches have different scales. The value of the atomic charge of one and the same atom in exactly the same molecule can differ significantly for different methods. The Li atomic charge is, for example, +0.85 au according to NPA,⁵ but according to Hirshfeld¹⁵ and our Voronoi deformation density (VDD) method,¹⁴ it amounts to +0.50 au and +0.39 au, respectively (see Table 6). These values are rather robust regarding the choice of exchange-correlation functional with fluctuations along LDA, BP86, BLYP, OLYP, PW91, and OPBE of 0.03 au for Hirshfeld and only 0.01 au for VDD. Our Hirshfeld and VDD values of +0.50 au and +0.39 au appear to be very close to the GAPT (generalized atomic polar tensors) charge of +0.4178 au computed by Cioslowski at the HF/6-31G** level.^{4k} Again, this does not justify an absolute valuation of 50, 39, or 42% “ionic”. The point is that atomic charges can *not* simply be interpreted as *absolute* bond polarity indicators.⁹ Atomic charges become meaningful, in principle, only through the comparison of trends computed with one and the same method. In this context, it is an asset of direct-space integration methods such as Hirshfeld and especially VDD that they provide a transparent picture of how the electronic density is redistributed among the atoms due to the formation of chemical bonds. Thus, while Hirshfeld and VDD atomic charges differ somewhat in their absolute values, they both indicate that the charge separation across the C–M bond in methylalkalimetal monomers decreases from Li to Na and increases thereafter along Na, K, and Rb (see Table 6; compare discussion about 5–7 in section 3.5). Both methods indicate also that tetramerization leads to a marked decrease of charge separation for M = Li (compare section 3.4).

It is thus desirable that a definition of the extent of ionicity *I* or covalency *C* involves a definition of both the purely ionic and covalent situation. In the context of MO theory, this can be achieved using the relative contribution *x* of the SOMO of one of the fragments, say the more electronegative one (here: the methyl radical), to the C–M electron-pair bonding MO. The purely ionic situation occurs for $x = x_I = 1$: the unpaired electron of the metal atom is completely transferred to the methyl SOMO which transforms, without admixture of the metal AO, into a lone-pair-like MO in the overall molecule. The purely covalent situation occurs for $x = x_C = 0.5$: the radical electrons of both fragments pair-up

Chart 2

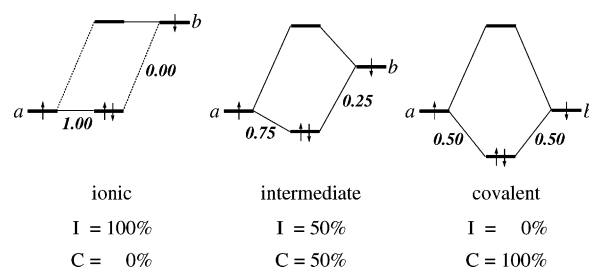


Table 7. Covalency *C* (in %) of Carbon–Metal Bonding in Methyl Alkalimetal Monomers and Tetramers^a

M	CH ₃ M	(CH ₃ M) ₄ ^b A ₁	(CH ₃ M) ₄ ^b T ₂
Li	60 (60)	94 (93)	52 (52)
Na	58 (58)	70 (71)	34 (34)
K	54 (52)	44 (45)	36 (35)
Rb	58 (55)	42 (42)	46 (45)

^a At BP86/TZ2P. *C* is computed with eq 7 using for *x* the Gross Mulliken contribution of the methyl 2a₁ SOMO to the electron-pair bonding combination (see values in Figures 3 and 4). Value in parentheses: idem, using for *x* the Gross Mulliken Population *P* that the methyl 2a₁ SOMO acquires in all occupied MOs of the overall molecule (see values in Tables 4 and 5) divided by 2, i.e., $x = P/2$.
^b Tetramer with methyl C–H bonds and metal atoms eclipsed (Li, Na) or staggered (K, Rb).

in an electron-pair bonding combination of the overall molecule that has equal contributions from either fragment SOMO. The percentage ionicity *I* and covalency *C* is then defined as in eqs 6 and 7, respectively, with $I + C = 100\%$.

$$I = \frac{x - x_C}{x_I - x_C} \times 100\% \quad (6)$$

$$C = \frac{x_I - x}{x_I - x_C} \times 100\% \quad (7)$$

This definition of ionicity and covalency has the advantage of a well-defined scale (i.e., the range of possible values is known), and the interpretation is firmly embedded in the MO model. This is illustrated in Chart 2, which shows the purely ionic and the purely covalent situation as well as a bond of intermediate polarity. We have used this notion of ionicity and covalency already earlier in the discussion in a qualitative fashion. As such, it is in fact widely used throughout MO theory.¹⁸

In Table 7, we have collected percentages of covalency *C* of the C–M bonds of our methylalkalimetal oligomers based on eq 7 using two ways of computing the fraction *x*. In the first one, *x* is the Gross Mulliken contribution of the methyl 2a₁ SOMO to the electron-pair bonding combination in the overall molecule (see values in Figures 3 and 4). Thus, the C–Li electron-pair bond in the methyllithium monomer is 60% covalent and that in A₁ symmetry of the tetramer is even 94% covalent! Covalency is reduced to 52% in the three C–Li electron-pair bonds in T₂ symmetry of the tetramer. The covalent character *C* of the C–M bond decreases from Li to the heavier alkalimetals, slightly so for the monomers, and more pronouncedly for the tetramers (in particular the A₁ component). Also, the difference in *C* between the more covalent A₁ and the more polar T₂ components in the tetramers decreases rapidly along Li–Rb.

There are still other possible ways to compute x , for example, on the basis of the Gross Mulliken population P (see values in Tables 4 and 5) that the methyl $2a_1$ SOMO acquires in all occupied MOs of the overall molecule ($x = P/2$: see values in parentheses in Table 7) or on the basis of fragment MO coefficients (not shown in Table 7). Note that the particular values of C and I depend on how x is computed. One must be aware that this introduces again a certain arbitrariness making C and I semiquantitative rather than quantitative. Nevertheless, any choice for x produces the same trends in C and I , i.e., a nearly constant extent of covalency of the C–M bond going from Li to the heavier alkalimetals, and the occurrence in the tetramer of methyl-lithium and to a lesser extent methylsodium of a more covalent component in A_1 and a more ionic component in T_2 symmetry. The quantities C and I as defined by eqs 6 and 7 also have a more practical disadvantage: they always require the analysis of the orbital electronic structure of the fragments as well as the bonding mechanism in the overall molecule. Thus, computing C and I is much less straightforward than the routine and automated computation of, e.g., VDD or Hirshfeld atomic charges.

Heterolytic Dissociation. So far, we have examined the extent of orbital mixing, its importance for trends in the bond strength and the polarity or charge separation in the C–M bond. Yet another criterion for classifying the C–M bond as covalent is its strong intrinsic preference for dissociating homolytically and not ionically or heterolytically (see section 3.2). To enable a quantitative comparison with other bonds, we have computed the ratio of $\Delta E_{\text{hetero}}/\Delta E_{\text{homo}}$ as a measure for this preference using bond energy values from Table 2. The $\Delta E_{\text{hetero}}/\Delta E_{\text{homo}}$ ratios of CH_3Li – CH_3Rb amount to 3.9, 5.0, 4.9, and 5.0, respectively. Thus, a methyl radical and alkalimetal atom are much closer in energy to the resulting methylalkalimetal molecule than the corresponding ionic fragments. Interestingly, the above $\Delta E_{\text{hetero}}/\Delta E_{\text{homo}}$ ratios of the C–M bonds are higher than the corresponding ratio of the C–H bond in methane which amounts to only 3.8. Apparently, the C–M bond has an even higher relative preference for homolytic dissociation than the slightly polar C–H bond, which is generally taken in organic chemistry as a covalent bond.

Yet, a number of observations has inspired a description of the C–M bond that corresponds to heterolytic or ionic dissociation, namely, in terms of the interaction between a methyl (or, more in general, an organyl) anion and an alkalimetal cation. One of these observations is that, according to IPP, NPA, and AIM, the alkalimetal obtains a large positive charge of 0.8–0.9 au (see, however, above in this section).^{2a,22} Furthermore, it appears that geometries of oligomers and trends in stability can be predicted assuming aggregates consisting of carbanions and metal cations,^{2b,3c,22} although only to some extent.^{3c,4j} And, finally, it is well established that organoalkalimetal compounds react in condensed-phase reactions through ionic mechanisms.¹

The problem with the above is that there is only an indirect relationship between these observations and the extent of polarity of the C–M bond or its preference for either homolytic or ionic dissociation. We have already pointed

Table 8. Analysis of the Carbon–Metal Bond between CH_3^- and M^+ in Methylalkalimetal Monomers^a

	$\text{CH}_3\text{--Li}$	$\text{CH}_3\text{--Na}$	$\text{CH}_3\text{--K}$	$\text{CH}_3\text{--Rb}$
ΔE_{A_1}	–15.2	–16.7	–14.5	–15.7
ΔE_{E_1}	–5.9	–3.8	–3.2	–3.3
ΔE_{oi}	–21.1	–20.5	–17.7	–19.0
ΔE_{Pauli}	44.0	42.5	44.5	55.5
ΔV_{elstat}	–197.4	–178.6	–158.4	–161.9
ΔE_{int}	–174.5	–156.6	–131.6	–125.4
ΔE_{prep}	0.2	0.8	0.6	0.8
ΔE_{hetero}	–174.3	–155.8	–131.0	–124.6

^a Bond energy decomposition (in kcal/mol) at BP86/TZ2P.

out above that the alkalimetal atomic charge is not an absolute bond polarity indicator. Furthermore, the ionic behavior of organoalkalimetal compounds in reactions is inherently a property of the entire reaction system. The latter comprises not only the organoalkalimetal molecule but also all other reactants involved, including the solvent. In general, interactions with solvent molecules promote heterolytic relative to homolytic dissociation because they stabilize situations involving charge separation. Thus, also bonds of which the covalent character is generally accepted, can behave ionically. The C–H bond, for instance, behaves (i.e., dissociates) ionically if a base abstracts a proton from a carbon acid.^{1b} Likewise, nucleophilic substitution is an example of a reaction system in which a (polar) covalent bond (e.g., carbon–halogen or carbon–oxygen) behaves ionically due to the interaction with a nucleophile.^{1b}

Nevertheless, it is instructive to carry out an ionic analysis of the C–M bond, that is, a bond energy decomposition of the interaction between CH_3^- and M^+ in CH_3M (see Table 8) and to compare this with the analysis of the interaction between CH_3^\bullet and M^\bullet in the same molecule (Table 4). In the ionic approach, the classical electrostatic interaction ΔV_{elstat} becomes the dominant bonding term with values that vary from –197.4 to –178.6 to –158.4 to –161.9 kcal/mol along Li–Rb (Table 8). On the other hand, the orbital interaction ΔE_{oi} becomes significantly smaller with values that vary from –21.1 to –20.5 to –17.7 to –19.0 kcal/mol along Li–Rb (Table 8). This has previously been interpreted as suggesting that, compared to the homolytic approach, the charge redistribution in the ionic analysis is smaller, that is, that the ionic fragments correspond more closely to the final charge distribution in the alkalimetal molecule than the neutral methyl and alkalimetal radical fragments.⁵ Another factor, not directly related to the extent of charge redistribution, that may cause the reduced ΔE_{oi} in the ionic analysis is the fact that we lose the stabilization associated with the electron dropping from the SOMO of the metal atom into the C–M bonding MO. The enormous increase in ΔV_{elstat} compared to the homolytic approach (compare Tables 4 and 8) is simply due to the energetically unfavorable charge separation that we enforce by our choice to completely transfer one electron from one of the constituting fragments of CH_3M to the other. It is perfectly valid to carry out such an analysis. Note however that the results refer to the high-energy process of heterolytic bond breaking (eq 5) and *not* to the energetically preferred homolytic bond dissociation (eq 4). Likewise, the ionic C–M interaction ΔE_{int} between

$(\text{CH}_3)_4^{4-}$ and M_4^{4+} in the tetramers $(\text{CH}_3\text{M})_4$ is significantly higher than that between the neutral $(\text{CH}_3)_4$ and M_4 mainly because of the much more stabilizing electrostatic interaction ΔV_{elstat} in the former (compare Tables 5 and S2 in the Supporting Information). The origin is again the energetically highly unfavorable charge separation associated with the complete transfer of four electrons from tetralithium to tetramethyl. The preparation energies $\Delta E_{\text{prep}}[(\text{CH}_3)_4^{4-}]$ and $\Delta E_{\text{prep}}[\text{M}_4^{4+}]$ are highly endothermic mainly because of electrostatic repulsion between the methyl anions and between the alkalimetal cations (see Table S2 in the Supporting Information). Note that the strongly stabilizing electrostatic interaction ΔV_{elstat} between $(\text{CH}_3)_4^{4-}$ and M_4^{4+} compensates for the highly destabilizing preparation energies $\Delta E_{\text{prep}}[(\text{CH}_3)_4^{4-}]$ and $\Delta E_{\text{prep}}[\text{M}_4^{4+}]$. This reflects that the C–M distances in $(\text{CH}_3\text{M})_4$ are shorter than the corresponding C–C and M–M distances (see Table 1).

Comparison with C–X Bond in Methyl Halides. To place our results into a broader chemical context, we have compared the C–M bond in methylalkalimetal monomers CH_3M with the C–X bond in methyl halides CH_3X with $\text{X} = \text{F}, \text{Cl}, \text{Br},$ and I . In a DFT study at BP86 and a basis set similar to ours, Deng et al.²³ found that the trend in C–X bond strength is governed by the difference in electronegativity between CH_3 and X and not by the bond overlap between the methyl $2a_1$ and halogen np_σ SOMOs ($n = 2-5$ along F–I). Thus, the C–X bond strength *decreases* (ref 23a: $\Delta E_{\text{homo}} = -119.4, -87.5, -75.8, -65.2$ kcal/mol) as the halogen atom becomes less electronegative along F–I (this work, Figure 2: $\epsilon(2p_\sigma) = -13.8, -10.1, -9.2, -8.3$ eV), even though the bond overlap increases (ref 23a: $\langle 1a_1 | np_\sigma \rangle = 0.26, 0.34, 0.35, 0.36$). This is highly interesting in the light of our present results. As pointed out earlier, the C–X bond is considered polar covalent and certainly not ionic. Yet, the trend in C–X bond strength of methyl halides along F–I suggests more polar character than the trend in C–M bond strength of methylalkalimetal molecules along Li–K. The explanation is not the absence in methyl halides of covalent features in the bonding mechanism, that is, interaction between and mixing of the methyl and halogen SOMOs.²³ The observed trend is caused instead by the fact that the electronegativity changes much more strongly along the halogen atoms than along the alkalimetal atoms (Figure 2: compare $\epsilon(np_\sigma)$ along F–I with $\epsilon(ns)$ along Li–Rb; see also ref 20a). This agrees well with the fact that the weight of the methyl $2a_1$ SOMO in the bonding $2a_1 + ns$ combination does actually not increase very much along the series (see Figure 3). Therefore, the trend in C–X bond strengths follows the electronegativity of the halogen atoms, whereas the trend in C–M bond strength correlates with the bond overlap.

4. Conclusions

The C–M bond in methylalkalimetal oligomers has substantial covalent character: it can well be viewed as an electron-pair bond between the SOMOs of the methyl radical and alkalimetal atom that gains substantial stabilization from the $\langle 2a_1 | ns \rangle$ bond overlap. This is not in contradiction with this electron-pair bond being highly polar, but it

disqualifies the current classification of the C–M bonding mechanism as “mainly ionic”.

These insights emerge from our quantum-chemical analyses of the methylalkalimetal monomers CH_3M and tetramers $(\text{CH}_3\text{M})_4$ with $\text{M} = \text{Li}, \text{Na}, \text{K},$ and Rb , at BP86/TZ2P. These analyses reveal significant orbital mixing in the C–M electron-pair bond of CH_3M between the methyl $2a_1$ and alkalimetal ns SOMOs (approximately 70% $2a_1 + 25\%$ ns). The C–M bond becomes longer and weaker, both in the monomers and tetramers, if one goes from Li to the larger and more electropositive Rb. Quantitative bonding analyses show that this trend is not only determined by decreasing electrostatic attraction but also, even to a larger extent, by the weakening in orbital interactions. The latter become less stabilizing along Li–Rb because the bond overlap $\langle 2a_1 | ns \rangle$ decreases as the metal ns atomic orbital (AO) becomes larger and more diffuse. Note that for a predominantly ionic bond, one would expect that the orbital interactions are *strengthened* along with the increasing difference in electronegativity between CH_3 and M along Li–Rb. Covalency of the C–M bond is further enhanced in the tetramers, especially for Li and to a lesser extent Na, because in the central M_4 cluster, the alkalimetal becomes effectively less electropositive. The C–M bond has furthermore a slightly stronger intrinsic preference for homolytic dissociation than, for example, the C–H bond, which is generally considered covalent.

Earlier evidence for classifying the C–Li bond as 80–90% ionic based on lithium atomic charges is not conclusive because atomic charges are *no absolute* bond polarity indicators. Different atomic charge methods have different scales and yield evidently different values for the same situation: *only trends* of atomic charges computed with one and the same method can be physically meaningful. Finally, while it is true that the polarity of a bond is the net result of the various features in the bonding mechanism, it is not true that this bonding mechanism and the relative importance of all its features (e.g., electrostatic attraction, bond overlap, charge transfer) can be deduced from the bond polarity in a straightforward manner.

Acknowledgment. Dedicated to Professor Gernot Frenking on the occasion of his 60th birthday. We thank the following organizations for financial support: the HPC-Europa program of the European Union, the Deutsche Akademische Austauschdienst (DAAD), The Netherlands Organization for Scientific Research (NWO), the Ministerio de Educación y Cultura (MEC), the Training and Mobility of Researchers (TMR) program of the European Union, the Dirección General de Enseñanza Superior e Investigación Científica y Técnica (MEC-Spain), and the DURSI (Generalitat de Catalunya). Excellent service by the Stichting Academisch Rekencentrum Amsterdam (SARA) and the Centre de Supercomputació de Catalunya (CESCA) is gratefully acknowledged.

Supporting Information Available: On-scale representation of CH_3M , $(\text{CH}_3\text{M})_4$ ecl, and $(\text{CH}_3\text{M})_4$ stag ($\text{M} = \text{Li}, \text{Rb}$) and additional analyses of the C–Li bond in CH_3Li_4 and monomer–monomer interactions as well as a

heterolytic approach to C–M bonding in methyl alkalimetal tetramers. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) See, for example: (a) Elschenbroich, Ch. *Organometallics*, 3rd ed.; Wiley-VCH: Weinheim: Germany, 2006. (b) March, J. *Advanced Organic Chemistry*, 4th ed.; Wiley-Interscience: New York, 1992.
- (2) (a) Streitwieser, A., Jr.; Williams, J. E., Jr.; Alexandratos, S.; McKelvey, J. M. *J. Am. Chem. Soc.* **1976**, *98*, 4778. (b) Streitwieser, A., Jr. *J. Organomet. Chem.* **1978**, *156*, 1.
- (3) (a) Kremer, T.; Harder, S.; Junge, M.; Schleyer, P. v. R. *Organometallics* **1996**, *15*, 585. (b) El-Nahas, A. M.; Schleyer, P. v. R. *J. Comput. Chem.* **1994**, *15*, 596. (c) Lambert, C.; Schleyer, P. v. R. *Angew. Chem.* **1994**, *106*, 1187; *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 1129. (d) Lambert, C.; Kaupp, M.; Schleyer, P. v. R. *Organometallics* **1993**, *12*, 853. (e) Kaufmann, E.; Raghavachari, K.; Reed, A. E.; Schleyer, P. v. R. *Organometallics* **1988**, *7*, 1597. (f) Bauer, W.; Winchester, W. R.; Schleyer, P. v. R. *Organometallics* **1987**, *6*, 2371. (g) Haeflner, F.; Brinck, T. *Organometallics* **2001**, *20*, 5134. (h) Gohaud, N.; Begue, D.; Pouchan, C. *Chem. Phys.* **2005**, *310*, 85. (i) Gohaud, N.; Begue, D.; Pouchan, C. *Int. J. Quantum Chem.* **2005**, *104*, 773. (j) Bushby, R. J.; Steel, H. L. *J. Organomet. Chem.* **1987**, *336*, C25. (k) Bushby, R. J.; Steel, H. L. *J. Chem. Soc., Perkin Trans 2* **1990**, 1143.
- (4) (a) Kwon, O.; Sevin, F.; McKee, M. L. *J. Phys. Chem. A* **2001**, *105*, 913. (b) Breidung, J.; Thiel, W. *J. Mol. Struct.* **2001**, *599*, 239. (c) Scalmani, G.; Brédas, J. L. *J. Chem. Phys.* **2000**, *112*, 1178. (d) Fressigné, C.; Maddaluno, J.; Giessner-Prettre, C. *J. Chem. Soc., Perkin. Trans 2* **1999**, 2197. (e) Tyerman, S. C.; Corlett, G. K.; Ellis, A. M.; Claxton, T. A. *J. Mol. Struct. (THEOCHEM)* **1996**, *364*, 107. (f) Wiberg, K.; Breneman, C. M. *J. Am. Chem. Soc.* **1990**, *112*, 8765. (g) Schiffer, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1986**, *124*, 172. (h) Dill, J. D.; Schleyer, P. v. R.; Binckley, J. S.; Pople, J. A. *J. Am. Chem. Soc.* **1977**, *99*, 6159. (i) Graham, G. D.; Marynick, D. S.; Lipscomb, W. N. *J. Am. Chem. Soc.* **1980**, *102*, 4572. (j) Sapse, A. M.; Raghavachari, K.; Schleyer, P. v. R.; Kaufmann, E. *J. Am. Chem. Soc.* **1985**, *107*, 6483. (k) Cioslowski, J. *J. Am. Chem. Soc.* **1989**, *111*, 8333. (l) Ponec, R.; Roithová, J.; Gironés, X.; Lain, L.; Torre, A.; Bochicchio, R. *J. Phys. Chem. A* **2002**, *106*, 1019.
- (5) Bickelhaupt, F. M.; van Eikema Hommes, N. J. R.; Fonseca Guerra, C.; Baerends, E. J. *Organometallics* **1996**, *15*, 2923.
- (6) Experimental structures of methyl alkalimetal monomers: (a) Grotjahn, D. B.; Pesch, T. C.; Brewster, M. A.; Ziurys, L. M. *J. Am. Chem. Soc.* **2000**, *122*, 4735. (b) Grotjahn, D. B.; Apponi, A. J.; Brewster, M. A.; Xin, J.; Ziurys, L. M. *Angew. Chem., Int. Ed. Engl.* **1998**, *37*, 2678. (c) Grotjahn, D. B.; Pesch, T. C.; Xin, J.; Ziurys, L. M. *J. Am. Chem. Soc.* **1997**, *119*, 12368. (d) Andrews, L. *J. Chem. Phys.* **1967**, *47*, 4834.
- (7) Experimental structures of methyl alkalimetal tetramers: (a) Weiss, E.; Lambertsen, T.; Schubert, B.; Cockcroft, J. K.; Wiedenmann, A. *Chem. Ber.* **1990**, *123*, 79. (b) Weiss, E.; Corbelin, S.; Cockcroft, J. K.; Fitch, A. N. *Angew. Chem.* **1990**, *102*, 728; *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 650. (c) Weiss, E.; Corbelin, S.; Cockcroft, J. K.; Fitch, A. N. *Chem. Ber.* **1990**, *123*, 1629. (d) Weiss, E.; Lambertsen, T.; Schubert, B.; Cockcroft, J. K. *J. Organomet. Chem.* **1988**, *358*, 1. (e) Weiss, E.; Hencken, G. *J. Organomet. Chem.* **1970**, *21*, 265.
- (8) (a) Günther, H.; Moskau, D.; Bast, P.; Schmalz, D. *Angew. Chem., Int. Ed. Engl.* **1987**, *26*, 1212. (b) Bauer, W.; Schleyer, P. v. R. *Adv. Carbanion Chem.* **1992**, *1*, 89. (c) Bauer, W. *Lithium Chemistry*; Wiley-Interscience: New York, 1995. (d) Fraenkel, G.; Martin, K. V. *J. Am. Chem. Soc.* **1995**, *117*, 10336. (e) Ebel, H. F. *Tetrahedron* **1965**, *21*, 699. (f) Armstrong, D. R.; Perkins, P. G. *Coord. Chem. Rev.* **1981**, *38*, 139.
- (9) Fonseca Guerra, C.; Handgraaf, J.-W.; Baerends, E. J.; Bickelhaupt, F. M. *J. Comput. Chem.* **2004**, *25*, 189.
- (10) Bickelhaupt, F. M.; Baerends, E. J. In *Rev. Comput. Chem.*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 2000; Vol. 15, pp 1–86.
- (11) (a) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; van Gisbergen, S. J. A.; Fonseca Guerra, C.; 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) Fonseca Guerra, C.; Visser, O.; Snijders, J. G.; te Velde, G.; Baerends, E. J. In *Methods and Techniques for Computational Chemistry*; Clementi, E., Corongiu, G., Eds.; STEF: Cagliari, 1995; pp 305–395. (d) Baerends, E. J.; Ellis, D. E.; Ros, P. *Chem. Phys.* **1973**, *2*, 41. (e) Boerrigter, P. M.; te Velde, G.; Baerends, E. J. *Int. J. Quantum Chem.* **1988**, *33*, 87. (f) te Velde, G.; Baerends, E. J. *J. Comput. Phys.* **1992**, *99*, 84. (g) Snijders, J. G.; Baerends, E. J.; Vernooijs, P. *At. Nucl. Data Tables* **1982**, *26*, 483. (h) Krijn, J.; Baerends, E. J. *Fit Functions in the HFS Method; Internal Report* (in Dutch); Vrije Universiteit: Amsterdam, 1984. (i) Slater, J. C. *Quantum Theory of Molecules and Solids Vol. 4*; McGraw-Hill: New York, 1974. (j) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200. (k) Becke, A. D. *J. Chem. Phys.* **1986**, *84*, 4524. (l) Becke, A. *Phys. Rev. A* **1988**, *38*, 3098. (m) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822 (Erratum: *Phys. Rev. B* **1986**, *34*, 7406). (n) Fan, L.; Ziegler, T. *J. Chem. Phys.* **1991**, *94*, 6057.
- (12) Atkins, P. W. *Physical Chemistry*; Oxford University Press: Oxford, 1982.
- (13) (a) Morokuma, K. *J. Chem. Phys.* **1971**, *55*, 1236. (b) Kitaura, K.; Morokuma, K. *Int. J. Quantum Chem.* **1976**, *10*, 325. (c) Bickelhaupt, F. M.; Nibbering, N. M. M.; van Wezenbeek, E. M.; Baerends, E. J. *J. Phys. Chem.* **1992**, *96*, 4864. (d) Ziegler, T.; Rauk, A. *Inorg. Chem.* **1979**, *18*, 1558. (e) Ziegler, T.; Rauk, A. *Theor. Chim. Acta* **1977**, *46*, 1.
- (14) The Voronoi deformation density (VDD) method was introduced in ref 5. See also: (a) ref 9. (b) Fonseca Guerra, C.; Bickelhaupt, F. M.; Snijders, J. G.; Baerends, E. J. *Chem. Eur. J.* **1999**, *5*, 3581. Voronoi cells are equivalent to Wigner-Seitz cells in crystals; for the latter, see: (c) Kittel, C. *Introduction to Solid State Physics*; Wiley: New York, 1986.
- (15) Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129.
- (16) Note that, for a proper comparison, we have to remove the zero-point vibrational energy (ΔZPE) correction (obtained at the HF level) that is contained in the UB3LYP bond energies of ref 3a using our own BP86 ΔZPE corrections, which yields UB3LYP values for ΔE_{homo} of -45.7 , -31.1 , -27.2 , and -24.6 kcal/mol along $M = \text{Li, Na, K, and Rb}$.
- (17) The description of the MO in terms of fragment MO coefficients instead of Gross Mulliken contributions yields the same picture, but it has the disadvantage of not being normalized, that is, the figures do not add up to 1 (or to 100%). Note that diffuse fragment MOs (often at higher energy) can make unphysical negative Gross Mulliken contributions, which are then compensated by positive Gross

Mulliken contributions that exceed 100%. This problem occurs in a slight form also in our systems in which the sum of the main positive Gross Mulliken contributions to the C–M electron-pair bonding combination are in some cases 101%–103% (M = Li and Na in Figures 3 and 4).

- (18) (a) Albright, T. A.; Burdett, J. K.; Whangbo, M.-H. *Orbital Interactions in Chemistry*; Wiley-Interscience: New York, 1985. (b) Gimarc, B. M. *Molecular Structure and Bonding – The Qualitative Molecular Orbital Approach*; Academic Press: New York, 1979.
- (19) The C–M bond distance also increases along Li–Rb because of the increasing number of metal core shells that enter into Pauli repulsion with closed shells on the methyl fragment. For a discussion on how the interplay of bonding and repulsive orbital interactions determines bond lengths, see, for example: Bickelhaupt, F. M.; DeKock, R. L.; Baerends, E. J. *J. Am. Chem. Soc.* **2002**, *124*, 1500.
- (20) (a) Mann, J. B.; Meek, T. L.; Allen, L. C. *J. Am. Chem. Soc.* **2000**, *122*, 2780. (b) Allen, L. C. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998; Vol. 2.
- (21) (a) Collins, J. B.; Streitwieser, A., Jr.; McKelvey, J. M. *J. Comput. Chem.* **1979**, *3*, 79. (b) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899. (c) Bader, R. W. F. *Atoms in Molecules, A Quantum Theory*; Clarendon Press: Oxford, U.K., 1990.
- (22) Wiberg, K. B.; Rablen, P. R. *J. Comput. Chem.* **1993**, *14*, 1504.
- (23) (a) Deng, L.; Branchadell, V.; Ziegler, T. *J. Am. Chem. Soc.* **1994**, *116*, 10645. For a related study, see: (b) Bickelhaupt, F. M.; Ziegler, T.; Schleyer, P. v. R. *Organometallics* **1996**, *15*, 1477.

CT050333S

JCTC

Journal of Chemical Theory and Computation

Heisenberg Exchange in Dinuclear Manganese Complexes: A Density Functional Theory Study

Elias Rudberg,* Paweł Sałek, Zilvinas Rinkevicius, and Hans Ågren

*Department of Theoretical Chemistry, Royal Institute of Technology,
SE-10691 Stockholm, Sweden*

Received December 21, 2005

Abstract: This work presents a systematic investigation of the performance of broken symmetry density functional theory for the evaluation of Heisenberg exchange constants. We study dinuclear $\text{Mn}^{\text{IV}}\text{--Mn}^{\text{IV}}$ complexes with bis(μ -oxo), bis(μ -oxo)(μ -carboxylato), and tris(μ -oxo) cores for this purpose, as these are of fundamental biological interest as well as being potential precursors for molecular magnets based on manganese complexes, the so-called Mn_{12} magnets. The obtained results indicate that quantitative agreement with available experimental data for the Heisenberg exchange constants can be achieved for most of the investigated complexes but also that there are significant failures for some compounds. We evaluate factors influencing the accuracy of obtained results and examine effects of different mappings between broken symmetry and Heisenberg Hamiltonian states in an attempt to formulate a reliable recipe for the evaluation of magnetic coupling in these complexes. An assessment of the bonding situation in the molecular system under investigation is found crucial in choosing the appropriate scheme for evaluation of the Heisenberg exchange constants.

I. Introduction

Contemporary research in magnetic properties of chemical compounds and solid-state materials often addresses the coupling between localized spins in terms of an empirical Heisenberg Hamiltonian.¹ In this description the magnetic spin coupling is parametrized in a pairwise manner via so-called Heisenberg exchange constants. These constants find applications primarily in mutually related fields of science devoted to paramagnetic compounds, such as (i) theory of magnetism, where the Heisenberg exchange constants are one of the key magneto-structural parameters of solid or molecular magnets required for characterization of their magnetization,² and (ii) electron paramagnetic resonance (EPR) and nuclear magnetic resonance (NMR) spectroscopies of compounds with multiple localized electronic spins, where the Heisenberg exchange constants enter the spin Hamiltonian and are determined along with other parameters unique for EPR or NMR spectra.^{3–5} The dependence of magnetic coupling on the electronic structure of molecular fragments carrying a localized spin and on localized spins arrangements in general has thus been of considerable interest in the field of theory of magnetism. This interest has also been greatly

spurred by the search for spin arrangements of single molecular magnets in which the magnetic coupling occurs between transition-metal ions in a complex with their ligand environments.^{2,6} Another factor stimulating research in this direction refers to attempts to synthesize organic magnets suitable for practical applications, as these require microscopic understanding of the mechanisms for the magnetic coupling between organic radicals in polymer hosts in order to guide synthesis efforts.⁷

In the field of EPR and NMR spectroscopies the interest in magnetic couplings in molecular systems with multiple localized spins is mainly motivated by the aid they may provide in interpretation of measurements and in determination of geometrical and electronic structure through the measurements.^{3–5} The need for microscopic understanding of magnetic coupling is particularly evident in applications of high field EPR spectroscopy on active sites of enzymes, where the interpretation of EPR spectra otherwise becomes more of an art than science. One notorious example of this kind is the oxygen evolving center in photosystem II, where different oxidation states have been assigned to manganese ions by various interpretations of EPR spectra (see for

example discussion in ref 8). The advances in these fields of research clearly make computational methods for the evaluation of the Heisenberg exchange constants highly desirable.

Computations of magnetic coupling as represented by Heisenberg exchange constants in molecular systems have posed a long standing problem in quantum chemistry, as calculations of this kind require an accurate description of electron correlation, both static and dynamic, as well as a reliable mapping between computed electronic states and states featured in the Heisenberg Hamiltonian.^{9–11} These requirements actually quite severely limit the choice of methodology that can be suitable for calculations. In the domain of ab initio methods various configuration interaction as well as multireference perturbation theory methods can be successfully applied to compute Heisenberg exchange constants in small molecules.^{9,10} For larger molecular systems, the broken symmetry density functional theory (BS-DFT) approach¹² has mostly been used for this purpose.^{6,9–11,13–15} However, the mapping between broken symmetry states and Heisenberg Hamiltonian states is nonobvious, especially if localized spin centers have more than one unpaired electron, as pointed out, for example, by L. Noodleman¹² and F. Neese.¹⁰ Despite this disadvantage of the BS-DFT approach, it is currently the only option for investigations of magnetic coupling in large molecular systems of experimental interest, and it has therefore been applied to a quite wide set of problems, ranging from investigations of molecular magnets to interpretation of EPR spectral parameters in paramagnetic transition-metal systems (see e.g. refs 14 and 16).

In the present paper we investigate the Heisenberg exchange constants in dinuclear manganese (Mn^{IV}–Mn^{IV}) complexes with bis(μ -oxo), bis(μ -oxo)(μ -carboxylato), and tris(μ -oxo) cores. The main focus is to assess the performance of the broken symmetry DFT formalism for evaluation of magnetic coupling, using for the purpose a system that is of fundamental biological interest, and to evaluate factors influencing the accuracy of obtained results. Apart from this, we also examine effects of different mappings between broken symmetry and Heisenberg Hamiltonian states in an aim to formulate a reliable recipe for evaluation of magnetic coupling in manganese complexes. The latter effort is a preparatory step for a future investigation of molecular magnets based on manganese complexes, the so-called Mn₁₂ magnets.

II. Computational Details

Magnetic coupling between localized spins in the Heisenberg Hamiltonian is described via empirical parameters; the Heisenberg exchange constants J_{AB} ¹

$$H = -2J_{AB}S_A \cdot S_B \quad (1)$$

where S_A and S_B are spins localized on centers A and B, respectively. A key to a successful computation of J_{AB} with BS-DFT is an appropriate mapping between broken symmetry, high spin, and Heisenberg Hamiltonian states. A theoretically well justified mapping scheme has been pro-

posed by L. Noodleman,¹² in which J_{AB} is defined as

$$J_{AB} = \frac{E_{BS} - E_{HS}}{S_{\max}^2}, \quad (2)$$

where E_{BS} and E_{HS} are the energies of the broken symmetry and high spin states obtained with the unrestricted DFT formalism, and where S_{\max} is the number of the unpaired electrons in the molecular fragment carrying spin S_A (assuming that the S_B spin related fragment has the same number of unpaired electrons as the S_A fragment). This evaluation scheme for J_{AB} is applicable for weakly bonded molecular fragments between which the magnetic orbital overlap is small.^{9,10,12} An alternative mapping scheme in the BS-DFT approach has been used by E. Ruiz and co-workers,¹⁷ following the work of Noodleman.^{12,18} In this scheme, J_{AB} is computed as

$$J_{AB} = \frac{E_{BS} - E_{HS}}{S_{\max}(S_{\max} + 1)} \quad (3)$$

This scheme assumes strong bonding between molecular fragments with localized spins i.e., when the overlap between magnetic orbitals of the spins is non-negligible. Consequently, this mapping scheme should probably be more acceptable for treating dinuclear manganese complexes than the one proposed by L. Noodleman (see eq 2). The above-described schemes for computation of Heisenberg exchange constants correspond in fact to two limiting cases of bonding situations between localized spins, namely weak and strong bonding between molecular fragments carrying localized S_A and S_B spins. Therefore, an assessment of the bonding situation in the molecular system under investigation is helpful in choosing between the J_{AB} evaluation schemes. Another way to compute J_{AB} with the BS-DFT approach, which is independent of the bonding situation in the molecule, has been proposed by M. Nishino et. al. as¹⁹

$$J_{AB} = \frac{E_{BS} - E_{HS}}{\langle S^2 \rangle_{HS} - \langle S^2 \rangle_{BS}} \quad (4)$$

Here $\langle S^2 \rangle_{HS}$ and $\langle S^2 \rangle_{BS}$ are the total spin angular momentum expectation values for high spin and broken symmetry states obtained with the unrestricted DFT formalism. This evaluation scheme indirectly accounts for the overlap between magnetic orbitals of localized spins by employing expectation values of the total spin angular momentum, $\langle S^2 \rangle_{HS}$ and $\langle S^2 \rangle_{BS}$. Here it is appropriate to mention that the S^2 operator expectation values are well defined for the unrestricted Hartree–Fock method, while in the context of unrestricted density functional theory the S^2 operator expectation values obtained from Kohn–Sham orbitals is not a well-defined procedure.²⁰ Despite this drawback, this approach is at the first glance most suitable for complexes such as the various dinuclear manganese complexes investigated in the present work. We nevertheless employ all three above-described J_{AB} evaluation schemes in order to assess their performance and suitability for computation of J_{AB} in general Mn^{IV}–Mn^{IV} complexes with different core arrangements. Previous theo-

Table 1: Results of Evaluation of Heisenberg Exchange Constants between Mn^{IV} Centers in Various Manganese Compounds (cm⁻¹) Using B3LYP and CAMB3LYP Exchange-Correlation Functionals^a

complex	B3LYP			CAMB3LYP			exp
	J_{AB}^b	J_{AB}^c	J_{AB}^d	J_{AB}^b	J_{AB}^c	J_{AB}^d	
Mn ₂ O ₂ (pic) ₄	-112.5	-84.4	-111.2	-99.1	-74.3	-98.2	-86.5 ^e
[Mn ₂ O ₂ Cl ₂ (bpea) ₄] ²⁺	-144.2	-108.1	-159.6	-130.3	-97.8	-129.1	-144 ^f
[Mn ₂ O ₂ (phen) ₄] ⁴⁺	-131.9	-98.9	-130.4	-124.3	-93.3	-123.4	-147 ^g
[Mn ₂ O ₂ (OAc)(bpea) ₂] ³⁺	-36.0	-27.0	-35.7	-23.8	-17.8	-23.7	-124 ^h
[Mn ₂ O ₂ (OAc)(Me ₄ dtne)] ³⁺	-37.5	-28.1	-37.2	-30.7	-23.0	-30.5	-100 ⁱ
[Mn ₂ O ₃ (Me ₃ tacn) ₂] ²⁺	-382.7	-287.0	-376.4	-370.7	-278.0	-367.9	-390 ^j

^a Ahlrich's VTZ basis set employed in all calculations. ^b J_{AB} evaluated using eq 2. ^c J_{AB} evaluated using eq 3. ^d J_{AB} evaluated using eq 4. ^e J_{AB} experimental data taken from ref 23. ^f J_{AB} experimental data taken from ref 24. ^g J_{AB} experimental data taken from ref 25. ^h J_{AB} experimental data taken from ref 26. ⁱ J_{AB} experimental data taken from ref 27. ^j J_{AB} experimental data taken from ref 28.

retical investigations of dinuclear manganese complexes can be found in the literature.^{9,21,22}

The selected test set of dinuclear manganese (Mn^{IV}–Mn^{IV}) compounds includes three molecules with bis(μ -oxo) core (Mn₂O₂(pic)₄, [Mn₂O₂Cl₂(bpea)₄]²⁺, [Mn₂O₂(phen)₄]⁴⁺), two molecules with bis(μ -oxo)(μ -carboxylato) core ([Mn₂O₂(OAc)(bpea)₂]³⁺, [Mn₂O₂(OAc)(Me₄dtne)]³⁺), and one molecule with tris(μ -oxo) core ([Mn₂O₃(Me₃tacn)₂]²⁺). We use the following notation for the ligands: bpea – *N,N*-bis(2-pyridylmethyl)ethyldiamine, Me₄dtne – 1,2-bis(4,7-dimethyl-1,4,7-triazacyclonon-1-yl)ethane, Me₃tacn – 1,4,7-trimethyl-1,4,7-triazacyclonane, OAcH – methanecarboxylic acid, phen – 1,10-phenanthroline, and picH – picolinic acid. The geometries of the enlisted compounds, employed in all calculations, have been obtained combining crystallographic data^{23–28} and molecular force field geometry optimization, where the positions of the heavy element atoms (C, N, O, Mn) have been taken from crystal structures and where the positions of the hydrogen atoms have been optimized (keeping heavy atoms positions fixed) using the MMFF94 force field²⁹ implemented in the Spartan program.³⁰ The only exception from this procedure is the Mn₂O₂(pic)₄ molecule for which the positions of the hydrogen atoms are available in the crystal structure data. Apart from building full size geometries for the above enumerated compounds, we also created reduced models of them in order to investigate the possibility to employ only rudiment ligand structures instead of full size ligands in the calculations of Heisenberg exchange constants as this would allow for a significant reduction of the computational cost. The reduced models were designed by substituting the ligands encountered in the manganese complexes by similar smaller ligands as bpea with (CHCH₂-NH)₂NH, Me₄dtne with 1,2-bis(1,4,7-triazacyclonon-1-yl)ethane, Me₃tacn with 1,4,7-triazacyclonane, phen with (CHNH)₂, and pic with NHCHCO₂. The position for heavy atoms (C, N, O, Mn) in the reduced ligands have been selected to be the same as in the nonreduced ligands, and the positions of the hydrogen atoms have been optimized employing an analogous procedure as in the case of optimizing the nonreduced compounds geometries.

The evaluation of J_{AB} constants for all compounds have been carried out employing the B3LYP exchange-correlation functional.^{31–34} This functional is found to perform relatively well in our calculations of the Heisenberg exchange constants for cases in which the BS-DFT approach is applicable. However, for large molecular systems such as Mn₁₂-type

single molecular magnets, the incorrect asymptotic behavior of the exchange part of this functional can become a potential problem. To sort out this problem, which in principle can hamper efforts of accurate evaluation of J_{AB} in large molecules, we investigated the suitability of the Coulomb attenuated model of the B3LYP functional (CAMB3LYP)³⁶ for computation of Heisenberg exchange constants. Results of this investigation are tabulated in Table 1 along with B3LYP and experimental results. CAMB3LYP predicts J_{AB} values to be on average 15 cm⁻¹ smaller (in terms of absolute values) compared to B3LYP calculation results. Consequently, the current CAMB3LYP exchange-correlation functional parametrization does not allow for the reproduction of the good performance of the ordinary B3LYP functional. Despite this deficiency of the CAMB3LYP functional in its current form, the in principle “more” asymptotically correct behavior of the exchange part of this functional might be advantageous in evaluation of J_{AB} in extended systems. To achieve better accuracy in these calculations one thus needs to make additional efforts to parametrize the CAMB3LYP functional.

The calculations reported in this work were carried out using our recently developed quantum chemistry program ErgoSCF,³⁵ in which we have implemented functionality for generating starting guesses and monitoring the electron spin density in unrestricted DFT calculations. The ErgoSCF program also includes an implementation of the CAMB3LYP functional.³⁶ We employed four different basis sets: 3-21G,³⁷ 6-31G,³⁸ 6-31G*,³⁸ and AhlrichsVTZ.³⁹ This selection is motivated mainly by the attempt to find the smallest possible basis set which still provides reliable J_{AB} constants for further large scale computations of the Heisenberg exchange constants in molecular magnets.

Before concluding the computational section we want briefly to discuss two technical details related to the evaluation of broken symmetry states in the unrestricted Kohn–Sham formalism. First, a reasonable starting guess for the broken symmetry state is crucial to achieve convergence in the unrestricted calculations. In our case the starting density for the broken symmetry state has been constructed in the following way: The alpha- and beta-densities of a converged HS solution are combined to form a total density matrix and a spin density matrix. The spin density matrix is then modified by changing the sign of the matrix elements that correspond to one of the Mn centers. Starting guesses for the alpha- and beta- density matrices for the BS state

Table 2: Heisenberg Exchange Constants between Mn^{IV} Centers in Mn₂O₂(pic)₄ in cm⁻¹^g

basis set	model ^a	J_{AB}^b	J_{AB}^c	J_{AB}^d	$\langle S^2 \rangle_{BS}$	$\langle S^2 \rangle_{HS}$
3-21G	full	-101.7	-76.3	-100.7	3.032	12.119
6-31G	full	-114.0	-85.5	-112.7	3.013	12.114
6-31G*	full	-110.3	-82.7	-109.1	3.009	12.109
AhlrichsVTZ	full	-112.5	-84.4	-111.2	3.005	12.111
3-21G	reduced	-104.5	-78.4	-103.3	3.026	12.130
6-31G	reduced	-117.9	-88.4	-116.4	3.007	12.123
6-31G*	reduced	-112.5	-84.4	-111.1	3.003	12.116
AhlrichsVTZ	reduced	-116.5	-87.4	-115.0	2.998	12.118
MIDI ^e	reduced	-126.0	-94.7	-124.0		
MIDI+pol(pdf) ^e	reduced	-117.0	-87.6	-115.0		
exp ^f			-86.5			

^a J_{AB} model defines geometry of compound used in calculations, where “full” denotes whole compound and “reduced” denotes the smaller size model of it. ^b J_{AB} evaluated using eq 2. ^c J_{AB} evaluated using eq 3. ^d J_{AB} evaluated using eq 4. ^e J_{AB} B3LYP calculations results taken from T. Soda et. al.⁹ ^f J_{AB} Experimental data taken from ref 23. ^g All calculations performed with B3LYP exchange-correlation functional.

are then formed using the total density matrix and the modified spin density matrix. Second, since broken symmetry calculations are notorious for their poor convergence, sometimes even with well designed starting guesses, one should always carefully examine the obtained state. One way of checking the validity of the result is to look at the spin density of the system; we here monitored the spin density on each Mn atom by means of numerical integration over a spherical region (radius 2.0 au) around each Mn atom. This procedure allowed us to ensure reliable control over the convergence of the broken-symmetry state in these unrestricted Kohn–Sham calculations.

III. Results and Discussion

A. Basis Set Dependence of Heisenberg Exchange Constants. The Mn₂O₂(pic)₄ compound with bis(μ -oxo) core that is a part of our selected test set was a subject of an earlier BS-DFT investigation by T. Soda et al.⁹ We therefore picked this complex for testing the suitability of various basis sets for evaluation of J_{AB} between the manganese centers. The Heisenberg exchange constants in the Mn₂O₂(pic)₄ compound computed using the B3LYP exchange-correlation functional and the various basis sets are presented in Table 2. The selection of the B3LYP functional for investigating the basis set dependence is motivated by the results of T. Soda et. al.,⁹ showing that this functional led to the best agreement between computational and experimental results. In the present investigation we employ two models of the Mn₂O₂(pic)₄ complex, denoted as “full” and “reduced” in Table 2, where the “full” model (see Figure 1) corresponds to the entire Mn₂O₂(pic)₄ molecule and the “reduced” model (see Figure 1) corresponds to a smaller molecule mimicking Mn₂O₂(pic)₄ in which the picolinic acid cation ligands are substituted by NHCHCO₂ ligands as proposed by T. Soda et. al.⁹ For both models the basis set dependence shows the pattern of a moderate change going from one tested basis set to another. The Heisenberg exchange constants obtained with the 6-31G, 6-31G* and Ahlrich’s VTZ basis sets are in agreement within a 5 cm⁻¹ range. An exception from this

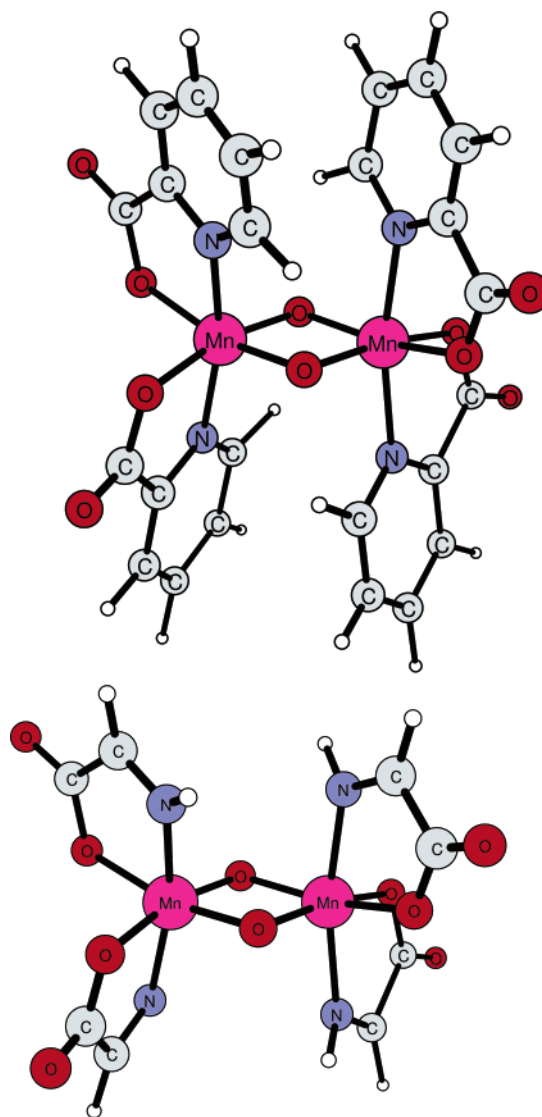


Figure 1. Mn₂O₂(pic)₄ compound and the reduced model of this compound.

trend is the small 3-21G basis set, which leads to a noticeable underestimation in terms of absolute values of the J_{AB} constants compared to results obtained with other basis sets. This behavior of the Heisenberg exchange constants with respect to various basis sets indicates that the basis sets suitable for magnetic coupling calculations should be flexible to capture essential features of the electron density distribution, especially around the Mn^{IV} centers in the broken symmetry and high spin states. At the same time, in our opinion, the use of large basis sets in calculations of this kind is unnecessary, since in the BS-DFT approach the J_{AB} constants are evaluated using energies differences between two states of the same molecule, thus canceling basis set incompleteness to a large extent. For example, improving from 6-31G to 6-31G* by adding polarization functions leads to a decrease of absolute values of the Heisenberg exchange constants only up to about 7%, depending on the J_{AB} evaluation scheme. At a first glance, this looks like a large effect, but in the context of the BS-DFT approach this effect is considerably smaller than the differences between the J_{AB} values computed using different mapping schemes between

Table 3: Heisenberg Exchange Constants between Mn^{IV} Centers in Various Manganese Compounds (cm⁻¹)^k

complex	model ^a	J_{AB}^b	J_{AB}^c	J_{AB}^d	$\langle S^2 \rangle_{BS}$	$\langle S^2 \rangle_{HS}$	exp
Mn ₂ O ₂ (pic) ₄	reduced	-116.5	-87.4	-115.0	2.998	12.118	
	full	-112.5	-84.4	-111.2	3.005	12.111	-86.5 ^e
[Mn ₂ O ₂ Cl ₂ (bpea) ₄] ²⁺	reduced	-148.5	-111.4	-146.3	2.989	12.122	
	full	-144.2	-108.1	-142.1	2.988	12.120	-144 ^f
[Mn ₂ O ₂ (phen) ₄] ⁴⁺	reduced	-141.7	-106.2	-139.9	3.006	12.116	
	full	-131.9	-98.9	-130.4	3.017	12.119	-147 ^g
[Mn ₂ O ₂ (OAc)(bpea) ₂] ³⁺	reduced	-24.5	-18.4	-24.3	3.028	12.098	
	full	-36.0	-27.0	-35.7	3.022	12.099	-124 ^h
[Mn ₂ O ₂ (OAc)(Me ₄ dtne)] ³⁺	reduced	-33.5	-25.1	-33.2	3.035	12.118	
	full	-37.5	-28.1	-37.2	3.032	12.117	-100 ⁱ
[Mn ₂ O ₃ (Me ₃ tacn) ₂] ²⁺	reduced	-381.7	-286.2	-375.5	2.963	12.110	
	full	-382.7	-287.0	-376.4	2.958	12.109	-390 ^j

^a J_{AB} model defines geometry of compound used in calculations, where "full" denotes whole compound and "reduced" denotes the smaller size model of it. ^b J_{AB} evaluated using eq 2. ^c J_{AB} evaluated using eq 3. ^d J_{AB} evaluated using eq 4. ^e J_{AB} experimental data taken from ref 23. ^f J_{AB} experimental data taken from ref 24. ^g J_{AB} experimental data taken from ref 25. ^h J_{AB} experimental data taken from ref 26. ⁱ J_{AB} experimental data taken from ref 27. ^j J_{AB} experimental data taken from ref 28. ^k Calculations carried out using B3LYP exchange-correlation functional and Ahlrich's VTZ basis set.

BS-DFT and Heisenberg Hamiltonian states and is thus of minor importance. With this reasoning we advocate the use of medium size basis sets, like Ahlrich's VTZ, in the investigation of magnetic coupling in larger molecular systems. Another practical argument for resorting to basis sets of this kind is that they are more suited for efficient calculations of large molecular systems with linear scaling methods than large basis sets with polarization and diffuse functions with small exponents which often lead to numerical problems in that context. Therefore, based on the discussion above we selected Ahlrich's VTZ basis set for all remaining calculations of the Heisenberg exchange constants in the dinuclear manganese complexes investigated in this paper.

B. General Trends in Heisenberg Exchange Constants.

The results of the calculations of the Heisenberg exchange constants between Mn^{IV} centers in dinuclear manganese complexes along with available experimental data are summarized in Table 3. Comparison of these results indicates that the J_{AB} computation schemes proposed by L. Noodleman (eq 2) and M. Nishino et. al. (see eq 4) allow for obtaining a good agreement between calculated and experimental J_{AB} values for [Mn₂O₂Cl₂(bpea)₄]²⁺ (Figure 2), [Mn₂O₂(phen)₄]⁴⁺ (Figure 3), and [Mn₂O₃(Me₃tacn)₂]²⁺ (Figure 6) compounds. The opposite situation is encountered for Mn₂O₂(pic)₄ (Figure 1), where the best agreement between calculations and experiment is obtained by employing the J_{AB} evaluation scheme given by eq 3. For the remaining two compounds with a bis(μ -oxo)(μ -carboxylato) core, namely [Mn₂O₂(OAc)(bpea)₂]³⁺ (see Figure 4) and [Mn₂O₂(OAc)(Me₄dtne)]³⁺ (see Figure 5), all three schemes for calculations of Heisenberg exchange constants perform poorly and severely underestimate the experimental J_{AB} numbers (in terms of absolute values). In line with previous observations, for positively charged complexes with bis(μ -oxo) and tris(μ -oxo) cores we observed a good performance of the J_{AB} evaluation scheme proposed by L. Noodleman (eq 2) as these compounds feature well localized unpaired electron orbitals on the Mn^{IV} centers in these complexes, which is a necessary condition for good performance of this computational scheme. For investigated neutral complexes with bis(μ -oxo) core, the localization of the unpaired electrons on manganese

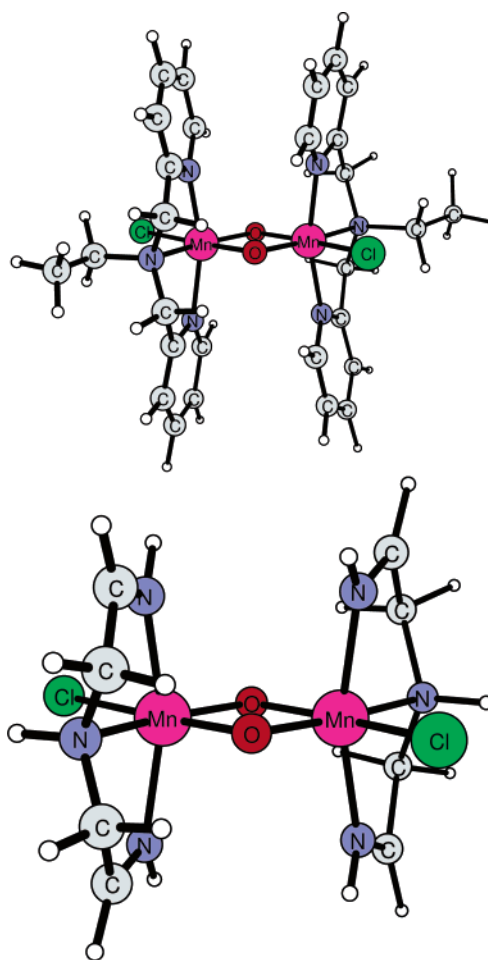


Figure 2. [Mn₂O₂Cl₂(bpea)₄]²⁺ compound and the reduced model of this compound.

centers are less pronounced, and consequently the overlap between magnetic orbitals of the Mn^{IV} centers is stronger. This in turn leads to good performance of the Heisenberg exchange computation scheme given by eq 3, which has been designed for this particular bonding situation between localized spins. The breakdown of all three J_{AB} constant schemes observed for manganese complexes with bis(μ -oxo)-(μ -carboxylato) core can probably be attributed to delocal-

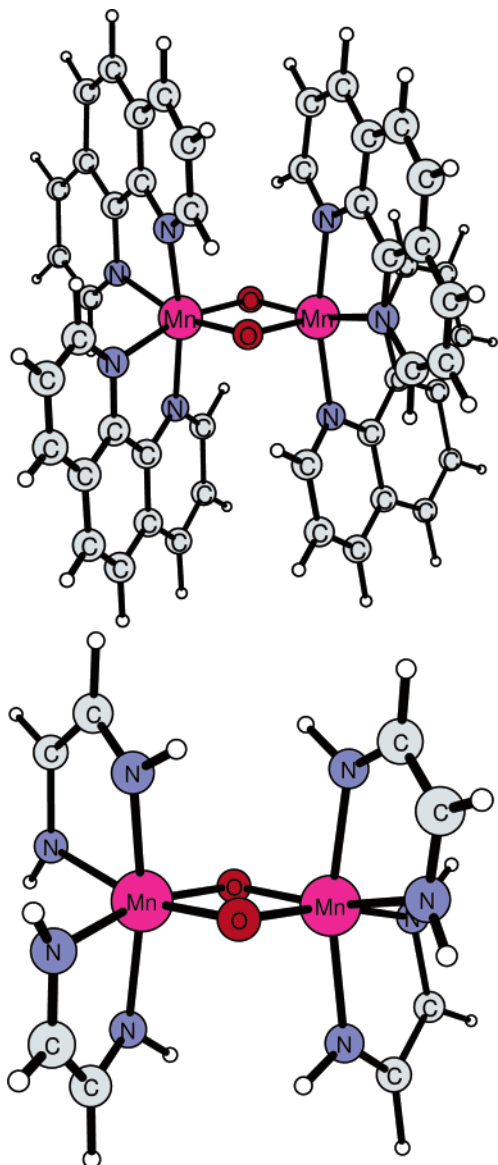


Figure 3. [Mn₂O₂Cl₂(phen)₄]⁴⁺ compound and the reduced model of this compound.

ization of the unpaired electrons over both Mn^{IV} centers in the high spin state, which is caused by the OAc ligand binding to the manganese centers. Limitations of J_{AB} evaluation schemes employed in our calculations are well-known for compounds with a delocalized high spin state, and detailed discussion of this problem can be for example found in ref 14. Finally, we note that the Heisenberg exchange constants evaluated using the scheme of state mapping proposed by M. Nishino et. al. closely follow the results obtained by the Noodleman scheme in agreement with the behavior witnessed in other investigations of J_{AB} . Here it is worthwhile also to note that this scheme, according to their authors, should be able to take into account the specifics of the bonding character in the molecule under investigation and therefore hypothetically should lead to results similar to the ones obtained with the third scheme (eq 3) for Mn₂O₂-(pic)₄. However, our calculation results presented in Tables 2 and 3 as well as the results of previous calculations by Soda et al. do not support this claim. This discrepancy can most likely be explained by the fact that the total spin angular

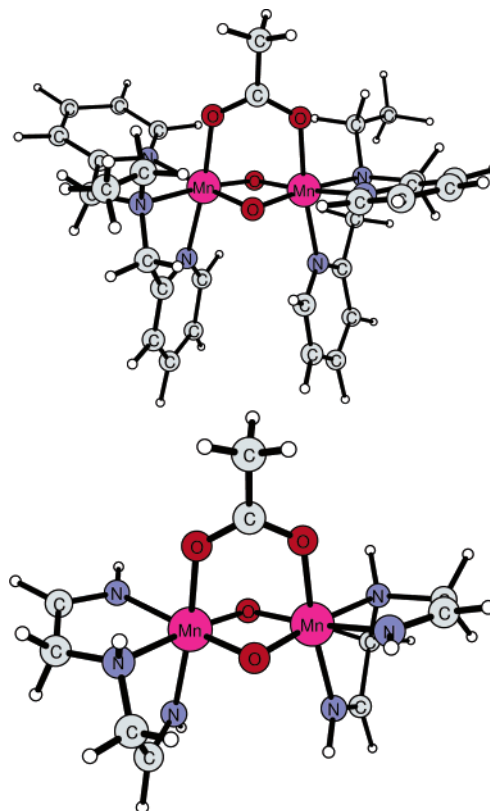


Figure 4. [Mn₂O₂(OAc)(bpea)₂]³⁺ compound and the reduced model of this compound.

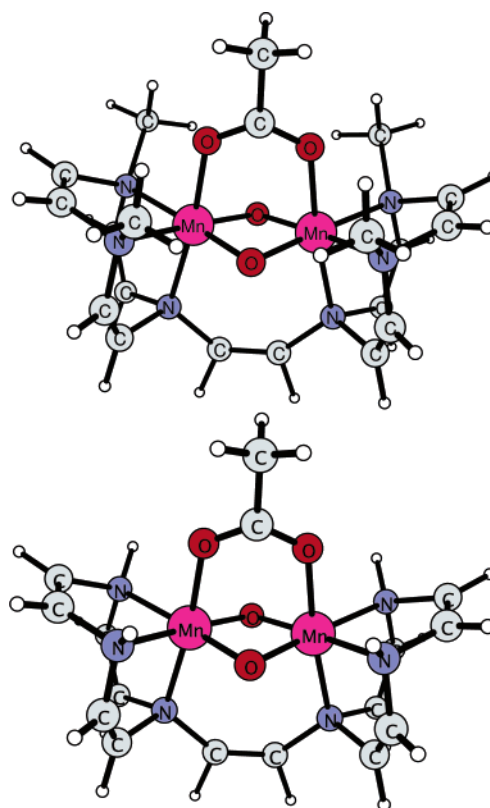


Figure 5. [Mn₂O₂(OAc)(Me₄dtne)]³⁺ compound and the reduced model of this compound.

momentum value evaluated using Kohn–Sham orbitals is rather inaccurate (a detailed discussion of this topic can be found in ref 20). Therefore, based on these arguments, and

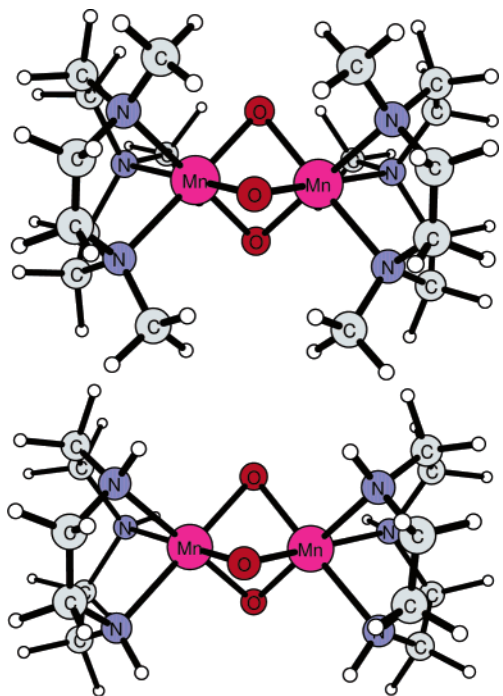


Figure 6. $[\text{Mn}_2\text{O}_3(\text{Me}_3\text{tacn})_2]^{2+}$ compound and the reduced model of this compound.

in our opinion, the use of mapping schemes proposed by Noodleman and by Ruiz are more justified than the one of M. Nishino et al. in connection with BS-DFT calculations. From a quantitative point of view, our numerical results for $\text{Mn}_2\text{O}_2(\text{pic})_4$, $[\text{Mn}_2\text{O}_2\text{Cl}_2(\text{bpea})_4]^{2+}$, $[\text{Mn}_2\text{O}_2(\text{phen})_4]^{4+}$, and $[\text{Mn}_2\text{O}_2(\text{phen})_4]^{4+}$ are in good agreement with experimental data, where the largest deviation is less than 10 cm^{-1} if one employs the most suitable J_{AB} computation scheme for each compound, i.e., eq 2 or eq 3. In the above-described results thus evidence that a reliable mapping scheme between the states obtained with the BS-DFT approach and the Heisenberg Hamiltonian states is a key required for accurate evaluations of J_{AB} constants. On the other hand, a selection of suitable mapping schemes for particular molecular systems can by no means be done automatically. It requires detailed information on electron density and orbital localization in order to deduct which of the schemes for the Heisenberg exchange constants will give accurate results and will be in line with the correct picture of the physical interaction in the molecular system.

Apart from the results of the calculations of the Heisenberg exchange constants for full size dinuclear manganese complexes, we also present in Table 3 calculation results for reduced model systems of these compounds. The model systems, which mimic the full size dinuclear manganese compounds, have been designed according to the recipe described in the Computational Details section and are in Table 3 marked as “reduced”. For the $\text{Mn}^{\text{IV}}-\text{Mn}^{\text{IV}}$ complexes with bis(μ -oxo) and tris(μ -oxo) cores, the differences between the Heisenberg exchange constant computed employing the full size compound (denoted in Table 3 as “full”) and its reduced model geometries are small, with the largest deviation not exceeding 5 cm^{-1} . Furthermore, the effect on J_{AB} going from the full scale compound to its model is more pronounced in the J_{AB} evaluation schemes proposed by

Noodleman and Nishino et al. This is expected as the denominator in these schemes is smaller than the one in the third case (see eqs 2–4). The differences between the Heisenberg exchange constants obtained using the full scale compound and its model geometries are severely pronounced for the complexes with a bis(μ -oxo)(μ -carboxylato) core. The use of the reduced compound model then leads to underestimation of the absolute J_{AB} values up to 40%. A behavior of this kind is consistent with the assumed delocalization of the unpaired electron orbitals over both Mn^{IV} centers, since in this case our procedure to build the reduced models leads to larger deviations between electronic structures of the real and the model compounds. Therefore, this finding indirectly supports our claim that the BS-DFT approach fails to predict the Heisenberg exchange constant related to the delocalization of unpaired electron orbitals over both manganese centers. It also follows that for dinuclear manganese complexes with well localized spins the use of smaller model compounds, which mimic the bonding situation of the manganese centers by replacing large ligands with suitable smaller ones, leads to only slight changes in the values of Heisenberg exchange constants. Consequently, one can recommend to employ such model compounds in order to reduce computational cost and to gain insight.

IV. Conclusion

This work presents a systematic investigation of the performance of broken symmetry density functional theory for evaluation of Heisenberg exchange constants in dinuclear $\text{Mn}^{\text{IV}}-\text{Mn}^{\text{IV}}$ compounds. We selected for exploration a number of complexes with three different arrangements of the Mn^{IV} centers, namely bis(μ -oxo), bis(μ -oxo)(μ -carboxylato), and tris(μ -oxo) cores, to cover different bonding characteristics occurring between these centers and in this way assess the performance of the BS-DFT approach with respect to these characteristics. Apart from investigating the suitability and accuracy of the BS-DFT approach for these manganese complexes we also aimed to design recipes for reliable calculations of Heisenberg exchange constants on molecular systems of this kind.

The results of our investigation emphasize the importance of the mapping between broken symmetry, high spin, states obtained with the unrestricted DFT formalism and states of the Heisenberg exchange constants, which has been observed in earlier works devoted to the evaluation of J_{AB} constants with the BS-DFT approach. A selection of an appropriate mapping scheme is found crucial for a reliable evaluation of Heisenberg exchange constants, where the scheme proposed by L. Noodleman¹² is well suited for weak bonding between the Mn^{IV} centers, while the scheme advocated by E. Ruiz¹⁷ based on Noodleman’s work is more acceptable for the opposite situation. However, as we witnessed in the case of compounds with a bis(μ -oxo)(μ -carboxylato) core none of the tested J_{AB} evaluation schemes is adequate, probably due to delocalization of the unpaired electrons over the two manganese centers. This example indicates that the BS-DFT approach cannot straightforwardly be applied in investigations of magnetic coupling if the unique features of electronic structure of each molecule under investigation

is not carefully considered. Based on the calculation results presented in this paper, we recommend using the Noodleman or Ruiz/Noodleman schemes for computations of Heisenberg exchange constants, while the scheme of Nishino does not seem to be well suited for unrestricted DFT due the general inability of the Kohn–Sham method to evaluate expectation values of the total spin angular momentum operator. Furthermore, we advocate the use of the B3LYP exchange–correlation functional in combination with Ahlrich’s VTZ basis set; this combination indeed allowed an accurate reproduction of experimental J_{AB} values for most of the investigated compounds.

Another important issue addressed in this paper is the effect of long-range interactions for evaluation of Heisenberg exchange constants. On one hand we showed that it is possible to build reduced model compounds of manganese complexes by substituting large ligands with suitable smaller ones without affecting the magnetic coupling between the Mn^{IV} centers. The success of this methodology indicates that only the closest environment of the Mn^{IV} centers have significant effects on their electronic structure due to a localized nature of the unpaired electron orbitals. However, in the case of extended molecular systems the electronic structure of the ligands can be significantly dependent on their surrounding, that also leads to changes in electronic structure of the manganese centers. The design of reliable reduced model compounds is evidently not straightforward in such cases, and one can recommend reduced model compounds only for molecular systems in which each localized spin center has a well distinguished set of ligands. Another aspect related to the evaluation of J_{AB} constants in large molecular systems is the improper asymptotic behavior of our recommended B3LYP exchange–correlation. To tackle this issue we investigated the performance of the CAMB3LYP functional, which has improved asymptotic behavior of the exchange part, in the evaluation of J_{AB} constants. The obtained results indicate that before this functional can be used for routine calculations of Heisenberg exchange constants it should be reparametrized in order to reproduce the performance of the B3LYP functional. In the future we plan to investigate new sets of parameters for CAMB3LYP functionals oriented for accurate prediction of J_{AB} constants as part of our efforts in the theoretical design of single molecular magnets.

Supporting Information Available: Computed total energies and spin populations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Heisenberg, W. Z. *Phys.* **1928**, *49*, 619.
- (2) Kahn, O. *Molecular magnetism*; VHC Publishers: New York, 1993.
- (3) Abragam, A.; Bleaney, B. *Electron Paramagnetic Resonance of Transition Ions*; Clarendon: Oxford, 1970.
- (4) Atherton, N. M. *Principles of Electron Spin Resonance*, 2nd ed.; Ellis Horwood, Prentice Hall: New York, 1993.
- (5) Bertini, I.; Luchinat, C. *Coord. Chem. Rev.* **1996**, *150*, 1–28.
- (6) Nagao, H.; Nishino, M.; Shigeta, Y.; Soda, T.; Kitagawa, Y.; Onishi, T.; Yoshioka, Y.; Yamaguchi, K. *Coord. Chem. Rev.* **2000**, *198*, 265–295.
- (7) Yakhmi, J. V. *Macromol. Symp.* **2004**, *212*, 141–158.
- (8) Mukhopadhyay, S.; Mandal, S. K.; Bhaduri, S.; Armstrong, W. H. *Chem. Rev.* **2004**, *104*, 3981–4026.
- (9) Soda, T.; Kitagawa, Y.; Onishi, T.; Takano, Y.; Shigeta, Y.; Nagao, H.; Yoshioka, Y.; Yamaguchi, K. *Chem. Phys. Lett.* **2000**, *319*, 223–230.
- (10) Neese, F. J. *Phys. Chem. Sol.* **2004**, *65*, 781–785.
- (11) Caballol, R.; Castell, O.; Illas, F.; Moreira, I. de P. R.; Malrieu, J. P. *J. Chem. Phys. A* **1997**, *101*, 7860.
- (12) Noodleman, L. *J. Chem. Phys.* **1981**, *74*, 5737–5743.
- (13) Ruiz, E.; Rodriguez-Fortea, A.; Tercero, J.; Cauchy, T.; Massobrio, C. *J. Chem. Phys.* **2005**, *123*, 074102.
- (14) Noodleman, L.; Peng, C. Y.; Case, D. A.; Mouesca, J. M. *Coord. Chem. Rev.* **1995**, *144*, 199–244.
- (15) Noodleman, L.; Lovell, T.; Liu, T.; Himo, F.; Torres, R. A. *Curr. Opin. Chem. Biol.* **2002**, *6*, 259–273.
- (16) Ciofini, I.; Daul, C. A. *Coord. Chem. Rev.* **2003**, *238–239*, 187–209.
- (17) Ruiz, E.; Cano, J.; Alvarez, S.; Alemany, P. *J. Comput. Chem.* **1999**, *20*, 1391–1400.
- (18) Noodleman, L.; Norman, J. G. *J. Chem. Phys.* **1979**, *70*, 4903–4906.
- (19) Nishino, M.; Yamanaka, S.; Yoshioka, Y.; Yamaguchi, K. *J. Phys. Chem. A* **1997**, *101*, 705–712.
- (20) Wang, J.; Becke, A. D.; Smith, V. H., Jr. *J. Chem. Phys.* **1995**, *102*, 3477–3480.
- (21) Barone, V.; Bencini, A.; Gatteschi, D.; Totti, F. *Chem. Eur. J.* **2002**, *8*, 5019–5027.
- (22) McGrady, J. E.; Stranger, R. *J. Am. Chem. Soc.* **1997**, *119*, 8512–8522.
- (23) Libby, E.; Webb, R. J.; Streib, W. E.; Folting, K.; Huffman, J. C.; Hendrickson, D. N.; Christou, G. *Inorg. Chem.* **1989**, *28*, 4037–4040.
- (24) Stebler, M.; Ludi, A.; Büergi, H. B. *Inorg. Chem.* **1986**, *25*, 4743–4750.
- (25) Pal, S.; Olmstead, M. M.; Armstrong, W. H. *Inorg. Chem.* **1995**, *34*, 4708–4715.
- (26) Pal, S.; Chan, M. K.; Armstrong, W. H. *J. Am. Chem. Soc.* **1992**, *114*, 6398–6406.
- (27) Schäfer, K. O.; Bittl, R.; Zweggart, W.; Lendzian, F.; Haselhorst, G.; Weyhermüller, T.; Wieghardt, K.; Lubitz, W. *J. Am. Chem. Soc.* **1998**, *120*, 13104–13120.
- (28) Wieghardt, K.; Bossek, U.; Nuber, B.; Weiss, J.; Bonvoisin, J.; Corbella, M.; Vitols, S. E.; Girerd, J. J. *J. Am. Chem. Soc.* **1988**, *110*, 7398–7411.
- (29) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490–519.
- (30) SPARTAN '02 program; Wavefunction, Inc.: Irvine, CA.
- (31) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

- (32) Vosko, S. J.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200. Parameterization V.
- (33) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (34) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (35) Rudberg, E.; Salek, P. *ErgoSCF version 1.0*; a quantum chemistry program supporting unrestricted DFT calculations, 2005.
- (36) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (37) H – Ne: Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939. Sc – Zn: Dobbs, K. D.; Hehre, W. J. *J. Comput. Chem.* **1987**, *8*, 861.
- (38) H–Ne: Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257. K – Zn: Rassolov, V.; Pople, J. A.; Ratner, M.; Windus, T. L. *J. Chem. Phys.* **1998**, *109*, 1223.
- (39) Schafer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.

CT050325B

Computational Characterization of Metal Binding Groups for Metalloenzyme Inhibitors

Kerwin D. Dobbs,[†] Amy M. Rinehart,[†] Michael H. Howard, Ya-Jun Zheng,* and Daniel A. Kleier*

DuPont Crop Protection, Stine-Haskell Research Center, P.O. Box 30,
Newark, Delaware 19714

Received August 3, 2005

Abstract: The mode of action of many pest or disease control agents involves inhibition of some metalloenzyme that is essential for the survival of the target organism. These inhibitors typically consist of a functional group that is capable of a primary binding interaction with the metal and a scaffold that is capable of secondary interactions with the remainder of the enzyme. To characterize the binding ability of various metal binding groups (BGs), we have performed electronic structure calculations on ligand displacement reactions in a model system related to the metalloenzyme, peptide deformylase: $E-M-R + BG \rightarrow E-M-BG + R$. Here E represents a model coordination environment for the metal M, and R is a reference ligand (e.g., water) that may be displaced by a metal binding group. Since the oxidation state of many of the metals considered allows for multiple spin states, we also studied the influence of spin state on the coordination environment. Qualitative considerations of electronic structure inspired by the calculations provide an understanding of binding energy trends across a variety of ligands for a given metal and across a variety of metals for a given ligand.

Introduction

Many biologically active molecules act by binding to a metal ion. Some act as ionophores by transporting metals across membranes.¹ Others act as inhibitors of metalloenzymes by binding a metal at the active site.^{2–6} In this report we describe the application of electronic structure calculations to study metal binding in model systems related to metalloenzymes such as peptide deformylase (PDF). PDF catalyzes the deformylation of the initial methionine of a nascent polypeptide and is a validated target for both antibiotics⁷ and herbicides.⁸ Crystal structures have been reported for the *E. coli*⁹ enzyme containing three different metal cofactors Fe(II),^{9,10} Ni(II),^{9,11,12} and Zn(II)⁹ and a variety of metal

coordinating ligands including water,⁹ the tripeptide, Met-Ala-Ser,⁹ hydroxamic acids such as β -sulfonyl- and β -sulfinylhydroxamic acids,¹¹ and actinonin^{12,10} as well as carboxylates exemplified by matlystatin analogues.¹⁰

We have been interested in designing selective PDF inhibitors as herbicides.^{13,14} To facilitate such a design effort, a better understanding of the nature of the interaction between metal center and the metal binding group of a potential inhibitor is critical. A close-up view of the active site of Zn–PDF⁹ from *E. coli* is illustrated in Figure 1. The enzyme contributes three ligands to the tetrahedral coordination sphere of the zinc: two histidines and a cysteine. The fourth ligand in this structure is a water molecule.

The calculations reported below address structural and spin state preferences for a simplified model of the active site of PDF in which the three amino acid residues are replaced by a tridentate ligand. These preferences are assessed as a function of metal type (Fe(II), Co(II) and Ni(II), Zn(II)). In addition, we evaluate the ability of other metal binding groups to displace the water from the model structures.

* Corresponding authors e-mail: YA-JUN.ZHENG@USA.dupont.com (Y.-J.Z.); e-mail: daniel.kleier@drexel.edu (D.A.K.) and phone: (215)895-1861. Corresponding author address: Department of Chemistry, Drexel University, Disque Hall, Room 305, 3141 Chestnut Street, Philadelphia, PA 19104-2875 (D.A.K.).

[†] Current address: DuPont Central Research and Development, Wilmington, DE.

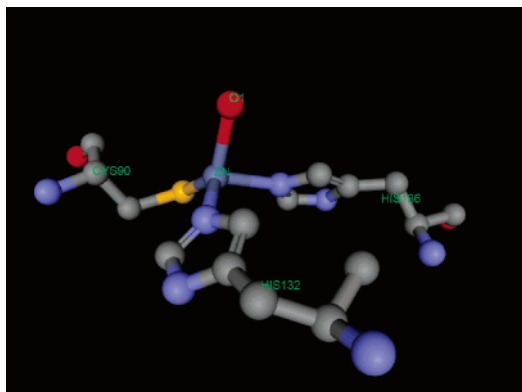


Figure 1. Detail of active site of Zn(II)PDF according to Becker and co-workers.⁹

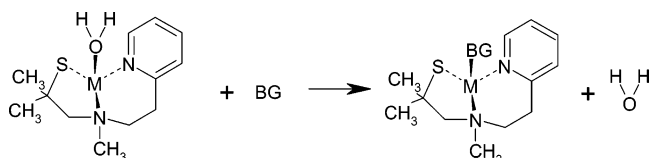


Figure 2. Water displacement reaction used to rank coordinating ability of binding groups, BG.

Methods

To rank order the coordinating ability of metal binding groups (BGs) we performed a series of computational studies of their ability to displace water from the complex shown in Figure 2. The complex consisted of a metal dication (iron(II), cobalt(II), nickel(II), or zinc(II)) wrapped in a tridentate spectator ligand 2-methyl-1-([methyl-(2-pyridin-2-ylethyl)-amino]propane-2-thiolate, referred to as PATH for short. We chose the PATH ligand as our model system not only for reasons of computational efficiency but also because of its promotion¹⁵ as a good structural mimic of the (His)(His)-(Cys) triad at the active site of metalloenzymes such as PDF and because direct comparison with experiment was possible.¹⁵

Density functional methods have been the method of choice for many modeling studies of metalloenzymes.¹⁰ All calculations reported in this paper were performed with the density functional theory (DFT) methods as implemented within the Gaussian 03 suite of programs.¹⁶ Each molecular structure was first optimized using the BP86/DGDZVP method. The optimized structure was then used in subsequent analytic vibrational frequency calculations at this same level of computation in order to ensure that the structure was indeed at a minimum on the potential energy surface. The pure BP86 functional^{17,18} was chosen mainly because of its enhanced performance for optimization and vibrational frequency calculations compared to B3LYP.¹⁹ Pure functionals are able to take advantage of using density fitting basis sets which expand the density in a set of atom-centered functions when computing the Coulomb interaction instead of computing all of the two-electron integrals.²⁰ DGDZVP basis sets²¹ are all-electron, double- ζ valence polarized basis sets which were optimized specifically for DFT methods.

From the outset of this study, we initially used the B3LYP functional for binding energy analysis, since no generally accepted protocols for transition-metal reaction energetics

had been reported in the literature²² at the time we began this study, and the B3LYP functional was known to do very well for the main group reactions.²³ The extensive experience of one of the authors (K.D.D.) with many different main group and organometallic systems led to the conclusion that differences between BP86 and B3LYP structures are minor while using the DGDZVP basis sets.²⁴ Building on this same experience, the reaction entropy, enthalpy, and free energy values reported below were determined from B3LYP/DGDZVP single-point energies on BP86/DGDZVP optimized structures in combination with the zero-point energy and vibrational thermal corrections (at 298.15 K) obtained from the BP86/DGDZVP vibrational frequencies. This same protocol was also used for determining the free energy differences, ΔG , between high- and low-spin states for (PATH)M(II)Br complexes in Table 1.

Results

Influence of Spin on Conformation of Coordination Sphere. The arrangement of ligands around the metal in the crystal structures of the Fe, Ni, and Zn forms of PDF is roughly tetrahedral in nature. A similar arrangement is observed in the crystal structures of the model (PATH)M(II)(BG) complexes,¹⁵ one of which is illustrated in Figure 3. In the model complexes the sulfide, tertiary amine nitrogen, and pyridine nitrogen serve as surrogates for the corresponding atoms in the cysteine and a pair of histidine residues of the PDF active site (see Figure 1).

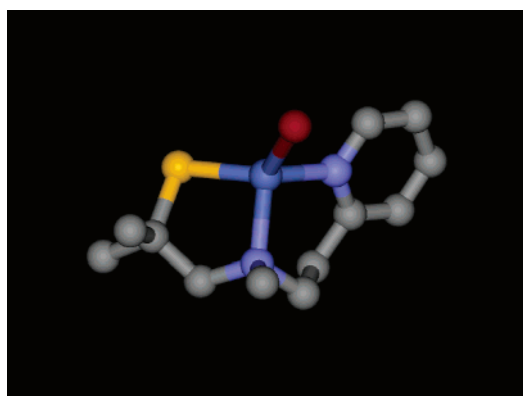
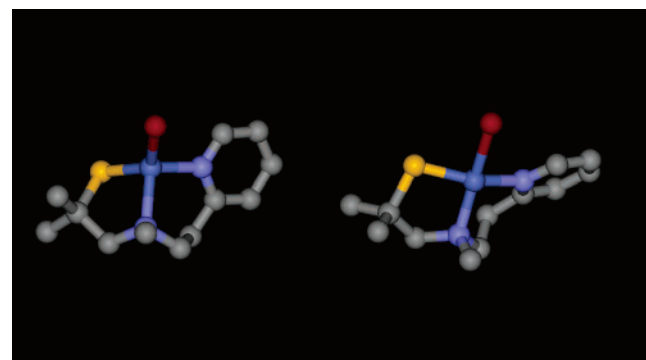
As shown in Table 1, our DFT calculations found the arrangement of ligands in the model (PATH)M(II)Br systems to be spin dependent. When the geometries of the high-spin states were optimized, a rough tetrahedral arrangement of ligands was generally found. In the low-spin states the optimized structures are better described as square planar. Table 1 also presents the energy difference between high- and low-spin states for (PATH)M(II)Br complexes. The high-spin tetrahedral conformation is favored according to the calculations for all the (PATH)M(II)Br complexes but only by a relatively modest 6.7 kcal/mol for the Ni(II) complex. Thus, it is, perhaps, not too surprising that a recent crystallographic investigation has revealed square planar Ni(II) centers in the *dimer* of (PATH)Ni(II)Br.^{15d} On the other hand, spectroscopic evidence for the (PATH)Co(II)Br complex supports a high-spin tetrahedral ground state¹⁵ consistent with predictions of the calculations.

The structures of the optimized quartet and doublet (PATH)Co(II)Br complexes are illustrated in Figure 4. The quartet state optimized to the *cis* diastereomer in which the N-CH₃ and bromide are on the same side of the fused 5,6-membered chelate ring system. Both the *N*-methyl and bromide groups are pointing toward the reader in Figure 4. In addition to a flatter arrangement of ligands about the cobalt, the optimized doublet structure has shorter bonds between the metal and the pyridine (N₁) and tertiary amine (N₂) nitrogens ($d[\text{N}_1\text{-Co}] = 1.96 \text{ \AA}$, $d[\text{N}_2\text{-Co}] = 2.03 \text{ \AA}$) as compared with the tetrahedral quartet structure ($d[\text{N}_1\text{-Co}] = 2.01 \text{ \AA}$, $d[\text{N}_2\text{-Co}] = 2.13 \text{ \AA}$).

Calculated geometric parameters for the high-spin (PATH)Co(II)Br and (PATH)Zn(II)Br complexes are compared with

Table 1. Arrangement of Ligands and Relative Free Energies, ΔG , of DFT Optimized Geometries for (PATH)M(II)Br Complexes in High- and Low-Spin States

metal	assigned multiplicity	optimized geometry	assigned multiplicity	optimized geometry	ΔG (high spin – low spin) kcal/mol
Fe(II)	quintet	tetrahedral	triplet	square planar	–18.5
Co(II)	quartet	tetrahedral	doublet	square planar	–16.6
Ni(II)	triplet	tetrahedral	singlet	square planar	–6.7
Zn(II)	singlet	tetrahedral			

**Figure 3.** Crystal structure of (PATH)CoBr from Chang et al.¹⁵ The positions of the hydrogen atoms have been suppressed for clarity.**Figure 4.** Comparison of optimized geometries for quartet (left) and doublet (right) (PATH)Co(II)Br.**Table 2.** Calculated and Experimental Geometric Parameters for Quartet (PATH)Co(II)Br^a

bond length	calcd	exptl	bond angle	calcd	exptl
Co–N ₁	2.01	2.04	S–Co–N ₂	91.1	91.9
Co–N ₂	2.13	2.09	N ₁ –Co–N ₂	101.0	100.2
Co–S	2.20	2.23	N ₁ –Co–Br	108.9	103.5
Co–Br	2.36	2.38	N ₂ –Co–Br	113.8	116.4
			S–Co–N ₁	112.0	119.7
			S–Co–Br	126.3	122.9

^a In this table N1 refers to the nitrogen of the pyridine ring, and N2 is the tertiary amine nitrogen.

those reported for the crystal structures in Tables 2 and 3, respectively. Key trends for distances and angles that involve the metal seem to be captured by the calculations (e.g., the ordering of bond distances and bond angles). The greatest discrepancies are the 8° difference for the S–Co–N₁ bond angle in the Co complex and the 12° difference in the S–Zn–Br angle in the Zn complex. Examination of the

Table 3: Calculated and Experimental^{15b} Geometric Parameters for (PATH)Zn(II)Br^a

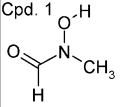
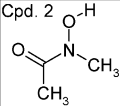
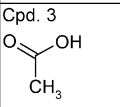
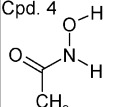
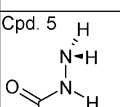
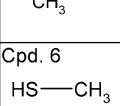
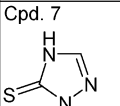
bond length	calcd	exptl	bond angle	calcd	exptl
Zn–N ₁	2.13	2.06	S–Zn–N ₂	90.2	92.9
Zn–N ₂	2.23	2.11	N ₁ –Zn–N ₂	96.7	99.3
Zn–S	2.28	2.26	N ₁ –Zn–Br	102.0	102.2
Zn–Br	2.39	2.38	N ₂ –Zn–Br	112.0	116.0
			S–Zn–N ₁	113.9	120.9
			S–Zn–Br	135.2	123.3

^a In this table N1 refers to the nitrogen of the pyridine ring, and N2 is the tertiary amine nitrogen.

calculated geometry revealed that the optimized conformation of the folded 5,6-membered chelate ring system is quite similar to that observed in the crystal structure. This conformational similarity can be appreciated by comparing the quartet structure on the left-hand side of Figure 4 with the experimental structure appearing in Figure 3. The conformation of the six-membered chelate ring in both the experimental and the optimized structure is a twist boat with the Co and an opposing methylene at the bowsprits. In both structures the conformation of the five-membered chelate ring places the exo methyl in an axial orientation and relatively close to the *N*-methyl group. It should be noted that at least one other energetically accessible conformation was discovered during the course of this work for the folded 5,6-membered chelate ring system. In this conformation the six-membered ring is closer to an idealized boat, while the five-membered ring assumes a conformation that places the exo methyl group in an equatorial orientation. Depending somewhat on the fourth ligand, this alternative conformation is calculated to be 3–4 kcal/mol higher in energy.

Metal Chelating Ability of Alternative Metal Binding Groups. Hydroxamic acids^{5,11} and *N*-acylhydroxylamines¹² have been the metal binding groups of choice for peptide deformylase inhibitors. We sought to understand the potential for alternative functional groups to substitute for hydroxamic acids by examining the optimized structures of their complexes with the metals in the model coordination systems and comparing computed enthalpies for the water displacement reaction shown in Figure 2. The water displacement calculations were performed on the high-spin complexes, since the tetrahedral geometries realized for this spin state are more representative of the coordination sphere observed by X-ray crystallography for both the (PATH)M(II)Br complexes and the active site of PDF. The 5- and 6-membered chelate rings maintained the optimized conformation described above for the (PATH)Co(II)Br high-spin complex.

Table 4. Calculated Thermodynamic Parameters for Displacement of Water from (PATH)M(II)(H₂O) Complexes by Various Metal Binding Groups (MBGs)^a

Metal→	Fe			Co			Ni			Zn		
Binding Group ↓	ΔS	ΔH	ΔG	ΔS	ΔH	ΔG	ΔS	ΔH	ΔG	ΔS	ΔH	ΔG
Cpd. 1 	-10.1	-12.7	-9.7	-12.0	-12.8	-9.2	-12.4	-9.8	-6.0	-12.7	-13.1	-9.3
Cpd. 2 	-6.6	-13.7	-11.8	-10.1	-14.2	-11.2	-11.0	-11.3	-8.1	-11.6	-14.1	-10.7
Cpd. 3 	-1.7	0.0	0.5	-5.5	-0.6	1.0	-5.5	0.0	1.6	-4.5	0.1	1.4
Cpd. 4 	-6.1	-13.0	-11.2	-9.4	-13.5	-10.7	-6.8	-10.8	-8.8	-9.6	-13.6	-10.7
Cpd. 5 	-10.1	-15.0	-12.0	-9.8	-13.7	-10.8	-13.8	-13.9	-9.8	-9.1	-13.0	-10.3
Cpd. 6 	-4.6	2.2	3.6	-6.2	1.8	3.7	-6.0	2.6	4.4	-5.7	2.3	4.0
Cpd. 7 	-9.7	-17.0	-14.1	-8.8	-17.8	-15.2	-9.7	-15.2	-12.3	-11.6	-19.5	-16.0

^a Entropy changes are in cal/mol. Enthalpy and free energy changes calculated at 298 K are in kcal/mol.

Calculated binding energies and entropies relative to water are presented in Table 4, and binding enthalpies are plotted as a function of metal in Figure 5. Of the four metals considered, binding is generally weakest to the (PATH)Ni(II) complex and similar in strength for coordination to the other three metals. Of the neutral complexes considered, the thiothiazolinone (compound 7 plotted in orange) generally binds best to all (PATH)M(II) complexes, followed closely by the *N*-acetylhydrazine (compound 5 in green), the *N*-acetylhydroxylamine (compound 4 in red), and the *N*-acetyl-*N*-methylhydroxylamines (compounds 1 and 2 in blue and black, respectively).

Although neutral acetic acid is not predicted to be a particularly strong binder, the anion is a different story. It is off scale because of the strongly stabilizing electrostatic interaction with the positively charged metal. It will be necessary to take account of desolvation in order to reliably

compare such anionic binding groups with the neutral ones described here.

Examination of the optimized structures for the *N*-acetylhydroxylamine and hydroxamic acid complexes revealed some unexpected results. Instead of bidentate chelation of the metal as observed in PDF cocrystals with ligands of this type, the calculations predict a single dative bond between the carbonyl oxygen of the binding group and the metal of the model system. Instead of forming the expected second dative bond with the metal, the hydroxyl group of these ligands spun around the N–O axis to donate a hydrogen bond to the nearby negatively charged sulfide of the PATH ligand (See Figure 6). Interestingly, a bidentate interaction with the metal is realized for the (PATH)Fe(II)(*N*-acetylhydrazine) complex as shown in Figure 7 but at the expense of losing the hydrogen bond with the sulfide.

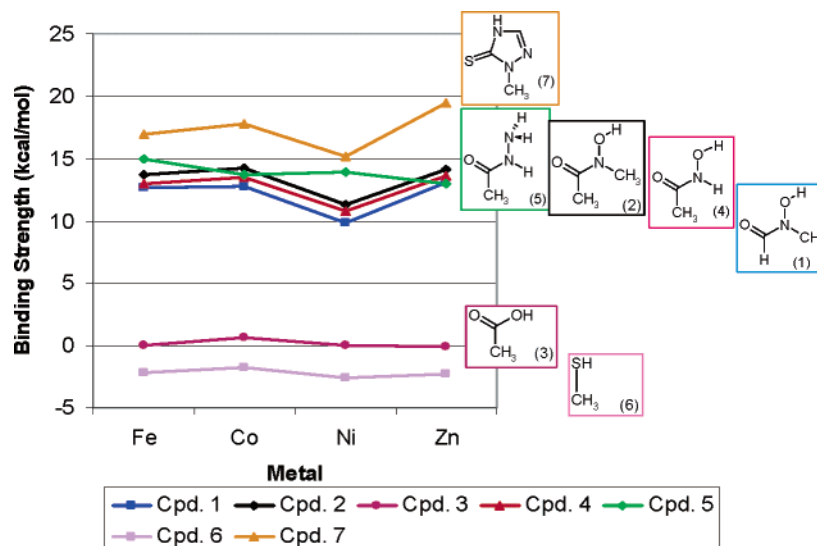


Figure 5. Binding strength ($-\Delta H$ for water displacement) for neutral binding groups as a function of metal.

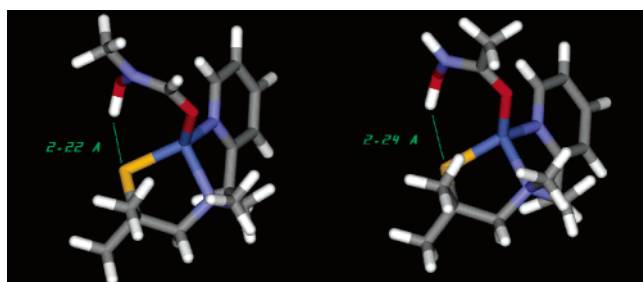


Figure 6. Optimized structures illustrating hydrogen bonding for (PATH)Co(II)(*N*-formyl,*N*-methylhydroxylamine) (compound 1, left) and *N*-acetylhydroxylamine (compound 4, right).

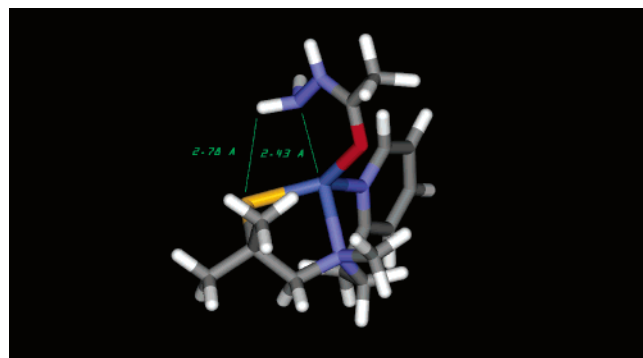


Figure 7. Optimized structure for (PATH)Fe(*N*-acetylhydroxylamine) complex. Bond distances shown in green are in angstroms. Bond distance to iron from the carbonyl oxygen of ligand is 2.12 Å.

Discussion

Geometries and Spin States. The general agreement between the calculated and experimentally determined geometries of the (PATH)M(II)Br complexes (Tables 2 and 3, Figures 3 and 4 left) builds confidence in the calculated binding trends reported here. Many of the predictions are also supported by qualitative considerations of molecular orbital interactions. For example, the preference for a roughly tetrahedral arrangement of ligands in the high-spin states of the Fe(II), Co(II), and Ni(II) complexes (see Table

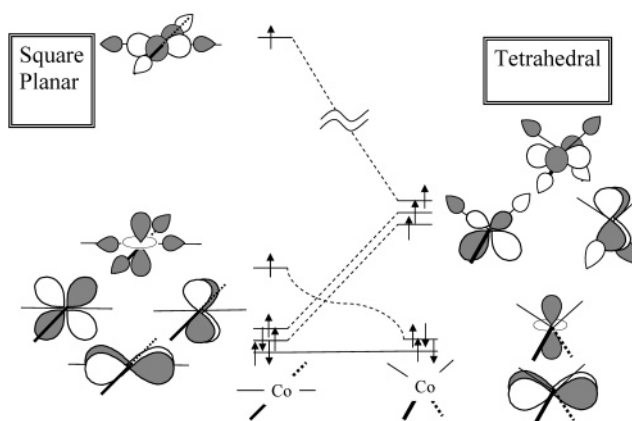


Figure 8. Walsh diagram for a tetrahedral to square planar conversion of a high-spin quartet Co(II) L_4 complex. The high-spin quintet state of the Fe(II) complex would have one less electron in the next to lowest of the depicted orbitals, whereas the high-spin triplet of the Ni(II) complex would have one additional electron in the middle or third level orbital. In all three high-spin cases the highest of these five orbitals is singly occupied.

1) can be understood on the basis of a Walsh diagram for the frontier orbitals²⁵ of a generalized M(II) L_4 complex (Figure 8). In this qualitative picture, the geometry of ML_4 systems is attributed in large part to the behavior of the highest of the singly occupied molecular orbitals (SOMOs). For the high-spin states, the energy of the highest SOMO is expected to rise dramatically as the tetrahedron is flattened due to increased antibonding character. This orbital is not occupied in the low-spin states, and the square planar structure is thus expected to be the more stable for the triplet Fe(II), doublet Co(II), and singlet Ni(II) complexes (see Table 1).

Alternatively, the Walsh diagram can be used to anticipate the spin states for scaffold enforced tetrahedral and square planar arrangements of ligands. In case of an enforced tetrahedral arrangement of ligands, Hund's rule would anticipate a high-spin configuration for the partially filled set of three degenerate orbitals (i.e., quartet state for d^7 Co-

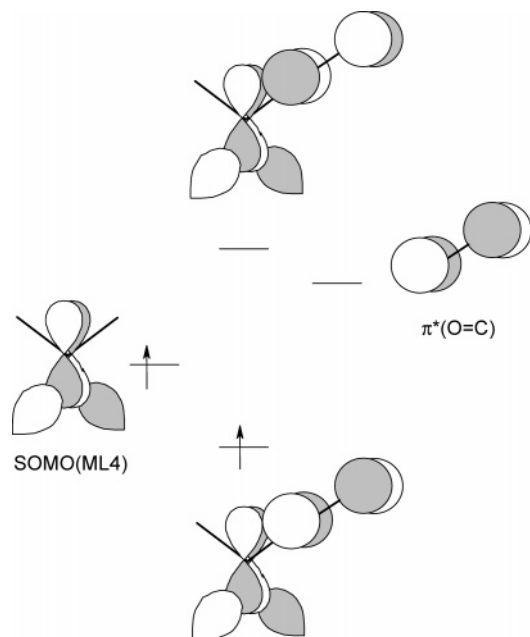


Figure 9. Interaction diagram for $\pi^*(\text{O}=\text{C})$ orbital of a π acceptor ligand and one of the metal SOMOs in a high-spin tetrahedral ML_4 complex. The energy gap between ligand $\pi^*(\text{O}=\text{C})$ orbital and SOMO(ML_4) will determine the degree of stabilization of SOMO of the complex due to back-bonding. In Zn(II) complexes, the lower of the orbitals of the complex would be doubly occupied.

(II) complexes and a triplet state for d^8 Ni(II) complexes). On the other hand, a square planar arrangement of ligands would favor low-spin d^7 (i.e., doublet Co(II)) and d^8 (i.e., singlet Ni(II)) complexes.

The conformation of ligands about the metals in (PDF)M(II)(BG) enzyme cocrystals is typically tetrahedral.⁹ The framework of the enzyme itself probably plays a role in enforcing this geometry and increasing the likelihood that the metal ions are in a high-spin state.

Binding Trends across Metals. A qualitative understanding of the trends in the calculated enthalpies of water displacement across the spectrum of metals (Table 4 and Figure 5) can be offered in terms of traditional concepts of metal to carbonyl back-bonding. Figure 9 is an orbital diagram for the stabilizing back-bonding interaction expected between the SOMO of a tetrahedrally coordinated high-spin metal and the LUMO of a π -acceptor ligand (e.g., carbonyl group). This stabilizing interaction is expected to weaken in proceeding across the transition metals from Fe(II) to Co(II) to Ni(II) due to the increasing energy gap between the $\pi^*(\text{O}=\text{C})$ orbital and the SOMOs as the latter drop in energy. The weakening of this back-bonding interaction may account for the decreasing exothermicity for water displacement by π -acceptor ligands in the order

$$-\Delta H(\text{Fe(II)}) \approx -\Delta H(\text{Co(II)}) > -\Delta H(\text{Ni(II)})$$

Although the energy gap is expected to widen even further upon passing to Zn(II) complexes, the presence of two rather than one electron in this back-bonding HOMO of the complex may account for the reversal in the downward trend in binding strength.

Binding Trends across Ligands. Trends in the calculated reaction enthalpy across the spectrum of ligands listed in Table 4 can also be understood in terms of traditional bonding concepts. For example, the electron releasing property of a methyl group relative to a hydrogen atom can account for enhanced stability predicted for compound 2 relative to compounds 1 and 4. The alternate metal coordination scheme found acetylhydrazine complexes (compound 5), compared with complexes formed with the acetylhydroxylamine ligand (compound 4), can be attributed to a combination of enhanced basicity and decreased hydrogen bond acidity of an amino group relative to a hydroxyl group. The terminal amino group of the acetylhydrazine thus forms a dative bond with the metal, while the corresponding acetylhydroxylamine donates a hydrogen bond to the sulfide of the PATH ligand rather than interact with the metal. It is interesting that metal dependent shifts between bidentate and monodentate metal binding have been observed for the formate ligand in PDF. In iron and cobalt PDF, bidentate metal binding that involves both formate oxygens is observed, while in Zn PDF monodentate binding of one oxygen to the metal and hydrogen bonding of the other oxygen to the protein backbone is observed.²⁶

It should be noted that the hydrogen bond predicted between the hydroxyl group of the hydroxamic acids and the basic sulfide of the PATH ligand in the (PATH)M(II)-(hydroxamic acid) complexes may not be present in the (PDF)M(II)(hydroxamic acid) complexes. In (PDF)M(II)-(hydroxamic acid) complexes, the hydroxyl group of a properly positioned hydroxamic acid can participate in a dative bond with the metal ion of the enzyme and simultaneously donate a hydrogen bond to a second base at the active site, which is not represented in our model system.

Despite the simplicity of our model system when compared with the actual (PDF)-Ni(II) active site, experimental measurements of PDF-Ni enzyme inhibition⁵ are consistent with many of our results including the high binding strength calculated for the *N*-formylhydroxylamine (compound 1, Table 4) and hydroxamic acid (compound 4) and the relatively low binding strength calculated for the carboxylic acid (compound 3) and thiol (compound 6). On the other hand, the *N*-acetylhydrazine (compound 5), which we calculated to be the penultimate metal binder among the BGs studied, is 200 times less active as an enzyme inhibitor than the corresponding hydroxamic acid (compound 4).

Conclusions

Discovering an effective alternative to the hydroxamate metal binding group is a goal of both medicinal and crop protection research. The results discussed demonstrate the value density functional methods as a tool to be used in this quest. When coupled with qualitative molecular orbital reasoning about binding interactions, DFT calculations provide both insight and numbers that are of use in our exploration for alternative metal binding groups. The calculations are also sensitive in a meaningful way to the spin state of the metal at the binding site, a feature that may well be critical to ligand design and the understanding of biochemical inhibition assay results.

Supporting Information Available: Tables of geometric parameters for quartet (PATH)Co(II)Br (Table S1) and singlet (PATH)Zn(II)Br (Table S2) calculated at both the BP86/DGDZVP and B3LYP/DGDZVP levels. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Baldwin, B. C.; Corran, A. J.; Robson, M. J. *Pestic. Sci.* **1995**, *44*, 81–83.
- (2) White, R. J.; Margolis, P. S.; Trias, J.; Yuan, Z. *Curr. Opin. Pharmacol.* **2003**, *3*, 502–507.
- (3) Wu, C.-S.; Huang, J.-L.; Sun, Y.-S.; Yang, D.-Y. *J. Med. Chem.* **2002**, *45*, 2222–2228.
- (4) Tobe, H.; Morishima, H.; Aoyagi, T.; Umezawa, H.; Ishiki, K.; Nakamura, K.; Yoshioka, T.; Shimauchi, Y.; Inui, T. *Agric. Biol. Chem.* **1982**, *46*, 1865–1872.
- (5) Smith, H. K.; Beckett, R. P.; Clements, J. M.; Doel, S.; East, S. P.; Launchbury, S. B.; Pratt, L. M.; Spavold, Z. M.; Thomas, W.; Todd, R. S.; Whittaker, M. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 3595–3599.
- (6) Chen, D. Z.; Patel, D. V.; Hackbarth, C. J.; Wang, W.; Dreyer, G.; Young, D. C.; Margolis, P. S.; Wu, C.; Ni, Z.-J.; Trias, J.; White, R. J.; Yuan, Z. *Biochemistry* **2000**, *39*, 1256–1262.
- (7) (a) Giglione, C.; Pierre, M.; Meinel, T. *Mol. Microbiol.* **2000**, *36*, 1197–1205. (b) Jain, R.; Chen, D.; White, R. J.; Patel, D. V.; Yuan, Z. *Curr. Med. Chem.* **2005**, *12* (14), 1607–1621.
- (8) (a) Dirk, L. M. A.; Williams, M. A.; Houtz, R. L. *Plant Physiol.* **2001**, *127*, 97–107. (b) Williams, M.; Hou, C.-X.; Dirk, L. M. A. *Abstract of Papers*, 227th ACS National Meeting, 2004; AGFD 41.
- (9) Becker, A.; Schlichting, I.; Kabsch, W.; Groche, D.; Schultz, S.; Wagner, A. F. V. *Nat. Struct. Biol.* **1998**, *5*, 1053–1058.
- (10) Madison, V.; Duca, J.; Bennett, F.; Bohanon, S.; Cooper, A.; Chu, M.; Desai, J.; Girijavallabhan, V.; Hare, R.; Hruza, A.; Hendrata, S.; Huang, Y.; Kravec, C.; Malcolm, B.; McCormick, J.; Miesel, L.; Ramamanathan, L.; Reichert, P.; Saksena, A.; Wang, J.; Weber, P. C.; Zhu, H.; Fischmann, T. *Biophys. Chem.* **2002**, *101–102*, 239–247.
- (11) Apfel, C.; Banner, D. W.; Bur, D.; Dietz, M.; Hirata, T.; Hubschwerlen, C.; Locher, H.; Page, M. G. P.; Pirson, W.; Rosse, G.; Specklin, J.-L. *J. Med. Chem.* **2000**, *43*, 2324–2331.
- (12) Clements, J. M.; Beckett, R. P.; Brown, A.; Catlin, G.; Lobell, M.; Palan, S.; Thomas, W.; Whittaker, M.; Wood, S.; Salama, S.; Baker, P. J.; Rodgers, H. F.; Barynin, V.; Rice, D. W.; Hunter, M. G. *Antimicrob. Agents Chemother.* **2001**, *45*, 563–570.
- (13) Coats, R. A.; Lee, S. L.; Davis, K. A.; Patel, K. M.; Rhoads, E. K.; Howard, M. H. *J. Org. Chem.* **2004**, *69*, 1734–1737.
- (14) (a) Howard, M. H.; Cenizal, T.; Gutteridge, S.; Hanna, W. S.; Tao, Y.; Totrov, M.; Wittenbach, V. A.; Zheng, Y.-J. *J. Med. Chem.* **2004**, *47*, 6669–6672. (b) Howard, M. H.; Cenizal, T. M.; Coats, R. A.; Samajdar, S. *Abstract of Papers*, 229th ACS National Meeting, 2005; AGRO 064. (c) Howard, M. H.; Cenizal, T. M.; Kucharczyk, R.; Samajdar, S. *Abstract of Papers*, 229th ACS National Meeting, 2005; AGRO 050.
- (15) (a) Chang, S.; Karambelkar, V. V.; Sommer, R. D.; Rheingold, A. L.; Goldberg, D. P. *Inorg. Chem.* **2000**, *41*, 239–248. (b) Chang, S.; Karambelkar, V. K.; diTargiani, R. C.; Goldberg, D. P. *Inorg. Chem.* **2001**, *40*, 194–195. (c) Chang, S.; Sommer, R. D.; Rheingold, A. L.; Goldberg, D. P. *Chem. Commun.* **2001**, 2396–2397. (d) Goldberg, D. P.; diTargiani, R. C.; Namuswe, F.; Minnihan, E. C.; Chang, S.; Zakharov, L. N.; Rheingold, A. L. *Inorg. Chem.* **2005**, *44*, 7559–7569.
- (16) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; 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.
- (17) “B” is the Becke 88 gradient-corrected exchange functional: Becke, A. D. *Phys. Rev.* **1988**, *A38*, 3098–3100.
- (18) “P86” is the Perdew 86 gradient-corrected correlation functional: Perdew, J. P. *Phys. Rev.* **1986**, *B33*, 8822–8824.
- (19) (a) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (20) (a) Dunlap, B. I. *J. Chem. Phys.* **1983**, *78*, 3140. (b) Dunlap, B. I. *J. Mol. Struct. (THEOCHEM)* **2000**, *529*, 37.
- (21) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560.
- (22) Even now, relative energetics for transition-metal complexes is a very active topic. For example, see: Fouqueau, A.; Casida, M. E.; Daku, L. M. L.; Hauser, A.; Neese, F. *J. Chem. Phys.* **2005**, *122*, 044110/1–13.
- (23) For example, see: Himo, F.; Siegbahn, P. E. M. *Chem. Rev.* **2003**, *103*, 2421–2456.
- (24) Dobbs, K. D.; Dixon, D. A. *J. Phys. Chem.* **1996**, *100*, 3965–3973. Also, B3LYP/DGDZVP structures for quartet (PATH)-Co(II)Br and singlet (PATH)Zn(II)Br are compared with BP86/DGDZVP and experimental structures in Supporting Information Tables S1 and S2.
- (25) Albright, T. A.; Burdett, J. K.; Whangbo, M.-H. *Orbital Interactions in Chemistry*; John-Wiley and Sons: New York, 1985; Chapter 19: The ML₂ and ML₄ Fragments, pp 358–380.
- (26) Jain, R.; Hao, B.; Liu, R.-p.; Chan, M. K. *J. Am. Chem. Soc.* **2005**, *127*, 4558–4559.
CT050192U

CuNO₂ and Cu⁺NO₂ Revisited: A Comparative ab Initio and DFT Study

Stepan Sklenak* and Jan Hrušák

J. Heyrovsky Institute of Physical Chemistry, Academy of Sciences of the Czech Republic, Dolejskova 3, 18223 Prague, Czech Republic

Received August 10, 2005

Abstract: We have reinvestigated CuNO₂ and Cu⁺NO₂ at ab initio as well as at pure and hybrid DFT levels of approximation employing large ANO basis sets. The systems were fully optimized using the CCSD(T), QCISD(T), BPW91, PBE, PBE0, and B3LYP methods. Several stationary points (minima and transition structures) were found on the related potential energy surfaces (PES). The C_{2v} bidentate η²-O,O isomer is calculated to be the most stable species on the CuNO₂ PES, followed by two monodentate isomers—the C_s η¹-O and C_{2v} η¹-N species which are higher in energy by 12 and 14 kcal/mol, respectively, at CCSD(T)/Basis-II (where Basis-II is 21s15p10d6f4g/8s7p5d3f2g for Cu; 14s9p4d3f/5s4p3d2f for O and N). On the Cu⁺NO₂ PES, the C_s monodentate η¹-O trans (0 kcal/mol) and cis (+3 kcal/mol at CCSD(T)/Basis-II) isomers are found, followed by the C_{2v} monodentate η¹-N isomer (+14 kcal/mol at CCSD/Basis-II). In contrast to the pure DFT, the hybrid DFT methods perform reasonably well for predicting the relative stabilities (except for η¹-N of CuNO₂) and structures; however, their predictions of the bond dissociation energies are less reliable (for CuNO₂ the difference is as much as 10 kcal/mol compared to the CCSD(T) values). The performance of the QCISD(T) method was analyzed, and, furthermore, the issue of symmetry breaking was investigated.

1. Introduction

Nitrogen oxides are important industrial pollutants which can be removed from air by a selective catalytic reduction¹ (SCR) on transition-metal zeolites. Copper is often employed in these processes.^{2–8} Furthermore, it was demonstrated in many studies that the monovalent Cu⁺ ion is the core of the active sites of copper zeolite catalysts.^{9–12} The mechanism of the SCR is not fully understood yet. However, it is plausible to assume that a key role is played by the CuNO₂ complex.

There are different ways¹³ in which NO₂ can coordinate to Cu or Cu⁺. NO₂ can act as a monodentate ligand and coordinate through either O (η¹-O coordination) or N (η¹-N coordination). It can also act as a bidentate ligand and interact with the copper via either two O atoms (η²-O,O coordination) or O and N atoms (η²-O,N coordination). Several theoretical studies of the CuNO₂ system in the gas phase^{14–16} and zeolites^{16–18} have been published.

Sodupe et al.¹⁵ studied the bonding of NO₂ to Cu and Ag using the MP2 and DFT methods in conjunction with moderate basis sets. The energy calculations were refined by MCPD, CCSD(T), and QCISD(T) single point calculations. Three isomers of CuNO₂ were found¹⁵—the most stable C_{2v} bidentate η²-O,O isomer, the C_s monodentate η¹-O isomer, and the least stable C_{2v} monodentate η¹-N isomer. Only moderately sized basis sets of DZ quality were used in the study,¹⁵ and thus the calculated relative energies of the isomers differed significantly depending on the levels of approximation used. In some cases, also sizable differences (up to 24 kcal/mol) between CCSD(T) and QCISD(T) were obtained and attributed to an unsound estimation of the triple excitations.¹⁵ Similar conclusions had already been drawn for CuCH₃ by Frenking et al.¹⁹ who reported “dramatic failure” of the QCISD(T) method. However, it was shown later²⁰ that this failure of the QCISD(T) method, which is reflected in the flawed bond energy, is due to the inferiority of the QCISD method itself rather than due to the failure of

Corresponding author e-mail: stepan.sklenak@jh-inst.cas.cz.

the perturbative estimate of connected triple excitation contributions (T). It will be discussed later in this paper that the CuNO_2 and Cu^+NO_2 systems suffer from similar problems, and, in some cases, symmetry breaking leads to further problems in evaluation of physical-chemical properties.

Sauer et al.¹⁶ studied the structure and stability of Cu^+NO_2 in the gas phase and in the ZSM-5 zeolite using the B3LYP method. In the gas phase, they found three minima and two transition states on the ground state ($^2A'$ and 2A_1) potential energy surface of Cu^+NO_2 . The $\eta^1\text{-O}$ trans isomer was calculated to be the most stable species. The $\eta^1\text{-O}$ cis and $\eta^1\text{-N}$ isomers are higher in energy by 2 and 10 kcal/mol, respectively. Sauer et al.¹⁶ concluded that the bonding in Cu^+NO_2 is mainly noncovalent and arises from the interaction of the $^1S(d^{10})$ state of Cu^+ and the 2A_1 ground state of NO_2 . Further information on Cu^+NO_2 can be extracted from the recently appeared comparative study of Ducere et al.¹⁴ on the binding of NO_2 , NH_3 , H_2O , NO , N_2O , N_2 , and O_2 to Cu^+ and Cu^{2+} at several DFT and ab initio levels.

In the present paper we recalculate the $[\text{Cu}, \text{N}, \text{O}_2]^{0/+}$ neutral and positively charged systems at the uniform CCSD(T) level of theory with large ANO basis sets.^{21,22} These calculations serve for evaluating reliable relative stabilities and interconversion profiles as well as benchmarks for the most common DFT methods.

2. Methods

All the studied species were fully optimized, and the vibrational frequencies were determined using the MOLPRO ab initio program package²³ employing the Roos augmented ANO basis sets^{21,22} in the contractions designated as Basis-I (Cu: $21s15p10d6f/6s5p4d2f$ and O,N: $14s9p4d/4s3p2d$) and Basis-II (Cu: $21s15p10d6f4g/8s7p5d3f2g$ and O,N: $14s9p4d3f/5s4p3d2f$) and obtained from the Extensible Computational Chemistry Environment Basis Set Database, Version 02/25/04.²⁴

The ab initio calculations were performed at the two correlated ab initio CCSD(T)^{25–29} and QCISD(T)^{25,26,29,30} levels of theory as implemented in the MOLPRO program. The open shell species were calculated using the spin unrestricted (UCCSD(T)/ROHF^{31,32} and UQCISD(T)/ROHF^{31,32}) methods. Some supporting calculations were performed with the GAUSSIAN03 program package³³ at the UCCSD(T)/UHF level.

It was pointed out by Urban et al.^{34,35} that for the $\text{Cu}\cdots\text{OH}_2$ complex the triple excitations which follow from correlating the $3p^6$ shell of Cu make a considerable contribution in the vicinity of the minimum of the interaction potential. To investigate the effect of the $3p^6$ shell of Cu on the relative energies of the CuNO_2 and Cu^+NO_2 species, we carried out single point CCSD and CCSD(T) calculations. However, the results showed that the effect of the $3p$ electrons on the relative energies is in the range of a few tenths of kcal/mol. Thus we decided to use the “frozen core” approximation as implemented in the MOLPRO program, i.e., only the copper 3d and 4s electrons as well as 2s and

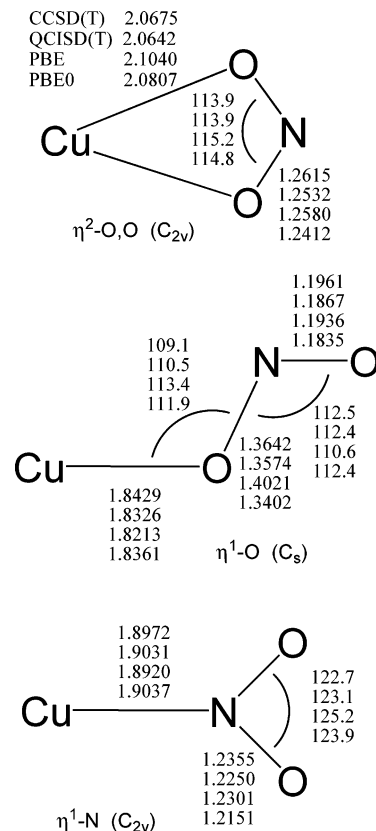


Figure 1. Optimized structures of the $\eta^2\text{-O,O}$ (a, top), $\eta^1\text{-O}$ (b, middle), and $\eta^1\text{-N}$ (c, bottom) isomers of CuNO_2 at CCSD(T)/Basis-II, QCISD(T)/Basis-II, PBE/Basis-II, and PBE0/Basis-II. Bond lengths are in Å and bond angles in deg.

2p electrons of N and O were correlated in all the CCSD(T) and QCISD(T) calculations.

In addition, we also performed calculations using two pure and two hybrid density function theory methods—BPW91,³⁶ PBE³⁷ and PBE0,³⁸ B3LYP,^{39–41} respectively. The implementations of the unrestricted DFT methods were used for the open shell species. Moreover, the ACESII⁴² program was employed to test the stability of HF solutions and to calculate the CCSD(TQ)⁴³ energies as well as to obtain the CCSD amplitudes which were checked for all the species to ensure that the systems are well described by a single reference configuration.

3. Results and Discussion

3.1. CuNO_2 . 3.1.1. Relative Stabilities and Structures. We found three minima and two transition states connecting these minima on the $[\text{Cu}, \text{N}, \text{O}_2]$ potential energy surface. The optimized structures of all the species of CuNO_2 as well as of NO_2 and NO_2^- are given in Figure 1a–c and Tables S1 and S2 of the Supporting Information.

The C_{2v} bidentate $\eta^2\text{-O,O}$ isomer (Figure 1a) is calculated to be the most stable isomer of CuNO_2 at all levels of theory (see also Figure 2 and Table 1) and represents a pronounced well on the related PES. Only slight differences in the geometrical parameters can be observed depending on the method used. Not surprisingly, there is good agreement between the QCISD(T) and CCSD(T) results, since both methods are assumed to be more or less identical.^{44,45} The

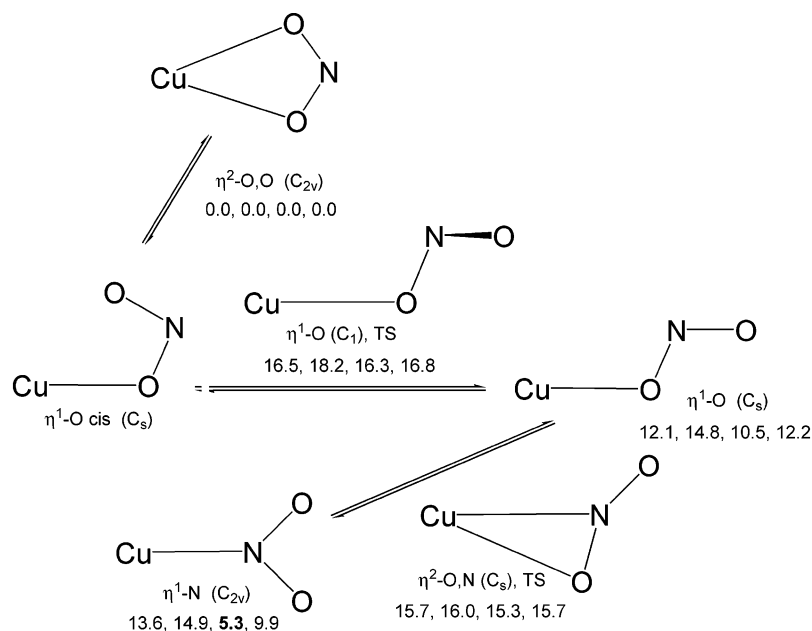


Figure 2. Relative energies (in kcal/mol) of the CuNO₂ isomers and transition states at CCSD(T)/Basis-II, QCISD(T)/Basis-II, PBE/Basis-II, and PBE0/Basis-II.

Table 1. Calculated Relative Energies (in kcal/mol) for All Minima and Transition States of CuNO₂^a

isomer	basis set	CCSD	CCSD(T)	QCISD	QCISD(T)	BPW91	PBE	PBE0	B3LYP
η^2 -O,O	Basis-I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
η^2 -O,O	Basis-II	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
η^1 -N	Basis-I	15.3	14.2	14.1	16.0	5.3	5.4	10.1	9.2
η^1 -N	Basis-II	14.8	13.6	13.8	14.9	5.2	5.3	9.9	9.1
η^1 -O	Basis-I	11.4	11.6	9.3	15.2	10.0	10.5	12.3	10.8
η^1 -O	Basis-II	11.8	12.1	10.2	14.8	10.1	10.5	12.2	10.7
η^2 -O,N (TS)	Basis-I	16.9	16.2	16.7	16.6	15.0	15.2	15.8	15.9
η^2 -O,N (TS)	Basis-II	16.4	15.7	16.3	16.0	15.0	15.3	15.7	15.9
η^1 -O (TS)	Basis-I	15.6	16.1	14.7	18.3	16.2	16.6	17.2	15.7
η^1 -O (TS)	Basis-II	15.9	16.5	15.3	18.2	15.9	16.3	16.8	15.4

^a The energy values include the electronic energy and zero point energy (ZPE). For the CCSD and QCISD levels, the ZPE values at CCSD(T) and QCISD(T), respectively, are used.

largest CCSD and QCISD amplitudes (0.09 and 0.11) as well as the values of the T1 diagnostic (0.030 and 0.037) are small.

It should be noted that the pure DFT calculated Cu–O bond lengths are slightly longer (0.04 Å) than that at the CCSD(T) level, and the inclusion of the “exact HF exchange” in the hybrid methods brings the ab initio and DFT results closer (difference of 0.02 Å). It has been stated earlier that the bonding between Cu and NO₂ in the η^2 -O,O isomer is mainly ionic.¹⁵ This ionic character of the metal–ligand bond is reflected in the structure of the NO₂ moiety that is very close to that of NO₂[−] ($r(\text{N–O})$: 1.262 Å; $\alpha(\text{O–N–O})$: 116.4° at CCSD(T)/Basis-II) rather than to that of NO₂ ($r(\text{N–O})$: 1.198 Å; $\alpha(\text{O–N–O})$: 134.1° at CCSD(T)/Basis-II).

The remaining two isomers on the neutral [Cu, N, O₂] PES are close in energy, and their relative order of stabilities depends strongly on the level of theory used (see Figure 2 and Table 1). The C_s monodentate η^1 -O isomer (Figure 1b) is the second most stable species at CCSD(T). The copper acts as a monodentate ligand, and it is coordinated only to the oxygen atom. The calculated Cu–N distance (2.627 Å at CCSD(T)/Basis-II) is significantly longer than the bonding

distance, and, furthermore, also the orbital analysis reveals that there is no significant contribution of the Cu–N overlap to the bonding (vide infra). All the methods provided similar structures. The calculated Cu–O bond is uniformly shorter than that in the C_{2v} bidentate η^2 -O,O isomer reflecting a larger covalent contribution to the bonding. The only geometry parameter which significantly varies at the different levels is the O_{Cu}–N bond distance which spans the interval from 1.340 Å (PBE0/Basis-II) to 1.413 Å (BPW91/Basis-I). The O_{Cu}–N bond is significantly longer than the N=O bond (by 0.17 Å at CCSD(T)/Basis-II) which, consistently with the valence bond picture, has a character of a double bond rather than a single bond. The calculated O–N–O bond angle (113° at CCSD(T)/Basis-II) is again much closer to that of NO₂[−] than to that of NO₂. Thus, also in this η^1 -O isomer the bonding is dominated by the ionic character.

The C_{2v} monodentate η^1 -N isomer (Figure 1c) is calculated at CCSD(T) to be the least stable CuNO₂ isomer (Figure 2 and Table 1). The calculated structures are very similar at all the levels used. The calculated N–O bond length (1.236 Å at CCSD(T)/Basis-II) is shorter and the O–N–O bond angle (123° at CCSD(T)/Basis-II) is larger than the corre-

sponding geometry parameters of the C_{2v} bidentate η^2 -O,O isomer, and their values are between those of NO_2 and NO_2^- . This fact reveals that the covalent contribution to the bonding is larger for the η^1 -N isomer than for the other two isomers.

Although the η^2 -O,O isomer is the most stable species at all computational levels (see Table 1), the order of the two less stable isomers is different at various levels of approximation. Let us first focus on the coupled cluster (CC) level. The η^1 -O isomer is calculated to be less stable than η^2 -O,O by 11–12 kcal/mol, while the η^1 -N isomer is higher in energy than η^2 -O,O by 14–15 kcal/mol. The effects of the perturbative contributions of connected triple excitations (hereafter (T)) as well as of the size of the basis set are negligible in both cases (smaller than 1 kcal/mol).

The influence of the perturbative contributions of connected quadruple excitations (hereafter (Q)) on the relative energies of the isomers of CuNO_2 was investigated as well. However, the CCSD(TQ)/Basis-I//CCSD(T)/Basis-I results reveal that the effect of (Q) on the relative energies is very small—a few tenths of kcal/mol. [η^1 -O and η^1 -N are less stable than η^2 -O,O by 12.0 and 14.7 kcal/mol, respectively, at CCSD(TQ)/Basis-I//CCSD(T)/Basis-I (plus the ZPE energy at CCSD(T)/Basis-I).] The negligible effect of (Q) is in agreement with already small effect of the triples (T).

All three isomers of CuNO_2 were also calculated employing the effective core potential of Hay and Wadt⁴⁶ and Basis-I at the CCSD(T) level. However, the relative energies of the three isomers as well as their optimized geometries were very close to those calculated at the CCSD(T)/Basis-II//CCSD(T)/Basis-I level. [η^1 -O and η^1 -N are less stable than η^2 -O,O by 12.8 and 15.5 kcal/mol, respectively, at CCSD(T)/ECP+Basis-I//CCSD(T)/ECP+Basis-I (plus the ZPE energy at CCSD(T)/Basis-I).]

The energy order of the isomers of CuNO_2 can be also rationalized using a simple concept of electronegativity. The copper atom which donates one s electron to the NO_2 moiety prefers to coordinate to a more electronegative element, i.e., oxygen. Thus η^2 -O,O, in which Cu coordinates to two oxygen atoms, is the most stable. Consequently, the η^1 -O species is less stable (Cu is ligated only to one oxygen atom) followed by η^1 -N (Cu coordinates to the nitrogen atom).

3.1.2. Bonding. The analysis of the orbitals involved in the formation of the bond between Cu and NO_2 in the η^2 -O,O isomer of CuNO_2 (Figure S1 of the Supporting Information) reveals that the bonding between Cu and NO_2 in CuNO_2 is mainly ionic, and it arises from the interaction of the $^1\text{S}(\text{d}^{10})$ state of Cu^+ and the $^1\text{A}_1$ ground state of NO_2^- . The 4s orbital of Cu, which is singly occupied in Cu, interacts with the SOMO orbital ($6a_1$) of NO_2 to form the HOMO orbital ($13a_1$) of CuNO_2 , which polarizes toward the NO_2 moiety. The $7b_2$ and $6b_2$ orbitals of CuNO_2 arise from the antibonding and bonding, respectively, combinations between the $3d_{yz}$ orbital of Cu and the $4b_2$ orbital of the NO_2 moiety. The remaining 3d orbitals of Cu do not significantly interact with the orbitals of NO_2 . The bonding in the other two isomers is very similar. The Mulliken populations calculated for all three isomers (Table 2) confirm an ionic character of all three isomers.

The bonding in all three isomers is driven by the donation

Table 2. Mulliken Populations in the s, p, and d Orbitals of Cu, N, and O of CuNO_2

isomer	atom	s	p	d	charge
η^2 -O,O	Cu	6.18	12.10	9.98	+0.73
η^2 -O,O	O	3.86	4.61	0.03	-0.51
η^2 -O,O	N	3.64	2.77	0.26	+0.29
η^1 -O	Cu	6.21	12.10	9.94	+0.73
η^1 -O	O ₁	3.88	4.70	0.03	-0.62
η^1 -O	N	3.65	2.85	0.24	+0.22
η^1 -O	O ₂	3.86	4.42	0.05	-0.34
η^1 -N	Cu	6.18	12.06	9.93	+0.82
η^1 -N	N	3.58	2.93	0.36	+0.09
η^1 -N	O	3.86	4.54	0.04	-0.45

(ca 0.8 e) of the 4s-electron on copper to the NO_2 fragment. The back-donation from NO_2^- into the 4p orbitals of Cu is sizably smaller. This back-donation is the largest for the η^2 -O,O isomer, about 0.08 e, and it is smaller for η^1 -O (0.04 e) and negligible for η^1 -N.

3.1.3. QCISD. When analyzing the QCISD and QCISD(T) relative energies, notable differences (1–4 kcal/mol) between the QCI and CC values are found for the η^1 -O and η^1 -N species. Moreover, the effect of (T), which is small at CC, is sizable at QCI especially for η^1 -O as it increases the relative energy by 5–6 kcal/mol with respect to η^2 -O,O. Surprisingly, the energy gap between the CC and QCI results for η^1 -O and η^1 -N as obtained by Sodupe et al.,¹⁵ when using a smaller [Cu: 8s6p4d] basis set, were substantially larger (up to 24 kcal/mol). In the manner of "dramatic failure of QCISD(T)"^{15,19,47} this effect was attributed to the unsound estimation of (T) i.e., the perturbative method was made responsible for the failure. These explanations ignore the fact that already the QCISD solution is severely flawed,^{20,48} and the omitted nonzero connected T_1 -terms in the QCISD equations are fully responsible for these irregularities. Furthermore, the QCISD method offers no significant computational advantages with respect to CCSD and should be avoided.

3.1.4. DFT. The results obtained at DFT depend on whether the functional employed is pure (BPW91 and PBE) or hybrid (PBE0 and B3LYP). η^1 -O is calculated to be 10 kcal/mol less stable than η^2 -O,O with the pure DFT, while the hybrid DFT values are very close to the 12 kcal/mol calculated at CCSD(T)/Basis-II. The pure DFT relative energies of η^1 -N with respect to η^2 -O,O (5 kcal/mol) and even the 10 kcal/mol calculated at PBE0 and B3LYP are in very poor agreement with the superior CCSD(T) values (14 kcal/mol) irrespective of the similar optimized geometries of η^1 -N.

3.1.5. Transition States. Two transition states were localized on the potential energy surface of CuNO_2 . The first one is the C_1 monodentate η^1 -O species and the second one is the C_s bidentate η^2 -O,N species. The calculated imaginary frequencies reveal that the isomerizations η^2 -O,O \rightarrow η^1 -O and η^1 -O \rightarrow η^1 -N proceed via the former and latter transition states, respectively.

The calculations showed that all three bond distances of the C_1 η^1 -O TS are close to those of the C_s η^1 -O isomer for all the methods used. The relative energy of the η^1 -O TS,

Table 3. Bond Dissociation Energies (in kcal/mol) of the η^2 -O,O Isomer of CuNO₂ with Respect to the Cu + NO₂ and Cu⁺ + NO₂⁻ Channels

channel	basis set	CCSD	CCSD(T)	HF	QCISD	QCISD(T)	BPW91	PBE	PBE0	B3LYP
Cu + NO ₂	Basis-I	55.3	54.7	48.7	56.3	53.8	44.5	48.0	48.2	46.3
Cu + NO ₂	Basis-II	55.9	55.2	47.0	56.7	54.3	43.6	47.1	47.3	45.5
Cu ⁺ + NO ₂ ⁻	Basis-I	172.8	176.0	157.4	174.8	176.1	185.6	189.0	179.6	179.4
Cu ⁺ + NO ₂ ⁻	Basis-II	172.8	176.3	156.7	174.8	176.5	186.2	189.4	179.8	179.8

Table 4. Vertical and Adiabatic Ionization Potentials (in kcal/mol) of the η^2 -O,O Isomer of CuNO₂

type	basis set	CCSD	CCSD(T)	QCISD	QCISD(T)	BPW91	PBE	PBE0	B3LYP
vertical	Basis-I	234.0	231.3	236.2	251.6	225.3	226.5	230.4	230.7
vertical	Basis-II	236.0	233.6	238.3	251.4	224.7	225.9	229.7	230.0
adiabatic	Basis-I	200.6	202.0	201.7	201.9	205.0	206.7	203.7	205.0
adiabatic	Basis-II	202.8	204.5	203.8	204.4	204.2	205.9	202.6	204.1

which also corresponds to the barrier of isomerization η^2 -O,O \rightarrow η^1 -O, is 16–17 kcal/mol at CCSD(T) and DFT. The imaginary frequency corresponds to the torsion mode, and thus the transition state connects the η^1 -O cis and trans species. However, all computational attempts to localize a η^1 -O cis species led to the η^2 -O,O isomer. Restricted optimization scans indicated that the η^1 -O cis species is rather a shoulder on the potential energy surface and the barrier for its isomerization into η^2 -O,O is most likely very small.

The calculated geometry parameters of the C_s bidentate η^2 -O,N TS depend significantly on the methods employed. The DFT schemes provide the structures having the Cu–O bond too short (by up to 0.20 Å) and the Cu–N bond too long (by up to 0.15 Å) with respect to the CCSD(T) results. In other words, the isomerization η^1 -O \rightarrow η^1 -N is found to have a late transition state at the CC and QCI levels, while it has an early TS at DFT. The imaginary frequency corresponds to C–N and C–O asymmetric stretching mode. Surprisingly, the calculated relative energies of the η^2 -O,N TS are within a small interval 15–17 kcal/mol for all the methods used.

3.1.6. Bond Dissociation Energies. In Table 3 we present the bond dissociation energies, hereafter D_e , of the η^2 -O,O isomer of CuNO₂ with respect to Cu and NO₂ as well as to Cu⁺ and NO₂⁻. The D_e values calculated at CCSD(T) are 55 and 176 kcal/mol for the Cu + NO₂ and Cu⁺ + NO₂⁻ channels, respectively. The effect of (T) is 3 kcal/mol for the latter channel and negligible for the former one. The QCI D_e values are rather close to the CC ones. The differences between the CCSD(T) and HF D_e values reveal the effect of electron correlation which is 6–8 and 19 kcal/mol for the Cu + NO₂ and Cu⁺ + NO₂⁻ channels, respectively. The D_e values calculated at the hybrid DFT are significantly smaller by (7–10 kcal/mol) than those calculated at CCSD(T) for the Cu + NO₂ channel. The main reason of the disagreement is the inability of DFT to correctly describe the copper atom (2A_g). The Cu-ionization potential calculated at PBE0/Basis-II and B3LYP/Basis-II is about 7 and 11 kcal/mol larger, respectively, than that calculated at CCSD(T)/Basis-II. On the other hand, for the Cu⁺ + NO₂⁻ channel the agreement between the hybrid DFT and CCSD(T) D_e values is significantly better as the difference is about 3.5 kcal/mol.

The effect of the size of the basis set is less than 1 kcal/mol for all the methods employed.

3.1.7. Ionization Potentials. To complete the figure and to make a bridge to the charged species we calculated the vertical (IP_v) and adiabatic (IP_a) ionization potentials of the η^2 -O,O isomer of CuNO₂. The individual values are revealed in Table 4. The IP_v values calculated at the CCSD, CCSD(T), QCISD, and hybrid DFT levels lie in a narrow interval 230–238 kcal/mol. It should be noted that the QCISD(T) values are significantly larger.

In contrast to IP_v, all the methods used provide very similar adiabatic ionization potentials (201–207 kcal/mol) since the geometries of Cu⁺NO₂ are relaxed and the corresponding energies are calculated at the minimum points of the energy potential surface.

3.2. Cu⁺NO₂. 3.2.1. Relative Stabilities and Structures. Let us turn our attention on the positively charged system. We found three minima and two transition states connecting these minima on the potential energy surface. The optimized structures of all the species of Cu⁺NO₂ are given in Figure 3a–c and Table S3 of the Supporting Information.

The C_s monodentate η^1 -O trans isomer ($^2A'$) (Figure 3a) is calculated to be the most stable isomer of Cu⁺NO₂ at all levels of theory, see Figure 4 and Table 5. The CCSD(T), QCISD(T), and hybrid DFT methods provide very similar structures. The Cu–O bond length is calculated to be 1.985 and 1.96 Å at CCSD(T)/Basis-II and hybrid DFT/Basis-II. Due to the missing bonding electron, the bond is longer than the corresponding Cu–O bond in the η^1 -O isomer of CuNO₂ by 0.14 Å. On the other hand, the lengths of the N–O bonds of η^1 -O trans of Cu⁺NO₂ are significantly shorter than those of η^1 -O of CuNO₂ (1.239 and 1.166 Å for Cu⁺NO₂; 1.364 and 1.196 Å for CuNO₂), and they are together with the value of the O–N–O bond angle (132°) close to the geometry parameters of NO₂ (1.198 Å and 134°). The pure DFT methods provided significantly shorter Cu–O bond lengths (~1.90 Å).

The C_s monodentate η^1 -O cis isomer ($^2A'$) (Figure 3b) is calculated to be the second most stable minimum lying 2–3 kcal/mol at all the levels used (see Figure 4 and Table 5) higher than η^1 -O trans. The calculated geometry parameters of η^1 -O cis are very close to those of η^1 -O trans possessing

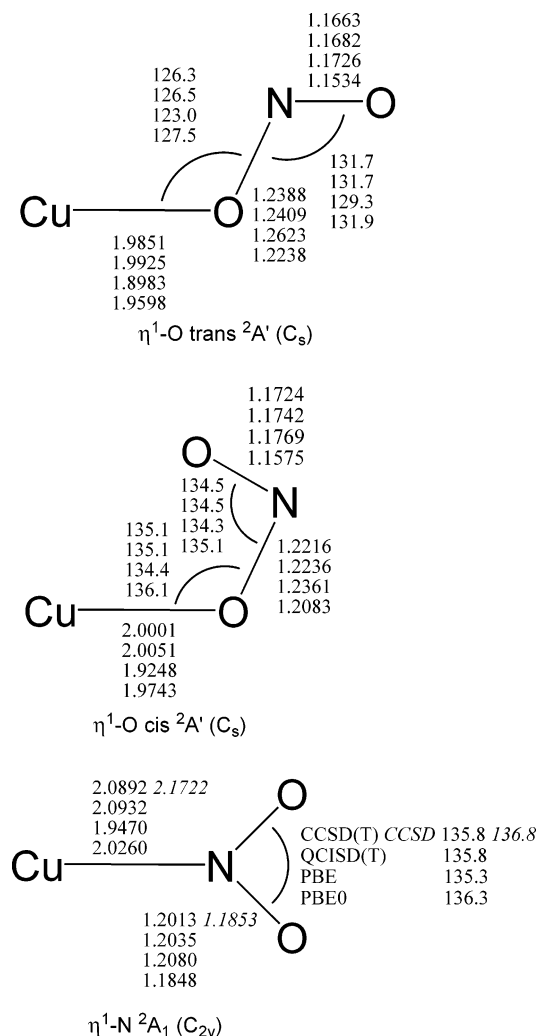
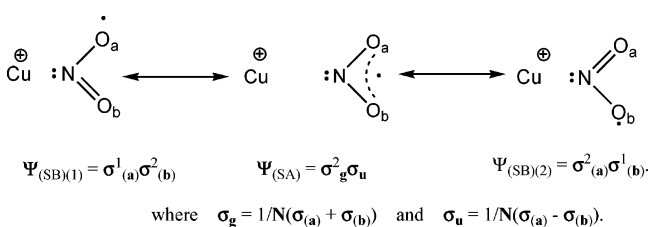


Figure 3. Optimized structures of the $\eta^1\text{-O trans}$ (a, top), $\eta^1\text{-O cis}$ (b, middle), and $\eta^1\text{-N}$ (c, bottom) isomers of Cu^+NO_2 at CCSD(T)/Basis-II, QCISD(T)/Basis-II, PBE/Basis-II, and PBE0/Basis-II. The values in italic are at CCSD/Basis-II (only for the $\eta^1\text{-N}$ isomer). Bond lengths are in Å and bond angles in deg.

the same trends for the methods used. It only might be mentioned that the O–N–O angle is slightly widened reflecting the steric (nonbonding) repulsion of the Cu^+ . The largest CCSD amplitudes (0.07) as well as the values of the T1 diagnostic (0.025) are very small for both isomers.

3.2.2. Symmetry Breaking. The C_{2v} monodentate $\eta^1\text{-N}$ isomer (2A_1) (Figure 3c) is calculated to be the least stable among the isomers of Cu^+NO_2 (see Figure 4 and Table 5) at all the levels used. This isomer can be described by two degenerate valence bond structures having the unpaired electron on either $\text{O}_{(a)}$ or $\text{O}_{(b)}$.



That indicates a possibility of symmetry broken Hartree–Fock (HF) solutions for this species.^{20,48–51} When the $\eta^1\text{-N}$

isomer (2A_1) is calculated in the C_{2v} symmetry, the wave function $\Psi_{(\text{SA})}$ is symmetry adapted (hereafter SA), and it belongs to the A_1 irreducible representation. $\Psi_{(\text{SA})}$ covers the resonance between two solutions bearing the unpaired electron on either $\text{O}_{(a)}$ or $\text{O}_{(b)}$. The symmetry adaptation is a further constrain in a variational calculation, and it might consequently lead to a higher energy. To investigate whether the symmetry adapted wave function of the $\eta^1\text{-N}$ isomer (2A_1) is stable, the stability of the HF solution was tested. We could not directly test the stability of the ROHF wave function (as used in the CCSD(T) and QCISD(T) calculations), but we tested the corresponding SA UHF wave function. The stability tests reveal that the SA UHF wave function, which is only slightly spin contaminated ($\langle S^2 \rangle = 0.78$), has several UHF \rightarrow UHF instabilities. When the orbital rotations corresponding to the instabilities were applied to the SCF eigenvectors and the SCF calculation was repeated with these rotated vectors as the starting guess, a UHF solution lower in energy by 6.3 kcal/mol was found. However, the price for lowering the energy is a heavy spin contamination ($\langle S^2 \rangle = 1.08$). Moreover, the corresponding UHF wave function does not transform as the A_1 irreducible representation of the C_{2v} point group.

The localized (symmetry broken; hereafter SB) solutions lead to a lower energy in a variational calculation, but the wave functions $\Psi_{(\text{SB})(1)}$ and $\Psi_{(\text{SB})(2)}$ do not transform as the totally symmetric irreducible representation of the molecular point group. The energy differences between the symmetry adapted (SA) and localized (SB) solutions for the $\eta^1\text{-N}$ isomer (2A_1) are negligible at CCSD despite the fact that the underlying ROHF wave function is heavily affected ($\Delta E^{(\text{SA}-\text{SB})} = 4$ kcal/mol). [The localized solution was obtained by running a calculation at the ROHF level with the $\eta^1\text{-N}$ isomer (2A_1) having two unequal N–O bond lengths and using that SCF solution as the guess in the subsequent calculations with the $\eta^1\text{-N}$ isomer (2A_1) possessing the optimized C_{2v} structure. The localized (SB) solution at the ROHF level leads to a lower energy than the SA solution by 4 kcal/mol. However, at CCSD both SA and SB solutions provide essentially the same energy due to the robustness of the CCSD method and its low energy sensitivity on the underlying SCF orbitals.] The largest CCSD amplitude (0.07) is rather small indicating that the effect is not due to a multireference character. Also the calculated T1 diagnostic of 0.025 is very small. On the other hand, the QCISD energy difference between the SA and SB solutions is sizable ($\Delta E^{(\text{SA}-\text{SB})} = 3.5$ kcal/mol) indicating that the orbital rotations could not be removed (the largest amplitude is 0.10). However, it is noteworthy that the inclusion of (T) for both CCSD and QCISD leads to the SA energy which is lower than the SB one. This indicates that both solutions (CCSD(T) and QCISD(T)) are not very reliable in these cases. To partially eliminate the effect of symmetry breaking, the geometry of the $\eta^1\text{-N}$ isomer (2A_1) was reoptimized at the CCSD level of theory. Sizable changes in geometry are observed. The Cu–N bond is calculated to be longer by 0.08 Å at CCSD ($r(\text{Cu-N})$ is 2.200 Å and 2.172 Å at CCSD/Basis-I and CCSD/Basis-II, respectively) than at CCSD(T). The $\eta^1\text{-N}$ isomer (2A_1) is higher in energy than the $\eta^1\text{-O}$

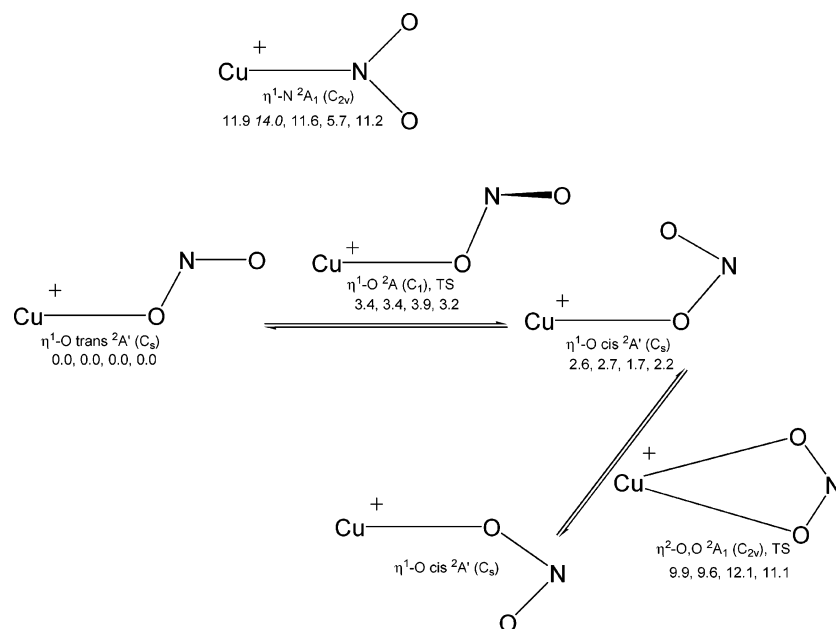


Figure 4. Relative energies (in kcal/mol) of the Cu⁺NO₂ isomers and transition states at CCSD(T)/Basis-II, QCISD(T)/Basis-II, PBE/Basis-II, and PBE0/Basis-II. The value in italic is at CCSD/Basis-II (only for the η^1 -N isomer).

Table 5. Calculated Relative Energies (in kcal/mol) for All Minima and Transition States of Cu⁺NO₂^a

isomer	state	basis set	CCSD	CCSD(T)	QCISD	QCISD(T)	BPW91	PBE	PBE0	B3LYP
η^1 -O trans	$^2A'$	Basis-I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
η^1 -O trans	$^2A'$	Basis-II	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
η^1 -O cis	$^2A'$	Basis-I	2.4	2.6 ^b	2.7	2.7	1.9	1.8	2.2	2.3
η^1 -O cis	$^2A'$	Basis-II	2.4	2.6	2.6	2.7	1.7	1.7	2.2	2.2
η^1 -N	2A_1	Basis-I	14.2	12.2	15.3	11.9	5.6	5.5	11.0	11.0
η^1 -N	2A_1	Basis-II	14.0	11.9	14.9	11.6	5.8	5.7	11.2	11.3
η^2 -O,O (TS)	2A_1	Basis-I	9.7	9.7	10.1	9.4	11.7	11.7	10.5	11.2
η^2 -O,O (TS)	2A_1	Basis-II	9.8	9.9	10.2	9.6	12.1	12.1	11.1	11.6
η^1 -O (TS)	2A	Basis-I	3.1	3.4	3.5	3.4	4.2	4.2	3.5	3.7
η^1 -O (TS)	2A	Basis-II	3.0	3.4	3.3	3.4	3.9	3.9	3.2	3.4

^a The energy values include the electronic energy and zero point energy (ZPE). For the CCSD and QCISD levels, the ZPE values at CCSD(T) and QCISD(T) are used. The CCSD energies of the η^1 -N isomer correspond to the reoptimized geometry at CCSD. ^b 2.5 kcal/mol at CCSD(T)/ECP+Basis-II/CCSD(T)/ECP+Basis-I (plus the ZPE energy at CCSD(T)/Basis-I).

trans one by 14.0 kcal/mol at CCSD/Basis-II (11.9 kcal/mol at CCSD(T)/Basis-II).

3.2.2.1. CASSCF and MR-SDCI. To shed further light on the problem described above, we carried out CASSCF^{52–60} and subsequently internally contracted MR-SDCI^{61,62} calculations of the η^1 -N isomer (2A_1) of Cu⁺NO₂. Employing multireference methods such as CASSCF might be a way to avoid symmetry breaking^{63–66} since these methods include more reference functions which are able to better describe several valence bond structures. On the other hand, there is only a small amount of dynamic electron correlation included in the CASSCF calculations, and, thus, we enhanced the treatment using the MR-SDCI method.

Our single point CASSCF/Basis-I/CCSD/Basis-I calculations employed four different active spaces (in the reduced C_s symmetry) as described in Table 6. The symmetry adapted (SA) and broken (SB) HF wave functions were used as the initial guess for the CASSCF calculations. The SA guess led to a lower CASSCF energy than the SB guess (see the $\Delta E_{(SB-SA)}$ values in Table 6). The energy gap between the SB and SA CASSCF solutions ($\Delta E_{(SB-SA)}$) decreased as the size of the active space increased indicating that even this

CASSCF method is unable to guarantee a single solution when it is started from the SA and SB guesses. A larger active space should lead to a single solution (in the full CI limit); however, such calculations became prohibited for technical reasons. The corresponding CI vectors reveal that for all the active spaces used the CASSCF wave function is strongly dominated by an SCF-like solution based on the leading ground-state electron configuration. This fact causes that CASSCF does not provide a single solution for the SA and SB guesses; however, on the other hand, it justifies the use of the single reference CCSD method which yields the same energy for both SA and SB solutions. The involvement of a low-lying excited state of the η^1 -N isomer (2A_1) of Cu⁺NO₂ could be ruled out since the first excited state is some 70 kcal/mol higher in energy.

Further, we applied the MR-SDCI method employing the results of the CASSCF(7,8) and CASSCF(7,7) calculations in order to investigate the effect of dynamic electron correlation. The energy gap between the SA and SB solutions is reduced by only 0.2 kcal/mol at MR-SDCI, and it is further reduced by 0.8–0.9 kcal/mol when the Davidson correction⁶⁷ (MRCI(Q)) is employed (Table 6). However, the MRCI

Table 6. Energy Differences (in kcal/mol) between the Symmetry Adapted (SA) and Localized (SB) Solutions at Different Levels of Approximation for the η^1 -N Isomer (2A_1)^f

method ^a	active space orbitals	$\Delta E_{(SB-SA)}$	method	$\Delta E_{(SB-SA)}$	method	$\Delta E_{(SB-SA)}$
CASSCF(13,13)	16a' - 23a', 4a'' - 8a'' ^b	1.51				
CASSCF(13,12)	16a' - 22a', 4a'' - 8a'' ^c	1.59				
CASSCF(7,8)	19a' - 22a', 5a'' - 8a'' ^d	3.29	MRCI	3.07	MRCI(Q)	2.14
CASSCF(7,7)	19a' - 22a', 6a'' - 8a'' ^e	3.48	MRCI	3.27	MRCI(Q)	2.49
HF		-1.39				
CCSD		0.02				

^a CASSCF(*n,m*) where *n* is number of electrons and *m* is number of orbitals. ^b Frozen orbitals: 1a' - 15a', 1a'' - 3a''. ^c Frozen orbitals: 1a' - 15a', 1a'' - 3a''. ^d Frozen orbitals: 1a' - 18a', 1a'' - 4a''. ^e Frozen orbitals: 1a' - 18a', 1a'' - 5a''. ^f The geometry optimized at CCSD/Basis-I is used.

method (based on the chosen active space), unlike the single reference CCSD approach, is unable to guarantee a single solution. MR-SDCI does not include the T_1 excitations in an exponential form and thus does not exhibit a low sensitivity on the underlying orbitals.

A conclusion can be drawn from the presented results that in the case of symmetry breaking the CCSD is the method of choice if the following three conditions are fulfilled: First, the CCSD energy gap between SA and SB solutions should be small. Second, the corresponding CASSCF wave function is strongly dominated by the leading ground-state electron configuration, and finally, no low-lying excited state of the same symmetry as the ground state is present.

3.2.2.2. Symmetry Breaking and Vibrational Frequencies. The existence of symmetry broken solutions apparently causes problems in the numerical calculations of vibrational frequencies. Namely, one small imaginary frequency corresponding to the Cu-N-O bending mode was obtained at all the ab initio levels but CCSD as a consequence of the numerical evaluation of the frequencies in lower symmetry point groups. The CCSD frequency of the Cu-N-O bending mode is a real number for the step larger than 0.03 Å indicating that the η^1 -N species (2A_1) is a minimum on the potential energy surface. A smaller step leads to an imaginary value of the Cu-N-O wavenumber. The other five frequencies do not significantly depend on the step size.

3.2.2.3. Symmetry Breaking and DFT. The performance of DFT for symmetry breaking cases was a subject of several studies.⁶⁸⁻⁷² Head-Gordon et al.⁶⁸ studied three open shell systems (NO_3 , O_4^+ , and O_2^+) for which the UHF wave function breaks spatial symmetry. It was concluded⁶⁸ that symmetry broken solutions were obtained with DFT only when unusually large fractions of HF exchange (above 70%) were included into the hybrid functionals. The exchange was found more important than correlation in determining the tendency to preserve or break symmetry in DFT.⁶⁸ However, even when the optimization of Kohn-Sham orbitals leads to a symmetric solution, there is no guarantee that the vibrational frequencies will be entirely free of the effects of symmetry breaking because the higher-lying asymmetric solutions might strongly interact with the symmetric solution.⁶⁸ In addition, the MOLPRO program calculates DFT second derivatives numerically, and thus the calculated frequencies can suffer from the same problems as those obtained at CCSD.

To test whether the DFT methods used suffer from symmetry breaking for the η^1 -N isomer (2A_1) of Cu^+NO_2 , a

symmetry broken UHF solution was obtained and used as the guess in the subsequent calculations employing the UBWP91, UPBE, UPBE0, and UB3LYP methods for η^1 -N possessing the optimized C_{2v} structure. The calculations led to the symmetric solutions for all four DFT methods employing both basis sets. The subsequent evaluation of the vibrational frequencies provided only positive values.

The relative energies of η^1 -N are 11 and 6 kcal/mol at the hybrid and pure DFT levels (Figure 4), respectively. The former value is in agreement with the CCSD one (14 kcal/mol); however, the latter energy is once again unrealistically low.

The Cu-N bond length is calculated to be significantly shorter at PBE0 and B3LYP than at CCSD by some 0.15 Å and extremely shortened at BPW91 and PBE by about 0.25 Å. These results indicate that the pure DFT methods fail to provide correct structures and relative energies of η^1 -N. The Cu-N bond is significantly longer than the corresponding bond in the neutral $CuNO_2$. The N-O bond lengths as well as the O-N-O bond angle of η^1 -N are calculated to be close to the corresponding geometry parameters of NO_2 .

3.2.3. Bonding. The analysis of the orbitals involved in the formation of the bond between Cu^+ and NO_2 in the η^1 -O trans isomer of Cu^+NO_2 (Figure S2 of the Supporting Information) reveals that the bonding between Cu^+ and NO_2 in Cu^+NO_2 is ionic, and it arises from the interaction of the ${}^1S(d^{10})$ state of Cu^+ and the 2A_1 ground state of NO_2 . The prevailing interaction between Cu^+ and NO_2 is the electrostatic interaction. The 4s orbital of Cu, which is empty in Cu^+ , interacts with the SOMO orbital (10a') of NO_2 to form the SOMO orbital (20a') of $CuNO_2$ which very strongly polarizes toward the NO_2 moiety. The 19a' and 16a' orbitals of Cu^+NO_2 arise from the antibonding and bonding combinations, respectively, between the $3d_{x^2-y^2}$ orbital of Cu and the 9a' orbital of the NO_2 moiety. The Cu $3d_{xz}$ and NO_2 2a'' orbitals interact to form the antibonding 6a'' and bonding 4a'' orbitals of Cu^+NO_2 . The remaining 3d orbitals of Cu do not significantly interact with the orbitals of NO_2 . The bonding in the other two isomers is very similar. The Mulliken populations calculated for all three isomers (see Table 7) predict the positive charge being located predominantly on the copper center.

The Mulliken populations of 6.08, 12.06, and 9.97 e in the s, p, and d orbitals, respectively, of Cu of η^1 -O trans show a back-donation of 0.11 e from NO_2 to Cu. The back-donation for η^1 -O cis is very close to that of η^1 -O trans. On the contrary, there is no back-donation for η^1 -N.

Table 7. Mulliken Populations in the s, p, and d Orbitals of Cu, N, and O of Cu⁺NO₂

isomer	atom	s	p	d	charge
η^1 -O trans	Cu	6.08	12.06	9.97	+0.88
η^1 -O trans	O ₁	3.83	4.40	0.03	-0.27
η^1 -O trans	N	3.50	2.78	0.25	+0.44
η^1 -O trans	O ₂	3.85	4.15	0.05	-0.05
η^1 -O cis	Cu	6.07	12.05	9.98	+0.90
η^1 -O cis	O ₁	3.82	4.39	0.01	-0.24
η^1 -O cis	N	3.50	2.75	0.27	+0.44
η^1 -O cis	O ₂	3.85	4.20	0.05	-0.10
η^1 -N	Cu	6.02	12.01	9.97	+1.00
η^1 -N	N	3.50	2.89	0.30	+0.26
η^1 -N	O	3.85	4.23	0.05	-0.13

3.2.4. Transition States. Two transition states were localized on the potential energy surface of Cu⁺NO₂. The first one is the C₁ monodentate η^1 -O species (²A), and the second one is the C_{2v} bidentate η^2 -O,O species (²A₁). The calculated imaginary frequencies reveal that the former transition state connects the η^1 -O trans and η^1 -O cis isomers, while the latter TS connects two η^1 -O cis isomers. All computational attempts to find a transition state connecting the η^1 -O trans and η^1 -N isomers led to a η^2 -O,N structure which is very close in energy and geometry to the η^1 -N isomer. We assume that the calculated structure is an artifact of symmetry breaking rather than a real transition state. None of the chosen computational method is able to correctly calculate the curvature of the Cu⁺NO₂ potential energy surface in the vicinity of the minimum corresponding to the η^1 -N isomer due to symmetry breaking. It should be noted that the η^2 -O,N structure is very close to that found by Sauer et al.¹⁶ at B3LYP.

The calculations also showed that all three bond distances of the η^1 -O TS are close to those of the C_s η^1 -O trans and cis isomers for all the methods used. The imaginary frequency corresponds to the torsion mode. The relative energies of the η^1 -O TS, which also correspond to the barrier of isomerization η^1 -O trans → η^1 -O cis, are 3–4 kcal/mol at all the levels employed.

Since the C_{2v} bidentate η^2 -O,O TS (²A₁) is an open shell species having two equivalent N–O bonds, there is a possibility of symmetry broken HF solutions for this species. The stability of the symmetry adapted UHF wave function, which is only very slightly spin contaminated ($\langle S^2 \rangle = 0.77$), was tested, and an UHF → UHF instability was found. When the orbital rotations corresponding to the instabilities were applied to the SCF eigenvectors and the SCF calculation was repeated with these rotated vectors as the starting guess, an UHF solution having essentially the same energy was found. The corresponding $\langle S^2 \rangle$ value is 0.82 indicating a low-spin contamination of the wave function without an UHF → UHF instability. Since there is no change in energy between the two UHF solutions, we assume that the energy of the symmetry adapted ROHF solution is the same as that of the symmetry broken ROHF solution. The question is whether the imaginary frequency of the η^2 -O,O species (²A₁) indicates that the species is a transition state or it is an artifact caused by symmetry breaking. We assume that the former is the

case since the imaginary frequency is significantly larger (e.g. 173 cm⁻¹ at CCSD(T)/Basis-II) than that of the η^1 -N species (²A₁). In addition, the imaginary frequency is not sensitive to the method used, basis set and step size employed in the numerical calculations. Therefore, the η^2 -O,O species is a transition state connecting two η^1 -O cis isomers since the imaginary frequency corresponds to the asymmetric Cu–O stretching mode. The corresponding barrier is calculated to be 7–10 kcal/mol.

3.2.5. Bond Dissociation Energies. In Table 8 we present the bond dissociation energies (D_e) of the η^1 -O trans isomer of Cu⁺NO₂ with respect to Cu⁺ and NO₂. The CC, QCI, and hybrid DFT values of D_e are 22–24 kcal/mol. The pure DFT schemes provide the values of D_e which are larger by 3–5 kcal/mol. The differences between the CCSD(T) and HF D_e values reveal the effect of electron correlation which is 9 kcal/mol.

3.3. Infrared Frequencies. The calculated infrared frequencies are revealed in Table 9 (selected species at CCSD(T)) and Tables S4 (all isomers of CuNO₂; all levels), S5 (NO₂ and NO₂⁻; all levels), and S6 (the η^1 -O trans and cis isomers of Cu⁺NO₂; all levels) of the Supporting Information.

3.3.1. IR Frequencies of the η^2 -O,O Isomer of CuNO₂. Let us discuss the infrared frequencies of the most stable C_{2v} η^2 -O,O isomer of CuNO₂. The wavenumber of the Cu–O asymmetric stretching mode is calculated to be around 210 cm⁻¹ at CCSD(T), 190 cm⁻¹ at QCISD(T), 120 cm⁻¹ at pure DFT, and 160–180 cm⁻¹ at hybrid DFT. These values are scattered over a wider range (120–210 cm⁻¹) as compared to the symmetric mode due to the discussed problems with symmetry breaking of the HF solution. On the other hand, the symmetric Cu–O stretching mode (similarly the other symmetric ones), which does not suffer from symmetry breaking, is calculated to lie in a narrow range 290–330 cm⁻¹ at all the levels used. The O,O out-of-plane mode is the same case, and thus the wavenumber values span a small interval 350–380 cm⁻¹. The remaining three modes are more interesting for experimentalists, since their wavenumbers lie in a region which is experimentally easily accessible. The wavenumber of O–N–O bending mode is calculated to be 865 and 875 cm⁻¹ at CCSD(T)/Basis-I and CCSD(T)/Basis-II, respectively. The QCISD(T) values are greater by some 20 cm⁻¹. The DFT infrared frequencies of the O–N–O bending mode are close to the ab initio ones. The pure DFT methods provided slightly smaller wavenumbers (855 cm⁻¹), while the hybrid DFT methods gave somewhat greater wavenumbers (890–910 cm⁻¹).

The asymmetric (as) and symmetric (ss) N–O stretching modes (1262 and 1287 cm⁻¹, respectively, at CCSD(T)/Basis-II) are much closer to those of NO₂⁻ (1273 cm⁻¹ (as) and 1303 cm⁻¹ (ss)) than to those of NO₂ (1345 cm⁻¹ (ss) and 1666 cm⁻¹ (as); all values at CCSD(T)/Basis-II). This fact indicates an ionic character of the η^2 -O,O isomer (Cu⁺NO₂⁻). It should be noted, that the N–O stretching modes calculated at DFT are not in agreement with the CCSD(T) values since the asymmetric stretching mode has a greater wavenumber than the asymmetric one by 5–50 cm⁻¹ depending on the functional. There is only one available experimental frequency (1220 cm⁻¹) of the stretching N–O

Table 8. Bond Dissociation Energies (in kcal/mol) of the η^1 -O Trans Isomer of Cu^+NO_2 with Respect to the $\text{Cu}^+ + \text{NO}_2$ Channel

basis set	CCSD	CCSD(T)	HF	QCISD	QCISD(T)	BPW91	PBE	PBE0	B3LYP
Basis-I	22.0	23.0	13.5	23.0	22.7	26.3	29.1	23.5	24.2
Basis-II	21.4	22.5	13.3	22.4	22.3	26.7	29.5	23.9	24.6

Table 9. Calculated CCSD(T) Infrared Frequencies (in cm^{-1})

species	symmetry	basis set	B ₂	A ₁	B ₁	A ₁	B ₂	A ₁
			Cu–O as	Cu–O ss	OO out	O–N–O b	N–O as	N–O ss
$\text{CuNO}_2 \eta^2\text{-O,O}$	C_{2v}	Basis-I	203.4	326.1	346.1	864.7	1216.9	1251.9
$\text{CuNO}_2 \eta^2\text{-O,O}$	C_{2v}	Basis-II	208.4	330.8	346.7	874.6	1262.0	1286.7
species	symmetry	basis set	A''	A'	A'	A'	A'	A'
			torsion	Cu–O–N b	Cu–O s	O–N–O b	O _{Cu} –N s	N–O s
$\text{CuNO}_2 \eta^1\text{-O}$	C_s	Basis-I	132.5	139.0	413.2	764.6	905.3	1560.2
$\text{CuNO}_2 \eta^1\text{-O}$	C_s	Basis-II	127.8	138.9	421.9	799.7	952.6	1584.4
species	symmetry	basis set	B ₂	A ₁	B ₁	A ₁	A ₁	B ₂
			Cu–N–O b	Cu–N s	OO out	O–N–O b	N–O ss	N–O as
$\text{CuNO}_2 \eta^1\text{-N}$	C_{2v}	Basis-I	129.8	325.3	375.9	806.1	1304.6	1412.0
$\text{CuNO}_2 \eta^1\text{-N}$	C_{2v}	Basis-II	144.6	329.5	378.3	817.1	1339.3	1457.5
species	symmetry	basis set	A ₁			A ₁	B ₂	
			O–N–O b			N–O ss	N–O as	
NO_2	C_{2v}	Basis-I	749.7			1316.9	1622.3	
NO_2	C_{2v}	Basis-II	758.8			1345.1	1665.5	
NO_2^-	C_{2v}	Basis-I	776.4			1267.4	1218.5	
NO_2^-	C_{2v}	Basis-II	787.9			1302.7	1273.1	
species	symmetry	basis set	A'	A''	A'	A'	A'	A'
			Cu–O–N b	torsion	Cu–O s	O–N–O b	O _{Cu} –N s	N–O s
$\text{Cu}^+\text{NO}_2 \eta^1\text{-O trans}$	C_s	Basis-I	135.5	121.2	269.6	811.4	1223.2	1768.0
$\text{Cu}^+\text{NO}_2 \eta^1\text{-O trans}$	C_s	Basis-II	121.8	124.7	268.5	801.7	1256.4	1755.5
$\text{Cu}^+\text{NO}_2 \eta^1\text{-O cis}$	C_s	Basis-I	96.0	214.5	291.9	745.3	1294.8	1674.6
$\text{Cu}^+\text{NO}_2 \eta^1\text{-O cis}$	C_s	Basis-II	98.2	214.4	295.3	749.9	1321.8	1711.7

mode of CuNO_2 , which was determined in Ar matrices⁷³ and assigned to the asymmetric stretching mode.

To further investigate the disagreement between the CCSD(T) and DFT frequencies of the N–O stretching modes of the $\eta^2\text{-O,O}$ isomer of CuNO_2 , we calculated the infrared frequencies of NaNO_2 (Table S7 of the Supporting Information) for which there are available experimental spectra⁷⁴ in solid Ar (1293 cm^{-1} ss, 1223 cm^{-1} as, and 826 cm^{-1} bending for the $\eta^2\text{-O,O}$ isomer of NaNO_2). We performed calculations on NaNO_2 , and the results reveal that the CCSD(T), QCISD(T), and all DFT methods reproduce the right order of the N–O stretching modes of NaNO_2 . Based on these results we firmly believe that most likely the symmetric N–O stretching mode of the $\eta^2\text{-O,O}$ isomer of CuNO_2 has a greater wavenumber than the asymmetric one as predicted by the CCSD(T) and QCISD(T) methods.

3.3.2. IR Frequencies of Cu^+NO_2 . Let us look briefly at the two most stable isomers of Cu^+NO_2 (Table 9 and Table S6 of the Supporting Information). The calculated wavenumbers of the Cu–O–N bending mode are 110–130 cm^{-1} and 80–110 cm^{-1} , for trans and cis, respectively. The wavenumber of the torsion mode is significantly lower for trans (120–150 cm^{-1}) than cis (210–270 cm^{-1}). The

wavenumber of the Cu–O stretching mode is about 300 cm^{-1} for both isomers. The three modes involving the NO_2 moiety lie in a region which is experimentally easily accessible. The O–N–O bending mode is calculated to have a greater wavenumber for trans (750–810 cm^{-1}) than cis (700–760 cm^{-1}). The wavenumbers of the O_{Cu}–N and N–O stretching modes are very scattered, and thus an eventual assignment of experimental bands will be difficult. However, all the methods indicate that the wavenumber of N–O stretching is significantly larger than that of O_{Cu}–N due to the electrostatic interaction between Cu^+ and O_{Cu}.

4. Conclusions

In this paper, we have presented a computational study of CuNO_2 and Cu^+NO_2 at the CCSD(T), QCISD(T), and DFT levels of approximation. Several stationary points (minima and transition states) were located on the CuNO_2 and Cu^+NO_2 potential energy surfaces. We investigated the performance of the two pure (BPW91 and PBE) as well as two hybrid (PBE0 and B3LYP) DFT methods with respect to the superior CCSD(T) method. The hybrid DFT methods are superior to the pure DFT and predict the geometries and relative stabilities which are close to the CCSD(T) results

for the most of the species. However, the PBE0 and B3LYP calculated relative energies of the η^1 -N isomer of CuNO₂ are smaller by 4–5 kcal/mol compared to the CCSD(T) value, and, moreover, both methods also predict the bond dissociation energies of CuNO₂ (for the Cu + NO₂ channel) which differ as much as 10 kcal/mol from the CCSD(T) values. The sizable differences between the CCSD(T) and QCISD(T) results were analyzed. We showed that the inferiority of the QCISD method itself with respect to CCSD is responsible for the failures not just the unsound estimation of the triple excitations (T). The issue of symmetry breaking was investigated, and it was demonstrated that in the case of symmetry breaking CCSD is the method of choice.

Acknowledgment. This work was financially supported by the project “Information Society” No. 1ET400400413 (“Development of a program environment for mathematic simulations and predictions in catalysis and electrocatalysis”). We would also like to thank Dr. Jiří Pittner for helpful discussions on quantum chemistry methods.

Supporting Information Available: Calculated bond lengths and angles of all calculated species of CuNO₂ (Table S1), of NO₂ and NO₂⁻ (Table S2), and of all calculated species of Cu⁺NO₂ (Table S3); schematic diagram of the most important orbitals involved in the formation of the η^2 -O,O isomer of CuNO₂ (Figure S1) and of the η^1 -O trans isomer of Cu⁺NO₂ (Figure S2); calculated infrared frequencies and intensities of CuNO₂ (Table S4), of NO₂ and NO₂⁻ (Table S5), and of Cu⁺NO₂ (Table S6); and calculated infrared frequencies of the η^2 -O,O isomer of NaNO₂ (Table S7). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Wichterlova, B.; Sobalik, Z.; Dedecek, J. *Appl. Catal., B* **2003**, *41*, 97–114.
- Shelef, M. *Chem. Rev.* **1995**, *95*, 209–225.
- Kuroda, Y.; Iwamoto, M. *Top. Catal.* **2004**, *28*, 111–118.
- Yahiro, H.; Iwamoto, M. *Appl. Catal., A* **2001**, *222*, 163–181.
- Iwamoto, M.; Yahiro, H.; Torikai, Y.; Yoshioka, T.; Mizuno, N. *Chem. Lett.* **1990**, 1967–1970.
- Kucherov, A. V.; Kucherova, T. N.; Slinkin, A. A. *Catal. Lett.* **1991**, *10*, 289–296.
- Petunchi, J. O.; Sill, G.; Hall, W. K. *Appl. Catal., B* **1993**, *2*, 303–321.
- Wichterlova, B.; Sobalik, Z.; Skokanek, M. *Appl. Catal. A* **1993**, *103*, 269–280.
- Iwamoto, M.; Yahiro, H.; Tanda, K.; Mizuno, N.; Mine, Y.; Kagawa, S. *J. Phys. Chem.* **1991**, *95*, 3727–3730.
- Li, Y. J.; Hall, W. K. *J. Catal.* **1991**, *129*, 202–215.
- Spoto, G.; Zecchina, A.; Bordiga, S.; Ricchiardi, G.; Martra, G.; Leofanti, G.; Petrini, G. *Appl. Catal., B* **1994**, *3*, 151–172.
- Wichterlova, B.; Dedecek, J.; Sobalik, Z.; Vondrova, A.; Klier, K. *J. Catal.* **1997**, *169*, 194–202.
- Hitchman, M. A.; Rowbottom, G. L. *Coord. Chem. Rev.* **1982**, *42*, 55–132.
- Ducere, J. M.; Goursot, A.; Berthomieu, D. *J. Phys. Chem. A* **2005**, *109*, 400–408.
- Rodriguez-Santiago, L.; Branchadell, V.; Sodupe, M. *J. Chem. Phys.* **1995**, *103*, 9738–9743.
- Rodriguez-Santiago, L.; Sierka, M.; Branchadell, V.; Sodupe, M.; Sauer, J. *J. Am. Chem. Soc.* **1998**, *120*, 1545–1551.
- Sierraalta, A.; Anez, R.; Brussin, M. R. *J. Phys. Chem. A* **2002**, *106*, 6851–6856.
- Sierraalta, A.; Anez, R.; Brussin, M. R. *J. Catal.* **2002**, *205*, 107–114.
- Bohme, M.; Frenking, G. *Chem. Phys. Lett.* **1994**, *224*, 195–199.
- Hrušák, J.; Tenno, S.; Iwata, S. *J. Chem. Phys.* **1997**, *106*, 7185–7192.
- Pouamerigo, R.; Merchan, M.; Nebotgil, I.; Widmark, P. O.; Roos, B. O. *Theor. Chim. Acta* **1995**, *92*, 149–181.
- Widmark, P. O.; Malmqvist, P. A.; Roos, B. O. *Theor. Chim. Acta* **1990**, *77*, 291–306.
- MOLPRO, a package of ab initio programs designed by H.-J. Werner and P. J. Knowles; version 2002.1; R. D. Amos; A. Bernhardsson; A. Berning; P. Celani; D. L. Cooper; M. J. O. Deegan; A. J. Dobyn; F. Eckert; C. Hampel; G. Hetzer; P. J. Knowles; T. Korona; R. Lindh; A. W. Lloyd; S. J. McNicholas; F. R. Manby; W. Meyer; M. E. Mura; A. Nicklass; P. Palmieri; R. Pitzer; G. Rauhut; M. Schütz; U. Schumann; H. Stoll; A. J. Stone; R. Tarroni, T. Thorsteinsson; H.-J. Werner.
- Extensible Computational Chemistry Environment Basis Set Database; Version 02/25/04; the Molecular Science Computing Facility; Environmental and Molecular Sciences Laboratory; the Pacific Northwest Laboratory; Richland, WA 99352.
- Hampel, C.; Peterson, K. A.; Werner, H. J. *Chem. Phys. Lett.* **1992**, *190*, 1–12.
- Scuseria, G. E.; Schaefer, H. F. *J. Chem. Phys.* **1989**, *90*, 3700–3703.
- Scuseria, G. E.; Janssen, C. L.; Schaefer, H. F. *J. Chem. Phys.* **1988**, *89*, 7382–7387.
- Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910–1918.
- Raghavachari, K. *J. Chem. Phys.* **1985**, *82*, 4607–4610.
- Pople, J. A.; Headgordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968–5975.
- Knowles, P. J.; Hampel, C.; Werner, H. J. *J. Chem. Phys.* **1993**, *99*, 5219–5227.
- Watts, J. D.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 8718–8733.
- Gaussian 03; Revision C.02*; M. J. Frisch; G. W. Trucks; H. B. Schlegel; G. E. Scuseria; M. A. Robb; J. R. Cheeseman; J. A. Montgomery, J.; T. Vreven; K. N. Kudin; J. C. Burant; J. M. Millam; S. S. Iyengar; J. Tomasi; V. Barone; B. Mennucci; M. Cossi; G. Scalmani; N. Rega; G. A. Petersson; H. Nakatsuji; M. Hada; M. Ehara; K. Toyota; R. Fukuda; J. Hasegawa; M. Ishida; T. Nakajima; Y. Honda; O. Kitao; H. Nakai; M. Klene; X. Li; J. E. Knox; H. P. Hratchian; J. B. Cross; V. Bakken; C. Adamo; J. Jaramillo; R. Gomperts; R. E. Stratmann; O. Yazyev; A. J. Austin; R. Cammi; C. Pomelli; J. W. Ochterski; P. Y. Ayala; K. Morokuma; G. A. Voth; P. Salvador; J. J. Dannenberg; V. G. Zakrzewski; S. Dapprich; A. D. Daniels; M. C. Strain;

- O. Farkas; D. K. Malick; A. D. Rabuck; K. Raghavachari; J. B. Foresman; J. V. Ortiz; Q. Cui; A. G. Baboul; S. Clifford; J. Cioslowski; B. B. Stefanov; G. Liu; A. Liashenko; P. Piskorz; I. Komaromi; R. L. Martin; D. J. Fox; T. Keith; M. A. Al-Laham; C. Y. Peng; A. Nanayakkara; M. Challacombe; P. M. W. Gill; B. Johnson; W. Chen; M. W. Wong; C. Gonzalez; J. A. Pople. Gaussian, Inc.: Wallingford, CT, 2004.
- (34) Antusek, A.; Urban, M.; Sadlej, A. J. *J. Chem. Phys.* **2003**, *119*, 7247–7262.
- (35) Urban, M.; Sadlej, A. J. *J. Chem. Phys.* **2000**, *112*, 5–8.
- (36) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 6671–6687.
- (37) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (38) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (39) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (40) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200–206.
- (41) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785–789.
- (42) Bartlett, R. J. ACESII is a program product of the Quantum Theory Project of University of Florida.
- (43) Bartlett, R. J.; Watts, J. D.; Kucharski, S. A.; Noga, J. *Chem. Phys. Lett.* **1990**, *165*, 513–522.
- (44) Paldus, J.; Cizek, J.; Jeziorski, B. *J. Chem. Phys.* **1989**, *90*, 4356–4362.
- (45) Paldus, J.; Cizek, J.; Jeziorski, B. *J. Chem. Phys.* **1990**, *93*, 1485–1486.
- (46) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270–283.
- (47) Luna, A.; Alcamí, M.; Mo, O.; Yanez, M. *Chem. Phys. Lett.* **2000**, *320*, 129–138.
- (48) Lynch, B. J.; Truhlar, D. G. *Chem. Phys. Lett.* **2002**, *361*, 251–258.
- (49) Hrušák, J.; Iwata, S. *J. Chem. Phys.* **1997**, *106*, 4877–4888.
- (50) Dunietz, B. D.; Head-Gordon, M. *J. Phys. Chem. A* **2003**, *107*, 9160–9167.
- (51) Einfeld, W.; Morokuma, K. *J. Chem. Phys.* **2000**, *113*, 5587–5597.
- (52) Werner, H. J.; Knowles, P. J. *J. Chem. Phys.* **1985**, *82*, 5053–5063.
- (53) Knowles, P. J.; Werner, H. J. *Chem. Phys. Lett.* **1985**, *115*, 259–267.
- (54) Feller, D. F.; Schmidt, M. W.; Ruedenberg, K. *J. Am. Chem. Soc.* **1982**, *104*, 960–967.
- (55) Siegbahn, P. E. M.; Almlof, J.; Heiberg, A.; Roos, B. O. *J. Chem. Phys.* **1981**, *74*, 2384–2396.
- (56) Roos, B. O.; Taylor, P. R. *Chem. Phys.* **1980**, *48*, 157–173.
- (57) Lam, B.; Schmidt, M. W.; Ruedenberg, K. *J. Phys. Chem.* **1985**, *89*, 2221–2235.
- (58) Ruedenberg, K.; Schmidt, M. W.; Gilbert, M. M.; Elbert, S. T. *Chem. Phys.* **1982**, *71*, 41–49.
- (59) Ruedenberg, K.; Schmidt, M. W.; Gilbert, M. M. *Chem. Phys.* **1982**, *71*, 51–64.
- (60) Ruedenberg, K.; Schmidt, M. W.; Gilbert, M. M.; Elbert, S. T. *Chem. Phys.* **1982**, *71*, 65–78.
- (61) Werner, H. J.; Knowles, P. J. *J. Chem. Phys.* **1988**, *89*, 5803–5814.
- (62) Knowles, P. J.; Werner, H. J. *Chem. Phys. Lett.* **1988**, *145*, 514–522.
- (63) Davidson, E. R.; Borden, W. T. *J. Phys. Chem.* **1983**, *87*, 4783–4790.
- (64) Engelbrecht, L.; Liu, B. *J. Chem. Phys.* **1983**, *78*, 3097–3106.
- (65) Feller, D.; Huyser, E. S.; Borden, W. T.; Davidson, E. R. *J. Am. Chem. Soc.* **1983**, *105*, 1459–1466.
- (66) McLean, A. D.; Lengsfeld, B. H.; Pacansky, J.; Ellinger, Y. *J. Chem. Phys.* **1985**, *83*, 3567–3576.
- (67) Langhoff, S. R.; Davidson, E. R. *Int. J. Quantum Chem.* **1974**, *8*, 61.
- (68) Sherrill, C. D.; Lee, M. S.; Head-Gordon, M. *Chem. Phys. Lett.* **1999**, *302*, 425–430.
- (69) Harju, A.; Rasanen, E.; Saarikoski, H.; Puska, M. J.; Nieminen, R. M.; Niemela, K. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2004**, *69*.
- (70) Corral, I.; Mo, O.; Yanez, M. *Int. J. Quantum Chem.* **2003**, *91*, 438–445.
- (71) Cohen, R. D.; Sherrill, C. D. *J. Chem. Phys.* **2001**, *114*, 8257–8269.
- (72) Russ, N. J.; Crawford, T. D.; Tschumper, G. S. *J. Chem. Phys.* **2004**, *120*, 7298–7306.
- (73) Worden, D.; Ball, D. W. *J. Phys. Chem.* **1992**, *96*, 7167–7169.
- (74) Lo, W. J.; Shen, M. Y.; Yu, C. H.; Lee, Y. P. *J. Chem. Phys.* **1996**, *104*, 935–941.

CT0502007

JCTC

Journal of Chemical Theory and Computation

Assessment of Model Chemistries for Noncovalent Interactions

Yan Zhao and Donald G. Truhlar*

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

Received February 9, 2006

Abstract: In the present study, we report tests of 57 model chemistry methods for calculating binding energies of 31 diverse van der Waals molecules arranged in five databases of noncovalent interaction energies. The model chemistries studied include wave function theory (WFT), density functional theory (DFT), and combined wave function-density-functional-theory (CWFDFDFT), and they include methods whose computational effort scales (for large systems) as N^7 , N^6 , N^5 , and N^4 , where N is the number of atoms. The model chemistries include 2 CWFDFDFT N^7 models, 4 multilevel WFT N^7 models, 5 single-level WFT N^7 models, 4 CWFDFDFT N^6 models, 3 multilevel WFT N^6 models, 11 single-level WFT N^6 models, 5 CWFDFDFT N^5 models, 10 single-level WFT N^5 models, 4 multilevel WFT N^5 models, 4 single-level DFT N^4 models, and 5 single-level WFT N^4 models. We draw the following conclusions based on the mean absolute errors in 31 noncovalent binding energies: (1) MCG3-MPW gives the best performance for predicting the binding energies of these noncovalent complexes. (2) MCQCISD-MPWB and MCQCISD-MPW are the best two N^6 methods. (3) M05-2X is the best single-level method for these noncovalent complexes. These four methods should facilitate useful calculations on a wide variety of practical applications involving hydrogen bonding, charge-transfer complexes, dipole interactions, weak (dispersion-like) interactions, and $\pi\cdots\pi$ stacking. If a user is interested in only a particular type of noncovalent interactions, though, some other methods, may be recommended for especially favorable performance/cost ratios. For example, BMC-CCSD has an outstanding performance for hydrogen bonding, and PWB6K has an outstanding cost-adjusted performance for dipole interaction calculations on very large systems. We also show that M05-2X performs well for interactions of amino acid pair residues.

1. Introduction

Noncovalent interactions play very important roles in many areas of science such as molecular recognition, protein folding, stacking of nucleobases, crystal packing, vapor–liquid condensation, polymer packing, soft materials design, self-assembly, supramolecular chemistry, solvation, and molecular scattering. It is especially noteworthy that noncovalent interactions underlie many complex biological functions including cell–cell recognition, intracellular signaling, and the regulation of gene expression. Understanding

various noncovalent interactions is a key to unraveling the mysteries of cellular function in health and disease and to developing new drugs as well as being a critical component in nanotechnological uses of soft materials.

Model chemistry is “an approximate but well-defined mathematical procedure of simulation”¹ of chemical phenomena. As pointed out by Pople,¹ there is a wide range of possible empiricism; model chemistry can even be *ab initio* (i.e., without parameters except for fundamental constants of physics). Several multilevel model chemistry methods, such as the Gaussian- n theories and their variants developed by Pople and co-workers,^{2–5} the related Weizmann- n methods,^{6,7}

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

the complete basis set (CBS) family of methods by Petersson and co-workers,^{8–10} the single-coefficient^{11–15} and multi-coefficient^{14–24} correlation methods (MCCMs) of our group, and the recent multilevel methods of Hu and co-workers,^{25,26} have been developed and validated for covalent interactions as required for application to thermochemistry (heats of formation, atomization energies, etc.) and kinetics (barrier heights). Although some research^{27–30} has employed these multilevel methods for hydrogen bonded clusters and ionic clusters, until now there has been only one systematic validation of multilevel methods for noncovalent interactions, and that study was limited to rare gas interactions.³¹ The lack of broader validation studies is partly due to the lack, until recently,^{32–34} of standard databases (analogous to the G3 database,^{5,35–37} Database/3,¹⁵ or a recent metal–ligand bond energy database³⁸) for nonbonded interactions. In a recent communication,³⁹ we compared several multilevel methods for the calculation of the stacking interaction energies in benzene dimers, and we found that the empirical hybrid of density functional theory (DFT) and wave function theory (WFT), also called multicoefficient extrapolated density functional theory,^{22,23} give the best results for benzene dimers. A key objective of the present article is to assess multilevel model chemistry methods against several recently developed databases^{32,33} for nonbonded interactions. We also present the results for several single-level methods for comparison. Both DFT and WFT are considered.

Section 2 describes the theories and databases used in the present work. Section 3 presents results and discussion, and section 4 has concluding remarks. The Appendix considers interaction energies of amino acid residues.

2. Theory and Databases

2.1. Theory. The levels of electron correlation used in the present paper include Møller–Plesset second-, third-, and fourth-order perturbation theory (MP2,⁴⁰ MP3,⁴¹ MP4⁴¹), Møller–Plesset fourth-order perturbation theory without singles and triples contributions⁴¹ (MP4DQ), Møller–Plesset fourth-order perturbation theory without triples contributions⁴¹ (MP4(SDQ)), quadratic configuration interaction with single and double excitations⁴² (QCISD), QCISD with quasiperturbative connected triples⁴² (QCISD(T)), coupled cluster with single and double excitations (CCSD),⁴³ and CCSD with quasiperturbative connected triples (CCSD(T)).⁴⁴ In general, core orbitals are doubly occupied in all configurations except for some MP2 calculations, and those are denoted MP2(full). We also present results for two hybrid meta-DFT methods, PWB6K³³ and M05-2X.⁴⁵ We note that PWB6K was found in a previous study³³ to perform best out of 25 density functionals tested against the databases employed here, and readers interested in the performance of other density functionals are referred to that study. For example, PWB6K was found to have an error three times lower than the popular B3LYP⁴⁶ functional. PWB6K was also found to have excellent performance for hydrogen bonds to π acceptor,⁴⁷ a type of interaction not present in the database. The M05-2X functional was not available (it had not yet been developed) at the time of those assessments,

but was subsequently shown to be very accurate,³⁴ and so it is included here.

Multicoefficient extrapolated DFT methods^{22,23} include both DFT and WFT components in the same calculation.²³ These calculations may be labeled as combined wave function density functional methods (abbreviated CWFDFT or WFT/DFT) or as fifth-rung methods on Jacob's ladder^{48,49} of density functionals. We compare the results obtained by multicoefficient extrapolated DFT methods (MC3BB,²² MC3MPW,²² MC3MPWB,²³ MCCO-MPW, and MPWB,²³ MCUT-MPW and MPWB,²³ MCQCISD-MPW and -MPWB,²³ and MCG3-MPW and -MPWB²³) to those obtained by pure-WFT-based multilevel methods, in particular, G3SX,⁵ CBS-QB3,⁹ G3SX(MP3),⁵ MCCM/3,¹⁵ and BMC-CCSD.²⁴ Within the MCCM/3 suite, we specifically consider MCG3/3, MC-QCISD/3, and MC-UT/3. We note that G3SX(MP3), MCG3/3, and MC-QCISD/3 were selected as particularly efficient methods in a previous systematic study of multilevel methods for thermochemistry.⁵⁰ Since then, though, the multicoefficient extrapolated DFT methods^{22,23} have been developed, and they show an even better performance. We note that the recently developed multilevel method BMC-CCSD²⁴ has similar cost to MC-QCISD/3 but improved performance for atomization energies, barrier heights, ionization potentials, and electron affinities. Thus it will be interesting to test BMC-CCSD for noncovalent interactions. BMC-CCSD and MC-QCISD/3 are considerably less expensive than G3SX and G3SX(MP3).

We also consider some examples of the older scaling-all-correlation (SAC) methods,^{12,14,15} which are single-coefficient correlation methods. Although previous tests of these methods for thermochemistry have shown worthwhile improvement over MP2 at essentially no additional cost, they are not as powerful as MCCMs, and the present tests will show if they are useful for noncovalent interactions. The three scaling-all-correlation (SAC) methods tested in this study use version-3s scaling coefficients.¹⁵

Note that the zero-point corrections were excluded from the G3SX, G3SX(MP3), and CBS-QB3 calculations (and all other methods) in this article since, in the standard spectroscopic notation, we are interested in predicting D_e ,⁵¹ not D_0 .

2.2. Noncovalent Interaction Databases. We tested all 57 considered methods (22 multilevel methods and 35 single-level methods) against five recently developed databases, in particular, HB6/04,³² CT7/04,³² DI6/04,³² WI7/05,³³ and PPS5/05,³³ 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 (NH₃)₂, (HF)₂, (H₂O)₂, NH₃⋯H₂O, (HCONH₂)₂, and (HCOOH)₂. The CT7/04 database consists of the binding energies of seven charge-transfer complexes, in particular C₂H₄⋯F₂, NH₃⋯F₂, C₂H₂⋯ClF, HCN⋯ClF, NH₃⋯Cl₂, H₂O⋯ClF, and NH₃⋯ClF. The DI6/04 database contains the binding energies of six dipole interaction complexes: (H₂S)₂, (HCl)₂, HCl⋯H₂S, CH₃Cl⋯HCl, CH₃SH⋯HCN, and CH₃SH⋯HCl. The WI7/05 database consists of the binding energies of seven weak interaction complexes, namely HeNe, HeAr, Ne₂, NeAr, CH₄⋯Ne, C₆H₆⋯Ne, and (CH₄)₂, all of which are bound by dispersion interactions.

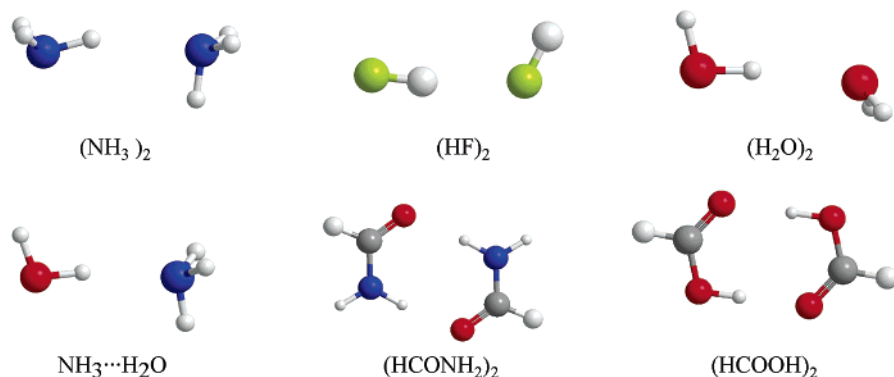


Figure 1. Geometries of the dimers in the HB6/04 database.

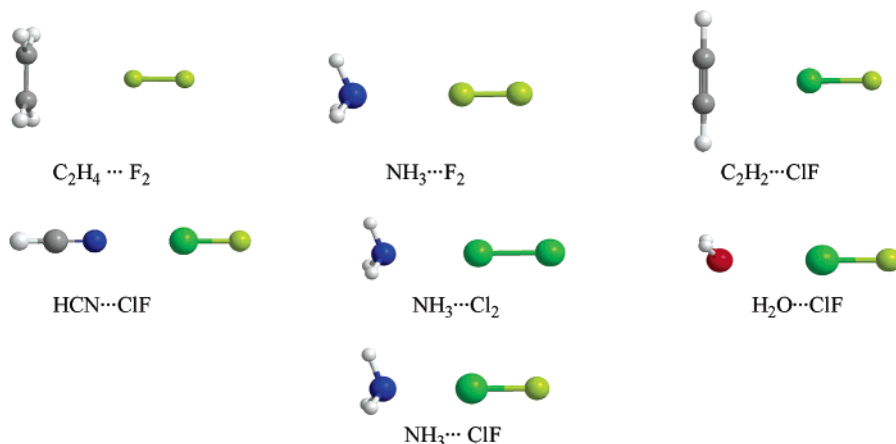


Figure 2. Geometries of the complexes in the CT7/04 database.

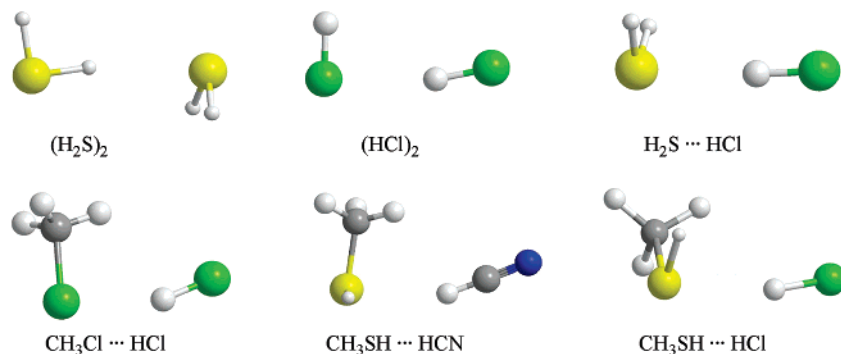


Figure 3. Geometries of the dimers in the DI6/04 database.

The PPS5/05 database consists of binding energies of five π - π stacking complexes, namely (C₂H₂)₂, (C₂H₄)₂, sandwich (C₆H₆)₂, T-shaped (C₆H₆)₂, and parallel-displaced (C₆H₆)₂.

Figures 1–5 depict the geometries of the noncovalent complexes in the present study.

2.3. Computer Programs, Geometries, Basis Sets, Counterpoise Correction, and Full Models. All the calculations in the present study are performed by using the locally developed program *MLGAUSS*⁵² in conjunction with *Gaussian03*.⁵³ The *MLGAUSS* program is available from the Truhlar group's software Web page.⁵⁴

The geometries for the benzene dimers are taken from Sinnokrot and Sherrill.⁵⁵ The geometries of all other complexes are optimized at the MC-QCISD/3¹⁵ level of theory. Note that these same geometries are used for all methods tested. For methods, namely G3SX(MP3), CBS-QB3, and

G3SX, that are ordinarily *defined* to use other geometries, we added the suffix “//Q” to denote this choice of geometries, which is used for *all* methods in this article.

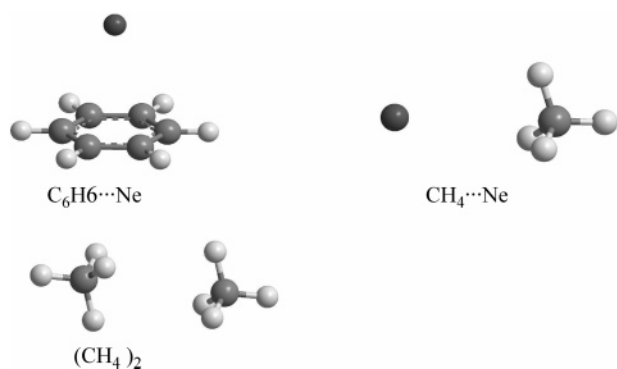
The basis sets used include the 6-31G(d),⁴¹ 6-31+G(d,p),⁴¹ 6-31G(2df,p),⁴¹ 6-31B(d),²⁴ G3Large,³ G3XLarge,⁵ modified Gaussian-3¹⁴ (MG3), and modified Gaussian-3 semidiffuse⁵⁶ (MG3S) basis sets. We note that the MG3 basis¹⁴ is also denoted G3LargeMP2.⁵⁷

For most of the tested methods, we perform calculations without the counterpoise corrections (CP)^{58,59} for basis set superposition error (BSSE). We do present, however, the CP-corrected results for the MP2/MG3S, M05-2X/MG3S, and PWB6K/MG3S levels of theory.

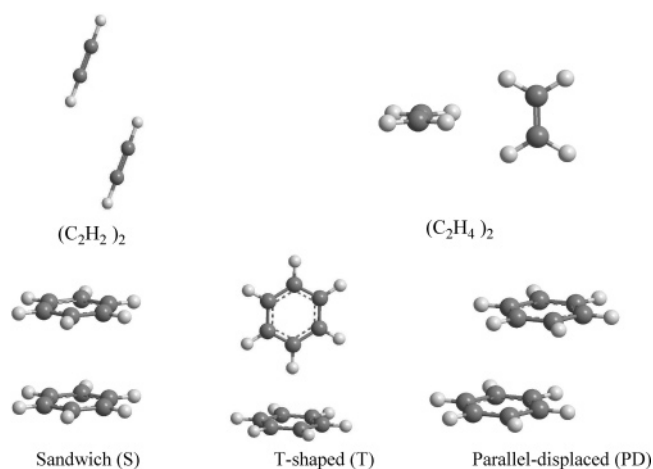
A comment on the distribution between all possible models and a “full theoretical model chemistry” is in order here. When special procedures for particular molecules or par-

Table 1. Binding Energies (kcal/mol) and Mean Errors (kcal/mol) for the HB6/04 Database by Multilevel Methods

	(NH ₃) ₂	(HF) ₂	(H ₂ O) ₂	NH ₃ ⋯H ₂ O	(HCONH ₂) ₂	(HCOOH) ₂	MSE	MUE
best estimate	3.16	4.57	4.97	6.41	14.94	16.15		
<i>N</i> ⁷ Methods								
G3SX(MP3)//Q	3.15	4.57	5.10	6.39	14.67	15.91	-0.07	0.11
MCG3	3.14	4.41	5.06	6.36	14.69	15.75	-0.13	0.16
CBS-QB3//Q	3.26	4.80	5.18	6.61	15.02	16.35	0.17	0.17
G3SX//Q	3.23	4.84	5.26	6.52	15.10	16.52	0.21	0.21
MCG3-MPW	2.91	4.48	4.86	6.09	14.39	15.72	-0.29	0.29
MCG3-MPWB	2.86	4.44	4.83	6.06	14.07	15.49	-0.41	0.41
<i>N</i> ⁶ Methods								
BMC-CCSD	3.25	4.87	5.22	6.43	14.94	16.08	0.10	0.12
MC-QCISD/3	3.16	4.48	5.07	6.34	14.25	15.22	-0.28	0.32
MCUT/3	3.22	4.60	5.14	6.41	14.20	15.12	-0.25	0.34
MCQCISD-MPW	2.87	4.45	4.83	6.07	13.98	15.35	-0.44	0.44
MCQCISD-MPWB	2.86	4.38	4.84	6.06	13.65	15.09	-0.55	0.55
MCUT-MPW	2.79	4.42	4.72	5.99	13.64	15.06	-0.59	0.59
MCUT-MPWB	2.80	4.41	4.79	6.02	13.37	14.95	-0.64	0.64
<i>N</i> ⁵ Methods								
SAC-MP2/MG3S	3.38	4.93	5.51	6.85	15.01	16.39	0.31	0.31
MC3MPW	3.26	4.76	5.52	7.04	14.17	16.21	0.13	0.39
MCCO/3	3.17	4.20	5.01	6.71	13.85	15.33	-0.32	0.44
MC3MPWB	3.20	4.80	5.45	6.82	13.80	15.73	-0.07	0.46
MCCO-MPW	2.76	4.17	4.82	6.24	13.45	15.62	-0.52	0.52
MC3BB	2.77	4.38	4.97	6.32	13.05	15.00	-0.62	0.62
MCCO-MPWB	2.69	4.24	4.77	6.06	13.03	15.19	-0.70	0.70
SAC-MP2/6-31+G(d,p)	4.28	5.02	6.61	8.25	15.00	15.97	0.82	0.88
SAC-MP2/6-31G(d)	4.83	7.76	7.54	8.93	18.85	19.28	2.83	2.83
average							-0.06	0.52

**Figure 4.** Geometries of selected dimers in the WI7/05 database.

particular symmetries are avoided, and a model is general and continuous, the model may be called a full theoretical model chemistry.¹ For this reason,^{17,18} we prefer SAC-,^{12,13} MCG3-,¹⁸ G3S-,⁴ and G3SX-type⁵ methods to G2-,² G3-,³ and G3X-type⁵ methods because the G2-, G3-, and G3X-type methods involve a discontinuous high-level correction, whereas scaling methods do not. Similarly, the use of CP corrections disqualifies a method as “full” because, for example, one needs special decisions such as whether to apply it only to van der Waals molecules but not (for example) to the O–H bond in water. Similarly, should Ne₂Be be treated as a complex of Ne₂ with Be or NeBe with Ne? Should NH₄Cl be treated as NH₃ complexed to HCl or NH₄⁺ complexed to Cl⁻? Finally, it is essentially impossible to apply CP corrections to amorphous solids and many other cases.

**Figure 5.** Geometries of the dimers in the PPS/05 database.

Nevertheless, CP corrections are often used for calculating van der Waals binding energies, so we do consider some non-“full” models employing CP corrections in this paper.

3. Results and Discussion

Results are given in Tables 1–3 and S1–S8, where tables with an S prefix are found in the Supporting Information. In particular, the binding energies and mean errors of the tested multilevel methods are listed in Tables 1, S1, S3, S5, S7, and 3, and results for the tested single-level methods are presented in Tables 2, S2, S4, S6, S8, S10, and 3. In these tables we classify the methods according to their scaling

Table 2. Binding Energies (kcal/mol) and Mean Errors (kcal/mol) for the HB6/04 Database by Single-Level Methods

method	(NH ₃) ₂	(HF) ₂	(H ₂ O) ₂	NH ₃ ...H ₂ O	(HCONH ₂) ₂	(HCOOH) ₂	MSE	MUE
best estimate	3.16	4.57	4.97	6.41	14.94	16.15		
<i>N</i> ⁷ Methods								
CCSD(T)/6-311+G(d,p)	3.66	4.66	5.91	7.17	13.48	14.51	-0.13	0.90
MP4/6-31G+(d)	4.25	5.51	6.90	8.31	14.89	15.80	0.91	1.05
MP4/6-31G(2df,p)	3.87	7.12	6.44	6.99	17.65	18.98	1.81	1.81
QCISD(T)/6-31G(d)	4.28	7.10	6.81	8.00	17.30	17.64	1.82	1.82
MP4/6-31G(d)	4.31	7.19	6.88	8.05	17.40	17.86	1.92	1.92
<i>N</i> ⁶ Methods								
MP3/6-311+G(d,p)	3.54	4.61	5.76	6.98	13.23	14.30	-0.30	0.89
MP3/6-31G+(d)	4.14	5.40	6.73	8.12	14.73	15.62	0.76	1.01
CCSD/6-311+G(d,p)	3.43	4.54	5.67	6.85	12.69	13.70	-0.55	1.03
MP4(SDQ)/6-311+G(d,p)	3.45	4.53	5.70	6.89	12.56	13.56	-0.58	1.09
MP4(DQ)/6-31B(d)	4.50	6.10	6.41	8.08	15.46	15.41	0.96	1.21
CCSD/6-31B(d)	4.51	6.14	6.43	8.06	15.82	15.77	1.09	1.21
MP4(SDQ)/6-31G(2df,p)	3.64	6.87	6.15	6.67	16.45	17.72	1.22	1.22
MP3/6-31G(2df,p)	3.74	6.83	6.20	6.82	16.85	18.09	1.39	1.39
QCISD/6-31G(d)	4.11	6.99	6.65	7.77	16.51	16.91	1.46	1.46
MP4(SDQ)/6-31G(d)	4.14	7.02	6.69	7.82	16.45	16.87	1.47	1.47
MP3/6-31G(d)	4.21	6.94	6.71	7.96	16.72	17.15	1.58	1.58
<i>N</i> ⁵ Methods								
MP2/MG3	3.32	4.92	5.43	6.74	14.87	16.29	0.23	0.25
MP2/MG3S	3.33	4.91	5.46	6.78	14.87	16.27	0.24	0.26
MP2(full)/G3Large	3.34	4.99	5.48	6.80	14.97	16.46	0.31	0.31
MP2/6-31+G(d,p)	4.02	4.91	6.37	7.89	14.40	15.51	0.48	0.88
MP2/MG3S-CP	2.87	3.94	4.53	5.96	13.22	14.10	-0.93	0.93
MP2/6-311+G(d,p)	3.73	4.66	5.99	7.35	13.31	14.45	-0.12	0.99
MP2/6-31G+(d)	4.27	5.51	7.00	8.49	14.87	15.96	0.98	1.07
MP2/6-31B(d)	4.73	6.35	6.78	8.59	16.57	16.89	1.62	1.62
MP2/6-31G(2df,p)	4.01	7.22	6.61	7.25	17.76	19.19	1.97	1.97
MP2/6-31G(d)	4.40	7.34	7.08	8.35	17.48	18.19	2.11	2.11
<i>N</i> ⁴ Methods								
M05-2X/MG3S-CP	3.01	4.78	5.13	6.40	14.33	16.22	-0.05	0.20
PWB6K/MG3S-CP	3.05	4.78	5.09	6.40	13.75	15.72	-0.23	0.34
M05-2X/MG3S	3.19	5.17	5.53	6.71	14.71	16.81	0.32	0.40
PWB6K/MG3S	3.23	5.19	5.51	6.73	14.13	16.43	0.17	0.44
HF/6-31G(d)	3.00	5.95	5.59	6.43	12.93	14.61	-0.28	0.95
HF/6-311+G(d,p)	2.25	4.17	4.75	5.42	10.41	12.17	-1.84	1.84
HF/MG3S	1.71	3.99	3.97	4.54	10.22	12.30	-2.24	2.24
HF/MG3	1.71	3.99	3.97	4.54	10.22	12.30	-2.24	2.24
HF/G3Large	1.71	3.99	3.96	4.54	10.21	12.29	-2.25	2.25
average							0.37	1.21

properties⁶⁰ (*N*⁷, *N*⁶, *N*⁵, or *N*⁴), where *N* is the number of atoms. The tables show the mean unsigned error (MUE, also called mean absolute deviation) and mean signed error (MSE). We use "CP" to denote calculations that do include the counterpoise correction for the BSSE.

To put the large number of results in this paper into perspective, we define an overall error quantity in Table 3, namely the mean MUE:

$$\text{MMUE} = [\text{MUE}(\text{HB}) + \text{MUE}(\text{CT}) + \text{MUE}(\text{DI}) + \text{MUE}(\text{WI}) + \text{MUE}(\text{PPS})]/5 \quad (1)$$

Our discussion will focus *mainly* on the highly averaged MUEs and MMUEs because they provide measures of the broad usefulness of the methods tested for various kinds of noncovalent interactions. The tables are arranged in such a

way that users interested in one or another subsets of the results may make their own comparisons and draw their own conclusions.

3.1. Hydrogen Bonding. Tables 1 and 2 summarize the results for hydrogen bonding calculations. Among the tested multilevel methods, G3SX(MP3) gives the lowest MUE for binding energies in the HB6/04 database. BMC-CCSD is the best *N*⁶ multilevel method, and it has an MUE only 10% larger than G3SX(MP3) with a cost more than five times lower, as well as having better scaling to large systems. MC3MPW is the best *N*⁵ multilevel method, with an MUE much larger than BMC-CCSD and a cost only slightly smaller, but better scaling. Table 2 shows that the best of the tested single-level methods for hydrogen bonding calculations are M05-2X/MG3S and MP2/MG3. MP2/MG3S

Table 3. Overall Results

method	type ^a	HB6/05 MUE	CT7/05 MUE	DI6/05 MUE	WI7/05 MUE	PPS5/05 MUE	MMUE	cost ^b
<i>N</i> ⁷ Methods								
MCG3-MPW	ML DFT/WFT	0.29	0.13	0.15	0.06	0.18	0.16	110
MCG3-MPWB	ML DFT/WFT	0.41	0.13	0.20	0.05	0.17	0.19	111
MCG3	ML WFT	0.16	0.19	0.29	0.05	0.80	0.30	104
G3SX(MP3)//Q	ML WFT	0.11	0.18	0.35	0.08	0.80	0.31	135
CBS-QB3//Q	ML WFT	0.17	0.38	0.36	0.11	0.57	0.32	204
G3SX//Q	ML WFT	0.21	0.26	0.39	0.11	0.84	0.36	1116
CCSD(T)/6-311+G(d,p)	SL WFT	0.90	0.62	0.47	0.07	0.95	0.60	409
MP4/6-31G(2df,p)	SL WFT	1.81	0.71	0.31	0.25	0.58	0.73	848
MP4/6-31G+(d)	SL WFT	1.05	0.96	0.50	0.14	1.06	0.74	150
MP4/6-31G(d)	SL WFT	1.92	0.75	0.69	0.10	0.28	0.75	61
QCISD(T)/6-31G(d)	SL WFT	1.82	0.83	0.74	0.10	0.46	0.79	71
<i>N</i> ⁶ Methods								
MCQCISD-MPWB	ML DFT/WFT	0.55	0.12	0.19	0.05	0.18	0.22	29
MCQCISD-MPW	ML DFT/WFT	0.44	0.12	0.18	0.07	0.33	0.23	29
MCUT-MPWB	ML DFT/WFT	0.64	0.14	0.25	0.06	0.52	0.32	22
MCUT/3	ML WFT	0.34	0.25	0.35	0.07	0.74	0.35	16
MCUT-MPW	ML DFT/WFT	0.59	0.18	0.22	0.09	0.75	0.37	22
MC-QCISD/3	ML WFT	0.32	0.23	0.37	0.07	0.85	0.37	23
BMC-CCSD	ML WFT	0.12	0.31	0.58	0.14	1.17	0.46	26
MP3/6-31G+(d)	SL WFT	1.01	0.86	0.56	0.12	0.47	0.60	5
MP3/6-311+G(d,p)	SL WFT	0.89	1.00	0.55	0.07	0.55	0.61	14
MP4(SDQ)/6-31G(2df,p)	SL WFT	1.22	0.80	0.66	0.24	0.24	0.63	18
MP4(SDQ)/6-311+G(d,p)	SL WFT	1.09	0.84	0.70	0.07	0.47	0.63	15
MP3/6-31G(2df,p)	SL WFT	1.39	0.85	0.45	0.23	0.26	0.63	17
CCSD/6-311+G(d,p)	SL WFT	1.03	0.99	0.74	0.07	0.46	0.66	68
MP3/6-31G(d)	SL WFT	1.58	0.90	0.73	0.09	0.69	0.80	2
MP4(SDQ)/6-31G(d)	SL WFT	1.47	0.85	0.87	0.10	0.76	0.81	1
QCISD/6-31G(d)	SL WFT	1.46	0.89	0.93	0.10	0.81	0.84	8
MP4(DQ)/6-31B(d)	SL WFT	1.21	1.99	0.86	0.14	0.69	0.98	1
CCSD/6-31B(d)	SL WFT	1.21	2.19	0.86	0.13	0.73	1.03	8
<i>N</i> ⁵ Methods								
MC3MPWB	ML DFT/WFT	0.46	0.46	0.38	0.07	0.48	0.37	7
MCCO-MPWB	ML DFT/WFT	0.70	0.27	0.33	0.07	0.69	0.41	21
MP2/MG3S-CP	SL WFT	0.93	0.26	0.25	0.18	0.48	0.42	22
MCCO-MPW	ML DFT/WFT	0.52	0.36	0.31	0.12	0.81	0.43	20
MC3MPW	ML DFT/WFT	0.39	0.52	0.36	0.13	0.93	0.47	6
MP2/MG3S	SL WFT	0.26	0.73	0.45	0.07	1.24	0.55	14
MP2/MG3	SL WFT	0.25	0.72	0.44	0.09	1.32	0.56	15
MC3BB	ML DFT/WFT	0.62	0.32	0.75	0.27	1.10	0.61	7
SAC-MP2/MG3S	ML WFT	0.31	0.86	0.53	0.08	1.38	0.63	14
MP2(full)/G3Large	SL WFT	0.31	0.80	0.57	0.10	1.40	0.63	36
MP2/6-31+G(d,p)	SL WFT	0.88	0.71	0.23	0.13	1.40	0.67	1
MP2/6-311+G(d,p)	SL WFT	0.99	0.47	0.29	0.08	1.69	0.70	5
MCCO/3	ML WFT	0.44	0.60	0.70	0.04	1.80	0.72	15
MP2/6-31G(d)	SL WFT	2.11	0.87	0.46	0.12	0.21	0.75	1
MP2/6-31G+(d)	SL WFT	1.07	1.01	0.29	0.14	1.37	0.78	1
SAC-MP2/6-31+G(d,p)	ML WFT	0.88	1.24	0.15	0.15	2.17	0.92	1
MP2/6-31G(2df,p)	SL WFT	1.97	1.18	0.18	0.26	1.16	0.95	3
SAC-MP2/6-31G(d)	ML WFT	2.83	1.67	0.17	0.14	1.04	1.17	1
MP2/6-31B(d)	SL WFT	1.62	3.35	1.13	0.13	0.31	1.31	1
<i>N</i> ⁴ Methods								
M05-2X/MG3S-CP	SL DFT	0.20	0.30	0.32	0.03	0.71	0.31	9
M05-2X/MG3S	SL DFT	0.40	0.46	0.27	0.09	0.49	0.34	6
PWB6K/MG3S	SL DFT	0.44	0.25	0.24	0.15	0.81	0.38	9
PWB6K/MG3S-CP	SL DFT	0.34	0.16	0.32	0.07	1.02	0.38	9
HF/6-31G(d)	SL WFT	0.95	2.39	2.12	0.23	2.69	1.68	0.1
HF/6-311+G(d,p)	SL WFT	1.84	3.09	2.17	0.31	3.42	2.17	1
HF/MG3S	SL WFT	2.24	3.77	2.40	0.30	3.39	2.42	5
HF/MG3	SL WFT	2.24	3.77	2.40	0.30	3.39	2.42	6
HF/G3Large	SL WFT	2.25	3.77	2.41	0.30	3.39	2.42	8
average		0.95	0.93	0.62	0.13	1.01	0.73	

^a ML denotes multilevel; SL denotes single-level. ^b The cost for each method is measured by the computer time for a single-point energy calculation of the T-shaped benzene dimer (at the fixed geometry of Sinnokrot and Sherrill⁵⁵) divided by the computer time for an MP2/6-31+G(d,p) energy calculation on the same dimer with the same computer program and same computer.

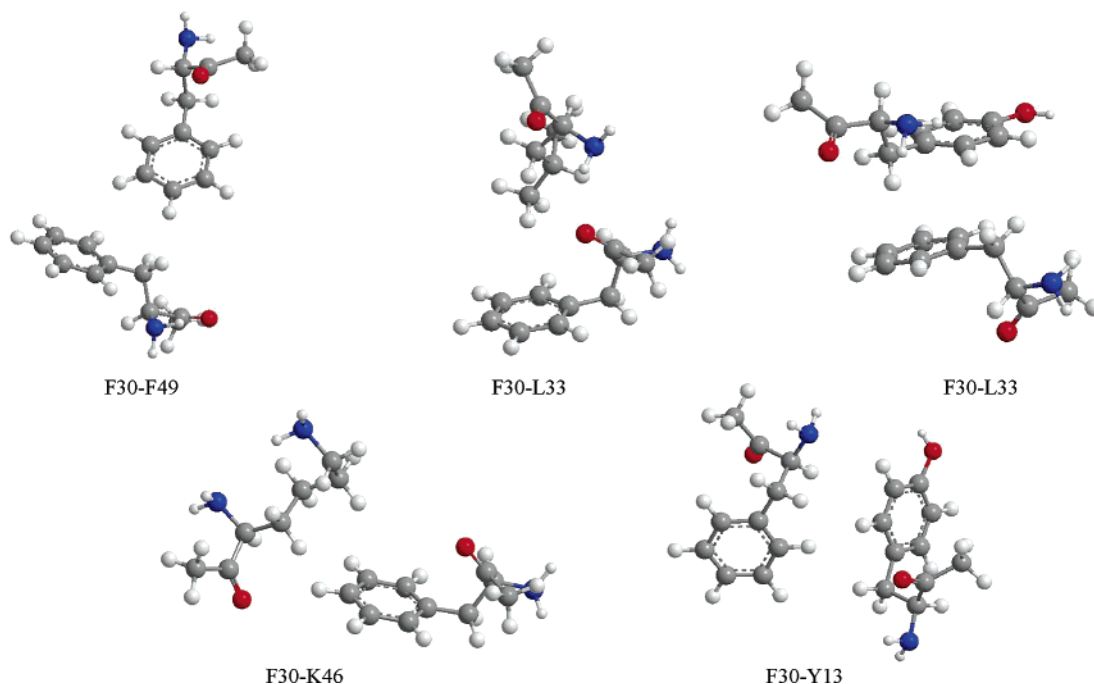


Figure 6. Geometries of the pairs of amino acid residues.

gives almost identical results to those obtained with MP2/MG3. Note that the only difference between MG3 and MG3S is that MG3S does not have diffuse basis functions on hydrogen atoms. The counterpoise correction improves the performance of the M05-2X/MG3S calculations, but it deteriorates the performance of the MP2/MG3S method by a large margin.

Evaluation of energies at a standard geometry to assess hydrogen bonding sometimes may lead to significantly different conclusions than would be obtained if the level of theory used for the energy calculation is also used for geometry optimization. This is especially true for methods that are defined to use inappropriately low levels of geometry. For example G3SX is defined to use B3LYP/6-31G(2df,p) geometries, which are not very good for hydrogen bonding because of the lack of diffuse functions.³⁷ CBS-QB3 suffers in the same way. In the present study, though, we use MC-QCISD/3 geometries, which are quite accurate. We have previously³² validated that mean errors are only slightly changed for hydrogen bonding, dipole interactions, and weak interactions when MC-QCISD/3 geometries instead of using consistently optimized geometries.

3.2. Charge-Transfer Complexes. Tables S1, S2, and 3 present the results for the charge-transfer complexes. Among the multilevel methods, MCQCISD-MPW and MCQCISD-MPW give the lowest MUE for calculating the binding energies in the CT7/04 database. MCG3-MPW and MCG3-TS are the two best N^7 methods, and MC3MPWB is the best N^5 multilevel method.

Tables S2 and 3 show that the PWB6K/MG3S method is the best single level method for calculating interaction energies in charge-transfer complexes.

3.3. Dipole Interaction. Tables S3, S4, and 3 summarize the results for the dipole interaction complexes. Among the multilevel methods, MCG3-MPW gives the lowest MUE for calculating the binding energies in the DI6/04 database.

MCQCISD-MPW is the best N^6 multilevel method, and SAC-MP2/6-31+G(d,p) is the best N^5 method.

From Tables S4 and 3, we can see that MP2/6-31G(2df,p) is the best single-level method, but this good performance is due to the error cancellation between the BSSE and the incomplete treatment of correlation, as can be ascertained by noticing that the MP2 method with larger basis sets gives worse results. PWB6K is the best N^4 method for dipole interactions.

3.4. Weak Interaction. Tables S5, S6, and 3 present the results for the weak interaction complexes. These complexes are bound by dispersion-like forces. MCCO/3 is the best multilevel method, whereas MCG3-MPW and MCQCISD-MPW are the best N^7 and N^6 methods, respectively.

Tables S6 and 3 show that M05-2X/MG3S (with CP correction) is the best single-level method for the calculations of binding energies of these weakly bound van der Waals complexes.

3.5. $\pi\cdots\pi$ Interaction. Tables S7, S8, and 3 summarize the results for the $\pi\cdots\pi$ stacking complexes. Among the multilevel methods, MCG3-MPW gives the lowest MUE for calculating the binding energies in the PPS5/05 database. MCQCISD-MPW is the best N^6 multilevel method, and MC3MPWB is the best N^5 multilevel method. Table S8 shows that MP2/6-31G(d) is the best single-level method, but this good performance is again due to error cancellation, since the MP2 method with larger basis sets give worse results. PWB6K was found to be the best density functional for stacking interactions in biological systems⁶¹ and tetramers⁶² of formic acid and formamide. Here we find that the new M05-2X functional is even better for $\pi\cdots\pi$ stacking.

3.6. Overall Results. Table 3 is a summary of the performance of the tested methods for noncovalent interactions. The rank order is according to the MMUE column, which is the average of the five database columns included in this table, as defined by eq 1. Clearly the exact position

Table 4. Binding Energies (kcal/mol) and Mean Errors for Amino Acid Residue Pairs^a

methods	F30-K46	F30-L33	F30-Y13	F30-F49	F30-Y4	MSE	MUE
best estimate	3.10 ^b	5.00 ^b	3.90 ^b	2.70 ^c	5.30 ^c		
M05-2X	2.53	4.47	3.41	2.07	3.62	-0.78	0.78
PWB6K	2.20	3.87	2.87	1.49	2.80	-1.35	1.35
PW6B95	1.82	2.91	2.36	1.04	2.03	-1.97	1.97
MPWB1K	1.55	2.76	2.05	0.85	1.81	-2.20	2.20
MPW1B95	1.47	2.41	1.93	0.72	1.53	-2.39	2.39
B97-1	1.69	1.65	2.17	1.08	0.72	-2.54	2.54
PBE	1.47	1.17	1.87	0.83	0.22	-2.89	2.89
TPSS	0.81	-0.35 ^d	0.82	0.07	-1.17 ^d	-3.97	3.97
B3LYP	0.42	-0.66 ^d	0.49	-0.24 ^d	-1.81 ^d	-4.36	4.36
O3LYP	-0.30 ^d	-4.13 ^d	-1.00 ^d	-1.41 ^d	-4.73 ^d	-6.31	6.31

^a Basis set: 6-31+G(d,p); geometries from ref 69. No counterpoise correction was made. MSE denotes mean signed error, i.e., mean signed deviation from the best estimate, and MUE denotes mean unsigned error. ^b MP2/CBS + Δ CCSD(T) from ref 69. ^c MP2/CBS + side chain Δ CCSD(T) from ref 69. ^d A negative number denotes that the interaction energy is repulsive at the geometry of ref 69.

in the ranking is not as meaningful as the general trends, but the MMUE provides a way to organize the discussion. In Table 3, we also tabulate the "cost" for each method, which is measured by the computer time for a single-point energy calculation of the T-shaped benzene dimer (at the fixed geometry of Sinnokrot and Sherrill⁵⁵) divided by the computer time for an MP2/6-31+G(d,p) energy calculation on the same dimer with the same computer program and the same computer. Although we are aware of the danger of timing algorithms with specific programs on specific computers, these costs (if not interpreted too finely), nevertheless help place the methods in a perspective of affordability.

From Table 3, we can see that the best performer for these noncovalent databases is MCG3-MPW, and its cost is much less than the G3SX, MP4/6-31(2df,p), and CCSD(T)/6-311+G(d,p) methods. The best N^6 method is MCQCISD-MPW. Note that three N^6 methods, namely MCQCISD-MPW, MCQCISD-MPW, and MCQCISD-TS, outperform the CBS-QB3 and G3SX N^7 methods with much less cost. The best N^4 method, M05-2X, outperforms the best N^5 method, MC3MPWB. M05-2X is also the best single-level method. To obtain better performance than M05-2X one must go to a method almost four times as expensive and with much worse scaling.

4. Concluding Remarks

It is clear that even the best single-level WFT methods are not competitive with the single-level DFT methods, multi-level WFT methods, and multilevel DFT/WFT methods, either for the kind of accuracy (MMUE of 0.31 kcal/mol) attainable with the low-cost methods or for the much higher standard of about half that error (MMUE of 0.16 kcal/mol). Although model chemistries were originally developed for covalent interactions and have been widely applied to such interactions, several of the modern model chemistries are sufficiently robust that they also give excellent results for noncovalent interactions, and they should be very useful for many important applications that require accurate models for noncovalent interactions. It is encouraging that the best performing methods in the current tests (MCG3-MPW,²³ MCQCISD-MPW,²³ MCQCISD-MPW,²³ and M05-2X³⁴) also show excellent performance^{22,23,34} for atomization ener-

gies, bond energies, barrier heights, ionization potentials, and electron affinities.

5. Software

M05-2X has been incorporated in *NWCHEM*⁶³ and *GAUSS-IAN03*⁵³ and will soon be available in release versions of these programs. All multilevel methods tested in this paper except CBS-QB3 are available in *MLGAUSS*,⁵² which requires *GAUSSIAN03* in order to execute. CBS-QB3 is available in *GAUSSIAN03*.⁵³

Acknowledgment. This work was supported by the Office of Naval Research under grant no. N00012-05-01-0538 and the U.S. Department of Energy, Office of Basic Energy Sciences.

Appendix

As an adjunct to this study, as requested by a referee, we calculated the binding energies of five pairs of amino acid residues by 10 different density functionals; the tested DFT methods are B3LYP,⁴⁶ PBE,⁶⁴ B97-1,⁶⁵ O3LYP,⁶⁶ TPSS,⁶⁷ MPW1B95,⁶⁸ MPWB1K,⁶⁸ PW6B95,³³ PWB6K,³³ and M05-2X.³⁴ The amino acid residues and the pair geometries are taken from ref 69, where a pair of residues was cut from a crystal structure, and each residue is modeled by an amino acid in which the -OH group is replaced by a -CH₃ group; see Figure 6. The results are given in Table 4. It is encouraging that the two best performing functionals are (in order) M05-2X and PWB6K because these two functionals (out of 14 tested) were also found³⁴ to be the two best functionals (in the same order) for both stacking and hydrogen bonding interactions in nucleobase pairs.

Supporting Information Available: Calculated binding energies and mean errors for the CT7/04, DI6/04, WI7/05, and PPS5/05 databases. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Pople, J. A. *Rev. Mod. Phys.* **1999**, *71*, 1267.
- (2) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221.

- (3) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764.
- (4) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 1125.
- (5) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **2001**, *114*, 108.
- (6) Martin, J. M. L.; Oliveira, G. d. *J. Chem. Phys.* **1999**, *111*, 1843.
- (7) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 4129.
- (8) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A. *J. Chem. Phys.* **1996**, 2598.
- (9) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822.
- (10) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **2000**, 6532.
- (11) Brown, F. B.; Truhlar, D. G. *Chem. Phys. Lett.* **1985**, *117*, 307.
- (12) Gordon, M. S.; Truhlar, D. G. *J. Am. Chem. Soc.* **1986**, *108*, 2.
- (13) Rossi, I.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *234*, 64.
- (14) Fast, P. L.; Corchado, J.; Sanchez, M. L.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 3139.
- (15) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 3898.
- (16) Fast, P. L.; Corchado, J.; Sanchez, M. L.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 5129.
- (17) Fast, P. L.; Corchado, J.; Sanchez, M. L.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *110*, 11679.
- (18) Fast, P. L.; Sanchez, M. L.; Truhlar, D. G. *Chem. Phys. Lett.* **1999**, *306*, 407.
- (19) Tratz, C. M.; Fast, P. L.; Truhlar, D. G. *PhysChemComm* **1999**, *2*, 14.
- (20) Fast, P. L.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 6111.
- (21) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2002**, *106*, 842.
- (22) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 4786.
- (23) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43.
- (24) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 1643.
- (25) Li, T.-H.; Mou, C.-H.; Hu, W.-P. *Chem. Phys. Lett.* **2004**, *397*, 364.
- (26) Li, T.-H.; Chen, H.-R.; Hu, W.-P. *Chem. Phys. Lett.* **2005**, *412*, 430.
- (27) Dunn, M. E.; Pokon, E. K.; Shields, G. C. *J. Am. Chem. Soc.* **2004**, *126*, 2647.
- (28) Day, M. B.; Kirschner, K. N.; Shields, G. C. *Int. J. Quantum Chem.* **2005**, *102*, 565.
- (29) Pickard, F. C.; Dunn, M. E.; Shields, G. C. *J. Phys. Chem. A* **2005**, *109*, 9183.
- (30) Dunn, M. E.; Pokon, E. K.; Shields, G. C. *Int. J. Quantum Chem.* **2004**, *100*, 1065.
- (31) Giese, T. J.; York, D. M. *Int. J. Quantum Chem.* **2004**, *98*, 388.
- (32) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415.
- (33) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656.
- (34) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.
- (35) Curtiss, L. A.; Redfern, P. C.; Rassolov, V.; Kedziora, G.; Pople, J. A. *J. Chem. Phys.* **2001**, *114*, 9287.
- (36) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *Chem. Phys. Lett.* **2002**, *359*, 390.
- (37) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2005**, *123*, 124107.
- (38) Schultz, N.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127.
- (39) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 4209.
- (40) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
- (41) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.
- (42) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968.
- (43) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.
- (44) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (45) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, submitted for publication.
- (46) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (47) Zhao, Y.; Tishchenko, O.; Truhlar, D. G. *J. Phys. Chem. B* **2005**, *109*, 19046.
- (48) 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; p 1.
- (49) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- (50) Lynch, B. J.; Truhlar, D. G. In *Recent Advances in Electron Correlation Methodology*; Wilson, A. K., Peterson, K. A., Eds.; American Chemical Society: Washington, DC, in press.
- (51) Herzberg, G. *Molecular Spectra and Molecular Structure. I. Spectra of Diatomic Molecules*, 2nd ed.; D. Van Nostrand: Princeton, NJ, 1950; p 437.
- (52) Zhao, Y.; Truhlar, D. G. *MLGAUSS-version 1.0*; University of Minnesota: Minneapolis, 2005.
- (53) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.;

- Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, revision C.01*; Gaussian, Inc.: Pittsburgh, PA, 2003.
- (54) Truhlar, D. G. <http://comp.chem.umn.edu/mccdir/software.htm>.
- (55) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200.
- (56) Corchado, J. C.; Chuang, Y.-Y.; Fast, P. L.; Villa, J.; Hu, W.-P.; Liu, Y.-P.; Lynch, G. C.; Nguyen, K. A.; Jackels, C. F.; Melissas, V. S.; Lynch, B. J.; Rossi, I.; Coitino, E. L.; Fernandez-Ramos, A.; Pu, J.; Albu, T. V.; Steckler, R.; Garrett, B. C.; Isaacson, A. D.; Truhlar, D. G. *POLYRATE, 9.1*; University of Minnesota: Minneapolis, MN, 2002.
- (57) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703.
- (58) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
- (59) Schwenke, D. W.; Truhlar, D. G. *J. Chem. Phys.* **1985**, *82*, 2418. **1987**, *86*, 3760 (E).
- (60) Raghavachari, K.; Anderson, J. B. *J. Phys. Chem.* **1996**, *100*, 12960.
- (61) Zhao, Y.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2701.
- (62) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 6624.
- (63) Aprà, E.; Windus, T. L.; Straatsma, T. P.; Bylaska, E. J.; de Jong, W.; Hirata, S.; Valiev, M.; Hackler, M.; Pollack, L.; Kowalski, K.; Harrison, R.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; V., T.; Krishnan, M.; Auer, A. A.; Brown, E.; Cisneros, G.; Fann, G.; Früchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers, 4.7*; Pacific Northwest National Laboratory: Richland, WA, 2005.
- (64) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett* **1996**, *77*, 3865.
- (65) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.
- (66) Hoe, W.-M.; Cohen, A. J.; Handy, N. C. *Chem. Phys. Lett.* **2001**, *341*, 319.
- (67) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (68) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.
- (69) Vondrášek, J.; Bendová, L.; Klusák, V.; Hobza, P. *J. Am. Chem. Soc.* **2005**, *127*, 2615.

CT060044J

On the Angular Dependence of the Vicinal Fluorine–Fluorine Coupling Constant in 1,2-Difluoroethane: Deviation from a Karplus-like Shape

Patricio F. Provasi*

*Department of Physics, University of Northeastern, Av. Libertad 5500,
W 3404 AAS Corrientes, Argentina*

Stephan P. A. Sauer†

*Department of Chemistry, University of Copenhagen, Universitetsparken 5,
DK-2100 Copenhagen Ø, Denmark*

Received March 20, 2006

Abstract: The angular dependence of the vicinal fluorine–fluorine coupling constant, ${}^3J_{\text{FF}}$, for 1,2-difluoroethane has been investigated with several polarization propagator methods. ${}^3J_{\text{FF}}$ and its four Ramsey contributions were calculated using the random phase approximation (RPA), its multiconfigurational generalization, and both second-order polarization propagator approximations (SOPPA and SOPPA(CCSD)), using locally dense basis sets. The geometries were optimized for each dihedral angle at the level of density functional theory using the B3LYP functional and fourth-order Møller–Plesset perturbation theory. The resulting coupling constant curves were fitted to a cosine series with 8 coefficients. Our results are compared with those obtained previously and values estimated from experiment. It is found that the inclusion of electron correlation in the calculation of ${}^3J_{\text{FF}}$ reduces the absolute values. This is mainly due to changes in the FC contribution, which for dihedral angles around the trans conformation even changes its sign. This sign change is responsible for the breakdown of the Karplus-like curve.

1. Introduction

The sensitivity of indirect spin–spin coupling constants (J) to structural changes is a powerful tool for determination of molecular structures and conformations. As an example we can mention the dependence of J on bond or dihedral angles which has been the object of many studies (see e.g., the review by Contreras and Peralta¹ and references therein). In particular, the study of the dependence of coupling constants on torsion angles became an important issue after Karplus presented his already classical equation.² In recent years some new attempts were made to explain the origin of this behavior.³

Coupling constants involving fluorine atoms have recently attracted much interest,^{4–12} due to the important biological activity of fluorinated organic compounds,¹³ their use in medicine for NMR imaging techniques,¹⁴ their possible use in quantum computers,¹⁵ and their unusual behavior such as e.g. long-range through-bond¹⁰ and through-space couplings.^{5,12} Another example for the unusual behavior of fluorine coupling constants is the dependence of vicinal fluorine–fluorine coupling constant ${}^3J_{\text{FF}}$ on the dihedral angle in 1,2-difluoroethane^{7,9,16} which differs greatly from the usual Karplus curve as found e.g. for the vicinal proton–proton coupling in ethane.^{17,18}

Kurtkaya et al.⁷ have calculated ${}^3J_{\text{FF}}$ using density functional theory (DFT)¹⁹ with the B3LYP functional²⁰ and the 6-311G(d,p) basis set.²¹ Their geometries were optimized at the same level of theory for each fixed F–C–C–F dihedral angle. They analyzed the dominating Fermi contact contribu-

* Corresponding author e-mail: patricio@unne.edu.ar.

† Present address: Max-Planck-Institut für Kohlenforschung, Kaiser Wilhelm-Platz 1, D-45470 Mülheim an der Ruhr, Germany.

tion to ${}^3J_{\text{FF}}$ in 1,2-difluoroethane in terms of the carbon–fluorine bond orbitals and the lone pairs orbitals of fluorine. However, their calculated vicinal coupling for the trans conformation is not in agreement with the known experimental value, which is not surprising since the B3LYP as well as most of the current functionals were shown^{5,6,8,18,22} to have problems reproducing coupling constants that involve at least one fluorine atom.

More recently, San Fabián and Westra Hoekzema⁹ presented Karplus curves for ${}^3J_{\text{FF}}$ in 1,2-difluoroethane calculated with the multiconfigurational random phase approximation (MCRPA)^{23,24} with various restricted active space self-consistent field (RASSCF) wave functions,²⁵ the second-order polarization propagator approximation (SOPPA),^{26,27} and DFT using the BLYP functional. Their geometries were either optimized at the B3LYP level using the cc-pVTZ basis set²⁸ for a fixed F–C–C–F dihedral angle or were kept fixed during rotation at the values of a standard C–C bond length and tetrahedral bond angles. They fitted their curves to truncated Fourier series in the torsion angle ϕ and found that the number of Fourier coefficients necessary for a proper representation of the Karplus curve is too large for an empirical parametrization based on experimental coupling constants. They conclude that the missing data have to be provided by high accuracy calculations. With respect to the different correlated methods, they find that SOPPA gives in general the best agreement with experimental values, but that important differences remain in particular for some of the Fourier coefficients and for the trans coupling, for which the largest differences between the various calculations are observed. Furthermore equally large changes in the Fourier coefficients are observed between the SOPPA calculations with standard and optimized geometries, and the authors concluded that good geometries must be used in the calculation of these couplings.

Several years ago a modification of the SOPPA method was introduced in which the Møller–Plesset correlation coefficients are replaced by the corresponding coupled cluster singles and doubles amplitudes in the SOPPA equations. This second-order polarization propagator approximation with coupled cluster singles and doubles amplitudes—SOPPA(CCSD)²⁹—called method was shown to give more accurate coupling constants than SOPPA.^{8,17,27,30–33} Furthermore tiny changes in the coupling constants such as temperature dependence and isotope effects, which originate in the geometry dependence of the coupling constants, could quantitatively be reproduced by SOPPA(CCSD) calculations.^{32,34}

In the present work we have therefore studied the large correlation and geometry effects on the vicinal fluorine–fluorine couplings in 1,2-difluoroethane and its four contributions using SOPPA(CCSD). Geometries optimized at the level of DFT/B3LYP and fourth-order Møller–Plesset perturbation theory (MP4)³⁵ using the 6-311G(d,p) basis set were employed in the calculations. In addition, calculations at the level of the random phase approximation (RPA),³⁶ the MCRPA with various RASSCF wave functions, and SOPPA were performed.

The paper is organized as follows: the next section gives a short review of the theory of spin–spin coupling constants and the quantum chemical methods for the calculation of J . The details of our calculations are explained in section 3. Section 4 is devoted to the presentation and discussion of our results, and finally in section 5 our conclusions are presented.

2. Theory

Ramsey³⁷ has explained the total nonrelativistic indirect nuclear spin–spin coupling constant between nuclei M and N as the sum of four contributions. They are the diamagnetic nuclear spin–electronic orbital (DSO), the paramagnetic nuclear spin–electronic orbital (PSO), nuclear spin–electronic spin dipolar (SD), and the Fermi contact (FC) contributions

$$J_{\text{MN}}^{\text{Tot}} = J_{\text{MN}}^{\text{DSO}} + J_{\text{MN}}^{\text{PSO}} + J_{\text{MN}}^{\text{SD}} + J_{\text{MN}}^{\text{FC}} \quad (1)$$

where the FC and SD terms account for the interaction of the nuclear spin with the spins of the electrons, and the PSO and DSO terms account for the interaction of the nuclear spin with the orbital angular momentum of the electrons.

The DSO contribution is a ground-state average value

$$J_{\text{MN}}^{\text{DSO}} = -\frac{1}{3} \frac{\gamma_{\text{M}}\gamma_{\text{N}}}{h} \left(\frac{\mu_0}{4\pi}\right)^2 \frac{e^2\hbar^2}{m_{\text{e}}} \sum_{\alpha=x,y,z} \left\langle \Psi_0 \left| \sum_i \frac{\vec{r}_{i\text{N}} \cdot \vec{r}_{i\text{M}} - (\vec{r}_{i\text{N}})_\alpha (\vec{r}_{i\text{M}})_\alpha}{|\vec{r}_{i\text{N}}|^3 |\vec{r}_{i\text{M}}|^3} \right| \Psi_0 \right\rangle \quad (2)$$

although it can also be expressed in a form which involves excited states.³⁸

The last three contributions can be expressed as a sum over excited states in the following way

$$J_{\text{MN}}^{\text{A}} = \sum_{\alpha=x,y,z} \frac{2}{3} \frac{\gamma_{\text{M}}\gamma_{\text{N}}}{h} \sum_{n \neq 0} \frac{\langle \Psi_0 | \hat{\mathcal{O}}_{\text{M}}^{\text{A}} | \Psi_n \rangle \langle \Psi_n | \hat{\mathcal{O}}_{\text{N}}^{\text{A}} | \Psi_0 \rangle}{E_0 - E_n} \quad (3)$$

where A = PSO, SD, FC. The explicit expressions of the above operators are

$$\hat{\mathcal{O}}_{\text{M}}^{\text{PSO}} = \frac{\mu_0}{4\pi} \frac{e\hbar}{m_{\text{e}}} \sum_i \frac{(\vec{l}_{i\text{M}})_\alpha}{r_{i\text{M}}^3} \quad (4)$$

$$\hat{\mathcal{O}}_{\text{M}}^{\text{FC}} = \frac{\mu_0}{4\pi} \frac{4\pi g_{\text{e}} e\hbar}{3m_{\text{e}}} \sum_i (\vec{s}_i)_\alpha \delta(\vec{r}_{i\text{M}}) \quad (5)$$

$$\hat{\mathcal{O}}_{\text{M}}^{\text{SD}} = \frac{\mu_0}{4\pi} \frac{g_{\text{e}} e\hbar}{2m_{\text{e}}} \sum_i \frac{3(\vec{s}_i \cdot \vec{r}_{i\text{M}})(\vec{r}_{i\text{M}})_\alpha - r_{i\text{M}}^2 (\vec{s}_i)_\alpha}{r_{i\text{M}}^5} \quad (6)$$

The gyromagnetic ratio of nucleus M is γ_{M} , $\vec{r}_{i\text{M}} = \vec{r}_i - \vec{r}_{\text{M}}$ is the difference of the position vectors of electron i and nucleus M, \vec{s}_i is the spin operator of electron i , $\vec{l}_{i\text{M}} = \vec{l}_i(\vec{R}_{\text{M}})$ is the orbital angular momentum operator of electron i with respect to the position of nucleus M (in SI units), $\delta(x)$ is the

Dirac delta function, and all other symbols in eqs 1–6 have their usual meaning.³⁹

The FC and SD contributions, that account for the interaction with the spin of the electrons, arise from the admixture of excited triplet states $|\Psi_n\rangle$ to the singlet ground state $|\Psi_0\rangle$, whereas the OP term only involves excited states $|\Psi_n\rangle$ of the same spin symmetry as the ground state $|\Psi_0\rangle$.

Using polarization propagator⁴⁰ or linear response function methods²⁴ all contributions to the coupling constants can be evaluated without explicit calculation of the excited states involved⁴¹

$$J_{MN}^A = - \left(\langle \Psi_0 | [(\hat{O}_M^A)_\alpha, \hat{h}_i] | \Psi_0 \rangle \quad \dots \right) \left(\begin{array}{ccc} \langle \Psi_0 | [\hat{h}_i, [\hat{H}, \hat{h}_j]] | \Psi_0 \rangle & \dots & \\ \vdots & & \ddots \end{array} \right)^{-1} \left(\begin{array}{c} \langle \Psi_0 | [\hat{h}_j, (\hat{O}_N^A)_\alpha] | \Psi_0 \rangle \\ \vdots \end{array} \right) \quad (7)$$

where \hat{H} is the electronic Hamiltonian of the system, and $\{\hat{h}_i\}$ is a complete set of operators. Different approximate propagator methods can then be derived by truncating the set of operators $\{\hat{h}_i\}$ and approximating the exact ground-state wave function Ψ_0 using either variational or perturbational approaches. Examples for the former are the random phase approximation³⁶ or self-consistent field (SCF) linear response function²⁴ and its multiconfigurational generalization MCRPA,^{23,24} where either only the molecular orbital coefficients or the molecular orbital coefficients and the determinant expansion coefficients are variationally optimized in the wave function Ψ_0 . The set of operators $\{\hat{h}_i\}$ consists then correspondingly of either only orbital rotation operators or orbital rotation operators and state transfer operators. To treat dynamic correlation properly with MCRPA very large determinant expansions have to be included in the wave function, which makes high accuracy MCRPA calculations prohibitively expensive.

Alternatively, dynamic correlation can be treated by methods based on Møller–Plesset perturbation⁴² such as SOPPA.^{26,27} From the viewpoint of the perturbation theory the polarization propagator, eq 7, is evaluated to the first order in the fluctuation potential⁴³ in RPA. Furthermore a closer analysis of the terms entering the RPA matrices⁴⁴ shows that the ground state in RPA is correlated by the inclusion of doubly excited determinants. For this reason RPA or coupled Hartree–Fock is sometimes considered to be a correlated method⁴⁵ contrary to the uncoupled Hartree–Fock approach. In SOPPA the matrix elements in eq 7 are evaluated through second order in the fluctuation potential. If the Møller–Plesset perturbation theory correlation coefficients in the SOPPA equations are replaced with coupled cluster single and double (CCSD) amplitudes, one obtains the SOPPA(CCSD) method.²⁹

The dihedral angle dependence of the vicinal fluorine–fluorine coupling constant is best represented by a truncated

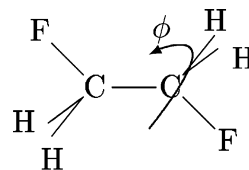


Figure 1. The optimization of structures was performed with the dihedral angle ϕ ($\angle\text{F}-\text{C}-\text{C}-\text{F}$) fixed at 0° , 15° , 30° , 45° , 60° , 80° , 90° , 100° , 115° , 135° , 150° , 165° , and 180° [see ref 7].

Fourier series in the dihedral angle ϕ

$${}^3J_{\text{F-F}}^{\text{Tot}}(\phi) = C_0 + C_1\cos(\phi) + C_2\cos(2\phi) + C_3\cos(3\phi) + C_4\cos(4\phi) + C_5\cos(5\phi) + C_6\cos(6\phi) + C_7\cos(7\phi) + \dots \quad (8)$$

Truncation of this series after the $C_2\cos(2\phi)$ term gives the original equation by Karplus.²

3. Details of Calculations

Geometry Optimizations. All geometry optimizations were performed with the Gaussian 98 program⁴⁶ at the Hartree–Fock (HF), DFT-B3LYP,²⁰ MP2,⁴² MP3, and full MP4³⁵ levels of theory using the 6-311G(d,p) basis set.²¹ In all structures the dihedral angle $\angle\text{F}-\text{C}-\text{C}-\text{F}$ was fixed to the values given in Figure 1 [see also ref 7]. The C and F 1s orbitals were kept frozen in the correlated calculations. The dihedral angle in the gauche conformation, i.e. the gauche angle, and the energy of the gauche conformation were obtained by fitting the rotamer energy curves with second-order splines.

J-Calculations. All J calculations were performed with the 1.2 version of the Dalton program package.⁴⁷ Locally dense basis sets (LDBS)^{11,48} were employed in order to keep the basis set size within the current limitations of the SOPPA implementation in the program. Hence, the aug-cc-pVTZ-J^{31,49} basis set, which ensures a very good description of the FC term [see ref 31 and therein cited references], was used for the fluorine and carbon atoms which define the coupling pathway, whereas the cc-pVTZ²⁸ basis set was employed for all hydrogen atoms.

In the MCRPA calculations we tested two different RASSCF wave functions (RAS-A and RAS-B), which differ in the number of orbitals included in RAS3. In all cases the 1s molecular orbitals of carbon and fluorine were kept frozen, and the remaining occupied Hartree–Fock orbitals were included in RAS2. Single and double excitations were allowed from RAS2 into RAS3. The RASSCF wave functions could therefore be described as truncated configuration interaction singles and doubles (CISD) wave functions with optimized orbitals. The nomenclature used for the RAS wave functions is $\text{inactive}_{\text{RAS1}}^{\text{RAS2}} \text{RAS3}$, where *inactive*, RAS1, RAS2, and RAS3 are the total numbers of orbitals in these spaces, as all RASSCF calculations were run without the use of symmetry. The precise details of the two RASSCF wave functions are given in Table 1. Compared with the active spaces employed by San Fabián and Westra Hoekzema⁹ we can see that our RAS-A is the same as R30, whereas RAS-B is larger than R45.

Table 1: Description of the RASSCF Wave Functions

label	active space ^{a,b}	N _{SD} ^c
RAS-A	⁴ RAS ₁₃ ¹³	14535
RAS-B	⁴ RAS ₃₁ ¹³	81810

^a The nomenclature for the active spaces is ^{inactive}RAS_{RAS1}^{RAS2}_{RAS3}, where *inactive*, RAS1, RAS2, and RAS3 are the total numbers of orbitals in these spaces, as all RASSCF calculations were run without the use of symmetry. ^b Only single and double excitations are allowed (0 → 2). ^c Number of determinants in the wave function.

Table 2: Relative Energies in kJ/mol of the Cis, Trans, and Gauche Conformations of 1,2-Difluoroethane and the Dihedral Angle in the Gauche Conformation Obtained at the HF, DFT-B3LYP, MP2, MP3, and MP4 Levels of Theory with the 6-311G(d,p) Basis Set

method	relative energies in kJ/mol			
	cis	trans	gauche	dihedral angle gauche
HF	33.11	-0.66	0.0	69.962°
DFT-B3LYP	32.58	1.55	0.0	71.683°
MP2	34.13	1.13	0.0	69.639°
MP3	31.31	0.77	0.0	69.523°
MP4	32.56	0.76	0.0	69.442°

4. Results and Discussion

In this section, we first discuss the energies of the three conformations obtained at the HF, DFT-B3LYP, MP2, and MP4 levels. We discuss then the dependence of the ³J_{FF} curves on the optimization of the geometries and the level of correlation included in the calculations. Finally, we compare our results with previous results and experimental values.

4.1. Rotamer Energies. A complete list of the HF, DFT-B3LYP, MP2, MP3, and MP4 energies for the optimized geometries is included in the Supporting Information.⁵⁰ In Table 2 we have collected the relative energies of the cis, trans, and gauche conformations of 1,2-difluoroethane as well as the values of the dihedral angle in the optimized gauche conformation, i.e. the gauche angle. All correlated methods predict the gauche conformation to be the absolute minimum

in agreement with the spectroscopic findings,^{51,52} whereas at the Hartree–Fock level the trans conformation is slightly lower in energy. B3LYP overestimates the relative energy of the trans conformation relative to the MP4 calculations, whereas both methods agree very well on the relative energy of the cis conformation. MP2, on the other hand, overestimates the relative energies of both rotamers. The shape of the rotamer potential energy curves are thus different in the various methods. This is also reflected in the predicted gauche angle which varies from 71.7° at B3LYP to 69.4° at MP4, i.e. by more than 2°. Compared with the uncorrelated HF calculation B3LYP predicts a larger angle, whereas all MP methods give smaller gauche angles.

4.2. Dependence of ³J_{FF} on the Optimized Geometries.

Vicinal fluorine–fluorine coupling constant curves ³J_{FF}(φ) have been calculated at the optimized MP4 and/or B3LYP geometries using RPA, RAS-A MCRPA, RAS-B MCRPA, SOPPA, and SOPPA(CCSD). The total coupling constants at both series of geometries obtained with the RAS-B wave function and at the SOPPA(CCSD) level are shown in Table 3. In the last three columns of Table 3 the changes in the total coupling constants and in the FC contribution (only at SOPPA(CCSD) level) due to the changes in the optimized geometries are given as well. Tables with the results for all four Ramsey components obtained with the four methods are given in the Supporting Information.⁵⁰

The effect of the changes in the geometry is largest around the cis conformation with ~1.5 Hz, whereas it becomes negligible for dihedral angles around 115° and somewhere between 45° and 60°. Close to the gauche conformation the differences become actually negative and go through a second maximum about 150°. We have to conclude therefore that the shape of the calculated Karplus curve depends on the method employed in the geometry optimization as can also be seen from the coefficients C_n in the Fourier series representation of the curves given in the Supporting Information.⁵⁰ From the last two columns in Table 3 we can see that for most dihedral angles the geometry induced changes are mainly due to the FC term, whereas around the gauche

Table 3: Calculated ³J_{FF} Karplus Curves as a Function of the Method and Optimized Geometry Used in the Calculations^a

φ [°]	B3LYP-geometry		MP4-geometry					Δ MP4-B3LYP geometry ^b		
	RAS-B J _{FF} ^{βTot}	SOPPA- (CCSD)J _{FF} ^{βTot}	RPA J _{FF} ^{βTot}	RAS-A J _{FF} ^{βTot}	RAS-B J _{FF} ^{βTot}	SOPPA J _{FF} ^{βTot}	SOPPA(CCSD) J _{FF} ^{βTot}	RAS-B J _{FF} ^{βTot}	SOPPA(CCSD)	
									³ J _{FF} ^{FC}	J _{FF} ^{βTot}
0	36.50	30.28	54.16	44.11	38.03	32.98	31.84	1.53	1.23	1.56
15	30.05	24.60	46.16	37.29	31.49	26.93	26.07	1.44	1.13	1.47
30	14.30	10.98	26.17	20.46	15.65	12.50	12.31	1.35	1.17	1.33
45	-2.12	-2.58	3.13	1.56	-1.35	-2.36	-1.87	0.77	0.82	0.71
60	-10.71	-8.96	-12.46	-10.21	-10.85	-9.94	-9.12	-0.14	0.05	-0.16
80	-10.01	-8.25	-15.89	-11.52	-10.56	-9.51	-8.77	-0.55	-0.29	-0.52
90	-8.39	-7.78	-13.52	-9.71	-8.89	-8.94	-8.26	-0.50	-0.18	-0.48
100	-7.50	-8.57	-10.21	-7.99	-7.82	-9.60	-8.91	-0.32	0.01	-0.34
115	-7.66	-12.18	-4.10	-6.13	-7.61	-13.13	-12.21	0.05	0.36	-0.03
135	-10.22	-20.54	4.51	-5.46	-9.67	-21.74	-20.11	0.55	0.81	0.43
150	-13.07	-27.45	9.42	-6.43	-12.39	-29.17	-26.86	0.68	0.98	0.59
165	-14.95	-32.24	12.89	-7.30	-14.33	-34.49	-31.67	0.62	1.01	0.57
180	-16.69	-33.83	14.29	-7.56	-16.15	-36.35	-33.34	0.54	1.02	0.49

^a Basis set: F and C, aug-cc-pVTZ-J; H, cc-pVTZ. ^b Difference between the ³J_{FF} calculated at the geometries optimized with DFT-B3LYP and MP4 with the 6-311G(d,p) basis set.

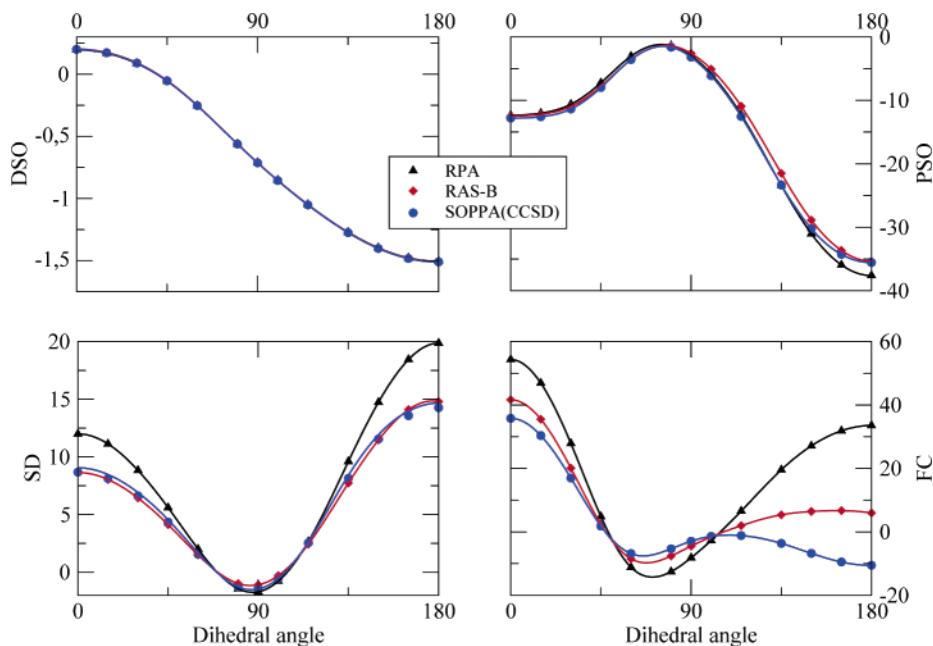


Figure 2. DSO, PSO, SD, and FC contributions to ${}^3J_{\text{FF}}$ at RPA, RAS-B, and SOPPA(CCSD) levels of approximation.

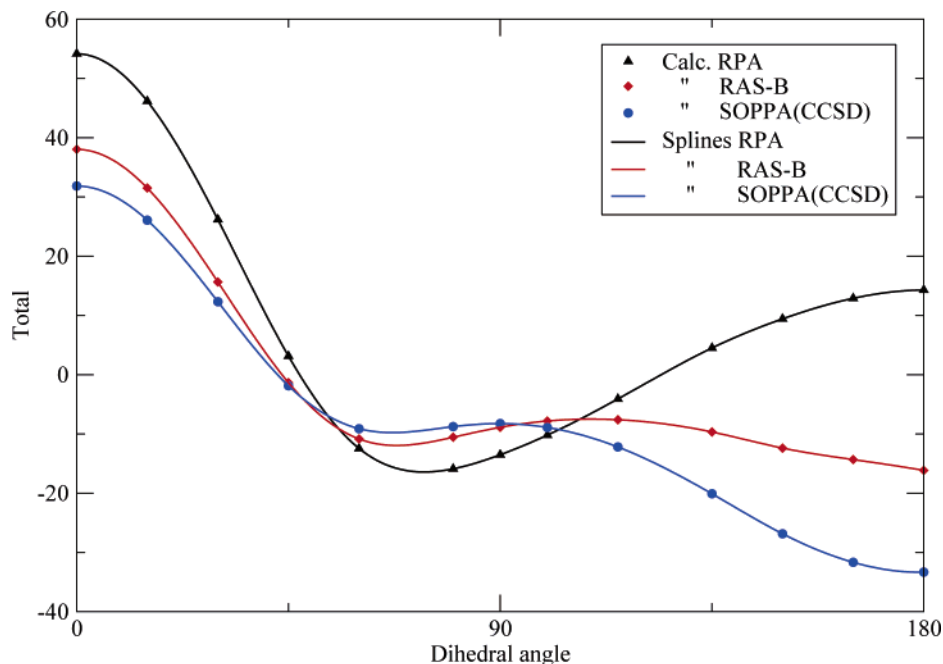


Figure 3. Total indirect nuclear spin–spin coupling constant ${}^3J_{\text{FF}}$ at RPA, RAS-B, and SOPPA(CCSD) levels of approximation.

conformation the smaller changes in the PSO term become dominant due to vanishing changes in the FC term.

4.3. Dependence of ${}^3J_{\text{FF}}$ on Electron Correlation. The total vicinal fluorine–fluorine coupling constants calculated with RPA, RAS-A MCRPA, RAS-B MCRPA, SOPPA, and SOPPA(CCSD) at the MP4 geometries are also presented in Table 3. The four contributions to ${}^3J_{\text{FF}}(\phi)$ and ${}^3J_{\text{FF}}^{\text{Tot}}(\phi)$ at the RPA, RAS-B MCRPA, and SOPPA(CCSD) levels are furthermore shown in Figures 2 and 3. Tables with all the results for both sets of geometries are available from the Supporting Information.⁵⁰

It is a well-known fact^{27,30,54,55} that the electron correlation is often irrelevant for the two singlet contributions, DSO and PSO, and very important for quantitative reproduction of the

two triplet contributions, SD and FC, and thus for the total indirect coupling constant, if the FC term is the dominant contribution. It is therefore not surprising that all four contributions to ${}^3J_{\text{FF}}(\phi)$ exhibit a very different dependence on electron correlation as shown in Figure 2. The DSO term is completely insensitive to electron correlation, while the PSO term changes only slightly around the trans conformation, where the RPA method underestimates the SOPPA(CCSD) value of -35.55 Hz by 2.05 Hz, i.e. $\sim 5.8\%$.

Larger changes, on the other hand, are observed for the two triplet contributions SD and FC. The effect of electron correlation on the SD contribution is recovered similarly by the RAS-B MCRPA and SOPPA(CCSD) calculations for the whole range of conformations. RPA overestimates the SD

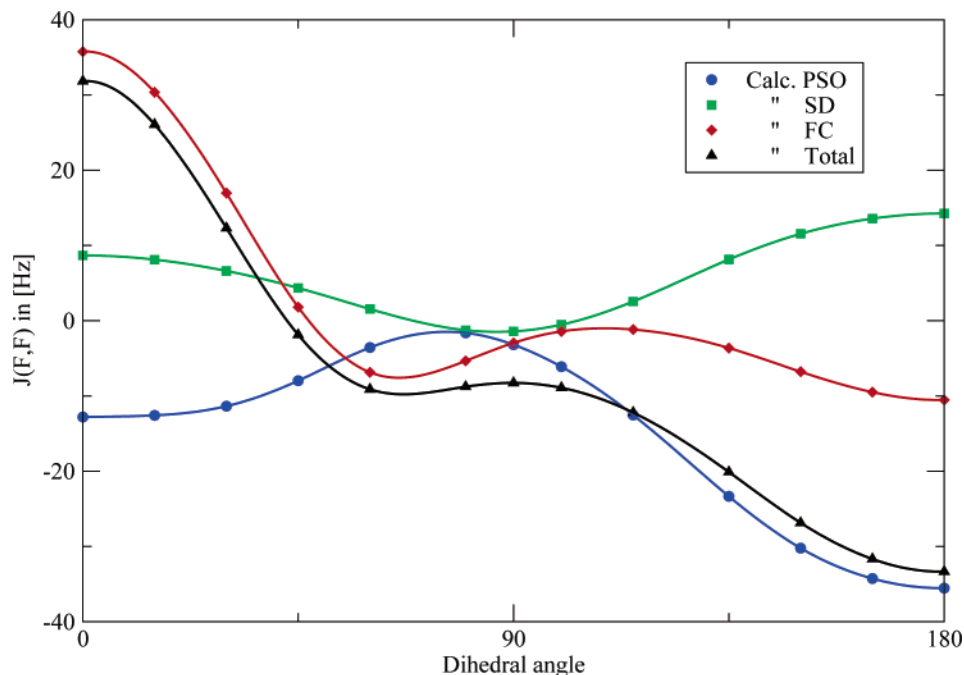


Figure 4. DSO, PSO, SD, and FC contributions to ${}^3J_{FF}$ at SOPPA(CCSD) level of approximation.

term compared to SOPPA(CCSD) and RAS-B MCRPA. The RPA values are about a factor of 1.3 too large. In absolute values this becomes most prominent for the *cis* and *trans* conformations, where the RAS-B MCRPA and SOPPA(CCSD) results differ from RPA by about 3.3 Hz for *cis*-1,2-difluoroethane and 5.6 Hz for *trans*-1,2-difluoroethane. However, it is important to point out that for all dihedral angles between 45° and 135° the deviation of the RPA results from SOPPA(CCSD) is smaller than 1.0 Hz. Hence, we can conclude that for most conformers of 1,2-difluoroethane the SD term is almost not affected by electron correlation.

The largest effect of the electron correlation can be observed for the FC contribution. It is changed for all values of the dihedral angle, but the changes for near-*trans* conformations are much larger than those for near-*cis* or near-*gauche* conformations. For the near-*trans* conformations, electron correlation reduces the FC term until it even changes sign. The corresponding changes in the FC term are as follows: ~ 27.6 Hz from RPA (33.53 Hz) to RAS-B (5.94 Hz) and ~ 16.5 Hz from RAS-B to SOPPA(CCSD) (-10.54 Hz). For *cis*-1,2-difluoroethane, on the other hand, the reduction in the FC term is less pronounced: 12.66 Hz from RPA (33.53 Hz) to RAS-B (5.94 Hz) and 5.88 Hz from RAS-B to SOPPA(CCSD) (35.76 Hz). Thus, the Karplus-like shape of the curve found at the RPA level of approximation is broken when the electron correlation is added in the calculation of the FC term.

Overall we find that the shape of the RAS-B MCRPA FC curve is more similar to the RPA curve than to the SOPPA curve. RAS-B overestimates also slightly the FC contribution around the *cis* and *gauche* conformations and predicts also the wrong sign of FC around the *trans* conformation. This holds even more for the RAS-A MCRPA curve. We observe thus a continuous change in the coupling constant curve on going from RPA through increasingly larger MCRPA

calculations to SOPPA(CCSD), i.e. with a better and better description of dynamic correlation.

Noteworthy is the fact that the change in the sign of the FC contribution for the near-*trans* conformations at SOPPA(CCSD) level does not agree with the Dirac vector model,⁵⁶ which predicts a positive three bond FF coupling. Furthermore, around the *gauche* conformations the Dirac vector model is not fulfilled for all level of approximations.

4.4. Dependence of ${}^3J_{FF}$ on the Dihedral Angle. In Figures 2 and 4 one can see that all four contributions exhibit also a very different dependence on the dihedral angle. We can furthermore see that the total vicinal coupling constant is dominated by the FC contribution in the range of dihedral angles from the *cis* to the *gauche* conformation. The always negative PSO contribution is almost canceled by the positive SD contribution in this range leading to a shift of about -4 Hz with respect to the FC term. Within ± 50 around the *trans* conformation, however, the total coupling constant is dominated by the PSO term, because here the positive SD contribution is almost compensated by the negative FC term, so that the curve is shifted by about 4 Hz compared to the PSO curve. This corresponds qualitatively to the findings by Kurtkaya et al.⁷ although quantitatively their B3LYP curve differs greatly from our SOPPA(CCSD) curve.

The Fourier analysis according to eq 8 of the dihedral angle dependence of ${}^3J_{FF}$ shows that for all levels of calculation the first five coefficients are larger than 1 Hz and are necessary for fitting the curves, see Table 4. At the SOPPA(CCSD) level the FC contribution follows the same scheme as the total coupling constant, and five coefficients are also necessary to fit the FC curve as well. For the SD and PSO contributions, on the other hand, only the first three and the fifth coefficients are necessary for fitting the curves. This is very much in contrast to the vicinal proton–proton

Table 4: Coefficients of the Cosine Series (in Hz) of ${}^3J_{\text{FF}}$ in 1,2-Difluoroethane at Various Levels of Approximation

method	contribution	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
RPA	${}^3J_{\text{FF}}^{\text{Tot}}$	6.999	9.702	24.204	10.118	3.277	0.318	−0.365	−0.279
RAS-A	${}^3J_{\text{FF}}^{\text{Tot}}$	1.111	15.609	14.294	10.053	3.123	0.362	−0.335	−0.259
RAS-B	${}^3J_{\text{FF}}^{\text{Tot}}$	−2.207	16.536	10.337	9.895	3.301	0.623	−0.393	−0.131
SOPPA	${}^3J_{\text{FF}}^{\text{Tot}}$	−8.695	24.396	3.760	9.854	3.369	0.609	−0.160	−0.251
SOPPA(CCSD)	${}^3J_{\text{FF}}^{\text{Tot}}$	−7.761	22.945	3.887	9.245	3.241	0.588	−0.155	−0.242
	${}^3J_{\text{FF}}^{\text{DSO}}$	−0.675	0.859	0.028	−0.001	−0.010	−0.003	0.001	0.001
	${}^3J_{\text{FF}}^{\text{PSO}}$	−14.610	11.170	−10.799	−0.049	0.981	0.312	0.318	−0.036
	${}^3J_{\text{FF}}^{\text{SD}}$	5.615	−2.747	6.410	−0.144	−0.606	0.086	0.030	0.014
	${}^3J_{\text{FF}}^{\text{FC}}$	1.909	13.664	8.248	9.440	2.876	0.193	−0.504	−0.221

Table 5: Calculated Gauche Angle and Corresponding ${}^3J_{\text{FF}}$ at RPA, RAS-B MCRPA, and SOPPA(CCSD) Levels^a

geometry	angle [°]	level	J^{Total} [Hz]
DFT	71.683	RPA	−15.68
		RAS-B	−11.22
		SOPPA	−9.87
		SOPPA(CCSD)	−9.07
MP4	69.442	RPA	−16.10
		RAS-B	−11.96
		SOPPA	−10.59
		SOPPA(CCSD)	−9.77

^a Values were obtained fitting the calculated curves using second degree splines.

couplings in ethane where only the three original Karplus coefficients are necessary (see e.g. ref 17).

4.5. Comparison with Previous Results. The couplings for the trans and gauche conformations have been estimated by Abraham and Kemp⁵³ to be −30 Hz and −10.9 Hz, respectively. Later on refs 51 and 52, the dihedral angle in the gauche conformation of 1,2-difluoroethane was estimated to be 71.0°–71.3°.

Using second degree splines to fit the calculated energy curves for the geometries optimized at the DFT-B3LYP and MP4 levels with the 6-311G(p,d) basis set, we found that the best estimate of the gauche angle is obtained at the DFT-B3LYP level with a value of ~71.7°, Table 5, which agrees with the results reported by Kurtkaya et al.⁷ of ~72.0 Hz. However, the best estimate of the vicinal coupling in the gauche conformation occurs at the SOPPA level with a value of ~10.6 Hz for the MP4-geometry, whereas RAS-B underestimates it and SOPPA(CCSD) overestimates it. Finally, for the trans conformation the best estimate comes from the SOPPA(CCSD) calculations which predicts ~−33.8 Hz for the DFT-geometry and ~−33.3 Hz for the MP4-geometry, whereas for this conformer SOPPA underestimates the coupling by −3.01 Hz and the RAS-B method overestimates it by 17.19 Hz.

The ${}^3J_{\text{FF}}(\phi)$ curve calculated at the DFT-D3LYP level by Kurtkaya et al.⁷ and of course also the SOPPA curve by San Fabián and Westra Hoekzema⁹ are similar to our SOPPA(CCSD) curve. However, the SOPPA(CCSD) couplings are smaller in absolute value than the DFT-D3LYP and SOPPA results. Previous experience with F–F coupling constant calculations^{8,10} showed that the SOPPA(CCSD) results are in general in better agreement with experimental couplings than SOPPA results, as it is the case also for most other

couplings studied so far^{17,27,30–34} and for the Karplus curve of the vicinal proton–proton couplings in ethane.¹⁷ We expect therefore that the SOPPA(CCSD) Karplus curve for the vicinal F–F couplings in 1,2-difluoroethane is also superior to a corresponding SOPPA curve. The largest differences between the SOPPA and SOPPA(CCSD) results are observed for the near-trans conformations with deviations of about 3.0 Hz in favor of the latter.

5. Summary

We have optimized the geometry of 1,2-difluoroethane for different dihedral angles $\angle\text{F–C–C–F}$ at two levels of approximation, DFT-B3LYP and MP4, using the 6-311G-(p,d) basis set. The calculated energies were interpolated with second-order splines in order to obtain the dihedral angle of the gauche conformation. For every optimized geometry ${}^3J_{\text{FF}}$ was calculated at different levels of theory: RPA, MCRPA (RAS-A and RAS-B), SOPPA, and SOPPA(CCSD). The obtained coupling constant curves were fitted to Fourier cosine series.

We find that the form of the Karplus curve depends on the method chosen in the geometry optimization because the changes are largest for the cis conformation. With the exception of dihedral angles close to the gauche angle it is mostly the FC term which is influenced by the changes in the geometry.

Electron correlation affects also mostly the FC contribution. However, these changes are larger for near-trans conformations than for near-cis or near-gauche conformations. For the near-trans conformations even the sign of the FC term is changed by electron correlation. The SD contribution, on the other hand, is affected almost equally for all dihedral angles, and the DSO and PSO terms are almost not affected by electron correlation at all. Hence, one can attribute the capricious form of the F–F Karplus curve in 1,2-difluoroethane to electron correlation effects on the FC contribution.

For dihedral angles in the range from the cis to the gauche conformation the total vicinal coupling constant is dominated by the FC contribution, whereas around the trans conformation it is dominated by the PSO term.

Comparison with previous DFT-B3LYP and SOPPA calculations shows that these follow the same trend as our SOPPA(CCSD) curve. However, along the whole range of dihedral angles the SOPPA(CCSD) couplings are smaller in absolute values than the results of the other two methods by

~10 Hz (in the cis conformation) to 25 Hz (in the trans conformation) for B3LYP⁷ and by 1 Hz (in the cis conformation) to 3 Hz (in the trans conformation) for SOPPA calculations with the MP4 geometry of this work.

Finally we note that the positive value for ${}^3J_{\text{FF}}^{\text{FC}}$ predicted by the Dirac vector model is not reproduced by all our calculated values around the gauche conformations and at the SOPPA(CCSD) level already from dihedral angles from ~45° on.

Acknowledgment. The authors want to thank Ruben H. Contreras, James P. Snyder, and their groups for providing us with the challenge of this project and the B3LYP molecular geometries. This research was supported financially by grants from SNF, FNU, and the Carlsberg Foundation and by computer time grants from DCSC. P.F.P. acknowledges support from CONICET and the CU-DNRC.

Supporting Information Available: Total energies of 1,2-difluoroethane as a function of the F-F dihedral angle obtained by geometry optimization at the HF, B3LYP, MP2, MP3, and MP4 levels; the four contributions to the vicinal F-F indirect nuclear spin-spin coupling constant at the RPA, RAS-A, RAS-B, SOPPA, and SOPPA(CCSD) levels as a function of the F-F dihedral angle for the B3LYP and MP4 optimized geometries; and Fourier coefficients of the dihedral angle dependence of the vicinal F-F indirect nuclear spin-spin coupling constant obtained at the SOPPA(CCSD) level for the B3LYP and MP4 optimized geometries are all given in the Supporting Information. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Contreras, R. H.; Peralta, J. E. *Prog. Nucl. Magn. Res. Spectrosc.* **2000**, *37*, 321–425.
- Karplus, M. *J. Chem. Phys.* **1959**, *30*, 11–15. Karplus, M. *J. Am. Chem. Soc.* **1963**, *85*, 2870–2871.
- Wilkens, S. J.; Westler, J. L.; Markley, J. L.; Weinhold, F. *J. Am. Chem. Soc.* **2001**, *123*, 12026–12036. Provasi, P. F.; Gómez, C. A.; Aucar, G. A. *J. Phys. Chem. A* **2004**, *108*, 6231–6238.
- Peruchena, N. M.; Aucar, G. A.; Contreras, R. H. *J. Mol. Struct. (THEOCHEM)* **1990**, *210*, 205–210. Lantto, P.; Kaski, J.; Vaara, J.; Jokisaari, J. *Chem. Eur. J.* **2000**, *6*, 1395–1406. Barone, V.; Peralta, J. E.; Contreras, R. H.; Snyder, J. P. *J. Phys. Chem. A* **2002**, *106*, 5607–5612.
- Peralta, J. E.; Barone, V.; Contreras, R. H.; Zaccari, D. G.; Snyder, J. P. *J. Am. Chem. Soc.* **2001**, *123*, 9162–9163.
- Lantto, P.; Vaara, J.; Helgaker, T. *J. Chem. Phys.* **2002**, *117*, 5998–6009.
- Kurtkaya, S.; Barone, V.; Peralta, J. E.; Contreras, R. H.; Snyder, J. P. *J. Am. Chem. Soc.* **2002**, *124*, 9702–9703.
- Barone, V.; Provasi, P. F.; Peralta, J. E.; Snyder, J. P.; Sauer, S. P. A.; Contreras, R. H. *J. Phys. Chem. A* **2003**, *107*, 4748–4754.
- San Fabián, J.; Westra Hoekzema, A. J. A. *J. Chem. Phys.* **2004**, *121*, 6268–6276.
- Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *J. Phys. Chem. A* **2004**, *108*, 5393–5398.
- Sanchez, M.; Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *Adv. Quantum Chem.* **2005**, *48*, 161–183.
- Contreras, R. H.; Esteban, Á. L.; Della, N. J.; Díez, E. W.; Head, E. *Mol. Phys.* **2006**, *104*, 485–492.
- Feeney, J.; McCormick, J. E.; Dauer, C. J.; Birdsall, B.; Moody, C. M.; Starkmann, B. A.; Young, D. W.; Francis, P.; Havlin, R. H.; Arnold, W. D.; Oldfield, E. *J. Am. Chem. Soc.* **1996**, *118*, 8700–8706. Colmenares, L. U.; Zou, X.; Liu, J.; Asato, A. E.; Liu, R. S. H. *J. Am. Chem. Soc.* **1999**, *121*, 5803–5804. Bilgiçer, B.; Fichera, A.; Kumar, K. *J. Am. Chem. Soc.* **2001**, *123*, 4393–4399. Bilgiçer, B.; Xing, X.; Kumar, K. *J. Am. Chem. Soc.* **2001**, *123*, 11815–11816. Dewel, H. S.; Daub, E.; Robinson, V.; Honek, J. F. *Biochemistry* **2001**, *40*, 13167–13176. Kitteringham, N. R.; O'Neill, P. M. Metabolism of fluorine-containing drugs. In *Annu. Rev. Pharmacol. Toxicol.* **2001**.
- Bachert, P. *Prog. Nucl. Magn. Res. Spectrosc.* **1998**, *33*, 1–56.
- Tei, M.; Mizuno, Y.; Manmoto, Y.; Sawae, R.; Takarabe, K. *Int. J. Quantum Chem.* **2003**, *95*, 554–557.
- Hirao, K.; Nakatsuji, H.; Kato, H.; Yonezawa, T. *J. Am. Chem. Soc.* **1972**, *94*, 4078–4087. Hirao, K.; Nakatsuji, H.; Kato, H. *J. Am. Chem. Soc.* **1973**, *95*, 31–41.
- Grayson, M.; Sauer, S. P. A. *Mol. Phys.* **2000**, *98*, 1981–1990.
- Helgaker, T.; Watson, M.; Handy, N. C. *J. Chem. Phys.* **2000**, *113*, 94022–9409.
- Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864–B871. Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789. Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200–206.
- Pople, J. A.; Hehre, W. J.; Ditchfield, R. *J. Chem. Phys.* **1972**, *56*, 2257–2261. Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. R. *J. Comput. Chem.* **1983**, *4*, 294–301.
- Malkin, V. G.; Malkina, O. L.; Salahub, D. R. *Chem. Phys. Lett.* **1994**, *221*, 91–99. Malkina, O. L.; Salahub, D. R.; Malkin, V. G. *J. Chem. Phys.* **1996**, *105*, 8793–8800.
- Dalgaard, E.; Jørgensen, P. *J. Chem. Phys.* **1978**, *69*, 3833–3844. Yeager, D. L.; Jørgensen, P. *J. Chem. Phys.* **1979**, *71*, 755–760. Yeager, D. L.; Jørgensen, P. *Chem. Phys. Lett.* **1979**, *65*, 77–80. Vahtas, O.; Ågren, H.; Jørgensen, P.; Aa. Jensen, H. J.; Padkjær, S. B.; Helgaker, T. *J. Chem. Phys.* **1992**, *96*, 6120–6125.
- Olsen, J.; Jørgensen, P. *J. Chem. Phys.* **1985**, *82*, 3235–3264.
- Olsen, J.; Roos, B. O.; Jørgensen, P.; Aa. Jensen, H. J. *J. Chem. Phys.* **1988**, *89*, 2185–2192. Malmqvist, P.-Å.; Roos, B. O. *Chem. Phys. Lett.* **1989**, *245*, 189–193.
- Nielsen, E. S.; Jørgensen, P.; Oddershede, J. *J. Chem. Phys.* **1980**, *73*, 6238–6246. Packer, M. J.; Dalskov, E. K.; Enevoldsen, T.; Aa, J. H. J.; Oddershede, J. *J. Chem. Phys.* **1996**, *105*, 5886–5900. Bak, K. L.; Koch, H.; Oddershede, J.; Christiansen, O.; Sauer, S. P. A. *J. Chem. Phys.* **2000**, *112*, 4173–4185.
- Enevoldsen, T.; Oddershede, J.; Sauer, S. P. A. *Theor. Chem. Acc.* **1998**, *100*, 275–284.
- Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

- (29) Sauer, S. P. A. *J. Phys. B: At., Mol. Opt. Phys.* **1997**, *30*, 3773–3780.
- (30) Kirpekar, S.; Sauer, S. P. A. *Theor. Chem. Acc.* **1999**, *103*, 146–153.
- (31) Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *J. Chem. Phys.* **2001**, *115*, 1324–1334.
- (32) Sauer, S. P. A.; Raynes, W. T.; Nicholls, R. A. *J. Chem. Phys.* **2001**, *115*, 5994–6006.
- (33) Krivdin, L. B.; Sauer, S. P. A.; Peralta, J. E.; Contreras, R. H. *Magn. Reson. Chem.* **2002**, *40*, 187–194. Sauer, S. P. A.; Krivdin, L. B. *Magn. Reson. Chem.* **2004**, *42*, 671–686.
- (34) Wigglesworth, R. D.; Raynes, W. T.; Sauer, S. P. A.; Oddershede, J. *Mol. Phys.* **1997**, *92*, 77–88. Wigglesworth, R. D.; Raynes, W. T.; Sauer, S. P. A.; Oddershede, J. *Mol. Phys.* **1998**, *94*, 851–862. Wigglesworth, R. D.; Raynes, W. T.; Kirpekar, S.; Oddershede, J.; Sauer, S. P. A. *J. Chem. Phys.* **2000**, *112*, 3735–3746. Wigglesworth, R. D.; Raynes, W. T.; Kirpekar, S.; Oddershede, J.; Sauer, S. P. A. *J. Chem. Phys.* **2001**, *114*, 9192.
- (35) Krishnan, R.; Pople, J. A. *Int. J. Quantum Chem.* **1978**, *14*, 91–100. Krishnan, R.; Frisch, M. J.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 4244–4245.
- (36) Rowe, D. J. *Rev. Mod. Phys.* **1968**, *40*, 153–166.
- (37) Ramsey, N. F. *Phys. Rev.* **1953**, *91*, 303–307.
- (38) Sauer, S. P. A. *J. Chem. Phys.* **1993**, *98*, 9220–9221.
- (39) Mills, I.; Cvitas, T.; Homann, K.; Kallay, N.; Kuchitsu, K. *Quantities Units and Symbols in Physical Chemistry*; Blackwell Scientific: Oxford, 1993.
- (40) Linderberg, J.; Öhrn, Y. *Propagator in Quantum Chemistry*; Academic Press: New York, 1973. Jørgensen, P.; Simons, J. *Second Quantization-Based Methods in Quantum Chemistry*; Academic Press: 1981.
- (41) Sauer, S. P. A.; Packer, M. J. *In Computational Molecular Spectroscopy*; Bunker, P. R., Jensen, P., Eds.; Wiley: London, 2000.
- (42) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.
- (43) Oddershede, J.; Jørgensen, P. *J. Chem. Phys.* **1977**, *66*, 1541–1556.
- (44) Hansen, Aa. E.; Bouman, T. D. *Mol. Phys.* **1979**, *37*, 1713–1724.
- (45) Parkinson, W. A.; Sabin, J. R.; Oddershede, J. *Theor. Chem. Acta* **1993**, *86*, 167–179.
- (46) 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.; Mennucci, B.; Cossi, M.; Adamo, C.; Jaramillo, J.; Cammi, R.; Pomelli, C.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Morokuma, K.; 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.; Gomperts, R.; 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.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian98, Revision A.11.2. Development Version, revision C.01*; Gaussian, Inc.: Pittsburgh, PA, 2001.
- (47) Helgaker, T.; Jensen, H. J. Aa.; Jørgensen, P.; Olsen, J.; Ruud, K.; Ågren, H.; Auer, A. A.; Bak, K. L.; Bakken, V.; Christiansen, O.; Coriani, S.; Dahle, P.; Dalskov, E. K.; Enevoldsen, T.; Fernandez, B.; Heattig, C.; Hald, K.; Halkier, A.; Heiberg, H.; Hettrema, H.; Jonsson, D.; Kirpekar, S.; Kobayashi, R.; Koch, H.; Mikkelsen, K. V.; Norman, P.; Packer, M. J.; Saue, T.; Sauer, S. P. A.; Taylor, P. R.; Vahtras, O. *DALTON, an electronic structure program, Release 1.2*; <http://www.kjemi.uio.no/software/dalton/dalton.html>, 2001.
- (48) Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *J. Chem. Phys.* **2000**, *112*, 6201–6208.
- (49) The aug-cc-pVTZ-J basis set can be found at <http://fyskem.ki.ku.dk/sauer/basissets>.
- (50) See the Supporting Information.
- (51) Friesen, D.; Hedberg, K. *J. Am. Chem. Soc.* **1980**, *102*, 3987–3994.
- (52) Takeo, H.; Matsumura, C.; Morino, Y. *J. Chem. Phys.* **1986**, *84*, 4205–4210.
- (53) Abraham, R. J.; Kemp, R. H. *J. Chem. Soc. B* **1971**, 1240–1245.
- (54) Scuseria, G. E. *Chem. Phys. Lett.* **1986**, *127*, 236–241.
- (55) Helgaker, T.; Jaszunski, M.; Ruud, K. *Chem. Rev.* **1999**, *99*, 293–352.
- (56) Harris, R. K. *Nuclear Magnetic Resonance Spectroscopy*; Pitman-London: 1983.

CT6000973

JCTC

Journal of Chemical Theory and Computation

Calculation of Nuclear Spin–Spin Coupling Constants of Molecules with First and Second Row Atoms in Study of Basis Set Dependence

Wei Deng

*Department of Chemistry, Yale University, 225 Prospect Street,
New Haven, Connecticut 06520*

James R. Cheeseman and Michael J. Frisch*

Gaussian Inc., 340 Quinnipiac St., Bldg 40, Wallingford, Connecticut 06492

Received March 24, 2006

Abstract: This paper proposes a systematic way to modify standard basis sets for use in NMR spin–spin coupling calculations, which allows the high sensitivity of this property to the basis set to be handled in a manner which remains computationally feasible. The new basis set series is derived by uncontracting a standard basis set, such as correlation-consistent aug-cc-pVTZ, and extending it by systematically adding tight s and d functions. For elements in different rows of the periodic table, different progressions of functions are added. The new basis sets are shown to approach the basis set limit for calculations on a range of molecules containing hydrogen and first and second row atoms.

Introduction

Nuclear magnetic resonance (NMR) is the most useful technique for chemical structure study in solution with extensive flexibility. Among all spectral information, spin–spin coupling constants are one of the most difficult to produce quantitatively.^{1,2} Advances in electronic structure theory, such as equation-of-motion coupled cluster theory^{3,4} or second-order polarization propagator approximations,⁵ are able to predict spin–spin coupling constants in good agreement with experiments. However, formidable computational cost prohibits the use of these methods for large systems.⁶

As an alternative, density functional theory (DFT) is computationally much less expensive with comparable accuracy.⁷ Recent studies have shown that DFT, particularly with the Becke three-parameter Lee–Yang–Parr (B3LYP) hybrid functional,⁸ provides promising and fast calculation of indirect nuclear spin–spin coupling constants on medium-sized^{9,10} and bulky molecules.¹¹ Evaluations have been made on the capability of B3LYP and linear response methods in spin–spin coupling calculations.^{12,13}

There are four isotropic contributions to the NMR coupling constants, Fermi contact (FC), spin-dipolar (SD), paramagnetic spin–orbit (PSO), and diamagnetic spin–orbit (DSO). Usually, the FC term is the major contribution among the four, and with standard basis sets, it has the largest error.

The accuracy of a spin–spin coupling constant calculation is highly dependent on the Gaussian basis set employed.^{14,15} The basis sets of quantum chemistry are well-developed for the valence electrons. However, NMR experiments probe the electron density closer to the nuclei, where many standard basis sets of ab initio theory will give erroneous results.^{16–18} Because the FC operator requires good characterization of core electrons, its contribution is highly dependent on details of the Gaussian basis sets which are relatively unimportant for most other properties. An analysis of basis set dependence in both complete active space self-consistent field wave functions and the DFT framework has been performed. Within the former, Helgaker et al. reported an extensive study of extending a correlation-consistent Gaussian basis set by uncontracting and augmenting the s-type functions at the tight end.¹⁶

Peralta et al.'s recent work investigated the basis set dependence within a DFT framework.¹⁹ Although their study

* Corresponding author e-mail: frisch@gaussian.com.

had shown the dependence of spin–spin coupling on the core basis set, the systematic examination of how to improve the basis set in an economical manner was not complete. Peralta et al.²⁰ proposed the use of the cc-pCVXZ-sd ($X = D$ and T) basis set, which is the cc-pCVXZ basis set with all s functions fully uncontracted. This basis set yields good results for the one-bond C–C coupling calculation and has been used successfully in bulky fullerene molecules, such as C70. However, the cc-pCVDZ-sd basis set has not shown promising results for coupling other than one-bond C–C coupling.

The aug-cc-pVTZ-J basis set, developed by Sauer and co-workers, gives a very good description of the FC term, with adequate treatment of the wave functions at the nucleus.²¹ The basis set fully uncontracts the aug-cc-pVTZ basis set, then augments it with four tight s -type functions and without the most diffuse second polarization function. Unfortunately, the aug-cc-pVTZ-J basis sets only contains H, four first row atoms (C, N, O, and F), and S (whose basis set includes, in addition to the modifications above, three additional tight d -type functions). Systematic studies for other first and second row atoms were not complete.

The large uncontracted universal Gaussian basis set (UGBS) was first introduced by Silver et al.²² and later generated by Jorge and de Castro.²³ It provides basis sets for all atoms which are at the basis set limit for valence angular momentum ($l = 0$ and 1). However, calculations using UGBS are very computationally demanding. In this paper, we present a systematic approach for expanding standard valence-oriented basis sets in order to compute spin–spin couplings without large errors arising from the basis set.

2. Theory and Computational Aspects

A variety of molecules containing hydrogen and atoms from the first two rows are studied in our work. All molecules are first optimized at the B3LYP/6-31G* level of theory, using the development version of the Gaussian computational program.²⁴ Because there are very few spin–spin coupling experimental values obtained in the gas phase, and the goal of this study is to reduce basis set errors, the results of the smaller basis sets derived in this study will be compared to results from very large basis sets for calibration.

The uncontracted UGBS2P basis sets are chosen to be the reference basis sets. UGBS2P includes two additional polarization functions for each function in the UGBS: one p and d function for each s function and one d and f function for each p function. Preliminary study showed that adding two tighter s functions than the tightest s functions in UGBS2P affected the results by less than 1%, compared to the unmodified UGBS2P basis. Therefore, the UGBS2P basis set was used as a reference in this study, and it was not considered necessary to examine the addition of functions tighter than those in UGBS2P to any of the smaller basis sets.

Our approach is to derive a basis set suitable for computation of the FC term from the original basis set, while using the original basis as-is for the rest of the terms in the spin–spin coupling. We uncontract the original basis set and then

Table 1. uTZ-Derived Basis Set Size and Maximum Number of Tighter s Functions To Be Added

atoms	number of basis functions in UGBS2P	number of basis functions in uTZ basis set	maximum number of tight s functions added
H	60	11	5
C	68		2
N	71	22	2
O	71		2
F	72		3
Si			2
P	84	31	3
S			2
Cl			2

add additional tight functions until the basis set error in the FC term is comparable to that of the other terms.

The added tight s functions have even-tempered exponents starting from the tightest s functions in the small basis set. For hydrogen and first row atoms, a ratio of 3 for successive exponents was used, while for second row atoms, a ratio of 2 was applied.

In section 3, we test this approach using the aug-cc-pVTZ basis sets²⁵ and determine how many and what type of functions should be added for the FC term. In section 4, we apply the rules developed in section 3 to two smaller basis sets, aug-cc-pVDZ and 6-311+G(d,p). In section 5, we compare our results to previous work on this problem.

3. Addition of Functions to aug-cc-pVTZ

3.1. Notation. We denote the uncontracted aug-cc-pVTZ basis by uTZ and use uTZ- sn to denote uTZ augmented by n tight s functions. uTZ- w will denote the uTZ basis augmented by the number of s functions found to be sufficient to saturate the core region. uTZ- wdn will denote a uTZ- w basis augmented by n d functions. A preliminary study indicated that tight p functions had virtually no effect on the FC term, and these are not considered further here.

The basis set sizes of the UGBS2P and uTZ- sn series for the first and second row atoms involved in this study are summarized in Table 1. It can be seen that the uTZ- sn sets are much smaller in size than the UGBS2P. The table also shows the number of s functions that could be added to the uTZ basis set before reaching the tightest s functions in UGBS2P.

3.2. H and First Row Atoms: HF, NH₃, HCN. The results for one-bond coupling $^1J(\text{H}^{19}\text{F})$ calculations in the HF molecule are shown in Table 2. The SD, PSO, and DSO terms are calculated using the contracted aug-cc-pVTZ basis set and remain unchanged for all uTZ- sn calculations. Compared to the UGBS2P, the PSO term contributes 1.5% of the error to the total contribution. In this and later tables, the optimal choice of ns is given in bold face.

As can be seen in Table 2, the FC terms increases in a regular pattern as s functions are added. Also, the third tighter s functions on fluorine and the fifth tighter s functions on hydrogen only have minor effects on the results. Therefore,

Table 2. $^1J(^1\text{H}^{19}\text{F})$ Coupling in the HF Molecule Evaluated Using the uTZ-*sn* Basis Set (Hz)

		F uTZ- <i>sn</i>				UGBS2P
		<i>n</i> = 0	1	2	3	
H uTZ- <i>sn</i>	<i>n</i> = 0	167.78 ^a	169.29	171.97	172.32	
		361.71 ^b	363.22	365.90	366.26	
	1	173.27	174.59	177.53	177.78	
		367.21	368.52	371.46	371.71	
	2	180.10	181.65	184.59	184.92	FC: 192.47
		374.04	375.58	378.52	378.86	SD: -3.14
	3	181.67	183.12	186.16	186.44	PSO: 201.31
		375.61	377.05	380.09	380.38	DSO: -0.20
	4	184.12	185.67	188.70	189.02	total: 390.44
		378.06	379.60	382.63	382.96	
5	184.49	185.99	189.06	189.36		
	378.43	379.92	382.99	383.30		

^a Fermi contact contribution. ^b Total spin-spin coupling. SD, PSO, and DSO are respectively -1.41, 195.10, and 0.25 (calculated using contracted aug-cc-pVTZ basis set). Expt.: 412–479 Hz (in various solvents).²⁷

Table 3. $^1J(^{14}\text{N}^1\text{H})$ Coupling in the NH₃ Molecule Evaluated Using the uTZ-*sn* Basis Set (Hz)

		N uTZ- <i>sn</i>			UGBS2P
		<i>n</i> = 0	1	2	
H uTZ- <i>sn</i>	<i>n</i> = 0	36.05 ^a	36.42	36.98	
		38.57 ^b	38.94	39.50	
	1	37.12	37.50	38.08	
		39.64	40.02	40.60	
	2	38.71	39.11	39.71	FC: 41.05
		41.23	41.63	42.23	SD: 0.17
	3	38.99	39.39	40.00	PSO: 2.34
		41.51	41.91	42.52	DSO: 0.04
	4	39.56	39.97	40.58	total: 43.60
		42.08	42.49	43.10	
5	39.62	40.02	40.64		
	42.14	42.54	43.16		

^a Fermi contact contribution. ^b Total spin-spin coupling. SD, PSO, and DSO are respectively 0.17, 2.30, and 0.05 (calculated using contracted aug-cc-pVTZ basis set). Expt.: 40 ± 1 Hz (neat liquid).²⁸

uTZ-s2 for fluorine and uTZ-s4 for hydrogen should sufficiently simulate the UGBS2P results.

This combination of a uTZ basis set (uTZ-s4 for hydrogen and uTZ-s2 for first row atoms) is named the “uTZ-w” basis set. Compared to the UGBS2P calculation, the uTZ-w basis set gave 2.0% error in the FC term and 2.1% error in the total contribution.

The results for one-bond coupling calculations in the NH₃ and HCN molecules and two-bond coupling $^2J(^1\text{H}^{14}\text{N})$ in HCN are shown in Tables 3–6. [The two-bond $^2J(\text{HH})$ couplings of NH₃ are listed in the next section.]

As with HF, the uTZ-w basis set (uTZ-s4 for hydrogen and uTZ-s2 for first row atoms) produces satisfactory results for NH₃ and HCN. The size of this uTZ-*sn* series and aug-cc-pVTZ is significantly smaller than that of UGBS2P, but the relative errors of total contribution obtained from the uTZ-w basis set and UGBS2P are small. In the case of

Table 4. $^1J(^{13}\text{C}^1\text{H})$ Coupling in the HCN Molecule Evaluated Using the uTZ-*sn* Basis Set (Hz)

		C uTZ- <i>sn</i>			UGBS2P
		<i>n</i> = 0	1	2	
H uTZ- <i>sn</i>	<i>n</i> = 0	248.21 ^a	250.86	254.70	
		248.52 ^b	251.17	255.01	
	1	255.44	258.18	262.13	
		255.75	258.49	262.44	
	2	266.53	269.38	273.50	FC: 283.08
		266.84	269.69	273.81	SD: 0.53
	3	268.36	271.24	275.39	PSO: -0.75
		268.67	271.55	275.70	DSO: 0.39
	4	272.38	275.30	279.51	total: 283.24
		272.69	275.61	279.82	
5	272.70	275.62	279.84		
	273.01	275.93	280.15		

^a Fermi contact contribution. ^b Total spin-spin coupling. Adding tight functions to N has no effect on $^1J(\text{CH})$ coupling calculation in HCN. SD, PSO, and DSO are respectively 0.61, -0.73, and 0.43 (calculated using contracted aug-cc-pVTZ basis set). Expt.: 261–274 Hz (in various solvents).²⁹

Table 5. $^1J(^{13}\text{C}^{14}\text{N})$ Coupling in the HCN Molecule Evaluated Using the uTZ-*sn* Basis Set (Hz)

		C uTZ- <i>sn</i>			UGBS2P
		<i>n</i> = 0	1	2	
N uTZ- <i>sn</i>	<i>n</i> = 0	6.30 ^a	6.37	6.46	
		11.69 ^b	11.76	11.85	FC: 6.76
	1	6.36	6.43	6.52	SD: 5.40
		11.75	11.82	11.91	PSO: 0.59
	2	6.46	6.53	6.63	DSO: -0.03
		11.85	11.92	12.02	total: 12.73

^a Fermi contact contribution. ^b Total spin-spin coupling. Adding tight functions to H has no effect on $^1J(\text{CN})$ calculation in HCN. SD, PSO, and DSO are respectively 4.77, 0.65, and -0.03 (calculated using contracted aug-cc-pVTZ basis set). Expt.: 26.4 Hz (neat liquid).²⁹ (The experimental spin-spin couplings have been converted from ^{15}N to ^{14}N , which is, within the Born–Oppenheimer approximation, related by the ratio of the nuclear magnetogyric ratios only if vibrational corrections are neglected.)

$^1J(^{13}\text{C}^{14}\text{N})$ and $^2J(^1\text{H}^{14}\text{N})$ of HCN, the relative error is a little higher than 5%, but the absolute error is within 0.8 Hz.

The SD, PSO, and DSO terms of these couplings calculated using the aug-cc-pVTZ basis set are also very close to the results of UGBS2P; the absolute errors are all within 0.7 Hz.

The study of molecules of hydrogen and the first row atoms shows that, compared to UGBS2P, the uTZ-*sn* basis set takes much less computation time for NMR spin-spin coupling constants, yet a specific combination of the uTZ-*sn* series (uTZ-w) can simulate UGBS2P’s results accurately.

3.3. H and Second Row Atoms: SiH₄, PH₃, and H₂S.

All of the $^1J(\text{XH})$ (X = Si, P, and S) coupling calculations are shown in Tables 7–9. In the three molecules, the SD, PSO, and DSO terms of $^1J(\text{XH})$ couplings are significantly smaller than the FC term. These three terms of the aug-cc-pVTZ basis set results are very close to the UGBS2P results. Hence, the basis set errors arise primarily from the FC term. Note that, for second row atoms, the added tight s functions

Table 6. ${}^2J({}^{14}\text{N}^1\text{H})$ Coupling in the HCN Molecule Evaluated Using the uTZ-*sn* Basis Set (Hz)

		N uTZ- <i>sn</i>			UGBS2P
		<i>n</i> = 0	1	2	
H uTZ- <i>sn</i>	<i>n</i> = 0	2.03 ^a	2.05	2.09	
		4.76 ^b	4.78	4.82	
	1	2.09	2.11	2.15	
		4.82	4.84	4.88	
	2	2.19	2.21	2.24	FC: 2.31
		4.92	4.94	4.97	SD: 0.65
	3	2.20	2.22	2.25	PSO: 2.77
		4.93	4.95	4.98	DSO: -0.44
	4	2.23	2.25	2.29	total: 5.29
		4.96	4.98	5.02	
5	2.23	2.26	2.29		
	4.96	4.99	5.02		

^a Fermi contact contribution. ^b Total spin–spin coupling. Adding tight functions to C has no effect on ${}^2J({}^{14}\text{N}^1\text{H})$ calculation in HCN. SD, PSO, and DSO are respectively 0.60, 2.56, and -0.43 (calculated using contracted aug-cc-pVTZ basis set). Expt.: 10.5–12.4 Hz (in various solvents).²⁹ (The experimental spin–spin couplings have been converted from ${}^{15}\text{N}$ to ${}^{14}\text{N}$, which is, within the Born–Oppenheimer approximation, related by the ratio of the nuclear magnetogyric ratios only if rovibrational corrections are neglected.)

Table 7. ${}^1J({}^{29}\text{Si}^1\text{H})$ Coupling in the SiH₄ Molecule Evaluated Using the uTZ-*sn* Basis Set (Hz)

		Si uTZ- <i>sn</i>			UGBS2P
		<i>n</i> = 0	1	2	
H uTZ- <i>sn</i>	<i>n</i> = 0	-188.82 ^a	-189.23	-190.42	
		-188.22 ^b	-188.63	-189.82	
	1	-194.65	-195.07	-196.30	
		-194.05	-194.47	-195.70	
	2	-202.87	-203.31	-204.59	FC: -210.04
		-202.27	-202.71	-203.99	SD: -0.17
	3	-204.42	-204.87	-206.15	PSO: 0.32
		-203.82	-204.27	-205.55	DSO: -0.02
	4	-207.38	-207.83	-209.13	total: -209.91
		-206.78	-207.23	-208.53	
5	-207.69	-208.14	-209.44		
	-207.09	-207.54	-208.84		

^a Fermi contact contribution. ^b Total spin–spin coupling. SD, PSO, and DSO are respectively 0.07, 0.55, and -0.02 (calculated using contracted aug-cc-pVTZ basis set). Expt.: -202.5 ± 0.2 Hz (gas phase).³⁰

have a ratio of 2 for successive exponents starting from the tightest *s* functions in uTZ basis sets. As for Si, a third additional *s* function would be tighter than the tightest functions in the reference basis, so we only considered adding two *s* functions. For S and P, a third *s* function could be tested.

As seen in the results, uTZ-*s*3 on P and S only gave a marginal improvement to the results compared to uTZ-*s*2, as was the case for uTZ-*s*5 on hydrogen to uTZ-*s*4. Therefore, uTZ-*s*2 on second row atoms and uTZ-*s*4 on hydrogen produced reasonable results for the uTZ-*sn* series. All of the ${}^1J(\text{XH})$ coupling calculations using this combination are within 2% accuracy to UGBS2P.

All ${}^2J(\text{HH})$ coupling calculation results (including NH₃) are listed in Table 10. Adding tighter *s* functions to the basis set on the center atoms does not affect the ${}^2J(\text{HH})$ calcula-

Table 8. ${}^1J({}^{31}\text{P}^1\text{H})$ Coupling in the PH₃ Molecule Evaluated Using the UTZ-*sn* Basis Set (Hz)

		P uTZ- <i>sn</i>				UGBS2P
		<i>n</i> = 0	1	2	3	
H uTZ- <i>sn</i>	<i>n</i> = 0	141.45 ^a	141.74	142.62	142.73	
		147.13 ^b	147.42	148.30	148.41	
	1	145.66	145.96	146.86	146.98	
		151.34	151.64	152.54	152.66	
	2	151.89	152.20	153.14	153.26	FC: 158.43
		157.57	157.88	158.82	158.94	SD: -0.82
	3	152.99	153.30	154.25	154.37	PSO: 6.35
		158.67	158.98	159.93	160.05	DSO: -0.01
	4	155.24	155.55	156.52	156.64	total: 163.96
		160.92	161.23	162.20	162.32	
5	155.44	155.75	156.72	156.85		
	161.12	161.43	162.40	162.53		

^a Fermi contact contribution. ^b Total spin–spin coupling. SD, PSO, and DSO are respectively -1.26, 6.92, and 0.02 (calculated using contracted aug-cc-pVTZ basis set). Expt.: 188.7 Hz (in complex solution)³¹ and 182.2 ± 0.3 Hz (neat liquid).³²

Table 9. ${}^1J({}^{33}\text{S}^1\text{H})$ Coupling in the H₂S Molecule Evaluated Using the uTZ-*sn* Basis Set (Hz)

		S uTZ- <i>sn</i>				UGBS2P
		<i>n</i> = 0	1	2	3 (over)	
H uTZ- <i>sn</i>	<i>n</i> = 0	17.07 ^a	17.11	17.22	17.24	
		21.57 ^b	21.61	21.72	21.74	
	1	17.51	17.54	17.66	17.67	
		22.01	22.04	22.16	22.17	
	2	18.30	18.34	18.46	18.48	FC: 19.25
		22.80	22.84	22.96	22.98	SD: -0.06
	3	18.40	18.44	18.56	18.58	PSO: 4.68
		22.90	22.94	23.06	23.08	DSO: -0.01
	4	18.69	18.73	18.85	18.87	total: 23.85
		23.19	23.23	23.35	23.37	
5	18.70	18.74	18.86	18.88		
	23.20	23.24	23.36	23.38		

^a Fermi contact contribution. ^b Total spin–spin coupling. SD, PSO, and DSO are respectively -0.18, 4.68, and 0.00 (calculated using contracted aug-cc-pVTZ basis set).

tions. The fifth tight *s* function on hydrogen is unnecessary, as uTZ-*s*5 on hydrogen gives very close total contribution results to those of uTZ-*s*4.

3.4. Second Row and First Row Atoms: SiF₄, PF₃, SF₆, and PCl₃. This section extends the basis set dependence study to molecules containing both first and second row atoms. Five molecules, SiF₄, PF₃, SF₆, Cl₂O, and PCl₃, are studied. The results of ${}^1J(\text{X}^{19}\text{F})$ (X = ${}^{29}\text{Si}$, ${}^{31}\text{P}$, and ${}^{33}\text{S}$) are shown in Tables 11–13, ${}^2J({}^{19}\text{F}^{19}\text{F})$ in Table 14, ${}^1J({}^{31}\text{P}^{35}\text{Cl})$ in Table 15, and ${}^2J({}^{35}\text{Cl}^{35}\text{Cl})$ in Table S1 in the Supporting Information.

For ${}^1J(\text{X}^{19}\text{F})$ coupling, as tight *s* functions are added to both F and second row atoms, the FC term increments are similar to those of previous results. The SD, PSO, and DSO terms are taken from the unmodified contracted aug-cc-pVTZ basis set results. Among the three terms, PSOs in three ${}^1J(\text{X}^{19}\text{F})$ couplings carry the largest error but are still below

Table 10. $^2J(^1H^1H)$ Coupling in NH_3 , SiH_4 , PH_3 , and H_2S Molecules Evaluated Using the uTZ-*sn* Basis Set (Hz)^a

		NH_3	SiH_4	PH_3	H_2S
H uTZ- <i>sn</i>	$n = 0$	-10.24 ^b	3.81	-10.74	-10.49
		-9.10 ^c	2.64	-10.83	-10.13
	1	-10.86	4.04	-11.37	-11.07
		-9.72	2.87	-11.46	-10.71
	2	-11.80	4.40	-12.38	-12.08
		-10.66	3.23	-12.47	-11.72
	3	-11.98	4.46	-12.55	-12.23
		-10.84	3.29	-12.64	-11.87
	4	-12.33	4.60	-12.93	-12.61
		-11.19	3.43	-13.02	-12.25
	5	-12.36	4.61	-12.96	-12.63
		-11.22	3.44	-13.05	-12.27
	SD	0.64	0.08	0.12	0.12
	PSO	5.50	1.09	1.18	2.08
	DSO	-5.00	-2.34	-1.39	-1.84
	FC	-12.51	4.92	-12.87	-12.48
SD	0.67	0.07	0.11	0.11	
UGBS2P	PSO	6.17	2.36	1.57	2.51
	DSO	-5.04	-2.35	-1.40	-1.86
	total	-10.71	5.00	-12.59	-11.71
expt.		-10.35 ± 0.80^d	2.75 ± 0.15^e	-13.2 ± 0.7^f	

^a Adding tight functions to N, Si, P, and S has no effect on $^2J(HH)$ coupling calculation. SD, PSO, and DSO remain unchanged (calculated using contracted aug-cc-pVTZ basis set). ^b Fermi contact contribution. ^c Total spin-spin coupling. ^d Neat liquid.³³ ^e Gas phase.³⁰ ^f Neat liquid.³²

Table 11. $^1J(^{29}Si^{19}F)$ Coupling in the SiF_4 Molecule Evaluated Using the uTZ-*sn* Basis Set (Hz)

		Si uTZ- <i>sn</i>			UGBS2P
		$n = 0$	1	2	
F uTZ- <i>sn</i>	$n = 0$	262.79 ^a	263.36	265.02	
		339.27 ^b	339.84	341.50	
	1	265.77	266.35	268.02	FC: 262.41
		342.25	342.83	344.50	SD: -4.45
	2	269.56	270.15	271.84	PSO: 85.67
		346.04	346.63	348.32	DSO: -0.55
	3	270.37	270.96	272.67	total: 343.08
		346.85	347.44	349.15	

^a Fermi contact contribution. ^b Total spin-spin coupling. SD, PSO, and DSO are respectively -4.00, 81.05, and -0.57 (calculated using contracted aug-cc-pVTZ basis set).

1.5%. The uTZ-*w* basis set shows good agreement with UGBS2P in total contribution calculation in PF_3 and SF_6 .

All $^2J(^{19}F^{19}F)$ coupling calculations are listed Table 14. There are two types of $^2J(^{19}F^{19}F)$ in SF_6 ; either the F-S-F is 90 or 180°. As seen in the table, $^2J(^{19}F^{19}F)$'s in SiF_4 , PF_3 , and SF_6 (90°) have similar values. As expected, the FC term increment percentages of all $^2J(^{19}F^{19}F)$'s are approximately the same and twice the $^1J(X^{19}F)$ FC increment of the uTZ-*sn* series on the F atom.

The uTZ-*w* basis set yields fairly similar values to the UGBS2P values of $^2J(^{19}F^{19}F)$ coupling in SiF_4 and SF_6 (90°). However, nontrivial errors remain in $^2J(FF)$ calculations of PF_3 and SF_6 (180°), which are 3.92 and 5.75 Hz, respectively, and 9.5% and 15% in relative error. In the case of PF_3 , this

Table 12. $^1J(^{31}P^{19}F)$ Coupling in the PF_3 Molecule Evaluated Using the uTZ-*sn* Basis Set (Hz)

		P uTZ- <i>sn</i>				UGBS2P
		$n = 0$	1	2	3	
F uTZ- <i>sn</i>	$n = 0$	-1261 ^a	-1264	-1272	-1273	
		-1524 ^b	-1527	-1535	-1536	
	1	-1274	-1277	-1285	-1286	FC: -1312
		-1537	-1540	-1548	-1549	SD: 38.78
	2	-1293	-1296	-1304	-1305	PSO: -322.38
		-1556	-1559	-1567	-1568	DSO: 0.76
	3	-1297	-1300	-1307	-1309	total: -1595
		-1560	-1563	-1570	-1572	

^a Fermi contact contribution. ^b Total spin-spin coupling. SD, PSO, and DSO are respectively 36.00, -299.57, and 0.85 (calculated using contracted aug-cc-pVTZ basis set). Expt.: -1441 Hz (neat liquid).³⁴

Table 13. $^1J(^{33}S^{19}F)$ Coupling in the SF_6 Molecule Evaluated Using the uTZ-*sn* Basis Set (Hz)

		S uTZ- <i>sn</i>				UGBS2P
		$n = 0$	1	2	3	
F uTZ- <i>sn</i>	$n = 0$	-284.37 ^a	-284.95	-286.63	-286.85	
		-306.91 ^b	-307.49	-309.17	-309.39	
	1	-287.41	-288.00	-289.70	-289.92	FC: -296.09
		-309.95	-310.54	-312.24	-312.46	SD: 9.99
	2	-291.63	-292.22	-293.95	-294.17	PSO: -34.75
		-314.17	-314.76	-316.49	-316.71	DSO: 0.39
	3	-292.43	-293.02	-294.75	-294.98	total: -320.46
		-314.97	-315.56	-317.29	-317.52	

^a Fermi contact contribution. ^b Total spin-spin coupling. SD, PSO, and DSO are respectively 9.17, -32.10, and 0.39 (calculated using contracted aug-cc-pVTZ basis set). Expt.: -250.1 Hz (gas phase).³⁵

Table 14. $^2J(^{19}F^{19}F)$ Coupling in SiF_4 , PF_3 , and SF_6 Molecules Evaluated Using the uTZ-*sn* Basis Set (Hz)

		SiF_4	PF_3	$SF_6(90^\circ)$	$SF_6(180^\circ)$	
F uTZ- <i>sn</i>	$n = 0$	-59.29 ^a	-50.57	-66.84	-21.27	
		-155.27 ^b	-34.72	-265.36	-42.81	
	1	-60.63	-51.83	-68.39	-21.75	
		-156.61	-35.98	-266.91	-43.29	
	2	-62.38	-53.24	-70.33	-22.38	
		-158.36	-37.39	-268.85	-43.92	
	3	-62.75	-53.62	-70.77	-22.51	
		-158.73	-37.77	-269.29	-44.05	
	SD	16.46	59.81	57.05	-7.90	
	PSO	-111.23	-43.26	-256.20	-9.72	
	DSO	-1.21	-0.70	0.63	-3.92	
	FC	-61.30	-54.17	-67.57	-18.19	
	SD	16.81	62.97	61.79	-8.47	
	UGBS2P	PSO	-115.49	-49.35	-267.89	-7.54
		DSO	-1.26	-0.76	0.57	-3.97
	total	-161.24	-41.31	-273.10	-38.17	

^a Fermi contact contribution. ^b Total spin-spin coupling. SD, PSO, and DSO remain unchanged (calculated using contracted aug-cc-pVTZ basis set).

is due to the basis set error in the aug-cc-pVTZ basis for the PSO and DSO terms.

The uTZ-*w* basis sets for the $^1J(^{31}P^{35}Cl)$ couplings in PCl_3 produce very good agreement with UGBS2p. Adding tight s functions to uTZ basis set had very little effect on the small two-bond $^2J(ClCl)$ calculations in PCl_3 .

Table 15. $^1J(^{31}\text{P}^{35}\text{Cl})$ Coupling in the PCl_3 Molecule Evaluated Using the $\text{uTZ-}sn$ Basis Set (Hz)

		P uTZ- <i>sn</i>				UGBS2P
		<i>n</i> = 0	1	2	3	
Cl uTZ- <i>sn</i>	<i>n</i> = 0	−100.72 ^a	−100.93	−101.55	−101.63	FC: −103.25
		−125.95 ^b	−126.16	−126.78	−126.86	SD: 11.06
	1	−100.91	−101.12	−101.74	−101.82	PSO: −37.54
		−126.14	−126.35	−126.97	−127.05	DSO: 0.05
	2	−101.49	−101.71	−102.32	−102.41	total: −129.67
		−126.72	−126.94	−127.55	−127.64	

^a Fermi contact contribution. ^b Total spin–spin coupling. SD, PSO, and DSO are respectively 9.86, −35.15, and 0.06 (calculated using contracted aug-cc-pVTZ basis set). Expt. $^1J(^{31}\text{P}^{35}\text{Cl})$ −120 Hz (neat liquid).³⁶

In some cases, when compared to the uTZ basis set, uTZ-*w* produced poorer results for the FC terms, such as $^1J(\text{SiF})$ in SiF_4 and both $^1J(\text{SF})$'s in SF_6 . In other words, adding tight s functions moves the FC term away from the UGBS2P results. Preliminary results demonstrated that this is because tight d functions are also important in these cases. Therefore, we generated uTZ-*wdn* basis sets by adding more d functions to the uTZ-*w* basis set. This problem only occurs in systems containing second row atoms; these additional d functions are tested for these atoms. The additional d functions have progressive exponents of 2 with reference to the tightest d function in the uTZ basis for that atom.

3.5. Overall Results for uTZ-Based Basis Sets. Table 16 lists all of the couplings in this study and the results of FC term and total contributions using UGBS2P, contracted

aug-cc-pVTZ, uTZ, uTZ-*w*, uTZ-*wd2*, and uTZ-*wd4* basis sets. The average absolute error (AAE) and maximum absolute error (MAE) are also listed. Table 17 shows the relative errors of the same coupling calculations using these basis sets and includes the average relative error (ARE) and maximum relative error.

Both absolute and relative errors are shown because each evaluates the calculation results from a different perspective. In some cases, the relative error is fairly high, while the actual difference between UGBS2P and a small basis set can be as low as 0.25 Hz [e.g., $^2J(^{35}\text{Cl}^{35}\text{Cl})$ in PCl_3]. In other cases, the absolute error is large [e.g., $^1J(\text{PF})$ of PF_3], but relative error is consistent with the other results.

Because the SD, PSO, and DSO terms are calculated using the unmodified aug-cc-pVTZ basis set, they remain unchanged for all aug-cc-pVTZ-derived basis calculations. The extended uTZ basis sets improve the total spin–spin coupling by improving the FC term. It is interesting to note that, in some cases, the FC term is improved but the total contribution is worsened. This is due to the cancellation of basis set errors between the FC term and the others. For some cases, in which the errors are modest in size and opposite in sign, improving just the FC term makes the overall results slightly worse.

From the tables, it can be seen that the unchanged aug-cc-pVTZ has very poor agreement with UGBS2P, with 23.65 Hz in AAE and 17.94% in ARE in total spin–spin coupling. The uTZ basis shows minor improvement yet still produces inadequate results, with 11.21 Hz in AAE and 10.77% in ARE. With tighter s functions added, the uTZ-*w* basis set

Table 16. Absolute Values of FC Term and Total Spin–Spin Coupling Calculations Using Different Basis Set Series (Hz)

		UGBS2P		contracted TZ		uTZ		uTZ- <i>w</i>		uTZ- <i>wd2</i>		uTZ- <i>wd4</i>	
		FC	total	FC	total	FC	total	FC	total	FC	total	FC	total
HF	$^1J(\text{HF})$	192.47	390.44	195.80	389.74	167.78	361.71	188.70	382.63	188.70	382.63	188.70	382.63
NH ₃	$^1J(\text{NH})$	41.05	43.60	36.57	39.09	36.05	38.57	40.58	43.10	40.58	43.10	40.58	43.10
	$^2J(\text{HH})$	−12.51	−10.71	−11.16	−10.02	−10.24	−9.10	−12.33	−11.19	−12.33	−11.19	−12.33	−11.19
HCN	$^1J(\text{HC})$	283.08	283.24	273.70	274.01	248.21	248.52	279.51	279.82	279.51	279.82	279.51	279.82
	$^2J(\text{HN})$	2.31	5.29	4.61	7.34	2.03	4.76	2.29	5.02	2.29	5.02	2.29	5.02
SiH ₄	$^1J(\text{SiH})$	−210.04	−209.91	−159.44	−158.85	−188.82	−188.22	−209.13	−208.53	−208.56	−207.96	−208.50	−207.90
	$^2J(\text{HH})$	4.92	5.00	3.33	2.17	3.81	2.64	4.60	3.43	4.84	3.67	4.88	3.71
PH ₃	$^1J(\text{PH})$	158.43	163.96	115.03	120.71	141.45	147.13	156.52	162.20	157.19	162.87	157.14	162.82
	$^2J(\text{HH})$	−12.87	−12.59	−10.51	−10.60	−10.74	−10.83	−12.93	−13.02	−12.67	−12.76	−12.63	−12.72
H ₂ S	$^1J(\text{SH})$	19.25	23.85	15.13	19.63	17.07	21.57	18.85	23.35	19.15	23.65	19.16	23.65
	$^2J(\text{HH})$	−12.48	−11.71	−10.15	−9.78	−10.49	−10.13	−12.61	−12.25	−12.31	−11.94	−12.26	−11.90
HCN	$^1J(\text{CN})$	6.76	12.73	6.70	12.10	6.30	11.69	6.63	12.02	6.63	12.02	6.63	12.02
SiF ₄	$^1J(\text{SiF})$	262.41	343.08	164.76	241.24	262.79	339.27	271.84	348.32	264.96	341.44	264.14	340.62
	$^2J(\text{FF})$	−61.30	−161.24	−70.52	−166.48	−59.29	−155.27	−62.38	−158.36	−61.12	−157.09	−60.97	−156.94
PF ₃	$^1J(\text{PF})$	−1312	−1595	−1165	−1427	−1261	−1524	−1304	−1567	−1301	−1564	−1300	−1563
	$^2J(\text{FF})$	−54.17	−41.31	−79.08	−63.24	−50.57	−34.72	−53.24	−37.39	−53.38	−37.54	−53.42	−37.57
SF ₆	$^1J(\text{SF})$	−296.09	−320.46	−249.95	−272.49	−284.37	−306.91	−293.95	−316.49	−293.46	−316.00	−293.32	−315.86
	$^2J(\text{FF})$ 90°	−67.57	−273.10	−84.64	−283.15	−66.84	−265.36	−70.33	−268.85	−66.98	−265.49	−66.54	−265.06
	$^2J(\text{FF})$ 180°	−18.19	−38.17	−19.06	−40.61	−21.27	−42.81	−22.38	−43.92	−18.45	−40.00	−17.98	−39.53
PCl ₃	$^1J(\text{ClP})$	−103.25	−129.67	−119.91	−145.14	−100.72	−125.95	−102.32	−127.55	−102.97	−128.20	−102.98	−128.21
	$^2J(\text{ClCl})$	−0.07	2.58	0.75	3.15	−0.06	2.33	−0.06	2.33	−0.07	2.32	−0.07	2.32
average absolute error				23.13	23.65	8.96	11.21	1.97	3.61	1.41	3.54	1.46	3.64
maximum absolute error				147.00	168.00	51.00	71.00	9.43	28.00	11.00	31.00	12.00	32.00
average H coupling absolute error				11.39	11.13	10.25	10.65	1.07	1.70	1.03	1.59	1.04	1.59
average no H coupling absolute error				36.04	37.41	7.55	11.83	2.96	5.71	1.84	5.69	1.92	5.89

Table 17. Relative Errors of the FC Term and Total Spin–Spin Coupling Calculations Using Different Basis Set Series

		UGBS2P		contracted TZ		uTZ		uTZ-w		uTZ-wd2		uTZ-wd4	
		FC (Hz)	total (Hz)	FC error	total error	FC error	total error	FC error	total error	FC error	total error	FC error	total error
HF	$^1J(\text{HF})$	192.47	390.44	1.73%	0.18%	12.83%	7.36%	1.96%	2.00%	1.96%	2.00%	1.96%	2.00%
NH ₃	$^1J(\text{NH})$	41.05	43.60	10.91%	10.34%	12.18%	11.54%	1.14%	1.15%	1.14%	1.15%	1.14%	1.15%
	$^2J(\text{HH})$	-12.51	-10.71	10.79%	6.44%	18.15%	15.03%	1.44%	4.48%	1.44%	4.48%	1.44%	4.48%
HCN	$^1J(\text{HC})$	283.08	283.24	3.31%	3.26%	12.32%	12.26%	1.26%	1.21%	1.26%	1.21%	1.26%	1.21%
	$^2J(\text{HN})$	2.31	5.29	99.57%	38.75%	12.12%	10.02%	0.87%	5.10%	0.87%	5.10%	0.87%	5.10%
SiH ₄	$^1J(\text{SiH})$	-210.04	-209.91	24.09%	24.32%	10.10%	10.33%	0.43%	0.66%	0.70%	0.93%	0.73%	0.96%
	$^2J(\text{HH})$	4.92	5.00	32.32%	56.60%	22.56%	47.20%	6.50%	31.40%	1.63%	26.60%	0.81%	25.80%
PH ₃	$^1J(\text{PH})$	158.43	163.96	27.39%	26.38%	10.72%	10.26%	1.21%	1.07%	0.78%	0.66%	0.81%	0.70%
	$^2J(\text{HH})$	-12.87	-12.59	18.34%	15.81%	16.55%	13.98%	0.47%	3.42%	1.55%	1.35%	1.86%	1.03%
H ₂ S	$^1J(\text{SH})$	19.25	23.85	21.40%	17.69%	11.32%	9.56%	2.08%	2.10%	0.52%	0.84%	0.47%	0.84%
	$^2J(\text{HH})$	-12.48	-11.71	18.67%	16.48%	15.95%	13.49%	1.04%	4.61%	1.36%	1.98%	1.76%	1.62%
HCN	$^1J(\text{CN})$	6.76	12.73	0.89%	4.95%	6.80%	8.17%	1.92%	5.58%	1.92%	5.58%	1.92%	5.58%
SiF ₄	$^1J(\text{SiF})$	262.41	343.08	37.21%	29.68%	0.14%	1.11%	3.59%	1.53%	0.97%	0.48%	0.66%	0.72%
	$^2J(\text{FF})$	-61.30	-161.24	15.04%	3.25%	3.28%	3.70%	1.76%	1.79%	0.29%	2.57%	0.54%	2.67%
PF ₃	$^1J(\text{PF})$	-1312	-1595	11.20%	10.53%	3.89%	4.45%	0.61%	1.76%	0.84%	1.94%	0.91%	2.01%
	$^2J(\text{FF})$	-54.17	-41.31	45.98%	53.09%	6.65%	15.95%	1.72%	9.49%	1.46%	9.13%	1.38%	9.05%
SF ₆	$^1J(\text{SF})$	-296.09	-320.46	15.58%	14.97%	3.96%	4.23%	0.72%	1.24%	0.89%	1.39%	0.94%	1.44%
	$^2J(\text{FF})$ 90°	-67.57	-273.10	25.26%	3.68%	1.08%	2.83%	4.08%	1.56%	0.87%	2.79%	1.52%	2.94%
	$^2J(\text{FF})$ 180°	-18.19	-38.17	4.78%	6.39%	16.93%	12.16%	23.03%	15.06%	1.43%	4.79%	1.15%	3.56%
PCl ₃	$^1J(\text{ClP})$	-103.25	-129.67	16.14%	11.93%	2.45%	2.87%	0.90%	1.63%	0.27%	1.13%	0.26%	1.13%
	$^2J(\text{ClCl})$	-0.07	2.58	1171.43%	22.09%	14.29%	9.69%	14.29%	9.69%	0.00%	10.08%	0.00%	10.08%
average relative error				76.76%	17.94%	10.20%	10.77%	3.38%	5.07%	1.06%	4.10%	1.07%	4.00%
maximum relative error				1171.43%	56.60%	22.56%	47.20%	23.03%	31.40%	1.96%	26.60%	1.96%	25.80%
average H coupling relative error				24.41%	19.66%	14.07%	14.64%	1.67%	5.20%	1.20%	4.21%	1.19%	4.08%
average no H coupling relative error				134.35%	16.06%	5.95%	6.52%	5.26%	4.93%	0.89%	3.99%	0.93%	3.92%

shows a decent improvement in total spin–spin coupling, with a 5.07% ARE and 3.61 Hz AAE compared to UGBS2P. The AAE of the FC term reduced 6.99 Hz from that of the uTZ basis set, and the AAE of total contribution reduced 7.60 Hz. The 0.6 Hz difference is from error cancellation. Going from the uTZ basis to uTZ-w produces a greater improvement in the couplings involving H atoms than in the couplings involving only heavy atoms.

For the contracted aug-cc-pVTZ basis set, the overall MAE occurs for the $^1J(\text{PF})$ in PF₃ and is 147 and 168 Hz, for the FC and total contribution, respectively. uTZ reduces these to 51 and 71 Hz (for FC and total, respectively). The uTZ-w basis again substantially improved these to 8 and 28 Hz, respectively, for the FC and total contribution. This makes the error in $^1J(\text{SiF})$ in SiF₄, 9.43 Hz, the MAE of the FC contribution for uTZ-w.

For a few couplings involving second row atoms, such as $^1J(\text{SiF})$ in SiF₄ and both $^2J(\text{FF})$'s in SF₆, adding tighter d functions made modest improvements in the FC term compared to uTZ-w. For both $^2J(\text{FF})$'s in SF₆, the absolute error reduced by about 2–4 Hz and the relative error was slightly reduced. In the case of $^2J(\text{FF})$ in SF₆ (90°), the FC term was improved while the total spin–spin coupling became slightly worse because of the cancellation of errors among four terms. The uTZ-wd2 and uTZ-wd4 basis sets also slightly improved the average error of the FC term for coupling without H atoms present: the AAE of coupling without H atoms, using the uTZ-wd2 basis set, improved by 1.12 Hz, and the ARE improved from 5.26% to 0.89%, compared to that in the uTZ-w basis set. However, the uTZ-

wdn basis sets provide very modest improvements to the overall results of spin–spin coupling calculations and, therefore, are not suggested by the authors for practical application.

4. Tests on Smaller Basis Sets

Two smaller basis sets, aug-cc-pVDZ and 6-311+G(d,p), have been tested using the same scheme.

Table 18 lists the results of all spin–spin couplings (the FC term and total coupling) using aug-cc-pVDZ-derived basis sets, compared to those using UGBS2P. The FC term was calculating using DZ (contracted aug-cc-pVDZ), uDZ, uTZ-w, and uTZ-wd2. The prefix “u” indicates an uncontracted basis set. The suffix “-w” stands for the same scheme of adding tight s functions as in uTZ-w, that is, four s functions to hydrogen with progressive exponents of 3, two s functions to first row atoms with exponents of 3, and two s functions to second row atoms with exponents of 2. The suffix “-wd2” means two tight d functions for second row atoms with progressive exponents of 2 are added, in addition to tight s functions. The SD, PSO, and DSO terms were calculated using the contracted aug-cc-pVDZ basis set. Table 19 shows the results for couplings using 6-311+G(d,p)-derived basis sets (uG, uG-w, etc.)

As can be seen from the AAEs and MAEs in the tables, the regular contracted basis sets produced generally poor results compared to those of UGBS2P. The uncontracted uDZ and uG basis sets both reduce the AAE to less than 20 Hz and MAE to less than 100 Hz for the total spin–spin

Table 18. FC Term and Total Spin–Spin Coupling Calculation Results Using aug-cc-pVDZ-Derived Basis Set (Hz)

		UGBS2P		aug-cc-pVDZ		uDZ		uDZ-w		uDZ-wd2	
		FC	total	FC	total	FC	total	FC	total	FC	total
HF	$^1J(\text{HF})$	192.47	390.44	406.18	579.05	149.24	322.11	181.80	354.67	181.80	354.67
NH ₃	$^1J(\text{NH})$	41.05	43.60	46.04	48.15	33.20	35.31	40.08	42.19	40.08	42.19
	$^2J(\text{HH})$	-12.51	-10.71	-12.12	-11.96	-8.76	-8.59	-11.98	-11.81	-11.98	-11.81
HCN	$^1J(\text{HC})$	283.08	283.24	312.25	312.39	230.30	230.44	277.67	277.81	277.67	277.81
	$^2J(\text{HN})$	2.31	5.29	0.66	2.63	1.92	3.89	2.30	4.27	2.30	4.27
SiH ₄	$^1J(\text{SiH})$	-210.04	-209.91	-174.39	-174.07	-178.00	-177.68	-211.02	-210.69	-208.32	-208.00
	$^2J(\text{HH})$	4.92	5.00	1.78	0.26	2.81	1.30	3.97	2.45	4.69	3.17
PH ₃	$^1J(\text{PH})$	158.43	163.96	138.34	141.34	134.12	137.11	158.84	161.84	158.16	161.16
	$^2J(\text{HH})$	-12.87	-12.59	-12.12	-12.36	-9.82	-10.06	-13.22	-13.46	-12.61	-12.85
H ₂ S	$^1J(\text{SH})$	19.25	23.85	13.75	17.42	16.33	20.00	19.40	23.07	19.69	23.36
	$^2J(\text{HH})$	-12.48	-11.71	-11.98	-11.86	-9.03	-8.90	-12.19	-12.07	-11.54	-11.42
HCN	$^1J(\text{CN})$	6.76	12.73	7.11	11.30	6.57	10.76	6.93	11.12	6.93	11.12
SiF ₄	$^1J(\text{SiF})$	262.41	343.08	165.32	228.55	297.02	360.26	310.56	373.79	286.16	349.39
	$^2J(\text{FF})$	-61.30	-161.24	-13.24	-102.31	-62.62	-151.69	-66.39	-155.46	-62.53	-151.59
PF ₃	$^1J(\text{PF})$	-1312	-1595	-1093	-1325	-1266	-1499	-1323	-1556	-1308	-1540
	$^2J(\text{FF})$	-54.17	-41.31	19.68	42.42	-48.36	-25.63	-51.33	-28.59	-52.31	-29.57
SF ₆	$^1J(\text{SF})$	-296.09	-320.46	-197.59	-218.56	-277.54	-298.51	-290.00	-310.97	-288.83	-309.79
	$^2J(\text{FF})$ 90°	-67.57	-273.10	-6.56	-194.83	-68.84	-257.13	-73.02	-261.30	-66.72	-255.01
	$^2J(\text{FF})$ 180°	-18.19	-38.17	-8.39	-26.99	-27.90	-46.50	-29.57	-48.17	-21.74	-40.34
PCl ₃	$^1J(\text{ClP})$	-103.25	-129.67	-79.25	-103.01	-93.02	-116.77	-95.79	-119.55	-97.03	-120.80
	$^2J(\text{ClCl})$	-0.07	2.58	1.02	3.37	0.05	2.40	0.05	2.40	0.04	2.39
average absolute error				45.16	49.70	14.46	19.27	5.64	8.74	3.35	8.41
maximum absolute error				219.00	270.00	52.78	96.00	48.15	39.00	23.75	55.00
H coupling average absolute error				28.69	26.93	15.99	18.63	1.88	4.74	1.95	4.76
non-H coupling average absolute error				63.28	74.74	12.78	19.97	9.78	13.14	4.90	12.43

Table 19. FC Term and Total Spin–Spin Coupling Calculation Results Using 6-311+G(d,p)-Derived Basis Set (Hz)

		UGBS2P		contracted G		uG		uG-w		uG-wd2	
		FC	total	FC	total	FC	total	FC	total	FC	total
HF	$^1J(\text{HF})$	192.47	390.44	132.02	320.15	146.41	334.54	165.55	353.68	165.55	353.68
NH ₃	$^1J(\text{NH})$	41.05	43.60	36.54	38.86	35.36	37.68	40.18	42.49	40.18	42.49
	$^2J(\text{HH})$	-12.51	-10.71	-11.43	-10.88	-10.11	-9.57	-12.06	-11.52	-12.06	-11.52
HCN	$^1J(\text{HC})$	283.08	283.24	266.67	267.00	246.19	246.52	279.78	280.11	279.78	280.11
	$^2J(\text{HN})$	2.31	5.29	2.55	5.02	2.04	4.50	2.30	4.77	2.30	4.77
SiH ₄	$^1J(\text{SiH})$	-210.04	-209.91	-188.24	-187.90	-189.08	-188.74	-210.95	-210.61	-210.02	-209.91
	$^2J(\text{HH})$	4.92	5.00	4.04	2.86	3.61	2.43	4.34	3.16	4.57	3.39
PH ₃	$^1J(\text{PH})$	158.43	163.96	131.39	136.26	140.32	145.18	156.22	161.09	157.36	162.23
	$^2J(\text{HH})$	-12.87	-12.59	-10.29	-10.58	-10.33	-10.63	-12.36	-12.66	-12.23	-12.53
H ₂ S	$^1J(\text{SH})$	19.25	23.85	14.13	18.47	16.45	20.80	18.20	22.54	18.75	23.10
	$^2J(\text{HH})$	-12.48	-11.71	-10.20	-10.16	-9.58	-9.53	-11.39	-11.35	-11.10	-11.06
HCN	$^1J(\text{CN})$	6.76	12.73	10.18	15.23	6.12	11.17	6.56	11.61	6.56	11.61
SiF ₄	$^1J(\text{SiF})$	262.41	343.08	327.35	409.59	308.53	390.77	324.23	406.47	316.85	399.10
	$^2J(\text{FF})$	-61.30	-161.24	-72.13	-179.06	-66.53	-173.45	-71.10	-178.02	-70.02	-176.94
PF ₃	$^1J(\text{PF})$	-1312	-1595	-1250	-1534	-1229	-1513	-1292	-1576	-1287	-1571
	$^2J(\text{FF})$	-54.17	-41.31	-62.98	-39.53	-49.60	-26.16	-53.03	-29.59	-52.89	-29.45
SF ₆	$^1J(\text{SF})$	-296.09	-320.46	-283.27	-304.67	-273.42	-294.82	-287.20	-308.60	-287.55	-308.95
	$^2J(\text{FF})$ 90°	-67.57	-273.10	-71.46	282.15	-70.37	-281.05	-75.24	-285.92	-70.71	-281.40
	$^2J(\text{FF})$ 180°	-18.19	-38.17	-32.34	-49.06	-34.42	-51.14	-36.80	-53.52	-30.11	-46.82
PCl ₃	$^1J(\text{ClP})$	-103.25	-129.67	-87.38	-114.17	-85.51	-112.30	-88.33	-115.13	-91.05	-117.84
	$^2J(\text{ClCl})$	-0.07	2.58	-0.17	2.45	0.13	2.74	0.13	2.74	0.13	2.75
average absolute error				16.15	42.84	16.15	17.76	8.63	10.30	7.67	9.35
maximum absolute error				64.94	555.25	83.00	82.00	61.82	63.39	54.44	56.02
H coupling average absolute error				12.94	13.86	12.72	13.65	3.45	4.50	3.23	4.28
non-H coupling average absolute error				19.68	74.72	19.92	22.27	14.33	16.67	12.56	14.92

coupling. The smaller uG basis sets generate slightly better results than uDZ.

The uDZ-w and uG-w basis sets, with tight s functions added, improved the results in most couplings. The uDZ-w basis sets considerably reduced the AAE to 8.74 Hz and the

MAE to 39 Hz, in total contribution. The uG-w basis sets had an AAE of 10.30 Hz and an MAE of 63.30 Hz, which is only a modest improvement from those of the uG basis sets. Adding d functions in both uDZ-w and uG-w basis sets produced no significant reduction in absolute errors.

In general, uTZ-derived basis sets produced the most accurate results, in comparison to UGBS2P. The uDZ- and uG-derived basis sets had only moderately good and barely adequate results, respectively. Hence, use of the uTZ-w basis set is strongly recommended. If the computational cost of uTZ-w is prohibitive, then uDZ-w and uG-w give qualitatively reasonable results at a substantially lower cost, although care should be used in making quantitative predictions using these basis sets.

The MAEs for aug-cc-pVTZ, uTZ, and uDZ-w basis sets are 168, 71, and 39 Hz, respectively, and so, uDZ-w is both much cheaper and more accurate than aug-cc-pVTZ or uTZ.

For all three basis sets examined in this paper, the procedure described here produces a substantial improvement in the reliability of the predicted spin–spin couplings over that of the unmodified basis. We expect the procedure to be applicable to other basis sets as well.

5. Comparison with cc-pCVXZ-sd and aug-cc-pVTZ-J Basis Sets

Different research groups have recently derived small basis sets for use in spin–spin coupling calculations. For comparison, results using two such basis sets, the cc-pCVXZ-sd ($X = D$ and T) basis set (developed by Peralta et. al.)¹⁹ and the aug-cc-pVTZ-J basis sets (developed by Sauer et. al.),²¹ have been examined for the molecules used in this study.

The cc-pCVXZ-sd basis sets have been successfully applied in one-bond C–C coupling calculations.²⁰ The results using cc-pCVXZ-sd ($X = D$ and T) basis sets are listed in Table S2 in the Supporting Information and are compared to those of the uXZ-w ($X = D$ and T) and UGBS2P basis sets. It can be seen that uXZ-w ($X = D$ and T) basis sets have smaller AAEs and MAEs in all couplings, by factors of 2–4.

Table S3 in the Supporting Information shows the results using aug-cc-pVTZ-J (TZ-J) basis sets, compared to those of uTZ-w and uTZ-wd2 basis sets. As can be seen in the table, TZ-J and uTZ-wd2 basis sets generate very comparable results; thus, both basis set series are appropriate for spin–spin coupling calculations. However, the TZ-J basis sets have been generated for only six elements (H, C, N, O, F, and S), while our uTZ-derived basis sets are applicable to at least all of the first and second row elements. Furthermore, our modification scheme is general and can be applied to other basis sets, such as aug-cc-pVQZ.

The scheme described in this paper has been implemented in the Gaussian 03 program, revision D,²⁶ with keyword “NMR = Mixed”. One should also specify “CPHF = Conv = 10” and “Int = ultrafine” as options.

6. Conclusion

In this paper, we have presented a general scheme of basis set modification in NMR spin–spin coupling constant calculation. The basis set used to compute the FC term is derived by uncontracting the original basis and then adding tighter s functions. The added tight s functions have even-tempered exponents starting from the tightest s functions in original basis set. For hydrogen and first row atoms, a ratio

of 3 for successive exponents was found to be optimal, while for second row atoms, a ratio of 2 was preferable.

Four tighter s functions are added for hydrogen and two tighter s functions for first and second row atoms. Tight d functions can also be added to second row atoms, in the same way as s functions, with a progressive exponent of 2 to the tightest d function in original basis sets, but produce marginal improvements. The SD, PSO, and DSO terms are calculated using an unmodified contracted basis set.

The three basis set series derived from aug-cc-pVTZ, aug-cc-pVDZ, and 6-311+G(d,p) have different accuracies in spin–spin coupling calculations. The uTZ-w basis sets produced an AAE of 3.61 Hz and an MAE of 28 Hz (total contribution), compared to 23.65 and 168 Hz, respectively, for the original aug-cc-pVTZ. The uDZ-w basis sets have an AAE and MAE of 8.74 and 39 Hz, respectively, compared to 49.70 and 270 Hz for the unmodified aug-cc-pVDZ. The AAE and MAE of uG-w are 10.30 and 63.39 Hz, respectively, compared to 42.84 and 555.25 Hz for the unmodified 6-311+G(d,p) basis set.

The uTZ-w basis produces spin–spin coupling constants, which are close to the basis set limit at moderate cost, and is the choice we strongly recommend. If uTZ-w is too expensive, then uDZ-w is a much better choice than the unmodified or uncontracted aug-cc-pVTZ basis set.

Acknowledgment. The work at Yale was supported by a Grant from Gaussian Inc.

Supporting Information Available: A summary of the calculated geometries and absolute energies of all molecules at the B3LYP/6-31G* level, tabulated calculation results of $^1J(^{35}\text{Cl}^{35}\text{Cl})$ coupling, and comparison results of uTZ-derived basis sets with cc-pCVXZ-sd ($X = D$ and T) and aug-cc-pVTZ-J (TZ-J) basis sets. This information is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Günther, H. *NMR Spectroscopy: Basic Principles, Concepts, and Applications in Chemistry*, 2nd ed.; John Wiley & Sons: New York, 1995.
- (2) Helgaker, T.; Jaszunski, M.; Ruud, K. *Chem. Rev.* **1999**, *99*, 293–352.
- (3) Sekino, H.; Bartlett, R. J. *Chem. Phys. Lett.* **1994**, *225*, 486–493.
- (4) Perera, S. A.; Sekino, H.; Bartlett, R. J. *J. Chem. Phys.* **1994**, *101*, 2186–2191.
- (5) Enevoldsen, T.; Oddershede, J.; Sauer, S. P. A. *Theor. Chem. Acc.* **1998**, *100*, 275–284.
- (6) Barone, V.; Peralta, J. E.; Contreras, R. H.; Snyder, J. P. *J. Phys. Chem. A* **2002**, *106*, 5607–5612.
- (7) Helgaker, T.; Watson, M.; Handy, N. C. *J. Chem. Phys.* **2000**, *113*, 9402–9409.
- (8) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (9) Sychrovsky, V.; Grafenstein, J.; Cremer, D. *J. Chem. Phys.* **2000**, *113*, 3530–3547.
- (10) Peralta, J. E.; Barone, V.; Contreras, R. H.; Zaccari, D. G.; Snyder, J. P. *J. Am. Chem. Soc.* **2001**, *123*, 9162–9163.

- (11) Jaszunski, M.; Ruud, K.; Helgaker, T. *Mol. Phys.* **2003**, *101*, 1997–2002.
- (12) Lutnaes, O. B.; Ruden, T. A.; Helgaker, T. *Magn. Reson. Chem.* **2004**, *42*, S117–S127.
- (13) Lantto, P.; Vaara, J.; Helgaker, T. *J. Chem. Phys.* **2002**, *117*, 5998–6009.
- (14) Kowalewski, J. *Prog. Nucl. Magn. Reson. Spectrosc.* **1977**, *11*, 1–78.
- (15) Kowalewski, J.; Laaksonen, A.; Roos, B.; Siegbahn, P. *J. Chem. Phys.* **1979**, *71*, 2896–2902.
- (16) Helgaker, T.; Jaszunski, M.; Ruud, K.; Gorska, A. *Theor. Chem. Acc.* **1998**, *99*, 175–182.
- (17) Oddershede, J.; Geertsen, J.; Scuseria, G. E. *J. Phys. Chem.* **1988**, *92*, 3056–3059.
- (18) Geertsen, J.; Oddershede, J.; Raynes, W. T.; Scuseria, G. E. *J. Magn. Reson.* **1991**, *93*, 458–471.
- (19) Peralta, J. E.; Scuseria, G. E.; Cheeseman, J. R.; Frisch, M. *J. Chem. Phys. Lett.* **2003**, *375*, 452–458.
- (20) Peralta, J. E.; Barone, V.; Scuseria, G. E.; Contreras, R. H. *J. Am. Chem. Soc.* **2004**, *126*, 7428–7429.
- (21) Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *J. Chem. Phys.* **2001**, *115*, 1324–1334.
- (22) Silver, D. M.; Wilson, S.; Nieuwpoort, W. C. *Int. J. Quantum Chem.* **1978**, *14*, 635–639.
- (23) de Castro, E. V. R.; Jorge, F. E. *J. Chem. Phys.* **1998**, *108*, 5225–5229.
- (24) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A. J.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; 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 Development Version*, rev. D.02; Gaussian Inc.: Wallingford, CT, 2004.
- (25) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (26) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A. J.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; 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*, rev. D.01; Gaussian Inc.: Wallingford, CT, 2004.
- (27) Martin, J. S.; Fujiwara, F. Y. *J. Am. Chem. Soc.* **1974**, *96*, 7632–7637.
- (28) Acrivos, J. V. *J. Chem. Phys.* **1962**, *36*, 1097–1098.
- (29) Dombi, G.; Diehl, P.; Lounila, J.; Wasser, R. *Org. Magn. Reson.* **1984**, *22*, 573–575.
- (30) Ebsworth, E. A.; Turner, J. J. *J. Chem. Phys.* **1962**, *36*, 2628.
- (31) Birchall, T.; Jolly, W. L. *Inorg. Chem.* **1966**, *5*, 2177–2179.
- (32) Lynden-Bell, R. *Trans. Faraday Soc.* **1961**, *57*, 888–892.
- (33) Bernheim, R. A.; Batizher, H. *J. Chem. Phys.* **1964**, *40*, 3446–3447.
- (34) Muetterties, E. L.; Phillips, W. D. *J. Am. Chem. Soc.* **1959**, *81*, 1084–1088.
- (35) Jackowski, K.; Wilczek, M.; Makulski, W.; Kozminski, W. *J. Phys. Chem. A* **2002**, *106*, 2829–2832.
- (36) Strange, J. H.; Morgan, R. E. *J. Phys. C: Solid State Phys.* **1970**, *3*, 1999–2011.

CT600110U

JCTC

Journal of Chemical Theory and Computation

Accurate Treatment of Energetics and Geometry of Carbon and Hydrocarbon Compounds within Tight-Binding Model

Alexander A. Voityuk*

Institució Catalana de Recerca i Estudis Avançats, Institute of Computational Chemistry, Universitat de Girona, 17071 Girona, Spain

Received February 16, 2006

Abstract: We show that a simple noniterative tight-binding model can provide reliable estimates of energetics and geometries of molecules with C–C and C–H bonds. The mean absolute error in heats of formation, ~ 4.6 kcal/mol, is essentially smaller than those found in previous tight-binding schemes. The internal consistency of the calculated heats of formation enables the reliable prediction of bond dissociation energies and isomerization enthalpies. The model gives accurate molecular geometries of hydrocarbons; the mean absolute errors in bond lengths and bond angles are 0.015 \AA and 1.4° , respectively. The calculated vibration frequencies agree reasonably well with experimental values. The method has proven to be transferable to complex carbon and hydrocarbon systems. The good performance of the model and its computational efficiency make it promising for simulations of carbon and hydrocarbon systems.

Introduction

While ab initio and DFT quantum chemical methods are widely used in molecular modeling, these techniques become inapplicable for extended systems containing hundreds of atoms, because of huge computational demands. Therefore, much less demanding semiempirical neglect of diatomic differential overlap (NDDO) methods are widely used for the quantum chemical treatment of such systems.¹ The schemes modified neglect of diatomic overlap (MNDO),² AM1,³ PM3⁴, and MNDO/d⁵ have been proven to give accurate estimates of ground-state energetics. New semiempirical NDDO models have been recently developed.^{6,7} In many cases, the results of semiempirical calculations are of the same quality as those of DFT calculations. Very recently, it has been shown that introducing the overlap matrix into the secular equations for MNDO-like methods leads to more accurate results.⁸ This study provides a new direction for the development of semiempirical NDDO schemes.

In the past three decade, tight-binding (TB) schemes have also been widely applied to a variety of chemical systems.^{9,10} The use of the parametrized TB approach for exploring the electronic properties of molecules and crystals was suggested

in the seminal work of Slater and Koster.¹¹ The TB models, which bear a close similarity to the extended Hückel method,¹² are computationally even more efficient than the semiempirical NDDO schemes. The limiting step of the semiempirical calculations of extended systems is the diagonalization of the Hamiltonian matrix. Usually, a single-point calculation requires about 20 iterations; it means that 20 diagonalizations of the Fock matrix are needed for a closed-shell system, and 40 such steps are required when an open-shell system is treated using the spin-unrestricted method. Unfortunately, the number of iterations for the self-consistent treatment of π -conjugated systems such as carbon nanotubes may remarkably increase with the size of the model. In such situations, the semiempirical calculations become very time-consuming. However, when a noniterative tight-binding method is used, only two matrix diagonalizations (for the Hamiltonian and overlap matrixes) are needed independent of whether a closed- or an open-shell system is considered. Therefore, the noniterative TB approach can be applied to systems containing up to a few thousand atoms. The models are intensively used in dynamic simulations of nanostructures.

Two approaches are employed to determine TB parameters. One is based on DFT calculations and usually referred

* Fax: 34 972418356. E-mail: alexander.voityuk@icrea.es.

to as DFTB.^{13,14} The DFTB method has been used for simulations of biological molecules, organic reactions, and nanostructures (see refs 15–18 and references therein). However, the performance of DFTB remains still not very clear, because no systematic assessment of the scheme has been published yet. Alternatively, the effective TB Hamiltonian can be parametrized using experimental data.^{19–25} While the semiempirical TB models allow one to obtain reasonable molecular geometries, the calculated heats of formation and reaction energies are found to be not very accurate. A typical error in atomization energies (and in formation enthalpies) of hydrocarbons is in the range of 30–50 kcal/mol, and therefore, it is too large as compared with those of the standard semiempirical methods. The main reason for the large errors is that the TB schemes have been parametrized with respect to the energetic and structural properties of various bulk phases. Also, this problem is closely related to the transferability of TB parameters and their dependence on the bonding environment of the systems.²⁶

The purpose of the present work is to describe a new noniterative TB scheme for the accurate treatment of molecules with C–C and C–H bonds. The model is referred to as the PNTB (parametrized noniterative tight-binding) scheme. The number of parameters for the short-range repulsion term is kept as small as possible to estimate the inherent accuracy of the effective Hamiltonian. The paper is organized as follows. In section 2, we briefly outline the tight-binding method and define the effective Hamiltonian and the short-range repulsive potential employed in the model. In section 3, we consider the performance of the proposed model by comparing the PNTB results with both experimental values and other calculations. In section 4, we give conclusions and outline possible extensions of the model.

Method

Within the tight-binding model, the total energy of the system can be expressed as

$$E = \sum_i^{\text{occ}} n_i \epsilon_i + E_{\text{rep}} = \sum_i^{\text{occ}} \langle \psi_i | H | \psi_i \rangle + E_{\text{rep}} \quad (1)$$

The first term, the electronic energy, is a sum of the orbital energies ϵ_i of all orbitals with the occupation number n_i . The second term is a short-range repulsion energy which is approximated by a sum of the interatomic potentials G_{AB} depending only on the distance between atoms A and B.

$$E_{\text{rep}} = \frac{1}{2} \sum_{A,B} G_{AB}(R_{AB}) \quad (2)$$

The effective one-electron Hamiltonian H is represented in a minimal basis of atomic orbitals (AOs). Because the AO basis is nonorthogonalized, the orbital energies ϵ_i are obtained by solving the generalized eigenvalue problem

$$\sum_i (H_{\mu\nu} - \epsilon_i S_{\mu\nu}) c_{\nu i} = 0 \quad (3)$$

The matrix elements of the Hamiltonian are defined as follows:

$$H_{\mu\mu} = U_{\mu}^A \quad (4)$$

$$H_{\mu\nu} = \beta_{\mu\nu} S_{\mu\nu} \quad (5)$$

Here, U_{μ}^A is the energy of an electron in AO φ_{μ} ($\mu = s, p, \text{ or } d$) at atom A, $\beta_{\mu\nu}$ is the resonance parameter which describes the two-center interaction of AOs μ and ν of atoms A and B, respectively.

$$\beta_{\mu\nu} = \beta_{\mu\nu}^{\text{AB}} \exp\left[-\frac{1}{2}(\lambda_A + \lambda_B)(R_{AB} - r_{AB}^0)\right] \quad (6)$$

where $\beta_{\mu\nu}^{\text{AB}}$ and λ_A are adjustable parameters and r_{AB}^0 is a scaling constant.

The repulsive potential G_{AB} includes a number of terms of different physical natures (the core–core repulsion, a correction due to double counting of the two-electron interaction, and the exchange–correlation energy). A simple exponential function is often used to approximate the potential:

$$G_{AB} = C_{AB} \exp[-\delta_{AB}(R_{AB} - r_{AB}^0)] \quad (7)$$

However, this function is unsuitable at small interatomic distances; at $R_{AB} = 0$, the potential remains finite instead of being infinite. Because at short distances the two-center matrix elements $H_{\mu\nu}$, eq 5, have large negative values, some computational problems may arise. We overcome the deficiency by using a scaling factor $\exp[1/2(r_{AB}^0/R_{AB} - 1)]$. Combining this factor with eq 7, one obtains

$$G_{AB} = C_{AB} \exp\left[-\delta_{AB}(R_{AB} - r_{AB}^0) + \frac{1}{2}(r_{AB}^0/R_{AB} - 1)\right] \quad (8)$$

The resulting potential should satisfactorily describe the short-range part of G_{AB} .

In opposition to the DFTB model, where the Hamiltonian matrix elements and the overlap integrals are defined only in a certain range of R_{AB} and assumed to be zero beyond this range,¹³ no such constraints are employed in our model. The overlap matrix for all atom pairs in eq 3 is calculated using Slater-type functions with exponents ζ_s and ζ_p for s and p AOs, respectively. The angular factors are determined by transformation of the atomic orbitals under rotation.¹¹

Because the energy of isolated atoms is calculated as

$$E^A = \sum_{\mu} n_{\mu} U_{\mu}^A \quad (9)$$

the total energy E of a system, eq 1, has a correct limit at large interatomic distances. The standard enthalpy of formation (at $T = 298$ K) is estimated as

$$\Delta H_f^0 = E - \sum_A E^A + \sum_A H_f^{0,298}(A) \quad (10)$$

where $\Delta H_f^{0,298}(A)$ is the experimental heat of formation of atoms A. The zero-point energy and the enthalpy term to heat the molecule from $T = 0$ K to $T = 298$ K are implicitly taken into account when one fits the semiempirical param-

Table 1. Parameters in the PNTB Hamiltonian

parameter	atomic parameters		parameter	bond-type parameters		
	H	C		H–H	C–H	C–C
U_s^A (eV)	–13.605	–16.960	β_{ss} (eV)	–21.100	–24.610	–27.535
U_p^A (eV)		–12.080	$\beta_{pp\sigma}$ (eV)			–21.596
ζ_s (au)	1.30	1.85	β_{sp} (eV)		–20.737	–24.5655 ^a
ζ_p (au)		1.60	$\beta_{pp\pi}$ (eV)			–18.979
λ_σ (Å ^{–1})	0.094	0.244	C_{AB} (eV)	1.043	0.963	0.903
λ_π (Å ^{–1})		0.058	δ_{AB} (Å ^{–1})	4.570	4.807	4.968

$$^a \beta_{sp} = 1/2(\beta_{ss} + \beta_{pp\sigma}).$$

eters to experimental $\Delta H_f^{0,298}$ values.¹ This approach is commonly used by the parametrization of semiempirical methods.

Parametrization. The parameters of the effective Hamiltonian and the repulsive potential G_{AB} were derived as follows. The parameter U_s for hydrogen was set to –13.605 eV (the negative of the ionization potential of the atom). The exponents ζ_s and ζ_p were fixed after preliminary test calculations. For the C–C pair, $\beta_{sp} = 1/2(\beta_{ss} + \beta_{pp\sigma})$. The scaling constants r_{AB}^0 for H–H, C–H, and C–C were set to 0.75, 1.10, and 1.45 Å, respectively. Note that the parameters r_{AB}^0 were introduced just to obtain similar values of the $\beta_{\mu\nu}^{AB}$ and C_{AB} parameters for different atom pairs and may be excluded from the scheme. All other parameters were fitted using experimental data of $\Delta H_f^{0,298}$ and structural parameters for several standard molecules. The experimental heats of formation were adopted from the NIST Chemistry WebBook.²⁷ The bond lengths and bond angles as well as the references to original sources can be found in refs 2–4. The molecules in a training set were chosen to represent the most common bonding situations in hydrocarbons. A non-linear least-squares method was used to optimize the semiempirical parameters. Several parametrization runs were carried out starting from different parameter values and using different training sets. The resulting parameters were tested in extensive survey calculations in order to choose the set which yields the most balanced results. Table 1 lists the final values of the parameters.

Results and Discussion

Heats of Formation. Table 2 contains the calculated and experimental heats of formation of several hydrocarbons which belong to different classes. A statistical evaluation for 83 molecules (see the Supporting Information) shows that the mean error is –0.6 kcal/mol (on average, the model slightly overestimated the stability of hydrocarbons) while the mean absolute error (MAE) is 4.6 kcal/mol. The corresponding errors of MNDO, AM1, and PM3 are 11.9, 11.0, and 7.7 kcal/mol, respectively. These large MAEs of the standard semiempirical schemes are mainly due to considerable overestimation of the heat of formation of C₆₀ (see below); they are reduced to 9.2, 7.0, and 5.6 kcal/mol, respectively, when the fullerene is excluded from the statistics. It should be emphasized that reparametrization of the MNDO-like methods for just CH compounds will essentially improve their performance. As seen from Table 2, small deviations of the calculated values of ΔH_f are obtained for both short and long alkanes. The method also

well reproduces the heats of formation of branched-hydrocarbon sterically crowded molecules with adjacent methyl groups. Small errors are found for cyclic and bicyclic molecules, for compounds with double and triple bonds, as well as for conjugated systems. Reliable estimates of ΔH_f are also predicted for aromatic compounds and the fullerene C₆₀. However, for some “difficult” molecules such as cubane and adamantane, the model provides less satisfactory data (the deviations from experimental results are found to be about 24 and 17 kcal/mol, respectively). The calculated heats of formation of radicals are in very good agreement with experimental data.

PNTB shows a considerable improvement over related TB schemes.^{19–25} While the method of Horsfield et al.²² gives accurate values of atomization energies for small alkanes, the error linearly increases with the size of the molecules (~5 kcal/mol per CH₂ group) and becomes 18 kcal/mol for C₅H₁₂. Then, while that scheme provides good results for compounds with double bonds, it considerably (by ~20 kcal/mol) underestimates the stability of molecules with triple bonds.²² The model of Zhao and Lu²⁵ overestimates atomization enthalpies with a mean absolute error of about 50 kcal/mol. The partial atomization enthalpy (the atomization enthalpy divided by the number of atoms in a molecule) can be calculated with PNTB with a MAE of ~0.5 kcal/mol, which is an order of magnitude smaller than the errors of the previous TB schemes.^{19–25}

Rotation Barriers. The energy barrier in the torsional motion about a single C–C bond arises from the steric interaction between the third nearest-neighbor atoms. In Table 3, we compare PNTB barriers with experimental data and ab initio values.^{28–31}

In ethane, the barrier is defined as the energy difference between eclipsed (D_{3h}) and staggered (D_{3d}) conformations. The experimental value of the rotational barrier is 2.9 kcal/mol. Usually, tight-binding schemes considerably underestimate rotational barriers; for instance, the model of Wang and Mak predicts free rotation of methyl groups in ethane.²¹ The PNTB calculated barrier, 1.3 kcal/mol, is half as large as the experimental value and close to the AM1 and PM3 estimates, 1.2 and 1.4 kcal/mol, respectively. Similarly, the PNTB barriers in propane and propene are too low. However, PNTB well reproduces the relative energies of butadiene and styrene conformations; the results are remarkably better than the corresponding energies calculated with MNDO, AM1, and PM3. While the PNTB scheme predicts a twisted structure of biphenyl, the twist angle is underestimated; it is found to be 24° instead of 44°. Because of that, the calculated

Table 2. Comparison with Experimental Results of Heats of Formation Calculated with PNTB, in kcal/mol

molecules	PNTB	exptl	molecules	PNTB	exptl
hydrogen	-5.3	0.0			
	alkanes			cyclic	
methane	-10.7	-17.8	cyclopropane	18.1	12.7
ethane	-16.6	-20.0	cyclopropene	70.3	66.2
propane	-22.7	-25.0	methylene-cyclopropane	48.5	47.9
<i>N</i> -pentane	-35.0	-35.1	cyclobutane	1.2	6.8
neopentane	-35.1	-40.2	cyclobutene	37.6	37.5
2,2,3,3-tetramethylbutane	-50.6	-56.2	cyclopentadiene	33.5	32.1
<i>N</i> -decane	-65.7	-59.6	cyclohexane, chair	-35.8	-29.5
	unsaturated		cyclohexene, half-chair	-6.3	-1.2
ethylene	12.6	12.5	bicyclobutane	66.0	51.9
propene	6.0	4.6	<i>trans</i> -bicyclopropyl	40.7	30.9
isobutene	-0.6	-4.0	bicyclo[2.1.0]pentane	38.3	37.8
1,3- <i>trans</i> -butadiene	24.5	26.3	adamantane	-48.9	-32.2
1,2-butadiene	34.6	38.8	cubane	124.6	148.7
acetylene	46.1	54.5		radicals	
propyne	39.1	44.2	methyl	39.6	34.8
allene	40.7	45.4	ethyl	32.6	28.0
	aromatic		<i>n</i> -propyl	27.0	24.0
benzene	19.8	19.7	isopropyl	25.7	22.3
fulvene	49.5	53.5	<i>n</i> -butyl	20.5	18.0
styrene	33.2	35.3	<i>s</i> -butyl	19.7	17.0
indene	40.7	39.1	<i>t</i> -butyl	19.2	11.0
mesitylene	1.9	-3.8	vinyl	69.2	63.4
naphthalene	35.5	35.9	HCC	114.7	123.0
azulene	62.0	73.5	allyl	38.9	39.0
anthracene	53.6	55.2	phenyl	80.6	79.0
phenanthrene	49.3	49.0	benzyl	47.5	49.0
biphenylene	102.4	100.5	cyclopropyl	76.3	66.9
fullerene C₆₀	630.5	634.8	cyclopentadienyl	59.0	58.0

Table 3. Conformational Energies for Prototypical Molecules, in kcal/mol

molecule	conformation	PNTB	MNDO	AM1	PM3	exptl (ab initio)
ethane (staggered)	eclipsed	1.3	1.0	1.2	1.4	2.9 ^a
propane (trans/trans)	cis/trans	1.4	1.2	1.3	1.5	3.3 (3.7) ^a
	cis/cis	2.9	2.8	3.0	3.2	8.8
propene (cis)	trans	0.6	0.2	0.6	0.6	2.0 (2.0) ^b
butadiene (trans)	cis	1.3	0.5	0.7	0.7	3.8 (3.6–4.1) ^c
	perpendicular	5.2	0.5	1.9	1.5	6.1 (4.9–6.1)
styrene (planar)	perpendicular	4.2	-1.4	1.4	1.4	3.0–3.3 (2.4–2.7) ^c
biphenyl (optimized)	twist angle	24°	90°	40°	0°	44° (44°) ^d
	planar	0.3	6.8	2.1	0.0	1.4 (3.1)
	perpendicular	3.1	0.0	1.1	1.0	1.6 (1.5)

^a Ref 28. ^b Ref 29. ^c Ref 30. ^d Ref 31.

energy of the coplanar conformation is lower and the energy of the perpendicular structure is higher than the reference values (Table 3).

Bond-Dissociation Energies. In Table 4, we compare energies for 20 bond-breaking reactions. Because the PNTB scheme provides accurate ΔH_f values for both open- and closed-shell systems (Table 2), the experimental C–H and C–C bond enthalpies are well reproduced by the calculation. The MAE is found to be 3.4 kcal/mol. Thus, PNTB can be applied to modeling bond-breaking processes.

Isomerization Reactions. The variety of hydrocarbons is based on the ability of carbon atoms to form single, double, and triple bonds. The changes in the valence state of carbon

atoms are associated with remarkable variations of atomic energies. Therefore, enthalpies of isomerization reactions can be considered as good test data to assess the performance of a computational method. Table 5 lists the calculated and experimental energies for 10 isomerization reactions. The mean absolute error amounts to 3 kcal/mol. The maximum deviation of 10 kcal/mol is found for the transformation of propyne into cyclopropene. The performance of PNTB becomes better for larger systems. The comparison suggests that PNTB provides consistent estimates of ΔH_f across different classes of hydrocarbons.

Fullerenes. Because fullerenes and carbon nanotubes play an important role in nanotechnology, computational modeling

Table 4. Bond Dissociation Enthalpies, in kcal/mol

bond		PNTB	exptl
CH ₄	→ CH ₃ + H	102.4	104.7
C ₂ H ₆	→ C ₂ H ₅ + H	101.2	100.1
	→ CH ₃ + CH ₃	88.8	82.8
C ₂ H ₄	→ C ₂ H ₃ + H	108.6	103.1
C ₂ H ₂	→ C ₂ H + H	120.6	120.6
C ₃ H ₈	→ <i>n</i> -C ₃ H ₇ + H	101.7	101.0
	→ <i>i</i> -C ₃ H ₇ + H	100.5	99.4
	→ C ₂ H ₅ + CH ₃	95.0	87.9
C ₃ H ₆	→ C ₃ H ₅ + H	85.1	86.2
C ₄ H ₁₀	→ <i>s</i> -C ₄ H ₉ + H	100.8	99.2
	→ <i>n</i> -C ₄ H ₉ + H	101.5	100.1
	→ 2C ₂ H ₅	94.1	86.0
<i>i</i> -C ₄ H ₁₀	→ <i>t</i> -C ₄ H ₉ + H	100.1	95.2
	→ <i>s</i> -C ₃ H ₇ + CH ₃	94.3	89.2
<i>c</i> -C ₃ H ₆	→ <i>c</i> -C ₃ H ₅ + H	110.3	106.3
Ph-H	→ Ph + H	112.8	111.4
Ph-CH ₃	→ PhCH ₂ + H	85.8	89.0
	→ Ph + CH ₃	106.3	101.7
Ph-C ₂ H ₅	→ PhCH ₂ + CH ₃	79.1	76.9
	→ Ph + C ₂ H ₅	105.4	99.9

Table 5. Enthalpies of Isomerization Reactions, in kcal/mol

Reactant	Product	PNTB	Exp.
	→	1.6	1.3
	→	31.7	22.0
	→	-8.2	-8.5
	→	2.0	4.0
	→	31.6	23.5
	→	4.9	2.7
	→	5.6	7.0
	→	-13.0	-10.1
	→	20.3	19.6
	→	7.2	10.8

of the carbon nanostructures has attracted much attention. To assess the performance of PNTB for such systems, we carried out calculations of several fullerenes. Note that large dispersions among the experimental heats of formation of C₆₀ and C₇₀ are found.³² For C₆₀, the PNTB calculated $\Delta H_f^{0,298}$, 630.5 kcal/mol, is close to the 634.8 kcal/mol adopted by NIST²⁷ while being 16 kcal/mol larger than the estimate obtained by Kolesov et al.³³ The B3LYP/6-31G* calculations predict $\Delta H_f^{0,298} = 618$ kcal/mol.³⁴ It should be noted that the standard semiempirical methods MNDO, AM1, and PM3 considerably underestimated the stability of C₆₀; the calculated $\Delta H_f^{0,298}$ values are 868.5, 972.6, and 811.0 kcal/mol, respectively. For C₇₀, PNTB gives 685 kcal/mol,

Table 6. Calculated and Experimental Bond Lengths in C₇₀, in Å

	PNTB	B3LYP6-31G*	experimental ³⁵	
			ND	GED
R _{aa}	1.443	1.452	1.460	1.46
R _{ab}	1.386	1.397	1.382	1.388
R _{bc}	1.442	1.44	1.449	1.453
R _{cc}	1.378	1.389	1.396	1.386
R _{cd}	1.440	1.449	1.464	1.468
R _{dd}	1.429	1.434	1.420	1.425
R _{de}	1.409	1.421	1.415	1.405
R _{ee}	1.452	1.471	1.477	(1.538)

Table 7. Comparison of Relative Energies of Isomers of C₃₀ and C₃₂ Fullerenes, in kcal/mol

molecule	PNTB	B3LYP6-31G* ¹⁷
C _{30_1} (C _{2v}) ^a	45.0	55.6
C _{30_2} (C _{2v})	6.5	4.0
C _{30_3} (C _{2v})	0.0	0.0
C _{32_1} (D ₂)	54.4	60.3
C _{32_2} (C ₂)	44.4	65.5
C _{32_3} (D _{3d})	58.4	73.9
C _{32_4} (C ₂)	22.2	26.0
C _{32_5} (D _{3h})	71.0	78.3
C _{32_6} (D ₃)	0.0	0.0

^a The numbering of isomers is the same as in ref 17.

which is in reasonable agreement with the B3LYP estimate of 658 kcal/mol³⁴ and the experimental values 666 and 658 kcal/mol tabulated in ref 32.

Also, PNTB calculations provide reliable estimates for structural parameters of fullerenes. In C₆₀, there are two types of C–C bonds with lengths 1.39 and 1.44 Å. The PNTB values, 1.395 and 1.445 Å, are in excellent agreement. In C₇₀ (D_{5h} symmetry), there are five circles of atoms labeled with a, b, c, d, and e from the capping pentagon to the equator.³⁵ Table 6 contains calculated and experimental bond lengths in C₇₀. As can be seen, the PNTB results are in good agreement with experimental and B3LYP data.

In Table 7, we compare relative energies of isomers of small fullerenes C₃₀ and C₃₂ calculated using the PNTB and B3LYP methods. Overall, the PNTB results agree satisfactorily with the B3LYP estimates. The PNTB energies are found to be similar to the data derived within the DFTB model.¹⁷ Note that the PNTB scheme reproduces the B3LYP data more accurately than the AM1 and PM3 semiempirical methods.

Geometries. Molecular geometries of some selected hydrocarbons are listed in Table 8. A statistical evaluation of structural parameters calculated by PNTB for 30 hydrocarbons (see the Supporting Information) shows that the mean absolute error in bond lengths is 0.015 Å (100 comparisons). The method systematically underestimate the C–C bond lengths, resulting in a mean sign error of –0.012 Å. Reliable results are also obtained for bond angles—the mean absolute error is 1.4° for 31 comparisons. Thus, the geometries of hydrocarbons can be well predicted by the PNTB model.

Table 8. Geometries of Selected Molecules^a

molecule	variable	calcd	exptl
hydrogen	H–H	0.760	0.741
methane	C–H	1.091	1.087
ethane	C–C	1.499	1.535
	C–H	1.093	1.094
	HCC	110.3	111.2
ethylene	C=C	1.319	1.339
	C–H	1.082	1.087
	CCH	122.3	121.3
acetylene	C≡C	1.206	1.202
	C–H	1.052	1.063
propene	C=C	1.321	1.336
	C–C	1.478	1.501
	C–H	1.082	1.081
propyne	C–H	1.097	1.098
	CCC	123.4	124.3
	CCH	123.1	121.5
	C≡C	1.207	1.206
	C–C	1.434	1.459
	C ₁ –H	1.053	1.056
allene	C ₃ –H	1.099	1.105
	CCH	111.2	110.2
	C=C	1.300	1.308
	C–H	1.087	1.087
neopentane	HCH	114.2	118.2
	C–C	1.504	1.539
	C–H	1.093	1.120
cyclopropane	HCC	110.1	110.0
	C–C	1.506	1.510
	C–H	1.076	1.074
cyclopropene	HCH	116.3	115.9
	C=C	1.302	1.296
	C–C	1.526	1.509
cyclopentadiene	HCC	153.1	149.9
	HCH	116.3	114.6
	C=C	1.339	1.345
benzene	C–C	1.462	1.468
	C–C	1.498	1.506
	C–C	1.385	1.397
naphthalene	C–H	1.083	1.083
	C ₁ –C ₂	1.368	1.381
	C ₂ –C ₃	1.403	1.417
	C ₁ –C ₉	1.407	1.422
	C ₉ –C ₁₀	1.419	1.412

^a Bond lengths are in Å; bond angles are in deg.

Vibrations. In Table 9, we compare the calculated and experimental vibrational frequencies.³⁶ While vibrational frequencies were not employed as reference functions by the fitting of the parameters, vibrational spectra are well reproduced by PNTB. The calculated frequencies tend to be predicted somewhat too low. The C–H stretching frequencies are about 4% smaller than the observed values. The PNTB predictions for C–C, C=C, and C≡C bond-stretching modes in C₂H₆, C₂H₄, and C₂H₂ are calculated to be 1266, 1499, and 1895 cm⁻¹, respectively. These values agree within 10% with the experimental frequencies, 1388, 1623, and 1974 cm⁻¹. Even for torsional motion in C₂H₆, PNTB gives 206 cm⁻¹, in agreement with the experimental value 289 cm⁻¹. The lowest-energy bending mode Π_g in C₂H₂ is found to be

Table 9. Comparison of PNTB Vibrational Frequencies (in cm⁻¹) with Experimental Values (in Parentheses)

molecule	symmetry of vibration	PNTB (exptl)
hydrogen	Σ _g	4118 (4401)
methane	A ₁	2928 (3158)
	T ₂	2869 (3019); 1252 (1357)
	E	1397 (1534)
ethane	A _{1g}	2883 (2896); 1266 (1388); 975 (995)
	A _{2u}	2904 (2915); 1314 (1370)
	E _u	2840 (2974); 1373 (1460); 750 (822)
	E _g	2832 (2969); 1357 (1468); 1121 (1190)
	A _{1u}	206 (289)
ethylene	A _{1g}	2959 (3026); 1499 (1623); 1253 (1342)
	A _{1u}	916 (1023)
	B _{1u}	2978 (2990); 1315 (1444)
	B _{2g}	783 (943)
	B _{2u}	2954 (3106); 704 (810)
	B _{3g}	2939 (3103); 1080 (1236)
acetylene	B _{3u}	950 (949)
	A _{2u}	916 (1023)
	Σ _g	3245.9 (3374); 1895 (1974)
	Σ _u	3225.2 (3289)
	Π _u	826 (747)
	Π _g	355 (624)

355 cm⁻¹. This value agrees well with the 324 cm⁻¹ from MP4/6311G* calculations.³⁶ The essential discrepancy with an experimental value, 624 cm⁻¹, appears to be mainly due to the anharmonicity of this mode.

Conclusions

We described a noniterative tight-binding model parametrized for calculating energetics and geometries of carbon and hydrocarbon systems. The scheme provides good estimates of the heats of formation and reaction enthalpies. The mean absolute error of the calculated Δ*H_f* is 4.6 kcal/mol, which is several times smaller than the errors found within related schemes. The method gives accurate estimates for C–H and C–C bond energies and isomerization reactions. This suggests the internal consistency of the predicted heats of formation, allowing a reliable analysis of trends across series of molecules. Molecular geometries of organic molecules are well predicted by PNTB; the mean absolute error of the bond lengths is 0.015 Å, and that of the bond angles is 1.4°. The calculated vibration frequencies reasonably agree well with experimental values. The method has proven to be transferable to complex carbon and hydrocarbon systems. Taking into account the computational efficiency and good performance of the model, we conclude that PNTB should be very promising for simulations of the formation of carbon nanostructures and the high-temperature degradation of hydrocarbons.

The PNTB results give impetus to the further development of parametrized tight-binding methods. We expect that the performance of the model can be improved (1) by adjusting the distance dependence of the resonance and overlap integrals and (2) by using more flexible short-range potentials (additional terms such as the pair-directed Gaussian functions⁶ may be introduced to describe the dispersion interaction).

Acknowledgment. This work has been supported by the Spanish Ministerio de Educación y Ciencia, Project No. CTQ2005-04563.

Supporting Information Available: Calculated and experimental heats of formation of 83 hydrocarbons (Table SI1); calculated and experimental geometries (Table SI2). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Thiel, W. *Adv. Chem. Phys.* **1996**, *93*, 703.
- (2) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
- (3) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (4) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
- (5) Thiel, W.; Voityuk, A. A. *J. Phys. Chem.* **1996**, *100*, 616.
- (6) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601.
- (7) Giese, T. J.; Sherer, E. C.; Cramer, C. J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 1275.
- (8) Sattelmeyer, K. W.; Tubert-Brohman, I.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2006**, *2*, 413.
- (9) Ohno, K.; Esfarjani, K.; Kawazoe, Y. *Computational Material Science*; Springer: Berlin, 1999.
- (10) Selvam, P.; Tsuboi, H.; Koyama, M.; Kubo, M.; Miyamoto, A. *Catal. Today* **2005**, *100*, 11.
- (11) Slater, J. C.; Koster, G. F. *Phys. Rev. B* **1954**, *94*, 1498.
- (12) Hoffman, R. *J. Chem. Phys.* **1963**, *39*, 1397.
- (13) Porezag, D.; Frauenheim, T.; Kohler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1995**, *51*, 12947.
- (14) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *58*, 7260.
- (15) Zheng, G. S.; Irle, S.; Elstner, M.; Morokuma, K. *J. Phys. Chem. A* **2004**, *108*, 3182.
- (16) Kruger, T.; Elstner, M.; Schiffels, P.; Frauenheim, T. *J. Chem. Phys.* **2005**, *122*, 114110.
- (17) Zheng, G. S.; Irle, S.; Morokuma, K. *Chem. Phys. Lett.* **2005**, *412*, 210.
- (18) Omata, Y.; Yamagami, Y.; Tadano, K.; Miyake, T.; Saito, S. *Physica E* **2005**, *29*, 454.
- (19) Xu, C. H.; Wang, C. Z.; Chan, C. T.; Ho, K. M. *J. Phys. Condens. Matter* **1992**, *4*, 6047.
- (20) Wang, C. Z.; Chan, C. T.; Ho, K. M. *Phys. Rev. Lett.* **1991**, *66*, 189.
- (21) Wang, Y.; Mak, C. H. *Chem. Phys. Lett.* **1995**, *235*, 37.
- (22) Horsfield, A. P.; Godwin, P. D.; Pettifor, D. G.; Sutton, A. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 15773.
- (23) Winn, M. D.; Rassinger, M.; Hafner, J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1997**, *55*, 5364.
- (24) Pan, B. C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2001**, *64*, 155408.
- (25) Zhao, J. J.; Lu, J. P. *Phys. Lett. A* **2003**, *319*, 523.
- (26) Lu, J. P.; Wang, C. Z.; Ruedenberg, K.; Ho, K. M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2005**, *72*, 205123.
- (27) *NIST Chemistry WebBook*, NIST Standard Reference Database, No. 69; Mallard, W. G., Linstrom, P. J., Eds.; National Institute of Standards and Technology: Gaithersburg, MD. <http://webbook.nist.gov> (accessed Oct 2005).
- (28) Epiotis, N. D.; Yates, D. L. *J. Am. Chem. Soc.* **1976**, *98*, 461.
- (29) Kundu, T.; Goodman, L.; Leszczynski, J. *J. Chem. Phys.* **1995**, *103*, 1523.
- (30) Head-Gordon, M.; Pople, J. J. *J. Phys. Chem.* **1993**, *97*, 1147.
- (31) Rubio, M.; Merchán, M.; Ortí, E. *Theor. Chim. Acta* **1995**, *91*, 17.
- (32) Yu, J.; Sumathi, R.; Green, W. H. *J. Am. Chem. Soc.* **2004**, *126*, 12685.
- (33) Kolesov, V. P.; Pimenova, S. M.; Pavlovich, V. K.; Tamm, N. B.; Kurskaya, A. A. *J. Chem. Thermodyn.* **1996**, *28*, 1121.
- (34) Cioslowski, J.; Rao, N.; Moncrieff, D. *J. Am. Chem. Soc.* **2000**, *122*, 8265.
- (35) Hedberg, K.; Hedberg, L.; Bühl, M.; Bethune, D. S.; Brown, C. A.; Johnson, R. D. *J. Am. Chem. Soc.* **1997**, *119*, 5314.
- (36) *NIST Computational Chemistry Comparison and Benchmark Database*, NIST Standard Reference Database Number 101 Release 12; National Institute of Standards and Technology: Gaithersburg, MD, Aug 2005; Johnson, R. D., III, Ed. <http://srdata.nist.gov/cccbdb> (accessed Oct 2005).

CT600064M

JCTC

Journal of Chemical Theory and Computation

Parametrization of Atomic Energies to Improve Small Basis Set Density Functional Thermochemistry

Edward N. Brothers* and Gustavo E. Scuseria

Department of Chemistry, Mail Stop 60, Rice University, Houston, Texas 77005-1892

Received March 23, 2006

Abstract: Enthalpies of formation predicted with density functional theory and small basis sets can be greatly improved by treating the atomic energies as empirical parameters. When a variety of functionals and small basis sets are used, the root-mean-square error in enthalpies of formation is reduced by a factor of approximately two for the least improved functional/basis set pair, with significantly larger reductions for other functionals, especially LSDA. When the 3-21G* and 3-21+G* basis sets are used with nonempirical functionals, it is possible to achieve accuracy greater than that of PM3, which was primarily designed to reproduce enthalpies of formation. In addition to decreasing statistical errors, our procedure can also remove qualitative errors in density functional/basis set pairs that fail for the prediction of enthalpies of formation.

I. Introduction

This paper shows that combining the improvement in enthalpies of formation by optimizing atomic energies such as was done recently by Csonka et al.¹ with the ability of density functional theory (DFT) to predict enthalpies of formation with small basis sets^{2,3} results in inexpensive density functional thermochemistry comparable to, or better than, semiempirical methods. This is notable in two regards; first, (very) small basis set density functional theory⁴ has never (to our knowledge) outperformed PM3,⁵ and second, it takes the corrective ability of parametrized atom energies beyond simply tightening error bars and actually removes qualitative errors.

The optimization of atomic energies to reduce thermochemical error was recently undertaken by Csonka et al.,¹ and they demonstrated that the errors in enthalpies of formation predicted by DFT can be partially attributed to problems with predicted atomic energies. When fixed geometries and tabulated frequency corrections are used with fairly large basis sets, this optimization substantially reduced the errors in enthalpies of formation for a variety of previously difficult molecules in the G3/99 set of compounds.⁶

In addition to improving enthalpies of formation, there is a second point that can be taken from that study which is

more subtle. It may be possible to create a set of atomic energies such that functionals heretofore considered unacceptable for thermochemistry because of large errors can in fact be useful. In other words, the atomic energy fitting procedure could correct qualitatively wrong results by removing the major impediment to calculating enthalpies of formation.

Concurrent with the atomic energy work cited above were two studies which demonstrated that DFT can predict enthalpies of formation accurately with some of the smallest common basis sets. In the first study, reasonable enthalpies of formation for many functionals were obtained, providing results comparable with semiempirical predictions while using geometries, energies, and frequencies all determined with the small basis sets.² In this case, all of the functionals that provided high-quality results were all based on both the density and the gradient of the density, that is, the generalized gradient approximation, or GGA. Meta-GGAs, which include terms based on the kinetic energy density, were not considered in that study, although they are included here.

The second study³ developed fully an idea first examined in the course of analyzing small basis density functional thermochemistry.² LSDA,⁷ which contains no gradient correction, was improved for a wide variety of properties through the use of an empirical parameter to scale the correlation, with special emphasis given to performance with small basis sets. This method was termed “cSVWN5”. It is

* Corresponding author e-mail: enb@rice.edu.

useful for investigating large systems because small basis set methods are intrinsically fast, density-only functionals are slightly faster than GGAs and meta-GGAs, and more complicated theories are CPU-intensive.

In this paper, atomic energy optimization is applied to small basis set DFT, including the functional developed especially to work with small basis sets, to greatly improve thermochemical prediction.

II. Test Set and Computational Method

The G3/99 set of Curtiss and co-workers was selected here for use as a test bed because it has become a common standard for determining the accuracy of quantum chemical approaches. In total, there are 13 atom types and 223 compounds used to examine enthalpies of formation in G3/99; however, five of these atom types appear in three or less compounds, specifically lithium, beryllium, sodium, aluminum, and boron. These atom types and the compounds containing them were thus removed from the set to avoid creating biased parameters for those atom types, resulting in a total test set of 213 compounds consisting of nine atom types.

The basis sets chosen for this study were STO-3G,⁸ 3-21G*, and 3-21+G*,⁹ which are the smallest basis sets in common use.¹⁰ It is important to note that the “*” on 3-21G* and 3-21+G* denote placing a *d* function on atoms heavier than neon and not on all non-hydrogen atoms, as is the case with other Pople basis sets. Also, these basis sets use the default Cartesian basis functions; that is, they use 6*d* rather than 5*d* components. For molecules containing atoms larger than neon, any basis-set-specific parameters, such as the atomic energies in this study, would have to be adjusted to compensate for the change in basis if 5*d* was desired.

Several categories of functionals are represented in this study. The two density-only functionals used in this study are LSDA, which uses the local correlation functional of Vosko, Wilk, and Nusair,⁷ and cSVWN5,³ which is equivalent to the LSDA used in this study with the local correlation scaled by 0.3. cSVWN5 was optimized by Riley et al. for use with 3-21G* and 3-21+G* and, thus, is neglected for STO-3G in this paper. The GGA PBE¹¹ developed by Perdew, Burke, and Ernzerhof, and the meta-GGA TPSS¹² created by Tao, Perdew, Staroverov, and Scuseria, are also examined. PBEh,^{13,14} which is PBE with 25% of the functional exchange replaced by exact exchange, and TPSSh,¹⁵ which uses 10% exact exchange, are also evaluated. While not a density functional, HF¹⁶ is included for comparison purposes. Note that, with the exception of cSVWN5, none of the functionals were developed by fitting internal parameters to enthalpies of formation or other experimental data; that is, they are nonempirical.

All calculations were performed in the Gaussian suite of programs.¹⁷ For all of the methods used in this study, geometries were optimized and frequencies calculated at the method of interest; that is, the energies were functional/basis//functional/basis throughout. Gaussian defaults were used in all of the calculations, with the exception of the integration grid, which was a pruned (99,590), or “ultrafine”, grid.

After the data was collected, three separate parametrizations were attempted. First, a single multiplicative scaling parameter was used with all calculated atomic energies to see if the fit could be accomplished with one parameter. This fit was also attempted starting from the correct total atomic energies.¹⁸ Finally, a full genetic algorithm¹⁹ (GA) optimization was undertaken in which each atomic energy was treated as an empirical parameter and all parameters were simultaneously fit versus the entire set of compounds in this study. This optimization procedure was selected because we desired to optimize all of the parameters simultaneously. With nine parameters for each functional/basis pair, a grid search would be out of the question. Also, it was unknown at the beginning of the study how close to the final parameters the initial values were, and thus, any method consisting of multiple line searches would have to be restarted at a variety of different inertial points, which is a problem easily avoided by using a GA. Thus, a GA was selected for this problem. Note that all of the compounds were used, rather than a “jack-knife” set, in which fitting would be conducted versus some compounds and evaluated versus a larger set which includes the set parametrized against. For the purposes of optimization, the root-mean-square (RMS) error was treated as the error to be minimized, as this biases the parametrization to pull in the furthest outliers first.

III. Results

Before discussing the results of the parametrization, it is necessary to briefly mention what the optimized atomic energies represent. Optimization does not necessarily move the atomic energies closer to the exact values. (A list of the difference between the exact energies¹⁸ and the calculated and parametrized energies using carbon as an example is available in the Supporting Information.) The difference between exact and calculated energies ranges up to 0.8 au, and the difference between the optimized and original atomic energies is small relative to the difference between the exact and calculated values. The parameters also compensate for issues with the basis set. There are several functionals which when used with large basis sets already produce very good enthalpies of formation,⁶ but by using small basis sets such as the ones considered in this study, the parameters are forced to deal with both functional shortcomings and the paucity of the basis set. Therefore, because the parameters do not represent an improvement in atomic energies versus exact values and because they are basis-set-specific, it would be erroneous to assign them any physical interpretation. They are simply empirical parameters whose strength is their efficacy.

Attempts were made to scale the atomic energies, meaning that all atomic energies were multiplied by a single parameter. Scaling exact atomic energies did not improve accuracy, and in fact, the results were worse than those using the original atomic energies. Scaling the calculated energies with a single parameter does improve the enthalpies of formation slightly, but the improvement is marginal at best. Thus, this data is not presented; it is mentioned to explain why using one parameter for each atomic energy was undertaken. The balance of this paper will deal with the outcome of the GA optimization.

Table 1. Mean Error (ME), Mean Absolute Error (MAE), Root-Mean-Squared Error (RMS), and Standard Deviation (SD) for Enthalpies of Formation in the G3/99 Set Using Original and Optimized Atomic Energies^a

basis	method	original				GA optimized			
		ME	MAE	RMS	SD	ME	MAE	RMS	SD
STO-3G	LSDA	-268.0	268.8	335.3	202.0	2.5	20.1	29.1	29.1
	PBE	-145.5	151.6	190.9	123.8	0.7	17.5	27.3	27.3
	PBEh	-111.7	126.7	165.1	121.9	0.5	18.7	29.1	29.1
	TPSS	-118.4	128.0	164.0	113.8	0.1	17.3	27.4	27.5
	TPSSh	-108.5	121.9	158.3	115.5	0.1	17.7	28.1	28.1
	HF	140.4	142.2	171.3	98.5	-1.7	22.0	36.1	36.2
3-21G*	LSDA	-113.5	113.5	136.1	75.2	2.8	7.3	9.5	9.1
	PBE	-13.2	15.5	19.8	14.9	0.9	4.3	6.1	6.1
	PBEh	15.0	15.2	19.8	12.9	1.0	4.4	5.8	5.7
	TPSS	11.5	13.5	17.2	12.8	0.3	4.1	6.1	6.1
	TPSSh	19.4	20.5	25.2	16.1	0.4	4.0	5.7	5.7
	HF	261.5	261.5	298.9	145.1	-1.5	7.8	11.7	11.7
3-21+G*	cSVWN5	-7.4	19.1	28.0	27.1	1.8	6.0	7.9	7.7
	LSDA	-100.8	100.8	121.7	68.4	2.1	6.3	8.1	7.8
	PBE	-1.0	7.5	9.9	9.9	0.3	3.6	5.7	5.7
	PBEh	24.2	24.2	29.0	16.0	0.5	3.8	5.1	5.1
	TPSS	21.5	22.5	27.2	16.8	-0.3	3.7	5.8	5.9
	TPSSh	28.5	29.2	35.0	20.5	-0.1	3.5	5.3	5.3
	HF	266.4	266.4	304.7	148.2	-1.8	7.9	11.7	11.6
	cSVWN5	7.9	16.3	19.3	17.7	1.1	4.9	6.6	6.5
	PM3	-1.0	6.9	9.5	9.4				

^a All values are in kcal/mol. PM3 is included on the last line for comparison purposes.

The results of the GA optimization listed in Table 1 demonstrate the effectiveness of the optimization, especially in light of how poor many of the original results are. By any reasonable criteria, the enthalpies of formation calculated for STO-3G are a failure, with enormous systematic errors due to overbinding for all functionals and underbinding for HF. Also, at any of the three basis sets considered in this study, LSDA fails. The best functional/basis set pair with regular energies is PBE/3-21+G*. This occurs because of a cancellation of errors, as the 3-21G* results show that PBE underbinds while PBEh, TPSS, and TPSSh overbind, and the addition of the diffuse function increases binding for all functionals, resulting in the good performance of PBE/3-21+G*.

In contrast, after the parametrization, the errors that result from systematic overbinding or underbinding (exhibited through nonzero mean errors) are nearly removed. The performance of LSDA is also brought much closer to the performance of the GGA and meta-GGA functionals, especially when the comparison is done with STO-3G. In fact, the errors after parametrization are small enough that no functional can be said to fail for enthalpies of formation. This is not to say that all basis sets are equally well suited for thermochemistry as long as parameters are present but rather that the parametrized methods at STO-3G are no longer qualitatively incorrect. This improvement from qualitatively incorrect to qualitatively correct can be most easily seen by examining Figures 1 and 2. With the original atomic energies, the enthalpies of formation are predicted to be hundreds of kilocalories per mole low and are shown in the graph to correspond very poorly to the experimental values, appearing almost randomly scattered. The parametrized results qualitatively correspond to the experimental values, albeit still with significant errors, similar in size to TPSS with 3-21+G* and no parameters.

Another measure of the success of the parametrization is the comparison with PM3 results.⁵ PM3 was parametrized primarily versus enthalpies of formation and is consistently more accurate than DFT with small basis sets and standard atomic energies.² PM3 still outperforms all of the STO-3G results, but functionals with split-valence basis sets are more accurate than PM3 after atomic energy optimization. Thus, using nonempirical functionals and small basis sets and correcting the errors in atomic energies allows thermochemical accuracy as high as that of semiempirical methods, which use far more parameters.

Performance is still determined by the basis set and functional, with the larger basis sets and functionals which consider more density-related quantities providing higher accuracy both before and after parametrization. For any optimized result in this study, the RMS for STO-3G is larger than that for 3-21G* and larger for 3-21G* than for 3-21+G*. The trend for mean absolute error follows the same trend for all methods except HF, and the deviation from the trend for HF is negligible. (More information on parametrization to improve HF results can be found in the work of Ruzsinszky and Csonka.²⁰) Thus, the new atomic energies do not alter the fact that bigger basis sets work better. Also, while the difference between density-only methods and the more modern GGAs and meta-GGAs is not particularly noticeable at the STO-3G level, with all methods providing errors of around 20 kcal/mol, there is a clear advantage to PBE, TPSS, and their hybrids with the split-valence basis sets.

It should be noted that the inclusion of exact exchange with small basis sets does not improve the thermochemistry without the atomic energy parametrization, and with parametrization, the pure functionals are outperformed by the hybrids only with the split-valence basis sets. Also, HF even after parametrization does worse than any of the DFT

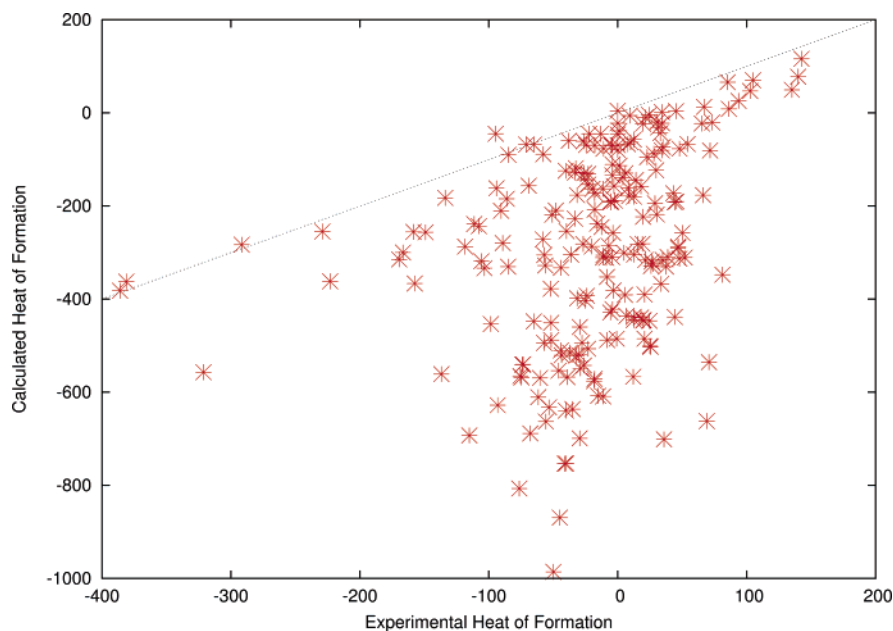


Figure 1. Enthalpies of formation calculated using LSDA/STO-3G. A line with a slope of unity is added to guide the eye.

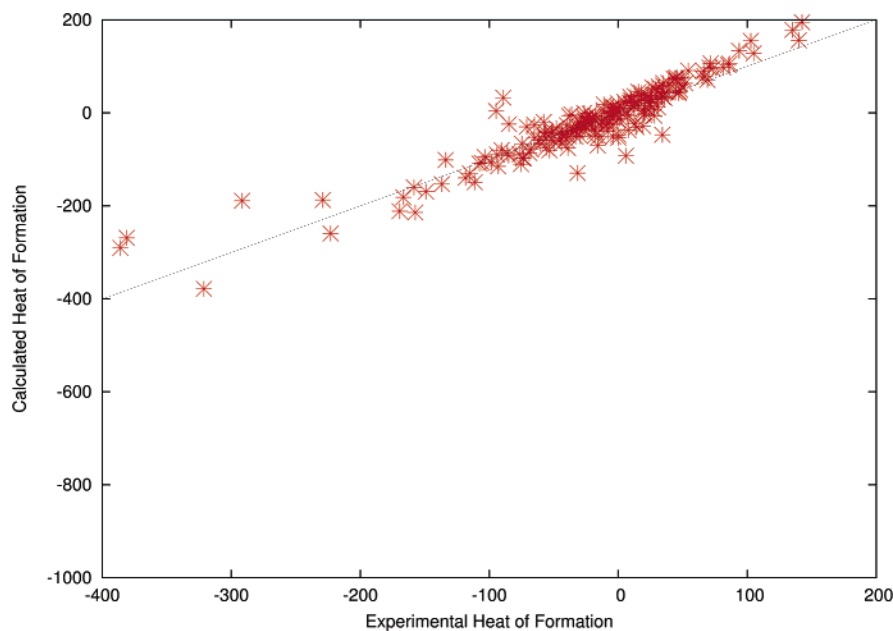


Figure 2. Enthalpies of formation calculated using LSDA/STO-3G and the optimized atomic energies. A line with a slope of unity is added to guide the eye.

methods. This can be attributed to the fact that, with small basis sets, exact exchange is a hindrance,² unlike in large basis sets, where it can improve enthalpies of formation greatly.⁶

The only empirical functional presented in this study (cSVWN5) has errors approximately halfway between the best functionals and LSDA after atomic energy optimization. It thus performs better than LSDA, which it was developed from, as well as PM3, and is slightly cheaper than functionals that include terms other than the density. This makes it ideal for investigating large systems.

To examine the size dependence of the errors, the error in the predicted enthalpy of formation was plotted versus the number of carbon atoms for the first eight alkanes. This curve was then fit to a line, and the slope of the line is

presented in Table 2. In this case, the larger the magnitude of the slope, the greater the size dependence of the error. With the split-valence basis sets and GGA, meta-GGA, hybrid functionals, and the empirical functional cSVWN5, the original size dependence of the error is small to begin with, and in most cases, the optimization reduces it further. The exceptions to this are cSVWN5/3-21G* and PBEh/3-21G*, and in both of those cases, the optimized values are still relatively small. For these functionals with STO-3G, the size dependence of the error is large and is reduced greatly by the parametrization. LSDA and HF also have extremely large initial slopes and are compensated for by the parametrization. For these methods, this improvement is probably due to the removal of large errors throughout the total test set. Finally, the large initial dependence of HF on size is

Table 2. Slope of the Line Created by Plotting the Number of Carbons versus the Difference between Experimental and Calculated Enthalpies of Formation for the First Eight Alkanes

method	original			GA optimized		
	STO-3G	3-21G*	3-21+G*	STO-3G	3-21G*	3-21+G*
LSDA	-111.6	-41.3	-37.8	-3.5	-2.0	-1.4
PBE	-66.0	-5.8	-2.3	-2.5	-0.9	-0.2
PBEh	-60.5	-0.4	2.4	-2.7	-0.9	-0.3
TPSS	-58.7	0.6	3.7	-2.0	-0.3	0.4
TPSSh	-57.7	1.6	4.5	-2.1	-0.3	0.3
HF	16.4	78.9	81.1	-2.0	0.8	1.5
cSVWN5		-0.8	3.3		-1.6	-1.0

due to the incompleteness of the these basis sets relative to the size necessary to converge exact exchange.

The differences between optimized and regular atomic energies are listed in the Supporting Information. Several of the series of corrections are all negative, in which case they are correcting overbinding, and the HF parameters are almost all positive, to correct underbinding. Methods without large mean errors, implying no large systematic errors, have a mixture of positive and negative values. As the values themselves are only interesting for implementation, they are omitted here.

IV. Conclusion

The use of optimized atomic energies to calculate enthalpies of formation calculated with small basis sets results in significant improvements when compared to experimental values. The improvements are large enough to allow small basis set density functional methods to achieve higher accuracy than PM3 for the first time. The improvement is even greater for STO-3G calculations, as using this basis previously gave enthalpies of formation that were off by hundreds of kilocalories per mole, and with atomic energy parameters, STO-3G now provides qualitatively correct values.

Acknowledgment. This work was supported by NSF-CHE-0457030 and the Welch Foundation. E.N.B. would like to acknowledge helpful comments from Nicole Brothers.

Supporting Information Available: The Supporting Information for this paper consists of the optimized parameters (three tables) and the difference of carbon atomic energies from the exact values both before and after parametrization (one table). This information is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Csonka, G. I.; Ruzsinszky, A.; Tao, J.; Perdew, J. P. *Int. J. Quantum Chem.* **2005**, *101*, 506.
- (2) Brothers, E. N.; Merz, K. M., Jr. *J. Phys. Chem. A* **2004**, *15*, 2904.
- (3) Riley, K. E.; Brothers, E. N.; Ayers, K. B.; Merz, K. M., Jr. *J. Chem. Theory Comput.* **2005**, *1*, 546.
- (4) Hohenberg, P.; Kohn, W. *Phys. Rev. B* **1964**, *136*, 864. Kohn, W.; Sham, L. *J. Phys. Rev. A* **1965**, *140*, 1133.

- (5) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
- (6) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374.
- (7) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200. LSDA is equivalent to the Gaussian keyword "SVWN5" and refers to using the fifth formula for local correlation proposed in the paper.
- (8) Hehre, W. J.; Stewart, R. F.; Pople, J. A. *J. Chem. Phys.* **1969**, *51*, 2657.
- (9) Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939.
- (10) Note that MIDI! (Easton, R. E.; Giesen, D. J.; Welch, A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta* **1996**, *93*, 281) could also have been selected as it is of comparable size and performance to these sets, although slightly larger than 3-21G* and 3-21+G*, and has approximately the same performance.
- (11) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (12) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (13) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029.
- (14) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.
- (15) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.
- (16) Roothan, C. C. *J. Rev. Mod. Phys.* **1951**, *23*, 69.
- (17) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.01; Gaussian, Inc.: Wallingford, CT, 2004.
- (18) Chakravorty, S. J.; Gwaltney, S. R.; Davidson, E. R.; Parpia, F. A.; Fischer, C. F. P. *Phys. Rev. A* **1993**, *47*, 3649.
- (19) Goldberg, D. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: San Mateo, CA, 1989.
- (20) Ruzsinszky, A.; Csonka, G. I. *J. Phys. Chem. A* **2003**, *107*, 8687, and references therein.

CT600109X

AM1/d Parameters for Magnesium in Metalloenzymes

Petra Imhof,[†] Frank Noé,[†] Stefan Fischer,[‡] and Jeremy C. Smith^{*†}

Computational Molecular Biophysics and Computational Biochemistry, IWR University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany

Received March 10, 2006

Abstract: AM1/d parameters are derived for magnesium, optimized for modeling reactions in metalloenzymes. The parameters are optimized with a Monte Carlo procedure so as to reproduce the geometries and energies of a training set calculated with density functional theory. The training set consists of compounds with magnesium coordinated to the oxygen atom of typical biological ligands. Optimization of AM1 parameters without extension to *d* functions leaves serious errors. The new AM1/d parameters provide a clear improvement in accuracy compared to the standard semiempirical methods AM1 and MNDO/d and will be particularly useful for modeling reactions in large biological systems at low computational cost.

1. Introduction

Magnesium is the metal cofactor of numerous metalloenzymes. A popular modeling approach to understanding such reactions in enzymes is the combined quantum mechanical/molecular mechanical (QM/MM) ansatz, where the region of interest (usually the active site) is treated quantum mechanically and the remainder of the enzyme is described with an empirical force field.^{1–3} Ab initio methods for the QM part are not only the most accurate but also the most computationally demanding and therefore used only in special cases. Alternatively, density functional (DFT) methods provide a more attractive balance of accuracy and computational cost than ab initio techniques and thus enjoy high popularity in the modeling of chemical reactions. However, although a single minimization step with DFT methods can be easily afforded, a complete optimization with thousands of such steps can become computationally costly. Especially when several of these minimizations are necessary, e.g., for the exploration of different reaction pathways, more economical methods are needed. Responding to this need, semiempirical methods provide a sufficiently accurate description of quantum regions in QM/MM setups of large systems for low computational cost.

Semiempirical methods derive their efficiency from explicit treatment of only valence electrons with a minimal basis set, the neglect of three- and four-center integrals, and the use of parametrized expressions for two-center integrals.^{4–8} The parameters are usually obtained by a fit of properties (e.g., heats of formation) to a variety of very small compounds. Often these training sets are not representative of reactions in biological systems. However, the situation can be improved by the development of reaction-specific parameters, which are tuned to most accurately describe the specific biological systems under study, at the expense of losing generality.

The AM1 model is at present one of the most suitable semiempirical methods for studying reactions,⁸ although it does have a tendency to predict bifurcated and too-weak hydrogen bonds.⁹

The standard AM1 parameters for magnesium have been developed for use in modeling the bacterial photosynthetic reaction center¹⁰ and were fitted to reproduce mainly properties of divalent magnesium compounds. These parameters work quite well for most of the compounds listed in ref 10, including magnesium porphyrin, but yield wrong angles for the geometry of 6-fold coordinated magnesium (e.g. $[\text{Mg}(\text{H}_2\text{O})_6]^{2+}$). The MNDO/d method¹¹ yields correct angle values but too long Mg–O bond lengths. Both methods use an *sp* basis for magnesium, and thus one cannot expect a proper description of hypervalent magnesium compounds.

In metalloenzymes, 6-fold coordinated magnesium is quite common (ref 12 provides a survey of the Brookhaven Protein

* Corresponding author phone: ++49 6221 8857; e-mail: biocomputing@iwr.uni-heidelberg.de.

[†] Computational Molecular Biophysics.

[‡] Computational Biochemistry.

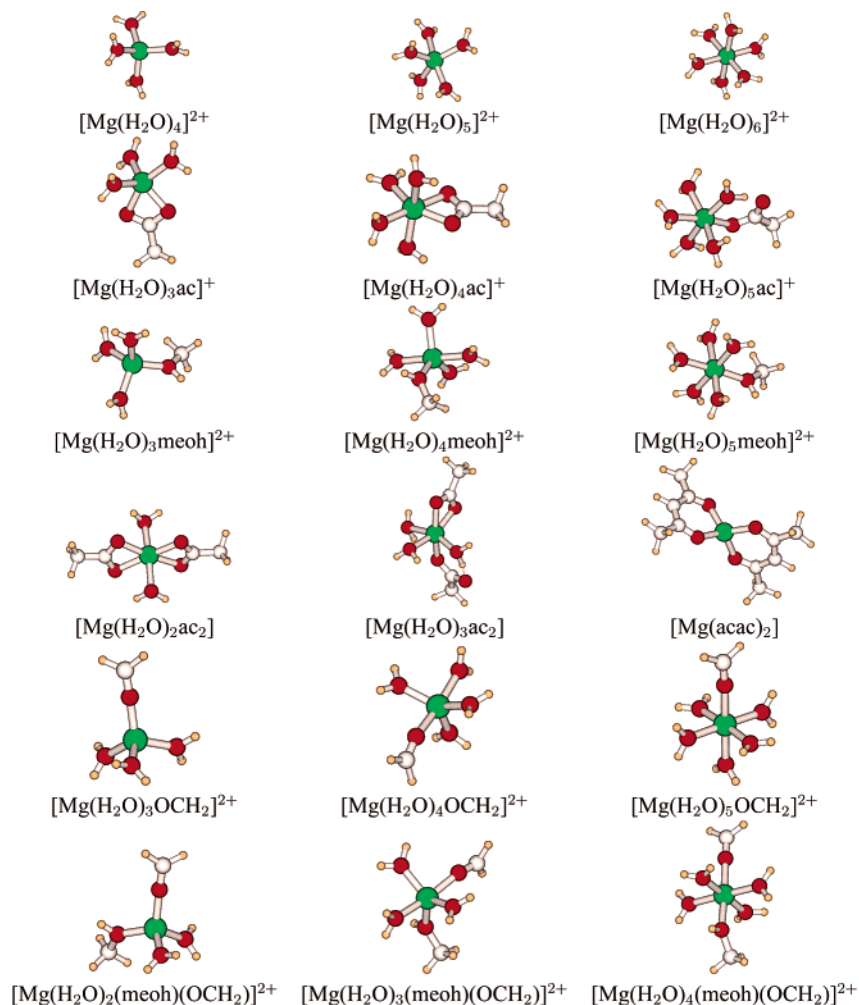


Figure 1. Compounds used in the training set for the magnesium parametrization.

Data Bank¹³ for X-ray and NMR structures of magnesium-bound proteins). To obtain a useful description of magnesium-containing active sites with different magnesium coordination spheres and magnesium-dependent reactions in metalloproteins at a semiempirical level, the present paper extends the AM1 parameters for magnesium to an *spd* basis in the AM1/d framework. The parameter set is derived specifically for oxygen-based ligands modeling magnesium coordination spheres that can be typically found in metalloproteins.

2. Methods

2.1. The Training Set. The AM1/d parameters for magnesium were derived by fitting properties of a set of magnesium compounds to a DFT training set consisting of model compounds for magnesium coordinated to the oxygen atom of typical biological ligands with coordination numbers 4, 5, and 6. These ligands are water, methanol (meoh), which models serine, threonine, and tyrosine amino acid side chains, acetate (ac) as a representative for aspartate and glutamate side chains, and formaldehyde (OCH₂) modeling the coordination by a backbone carbonyl oxygen atom. The compounds used in the training sets are shown in Figure 1. The Cartesian coordinates of the DFT optimized structures used as a training set are given as Supporting Information.

The DFT data set was obtained by geometry optimization with the B3LYP functional^{15,16} and a 6-31++G(d,p) basis set with subsequent single-point energy calculations using a 6-311++G(3df,2p) basis set. Normal-mode analysis on the optimized geometries was carried out to verify that a minimum energy structure has been obtained. All DFT calculations were performed using the Turbomole program package.¹⁷ The B3LYP/6-311++G(3df,2p)//B3LYP/6-31++G(d,p) procedure is abbreviated as DFT in the remainder of this paper. This procedure followed here is similar to that described in ref 14 for the development of AM1/d parameters for phosphorus reaction-specific for nucleophilic attacks on biological phosphates.

In ref 14, *d*-orbitals are introduced only where necessary, e.g. on the phosphorus, while treating C, H and O atoms with standard AM1 parameters. We follow a similar approach, by extending the AM1 basis set to *d*-orbitals where necessary (here for magnesium) while keeping as much of the standard AM1 model as possible. Thus, the magnesium complexes are composed of the ligand molecules, which are treated with standard AM1, and the additional Mg²⁺ ion, which is treated with the more extended AM1/d.

In a molecular orbital picture the basis functions of all atoms together form the molecular orbitals. Since mixed basis sets have to be used with care this would mean that, in a

Table 1. Optimized AM1/d and AM1 Parameters for Magnesium

parameter	AM1/d	AM1'
U_{ss}/eV	-16.63758	-12.83615
U_{pp}/eV	-11.97469	-9.51125
U_{dd}/eV	-10.90361	
β_s/eV	-3.60785	-1.26808
β_p/eV	-2.07794	-0.93230
β_d/eV	-3.30858	
ζ_s/au	1.16850	1.57114
ζ_p/au	1.07072	1.25833
ζ_d/au	0.93469	
$\alpha/\text{\AA}^{-1}$	1.28263	1.80310
a_1 (dimensionless)	1.84869	1.99069
$b_1/\text{\AA}^{-2}$	4.22931	3.80477
$c_1/\text{\AA}$	0.66917	0.66033
a_2 (dimensionless)	0.03381	-0.00626
$b_2/\text{\AA}^{-2}$	3.57399	3.06817
$c_2/\text{\AA}$	2.33163	1.53666
a_3 (dimensionless)	0.02860	-0.00581
$b_3/\text{\AA}^{-2}$	2.27472	2.33455
$c_3/\text{\AA}$	2.89337	2.42691
$\rho^{\text{core}}/\text{au}$	0.94048	
g_{sp}/eV	7.48305	8.29115
h_{sp}/eV	0.67433	0.53547
$\bar{\zeta}_s/\text{au}$	1.61862	
$\bar{\zeta}_p/\text{au}$	1.48840	
$\bar{\zeta}_d/\text{au}$	1.07347	

semiempirical framework, all parameters used must be reoptimized. In trial calculations, further reoptimization of the AM1 parameters was performed (data not shown). Only changes of the parameters for oxygen, which is directly bound to magnesium and thus should be most affected, resulted in any significant influence on the energy and geometry data. However, no significant improvement was obtained, and thus, for simplicity, standard parameters were retained for all elements other than magnesium.

Properties used for the fitting reported here include geometries, Mg–O bond distances and O–Mg–O angles, and reaction energies for ligand exchange (see Tables B in the Supporting Information). The reaction energies included in the fits are listed in the results section (see Table 2 and Table C in the Supporting Information).

Although the aim of the fitting is to obtain parameters that reproduce DFT geometries and *relative* DFT energies (reaction and protonation energies), the absolute heat of formation of $[\text{Mg}(\text{acac})_2]$ is included as a reference to keep the shift of the absolute energies moderate.

2.2. The Error Function. For AM1/d there are 25 adjustable parameters: U_{ss} , U_{pp} , and U_{dd} for the one-electron integrals; ζ_s , ζ_p , ζ_d , β_s , β_p , and β_d for the resonance integrals; and α , a , b , c , and ρ_{core} for the core–core interaction.⁶ For one-center two-electron integrals only the parameters g_{sp} and h_{sp} are given explicitly in the implementation of the MNDO program which was employed here,¹⁸ and the other one-center Coulomb integrals g_{ss} , g_{pp} , and g_{dd} are calculated from orbital exponent parameters $\bar{\zeta}_s$, $\bar{\zeta}_p$, and $\bar{\zeta}_d$.^{11,19,20}

In the optimization procedure, the AM1/d parameter set $\lambda = (U_{ss}, U_{pp}, \dots, \bar{\zeta}_d)$ was varied so as to minimize the deviation of geometries, reaction energies, and heats of formation with respect to the reference values. This deviation is measured by the following error function

$$\chi^2 = \sum_i \sum_a^{\text{compprop}} w_a [Y_{ia}^{\text{AM1/d}}(\lambda) - Y_{ia}^{\text{DFT}}]^2$$

where Y_{ia}^{DFT} is the DFT, and $Y_{ia}^{\text{AM1/d}}$ is the AM1/d value for property a of compound i . w_a is the weighting factor used for each property: bond lengths, bond angles, reaction energies, and heats of formation.

As start parameters the standard sp MNDO/d parameters²⁰ and the standard AM1 core–core parameters¹⁰ were taken. For the additional d specific start parameters were set: $U_{dd} = U_{pp}$, $\zeta_d = \zeta_p$, $\beta_d = \beta_p$ and $\bar{\zeta}_s = \zeta_s$, $\bar{\zeta}_p = \zeta_p$, $\bar{\zeta}_d = \zeta_d$. The weighting factors used were as follows: absolute energies 0.1 (kcal/mol)⁻², relative energies 1 (kcal/mol)⁻², bond distances 100 \AA⁻², bond angles $10^{\circ-2}$.

In each iteration of the optimization procedure, the properties on the semiempirical level were computed for fully geometry-optimized structures using a prerelease version of the MNDO99 program.¹⁸

2.3. Optimization. The error function χ^2 was minimized using a Monte Carlo procedure. This was initialized with the starting parameters λ_0 . At each step $t + 1$, a new parameter set λ_{t+1} was generated by randomly perturbing the previous parameter set λ_t

$$\lambda_{j,t+1} := \lambda_{j,t} + s(r - 0.5)\sigma_j$$

where s is the step length, $r \in [0, 1]$ is a random number, the index j runs over the parameters, and the standard deviations σ are identical to the initial parameter set $\sigma = \lambda_0$. A step and the new parameter set were accepted, if the new error function had a lower value than previously. Otherwise, it was rejected, and the old parameter set was kept. A step is also rejected, if one of the minimizations does not yield a true minimum (only positive vibrational frequencies).

The error function above was evaluated for each compound in each step, i.e. when the result for a compound produced terms whose sums were already larger than the old error value, the step was rejected immediately. The step length was changed adaptively. Upon an accepted step, the step length was multiplied by a factor of 1.5, otherwise it was divided by a factor of 2, while always remaining within a set of bounds, here: $s \in [0.05, 0.3]$.

3. Results and Discussion

Table 2 shows Mg–O bond distances, O–Mg–O angles, and the reaction energies of ligand substitution at the central magnesium. The optimized parameters are listed in Table 1. AM1' denotes the adjusted sp parameters, and AM1/d has fitted parameters for a spd basis.

Energies. Figure 2 shows reaction energies for ligand exchange reactions at the magnesium center calculated at the different semiempirical levels (AM1/d, AM1', AM1, and MNDO/d) plotted versus the DFT reference. Table 2 lists the respective values.

Table 2. Reaction Energies in kcal/mol for Magnesium Compounds^a

no.	reaction	CN	charge	DFT	AM1/d	AM1'	AM1	MNDO/d
1	[Mg(H ₂ O) ₆] ²⁺ + meoh → [Mg(H ₂ O) ₅ meoh] ²⁺ + H ₂ O	6 → 6	2 → 2	-3	3	1	-2	2
2	[Mg(H ₂ O) ₄] ²⁺ + meoh → [Mg(H ₂ O) ₃ meoh] ²⁺ + H ₂ O	4 → 4	2 → 2	-6	-1	-1	-4	-1
3	[Mg(H ₂ O) ₅] ²⁺ + meoh → [Mg(H ₂ O) ₄ meoh] ²⁺ + H ₂ O	5 → 5	2 → 2	-4	1	1	-3	0
4	[Mg(H ₂ O) ₅ ac] ⁺ → [Mg(H ₂ O) ₄ ac] ⁺ + H ₂ O	6 → 6	1 → 1	15	7	23	27	8
5	[Mg(H ₂ O) ₄] ²⁺ + OCH ₂ → [Mg(H ₂ O) ₃ OCH ₂] ²⁺ + H ₂ O	4 → 4	2 → 2	-5	-12	-4	-7	-16
6	[Mg(H ₂ O) ₅] ²⁺ + OCH ₂ → [Mg(H ₂ O) ₄ OCH ₂] ²⁺ + H ₂ O	5 → 5	2 → 2	-3	-8	-3	-5	-13
7	[Mg(H ₂ O) ₆] ²⁺ + OCH ₂ → [Mg(H ₂ O) ₅ OCH ₂] ²⁺ + H ₂ O	6 → 6	2 → 2	-1	-6	-2	-5	-11
8	[Mg(H ₂ O) ₄] ²⁺ + meoh + OCH ₂ → [Mg(H ₂ O) ₂ (meoh)(OCH ₂)] ²⁺ + 2H ₂ O	4 → 4	2 → 2	-11	-12	-5	-10	-16
9	[Mg(H ₂ O) ₅] ²⁺ + meoh + OCH ₂ → [Mg(H ₂ O) ₃ (meoh)(OCH ₂)] ²⁺ + 2H ₂ O	5 → 5	2 → 2	-7	-7	-3	-8	-13
10	[Mg(H ₂ O) ₆] ²⁺ + meoh + OCH ₂ → [Mg(H ₂ O) ₄ (meoh)(OCH ₂)] ²⁺ + 2H ₂ O	6 → 6	2 → 2	-4	-3	0	-6	-9
11	[Mg(H ₂ O) ₄] ²⁺ + 2H ₂ O → [Mg(H ₂ O) ₆] ²⁺	4 → 6	2 → 2	-57	-56	-65	-69	-59
12	[Mg(H ₂ O) ₅ ac] ⁺ → [Mg(H ₂ O) ₃ ac] ⁺ + 2H ₂ O	6 → 5	1 → 1	31	24	49	55	26
13	[Mg(H ₂ O) ₄ ac] ⁺ → [Mg(H ₂ O) ₃ ac] ⁺ + H ₂ O	6 → 5	1 → 1	17	17	27	28	18
14	[Mg(H ₂ O) ₅] ²⁺ + ac → [Mg(H ₂ O) ₅ ac] ⁺	5 → 6	2 → 1	-231	-225	-256	-265	-235
15	[Mg(H ₂ O) ₄] ²⁺ + ac → [Mg(H ₂ O) ₄ ac] ⁺	4 → 6	2 → 1	-246	-250	-268	-275	-260
16	[Mg(H ₂ O) ₅] ²⁺ + ac → [Mg(H ₂ O) ₄ ac] ⁺ + H ₂ O	5 → 6	2 → 1	-216	-218	-233	-238	-228
17	[Mg(H ₂ O) ₆] ²⁺ + ac → [Mg(H ₂ O) ₃ ac] ⁺ + 3H ₂ O	6 → 5	2 → 1	-173	-178	-176	-178	-183
18	[Mg(H ₂ O) ₄] ²⁺ + 2ac → [Mg(H ₂ O) ₂ ac ₂] ⁺ + 2H ₂ O	4 → 6	2 → 0	-354	-362	-374	-45	-379
19	[Mg(H ₂ O) ₆] ²⁺ + ac → [Mg(H ₂ O) ₅ ac] ⁺ + H ₂ O	6 → 6	2 → 1	-204	-201	-225	-233	-209
20	[Mg(H ₂ O) ₆] ²⁺ + ac → [Mg(H ₂ O) ₄ ac] ⁺ + 2H ₂ O	6 → 6	2 → 1	-189	-194	-203	-206	-201
21	[Mg(H ₂ O) ₄ ac] ⁺ + ac → [Mg(H ₂ O) ₂ ac ₂] ⁺ + 2H ₂ O	6 → 6	1 → 0	-108	-112	-106	230	-119
22	[Mg(H ₂ O) ₄ ac] ⁺ + ac → [Mg(H ₂ O) ₃ ac ₂] ⁺ + H ₂ O	6 → 6	1 → 0	-122	-124	-119	-140	-123
23	[Mg(H ₂ O) ₅ ac] ⁺ + ac → [Mg(H ₂ O) ₃ ac ₂] ⁺ + 2H ₂ O	6 → 6	1 → 0	-108	-117	-96	-112	-115
24	[Mg(H ₂ O) ₆] ²⁺ + 2ac → [Mg(H ₂ O) ₂ ac ₂] ⁺ + 4H ₂ O	6 → 6	2 → 0	-297	-306	-309	24	-320
25	[Mg(H ₂ O) ₆] ²⁺ + 2ac → [Mg(H ₂ O) ₃ ac ₂] ⁺ + 3H ₂ O	6 → 6	2 → 0	-312	-318	-322	-345	-324

^a AM1/d values are calculated with fitted magnesium *spd* parameters, AM1' with fitted *sp* parameters. Column CN gives the change in coordination number, charge lists the change in charge of the magnesium complex.

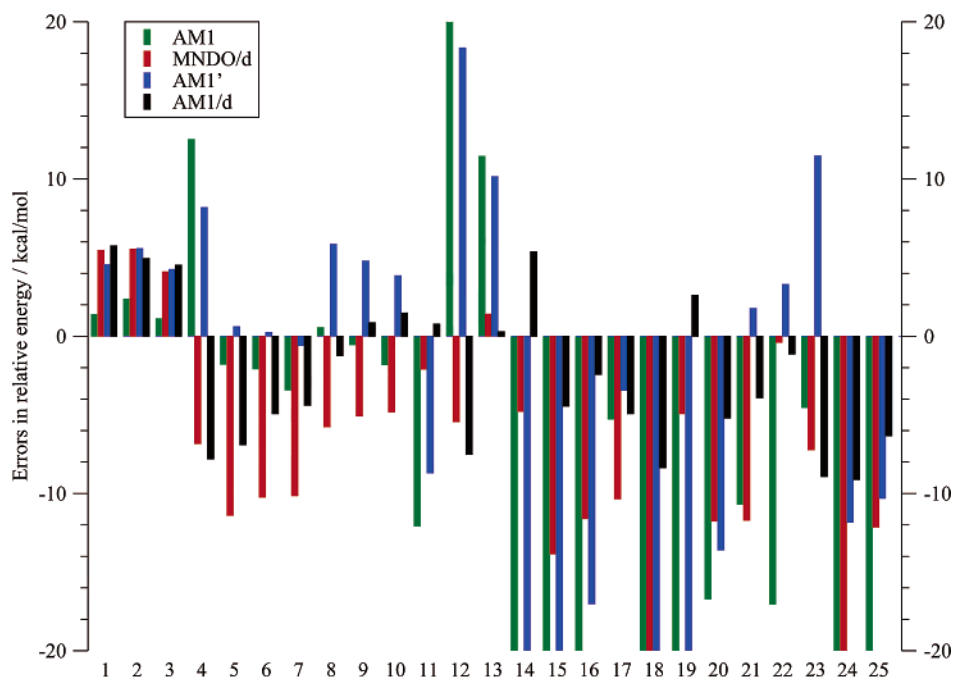


Figure 2. Errors of semiempirical reaction energies compared to DFT. AM1/d values are calculated with fitted magnesium *spd* parameters, AM1' with fitted *sp* parameters. The plot for energy errors is cropped at ± 20 kcal/mol, and some AM1 values exceed this range. The respective reactions are listed in Table 2.

The ligand exchange reactions can be separated into four classes: 1. reactions with change in coordination number (reactions 11–13), 2. reactions with change of the charge of the magnesium compound (19–25), 3. both of the above (14–18), and 4. neither of the above (1–10). As anticipated, the largest reaction energies are found for those reactions in

which the charge of the magnesium complex is decreased, i.e. opposite charges are brought together from an infinite distance. These reaction energies are less in solution^{12,21} or in a protein environment than in vacuo. One effect which leads to a reduction of the reaction energies is charge screening by the solvent. In addition, in a protein environ-

Table 3. Performance of the Semiempirical Methods AM1/D, AM1', AM1, and MNDO/d for the Magnesium Complexes in Figure 1^a

property (number of comparisons)	AM1/d	AM1'	AM1	MNDO/d
relative energies (25): mean abs. error/kcal/mol	5	10	14	9
relative energies: max abs. error/kcal/mol	9	25	39	26
bond lengths (93): mean abs. error/Å	0.02	0.15	0.07	0.07
bond lengths: max abs. error/Å	0.07	0.21	0.18	0.23
angles (199): mean abs. error/degree	4	11	15	4
angles: max abs. error/degree	24	40	93	31

^a AM1/d values are calculated with fitted magnesium *spd* parameters, AM1' with fitted *sp* parameters.

ment reaction energies would not be calculated as the difference between infinitely separated reactants and products but rather would include electrostatic interaction of the reacting partners in both reactant and product states. For this type of reaction, the semiempirical methods show the largest difference from the DFT reference. MNDO/d is closer than standard AM1 but still deviates 10–20 kcal/mol from the DFT reaction energies. Fitting of the *sp* parameters brings the AM1' values close to those of MNDO/d and is even better in two cases. With inclusion of *d* parameters, however, the reaction energy errors are significantly reduced, to at most 9 kcal/mol.

Changes in the coordination number with conservation of the charge are reproduced better by MNDO/d than standard AM1. Optimized *sp* parameters (AM1') do not significantly improve the results. An extension to *d* functions is clearly necessary for a proper energetic description of reactions with hypervalent magnesium compounds.

For those reactions in which neither the coordination number nor the charge of the magnesium complex change, all semiempirical methods perform quite well.

The average absolute error of all reactions evaluated is 5 kcal/mol for AM1/d and is significantly lower than those for AM1', AM1, and MNDO/d, see Table 3. The larger and thus more flexible *spd* basis clearly provides a more balanced description of the different types of ligand exchange reaction.

Geometries. Mg–O distances in magnesium compounds calculated with AM1/d deviate by at most 0.07 Å, i.e. 3%, from the DFT values to larger and smaller distances, the mean absolute error being 0.02 Å. With both MNDO/d and standard AM1 the distances are too long, by 0.07 Å on average (see Table 3). AM1' with fitted *sp* parameters strongly underestimates the Mg–O bond lengths, which are uniformly shifted by –0.15 Å relative to the AM1/d bond lengths. The improvement in bond distances by AM1/d is due to the fitting procedure, in which specific parameters have been derived for a class of compounds in which the magnesium atom is directly bound only to oxygen atoms. Standard parameters derived for more general applicability must simultaneously represent other bond types such as Mg–C, Mg–H, or Mg–X (X = halogen), which is a more difficult task. In MNDO/d a partial tuning is achieved by interaction specific core parameters α for Mg–H, Mg–C, and Mg–S.²⁰

As shown in Figure 4 O–Mg–O bond angles range from about 60° (at the bidentate acetate ligand) to linear (180°). The standard AM1 values strongly deviate from the angles calculated with DFT (mean absolute error: 15°, see Table

3) and cannot be improved significantly by fitting the *sp* parameters. MNDO/d shows an average error in bond angles of only 4°, the same as is achieved with fitted *spd* AM1 parameters. The maximum AM1/d error for O–Mg–O angles is 24° compared to 31° calculated for MNDO/d. For both methods, the largest angle errors can be attributed to errors in the treatment of intramolecular hydrogen bonds, rather than inaccuracies in the magnesium parameters: a too weak O_a–H···O_b interaction leads to a too large O_a–Mg–O_b angle. This effect is particularly pronounced for those complexes including acetate. The structure of [Mg(H₂O)₃-ac₂] is the worst case in this regard: the hydrogen atoms point in different directions compared to the DFT optimized structure, leading to bifurcated hydrogen bonds, to which AM1 is known to be prone.⁹ Interestingly, the bite angle of the acetate ligand (ca. 119°, which is not included in the training set properties) is also reproduced best for all complexes with AM1/d. However, the improvement on standard AM1 is marginal. This may be attributed to the fact that the use of standard parameters for first row elements (H, C, O) leads to well-reproduced O–C–O angle values. However, this agreement shows that the presented optimized magnesium parameters indeed work in concert with these standard parameters and lead to an overall improvement. As an additional test, we combined the AM1/d parameters for phosphorus from ref 14 with our AM1/d parameters for magnesium (and standard parameters for H, C, and O) and evaluated the reaction of pentaquomagnesium dimethyl phosphate with water to pentaquomagnesium methyl phosphate plus methanol (for structures see Figure 1 in the Supporting Information). The hydrolysis of dimethyl phosphate has been taken into account in the parametrization for phosphorus in ref 14. The Mg–O and P–O distances agree on average within 0.02 Å with the DFT-optimized distances (the maximal error is 0.07 Å in 20 distances), and the O–Mg–O and O–P–O angles differ on average 7° (42 angles). The largest geometric differences from DFT optimized values for the magnesium phosphates is 26° for one O–Mg–O angle. The AM1/d calculated reaction energy of –4 kcal/mol agrees well with the DFT value of 1 kcal/mol. This shows that a combination of specific AM1/d parameters for phosphorus and magnesium can be combined together and with standard (C, H, O) AM1 parameters to give sufficiently reliable results.

4. Conclusions

The present paper presents the results of the development of AM1/d parameters for magnesium. These parameters

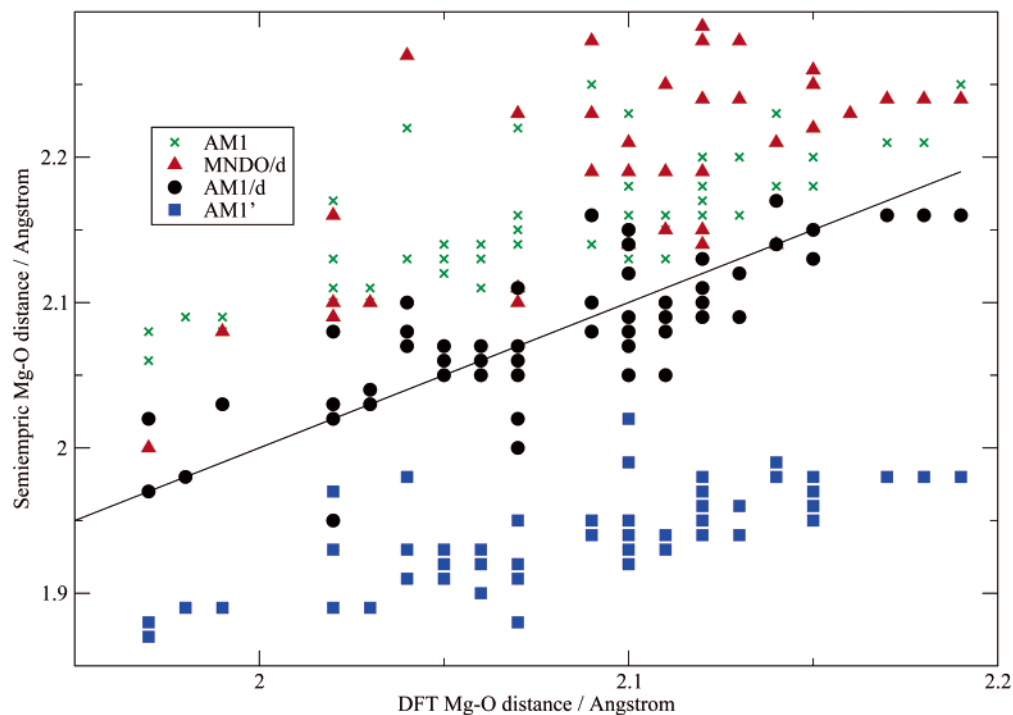


Figure 3. Semiempirical vs DFT Mg–O bond distances. AM1/d values are calculated with fitted magnesium *spd* parameters, AM1' with fitted *sp* parameters.

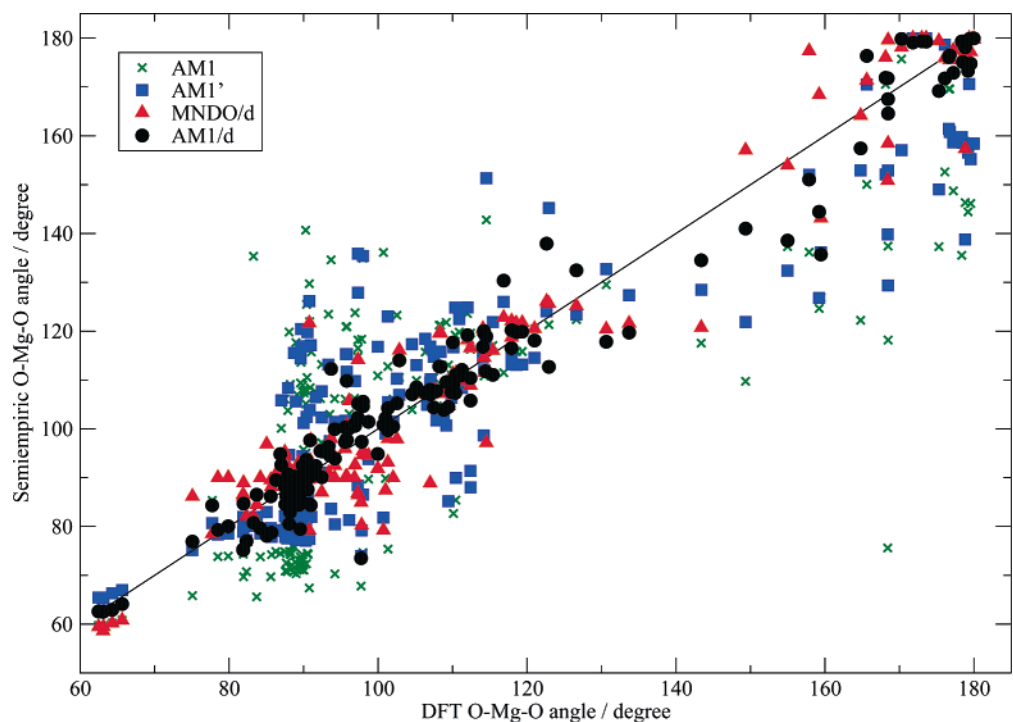


Figure 4. Semiempirical vs DFT O–Mg–O angles. AM1/d values are calculated with fitted magnesium *spd* parameters, AM1' with fitted *sp* parameters.

provide a significantly improved description of biologically important magnesium complex geometries and reaction energies on a semiempirical level relative to standard semiempirical methods. Attempts to fit AM1 parameters for an *sp* basis are of limited success, showing, that for a proper semiempirical description of hypervalent compounds, the extension of the basis to *d* orbitals is necessary.

For the systems investigated in this work MNDO/d turns out to be superior to standard AM1. The quality of the

specifically parameterized AM1/d results, however, is clearly superior to that of the standard methods. This shows that the effort of developing reaction- or system-specific parameters is worthwhile when high accuracy is desired, rather than covering a large variety of compounds.

Remaining deviations from the DFT values can be traced back to the underestimation of hydrogen-bond strengths on the AM1 level. The compounds used in the magnesium training set cover a variety of possible coordination spheres

for biological magnesium and may thus be used in applications to a broad range of magnesium-containing proteins. They also work well for magnesium phosphates, when combined with the phosphorus parameters reported in ref 14. However, to cover all possible magnesium-coordination partners in proteins, the parametrization has to be extended to include Mg–N bonds such as magnesium–histidine complexes. This is subject of ongoing work.

Particularly when used in combined QM/MM calculations the new AM1/d parameters reported here furnish a method for modeling magnesium-containing biological systems with reasonable accuracy at low computational cost.

Acknowledgment. We thank Walter Thiel for support with a prerelease version of the MNDO program. This work was funded by the Deutsche Forschungsgemeinschaft as part of the SFB 623.

Supporting Information Available: Cartesian coordinates of the B3LYP/6-31++G(d,p) optimized structures of the training set and Mg–O distances and O–Mg–O angles derived therefrom as well as reaction energies and total energies. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Gao, J.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467–505.
- (2) Warshel, A. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425–443.
- (3) Åqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523–2544.
- (4) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.
- (5) Thiel, W.; Voityuk, A. A. *Adv. Chem. Phys.* **1996**, *93*, 703–757.
- (6) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209–220.
- (7) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (8) Bredow, T.; Jug, K. *Theor. Chim. Acta* **2005**, *113*, 1–14.
- (9) Dannenberg, J. J.; Evleth, E. M. *Int. J. Quantum Chem.* **1992**, *44*, 869–885.
- (10) Hutter, M. C.; Reimers, J. R.; Hush, N. S. *J. Phys. Chem. B* **1998**, *102*, 8080–8090.
- (11) Thiel, W.; Voityuk, A. A. *Int. J. Quantum Chem.* **1992**, *44*, 807–829.
- (12) Dudev, T.; Cowan, J. A.; Lim, C. *J. Am. Chem. Soc.* **1999**, *121*, 7665–7673.
- (13) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (14) Lopez, X.; York, D. M. *Theor. Chim. Acta* **2003**, *109*, 149–159.
- (15) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040–1046.
- (16) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (17) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.
- (18) Thiel, W. *MNDO Version 6.1*; Max-Planck-Institut fuer Kohlenforschung, Mülheim a.d. Ruhr, Germany, 2004.
- (19) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1992**, *81*, 391–404.
- (20) Thiel, W.; Voityuk, A. A. *J. Phys. Chem.* **1996**, *100*, 616–626.
- (21) Mayaan, E.; Range, K.; York, D. M. *J. Biol. Inorg. Chem.* **2004**, *9*, 807–817.

CT600092C

JCTC Journal of Chemical Theory and Computation

Assigning the Protonation States of the Key Aspartates in β -Secretase Using QM/MM X-ray Structure Refinement

Ning Yu,^{†,‡} Seth A. Hayik,[†] Bing Wang,[†] Ning Liao,[†] Charles H. Reynolds,[§] and Kenneth M. Merz, Jr.^{*,†,||}

*Department of Chemistry, The Pennsylvania State University,
104 Chemistry Research Building, University Park, Pennsylvania 16802, and
Johnson & Johnson Pharmaceutical Research and Development, L.L.C.,
P.O. Box 776, Welsh and McKean Roads, Spring House, Pennsylvania 19477-0776*

Received January 4, 2006

Abstract: β -Secretase, aka β -APP cleaving enzyme (BACE), is an aspartyl protease that has been implicated as a key target in the pathogenesis of Alzheimer's disease (AD). The identification of the protonation states of the key aspartates in β -secretase is of great interest both in understanding the reaction mechanism and in guiding the design of drugs against AD. However, the resolutions of currently available crystal structures for BACE are not sufficient to determine the hydrogen atom locations. We have assigned the protonation states of the key aspartates using a novel method, QM/MM X-ray refinement. In our approach, an energy function is introduced to the refinement where the atoms in the active site are modeled by quantum mechanics (QM) and the other atoms are represented by molecular mechanics (MM). The gradients derived from the QM/MM energy function are combined with those from the X-ray target to refine the crystal structure of a complex containing BACE and an inhibitor. A total number of 8 protonation configurations of the aspartyl dyad were considered, and QM/MM X-ray refinements were performed for all of them. The relative stability of the refined structures was scored by constructing the thermodynamic cycle using the energetics calculated by fully quantum mechanical self-consistent reaction field (QM/SCRF) calculations. While all 8 refined structures fit the observed electron density about equally well, we find the monoprotonated configurations to be strongly favored energetically, especially the configuration with the inner oxygen of Asp32 protonated and the hydroxyl of the inhibitor pointing toward Asp228. It was also found that these results depend on the constraints imposed by the X-ray data. We suggest that one of the strengths of this approach is that the resulting structures are a consensus of theoretical and experimental data and remark on the significance of our results in structure based drug design and mechanistic studies.

Introduction

Amyloid plaques are extracellular deposits of β -amyloid proteins ($A\beta$) which accumulate outside the brain's nerve

cells. The presence of amyloid plaques in the brain is a characteristic feature of Alzheimer's disease (AD).¹ $A\beta$ is derived in vivo from proteolytic cleavage of the membrane-anchored amyloid precursor protein (APP) by β - and γ -secretases.^{2–5} Therefore, designing small molecule drugs that can inhibit $A\beta$ production constitutes a promising strategy for treating AD, especially for patients who are still in the early clinical phases of the disease with minimal cognitive impairment.

One of the therapeutic targets in these drug design efforts, β -secretase or β -APP cleaving enzyme (BACE), belongs to

* Corresponding author e-mail: merz@qtp.ufl.edu.

[†] The Pennsylvania State University.

[‡] Present address: GlaxoSmithKline Pharmaceuticals, 1250 South Collegeville Road, Collegeville, PA 19426.

[§] Johnson and Johnson Pharmaceutical Research and Development.

^{||} Present address: Department of Chemistry, Quantum Theory Project, University of Florida, 2328 New Physics Building, P.O. Box 118435, Gainesville, FL 32611-8435.

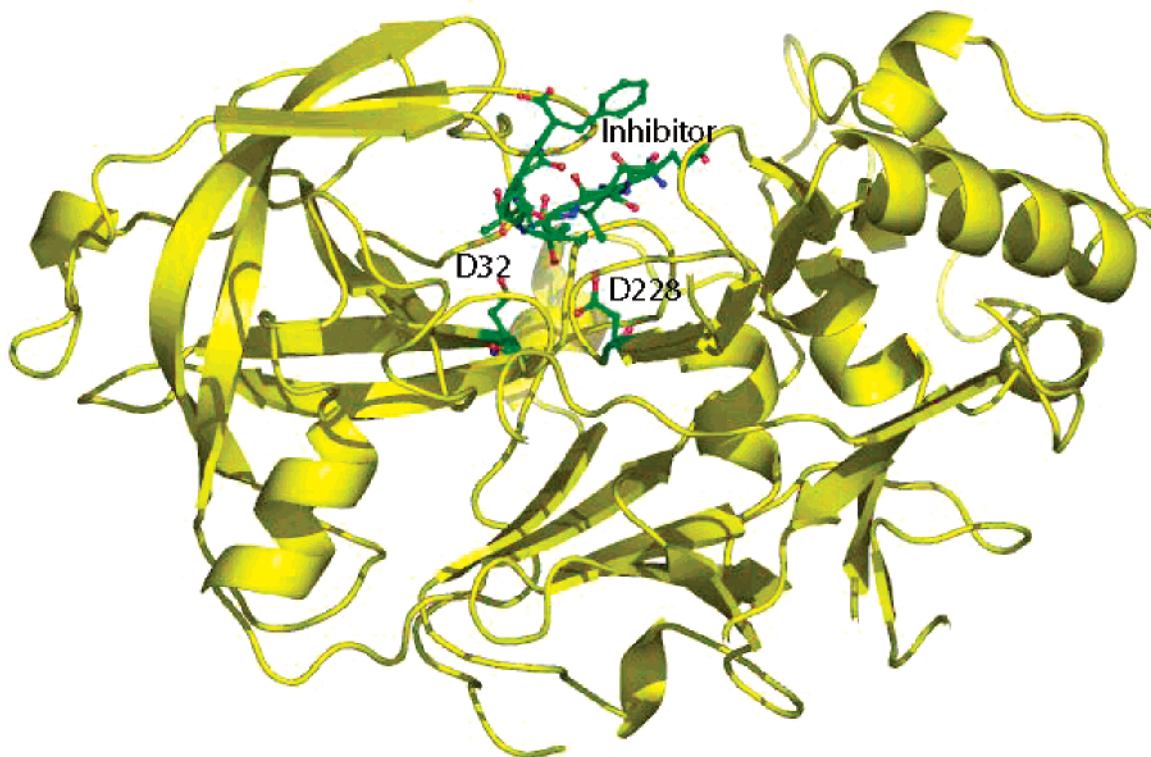


Figure 1. The 1FKN structure⁶ where the protein is rendered with ribbon representation and the catalytic aspartates and the inhibitor are rendered with ball-and-stick models.

the family of aspartyl proteases. A number of crystal structures have been solved which provide direct structural information on this important enzyme. The first X-ray structure of β -secretase in complex with an inhibitor OM99-2 (1FKN) was solved by Hong et al. at a resolution of 1.9 Å.⁶ Hong and co-workers later determined the crystal structure of BACE bound with a more potent inhibitor OM00-3 (1M4H) at a resolution of 2.1 Å⁷ and the crystal structure of apo BACE (1SGZ) at a resolution of 2.0 Å.⁸ Patel and colleagues solved the structure of apo-BACE (1W50) at 1.75 Å and the inhibitor bound BACE (1W51) at 2.55 Å.⁹ Figure 1 shows the 1FKN structure solved by Hong et al.⁶ A common structural feature among aspartyl proteases is the presence of two aspartates near the active sites,^{10,11} which, in the case of β -secretase, are Asp32 and Asp228.

The reaction mechanism shown in Figure 2 was proposed by Andreeva et al.,¹⁰ which involves a base-catalyzed attack of a water molecule on the scissile amide carbonyl to form a tetrahedral intermediate. The reaction then proceeds via C–N bond cleavage, yielding the products of proteolysis. This mechanism implies a monoprotonated arrangement of the catalytic aspartates^{12–16} and a complex network of hydrogen bonds at the active site. Figure 3 displays the definitions of the possible protonation states as well as the inner and outer oxygen atoms of Asp32 and Asp228 that will be considered in this work. This mechanism also suggests that if β -secretase starts a reaction cycle with Asp32 protonated and Asp228 ionized, the key aspartates will be in the opposite protonation states after formation of the tetrahedral intermediate, namely Asp228 will be protonated as a result of abstraction of a proton from the attacking water molecule and Asp32 will in turn be deprotonated after

delivering its proton to the carbonyl of the substrate. The fact that one of the aspartates is believed to be protonated suggests a highly hydrophobic active site environment, which causes a large pK_a shift on one or both of the side chains of the key aspartates. On the other hand, if the monoprotonated configuration is assumed by β -secretase under physiological conditions, a further question arises regarding the respective protonation states of Asp32 and Asp228 at each step in the enzyme's catalytic cycle.

The issue of the protonation patterns of the key aspartyl groups in β -secretase has spurred widespread interests from both the experimental and theoretical perspectives. In principle, the protonation states of buried ionizable residues can be directly probed by locating the coordinates of hydrogen atoms in diffraction experiments. However at the resolutions that the β -secretase crystal structures were solved,^{6–9} hydrogen atom coordinates cannot be determined. Neutron scattering is another diffraction technique and can locate hydrogen atoms directly. It has been applied in a study by Coates et al. to determine the protonation state of the critical residues in endothiapepsin.¹⁷ Nevertheless, the resolutions and R values of neutron structures are often much poorer than those of similar X-ray structures, and currently the difficulties associated with this technique have limited its application to a broader range of systems. Despite the lack of unequivocal structural evidence, much of the experimental work on this subject has based their analyses on the assumption of a monoprotonated configuration. For example, Touloukhonova et al.¹⁸ employed peptide inhibition data, solvent kinetic isotope effects, and proton NMR spectroscopy to study the steady-state kinetics mechanism of the proteolysis reaction of BACE in the presence of its

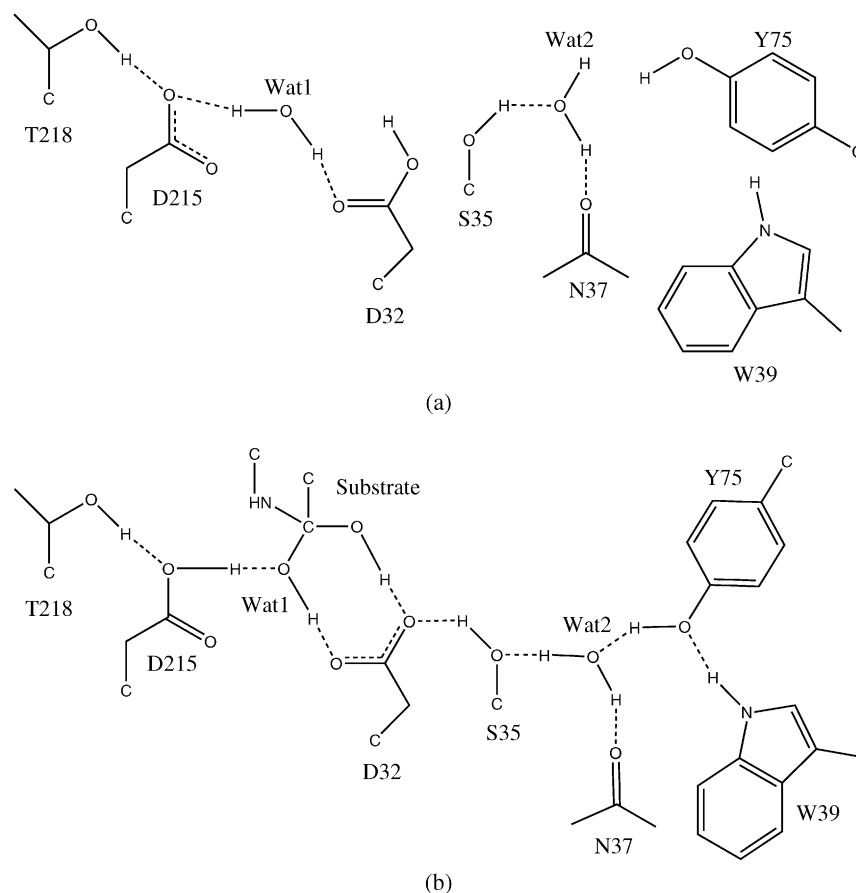


Figure 2. Schematic representation of the proposed acid–base mechanism proposed by Andreeva et al.¹⁰ for the proteolysis reactions catalyzed by aspartyl proteases. (a) The initial step before substrate binding. (b) The tetrahedral intermediate and its interactions with the key residues. The numbering scheme is for pepsins.

substrates and inhibitors and pointed to the steps following the collapse of the tetrahedral intermediate as rate limiting. On the theoretical side, Park and Lee utilized molecular dynamics (MD) simulations and computational docking experiments to assess the relative stability of the protonation states when the enzyme was bound to an 8-residue peptidomimetic inhibitor (OM99-2),¹⁹ where the scissile amide carbonyl was replaced by a hydroxyethylene fragment. Two monoprotinated configurations were considered in this study, one of them being the state where Asp32 was protonated and the hydroxyl of the hydroxyethylene pointed toward Asp228 (32i in Figure 3) and vice versa (228i in Figure 3). Although the simulations suggested the former to be the favored arrangement, the possibilities of diprotonated and dideprotonated configurations and alternative monoprotinated states were not addressed. Rajamani and Reynolds employed a quantum mechanical description of the same OM99-2 bound enzyme and studied all the protonation states in Figure 3.²⁰ They performed energy minimization on the coordinates of the atoms in the key region in the presence of the protein environment approximated by a truncated system and concluded that the energetically favored configuration was the monoprotinated 228i state. Polgar and Keseru performed pK_a calculations also on the 1FKN structure, and the titration curves they computed suggested that the 32i state was most probable when the inhibitor was bound to the enzyme.²¹

Studies of the protonation states on a related system in the aspartyl protease family, HIV-1 protease, have not generated a consistent answer either. Yamazaki and co-workers showed NMR and X-ray evidence that the aspartyl dyad adopted a diprotonated configuration in a complex of the enzyme bound to a non-peptide cyclic urea-based inhibitor.²² Hyland et al. investigated the kinetic mechanism of the reaction for 4 oligopeptide substrates and 2 competitive inhibitors²³ and proposed that substrates should only bind to HIV-1 protease in the monoprotinated state.²⁴ Piana and colleagues carried out a series of MD simulations using the Car-Parrinello (CP) method^{25,26} and the QM/MM approach²⁷ to assign the protonation state and to characterize the free energy profile of the catalytic reaction. The CPMD results on the free enzyme suggested that the monoprotinated state was most likely,²⁵ which was used as the basis to compute the reaction free energies.²⁷ In the complex formed by HIV-1 protease bound to pepstatin A, however, the diprotonated configurations were calculated to be more stable, and the ¹³C NMR chemical shifts and isotopic shifts for the diprotonated configurations simulated in the CPMD calculations were shown to be consistent with experiments.²⁶ A rationalization was provided which pointed to the polarity of the ligand as an important determinant of the protonation state of the receptor.²⁶

In this paper, we apply a new technique, quantum mechanical/molecular mechanical (QM/MM) X-ray structure

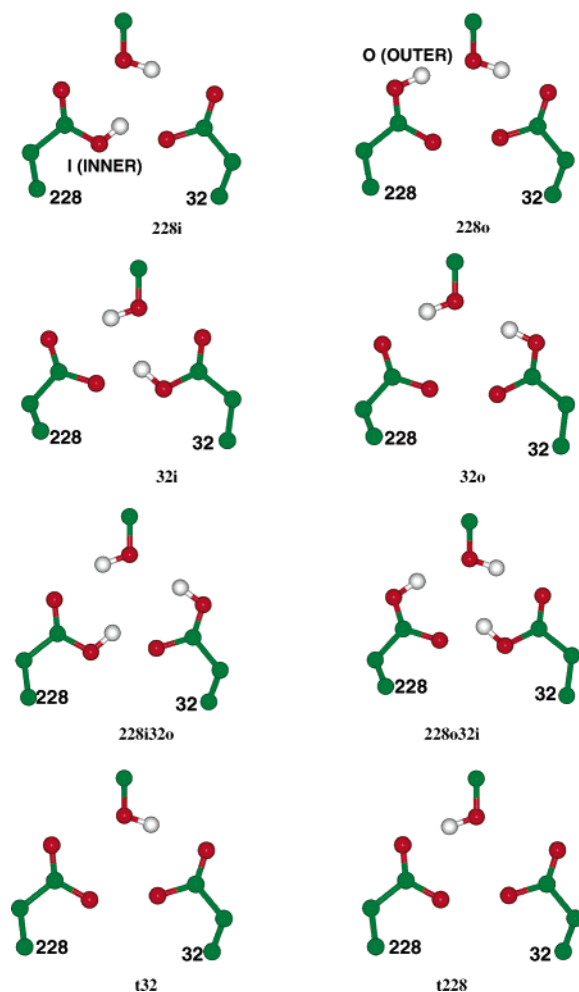


Figure 3. Definitions of the monoprotonated, diprotonated, and dideprotonated states considered in this work and the inner and outer oxygen atoms of the key aspartates.

refinement, to assign the protonation states of the key aspartates in a BACE crystal. This approach constrains the QM/MM calculations to be consistent with the X-ray diffraction data. We have chosen the 1FKN structure as the basis for our calculations because it is a relatively high-resolution crystal structure and has been studied very extensively in previous theoretical investigations.^{19–21} Under the constraint of the X-ray diffraction data, we will construct a set of all-atom models containing the coordinates for hydrogen atoms in the 8 protonation states defined in Figure 3 and use an accurate energy function to identify the most probable one. Indeed, while much of the previous modeling on aspartyl proteases has involved energy minimization of X-ray coordinates, we are carrying out modeling constrained by experiment and are asking specifically what the protonation pattern of the observed X-ray experiment is. Energy minimization of protein structures starting from X-ray coordinates could introduce computational artifacts (e.g., local groups undergoing significant conformational changes, alternate protonation patterns for local groups, etc.), which could alter the prediction of the preferred protonation state of the crystal structure and ultimately affect the outcome of calculations on mode of action and inhibition. It is also possible that the crystal structure imposes constraints

on the structure that will not be considered in unconstrained models. The 1FKN structure was crystallized in 0.2 M ammonium sulfate and 22.5% PEG 8000 buffered with 0.1 M Na-cacodylate at a pH of 7.4 and a temperature of 20 °C.⁶ While these conditions may be quite different from those under which the experimental mechanistic studies and binding assays are performed, it is generally expected that the physiologically relevant conformations of proteins are not substantially altered by their incorporation into crystal lattice. Furthermore, considering that the solvent content of the 1FKN structure is 56%,⁶ we hypothesize that the crystal structure is a reasonable model for studying the protonation state of BACE.

Theoretical Background

In protein crystallography, a model describing the electron density distribution within the unit cell is constructed with atomic coordinate parameters and vibrational parameters. These parameters are adjusted so that the predicted structure factors computed from the model electron density by Fourier transform best fit the experimental signals. In practice, refining the coordinate and vibrational parameters for all the atoms in a protein is often an underdetermined process since the amount of diffraction data is usually not sufficient to fully determine all the parameters. Consequently, the energy refinement formalism (EREF)²⁸ was introduced to remedy this problem, giving rise to the following equation

$$E_{\text{total}} = E_{\text{chem}} + w_{\text{X-ray}} E_{\text{X-ray}} \quad (1)$$

where E_{total} is the function to be minimized in structure refinement, E_{chem} is a stereochemical energy function, $E_{\text{X-ray}}$ is an X-ray target function describing the discrepancy between the observed and predicted structure factors, and $w_{\text{X-ray}}$ is the weight at which the dimensionless quantity, $E_{\text{X-ray}}$, is factored into E_{total} . The purpose of EREF is to overcome the problem of a poor data-to-parameter ratio by supplementing the observations, $E_{\text{X-ray}}$, with a restraining energy function, E_{chem} .^{28–30} Due to various difficulties discussed in earlier papers on this subject, in routine crystal structure refinement E_{chem} has evolved into an incomplete molecular mechanics energy function without the electrostatics and attractive van der Waals terms.^{31,32} Moreover, the bond-length and bond-angle parameters employed in commonly used refinement programs such as Crystallography and NMR System (CNS) were derived from a statistical analysis of high-resolution crystal structures of small molecules by Engh and Huber.³³ Utilization of an incomplete energy function, despite its necessities, has a few important limitations: (1) It makes orienting hydrogen atoms very difficult if their coordinates are desired. (2) The value computed with E_{chem} does not provide an estimate of the true energy of the system because of the omission of some of the key contributions. Hence, the motivation for the present work is obvious: we wish to see whether replacing the incomplete energy function with a physical potential will lead to final structures that both satisfy the experimental signals and provide a basis for an analysis of the energetics

of the system. The use of QM for the active site not only provides a description with enhanced accuracy but also allows possible electronic structure changes to take place. In a recent contribution, we presented a study where we combined X-ray reflection data with linear-scaling QM calculations to refine the crystal structure of bovine pancreatic trypsin inhibitor (BPTI).³⁴ Through comparisons with the structures refined with the MM potential in CNS, we demonstrated that the QM energy restraints were capable of maintaining reasonable stereochemistry to the extent that the R and R_{free} values of the QM refined structures are comparable to those of the CNS ones.

Our method is similar to the one pioneered by Ryde et al., who utilized the energy restraints derived from QM/MM calculations to refine the crystal structures of several protein–ligand cocrystals.^{35–41} Major differences include the following: (1) We only apply the QM/MM X-ray refinement to construct a set of structures that satisfy the X-ray data equally well and score their relative stability with a more accurate energy function with a continuum description of the bulk solvent. (2) Our implementation is different, which will be discussed in more detail in the following section.

It is also important to highlight the differences between this approach and the calculations used to estimate the $\text{p}K_{\text{a}}$ shifts of ionizable residues:^{42–45} since our structures are refined within the context of X-ray data, which is largely dominated by the configurations that are accessible energetically, the computed energy differences between the favored and disfavored states are inevitably larger than the true values as the structures of the disfavored states are also constrained to fit the X-ray data. This is necessarily so because when the X-ray data are used to constrain the structure it forces a state to occupy a region of geometric space that could be highly destabilizing to its specific protonation pattern. In other words, by using experimental constraints we are disallowing unstable structures to relax into geometries that while lower in energy represent structures that poorly reproduce the observed experimental data. In a way this exaggeration may be regarded as an amplification of the signal of interest, which can be helpful in cases where the energy differences between states are close to the expected margin of error of the energy function.

Finally, it should be noted that one of the greatest advantages of our method is that it provides an escape from the vagaries of energy minimization. Because the structures that we perform calculations on satisfy the constraints imposed by experimental signals, they should be in principle physically realistic and free of the artifacts that may be present when the structures are solely determined by approximate models of protein structure. On the other hand, this method can overcome the potential biases caused by systematic and random errors in crystal structures. As Ryde et al. pointed out,⁴¹ the QM/MM refinement approach allows all the atoms in the system to relax as the structure of the active site is refined, subject to the constraint imposed by the diffraction data, which is important in preventing errors in the coordinates of the atoms outside the active site from propagating into the structure of the region in question.

Computational Procedures

All refinements were performed with the AMBER^{8,46} and CNS⁴⁷ software packages via an interface that coupled the two programs together. The SANDER module is the main energy minimization/molecular dynamics driver in the AMBER package. The component that handles the QM calculations in SANDER is our linear-scaling semi-empirical electronic structure program DivCon.^{48–51} We modified the routines in SANDER that compute forces to make an additional call to the interface, where the atomic coordinates were output to a scratch file. CNS is then invoked via a system call to calculate the X-ray target function and its Cartesian-space gradient based on the coordinates in the scratch file. In practice, this was accomplished by modifying the CNS input script, minimize.inp, in the same manner as Ryde et al.⁴¹ Next the X-ray target function and the gradient deposited in the scratch files were read into SANDER and added to the physical energy and gradient according to eq 1. SANDER X-ray structure refinement proceeds by minimizing the total target function using either the steepest descent or the conjugate gradient method.

The structure and X-ray data were taken from the crystal structure of OM99-2 bound β -secretase⁶ available in the protein data bank (PDB ID: 1FKN). All the atoms in the crystallographic model were retained in our simulations in order to ensure full reproduction of the observed electron density. Because of the intrinsic symmetry in the crystallographic model, only one of the two protein chains, one of the two inhibitor chains, and the waters were refined. Initially, when adding hydrogen atoms to the crystal structure with the LEaP program in AMBER,⁴⁶ we assumed that all the titratable groups adopted their most common protonation states at $\text{pH} = 7.0$. The all-atom model constructed in this way contains 13 915 atoms with a net charge of -24 . This starting structure was subject to 1000–1500 steps of refinement using a MM potential and restraints derived from the X-ray data. In all the refinements discussed in this work, the atomic B factors were held fixed at their values in the crystal structure, and hydrogen atoms were neglected in the calculation of the structure factors. The reflection data file that we obtained from the PDB contains 69 056 reflections between the resolution limits of 1.90 and 24.90 Å, in which 6748 reflections were marked with the free R flag. The experimental paper reports an R value of 0.180 and a R_{free} value of 0.228, which were reproduced successfully with our version of CNS.

After preprocessing, the structure refined with $w_{\text{X-ray}} = 0.4$, which had an R value of 0.186 and an R_{free} value of 0.222, was selected for further processing. The initial structures for the 8 protonation configurations as defined in Figure 3 were constructed in the same manner as in ref 20. These 8 initial structures were refined with our QM/MM X-ray refinement method for another 700 to 1000 steps. Figure 4 shows the chemical structure of the inhibitor as well as the partitioning scheme in the QM/MM calculations. The QM region consists of the two key aspartates and the nonstandard residues Lol and Alq of the inhibitor, making the total number of QM atoms between 68 and 70. The

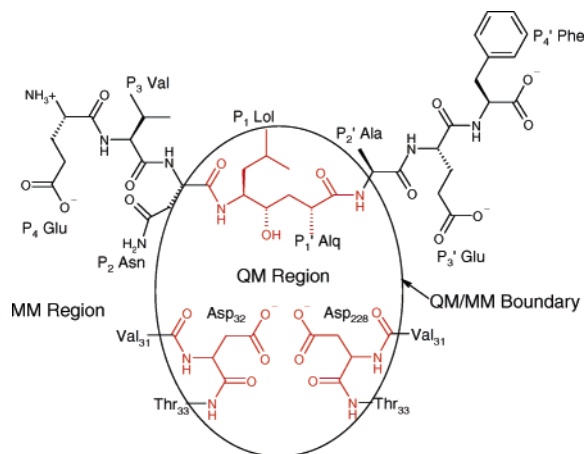


Figure 4. Chemical structure of OM99-2 and the partitioning of QM and MM regions. Atoms colored in red are in the QM region, whereas atoms colored in black are in the MM region.

decision to adopt such a partitioning scheme was motivated not only by computational efficiency but also by necessity to avoid computational artifacts as a result of our present inability to model bulk solvent in QM/MM calculations. In fact, we have found that minimization of a structure with charged surface groups in gas-phase QM calculations can lead to spurious proton transfers and bond breaking.⁵² Since the boundary between the QM and MM regions bisects some chemical bonds, we introduced hydrogen link atoms to cap the open valences of the QM region. It should be noted that there are several versions of the link atom approach which differ in their handling of the QM/MM boundary by either making link atoms interact with MM atoms (the HQ scheme) or making them unaware of the MM atoms (the QQ scheme).^{53,54} Because the standard AMBER 8 release only supports the QQ scheme, it was modified to implement HQ as well. Through extensive computational experiments, we found that the HQ scheme yielded much better geometries at the linkage region for our system. Thus in all the calculations presented here, the HQ link atom scheme was employed to treat the QM/MM boundary.

Upon completion of these refinements, the unrefined protein and inhibitor chains and the crystallographic water surrounding them were stripped off using the SORTWATER utility in the Collaborative Computational Project, Number 4 (CCP4) software suite.⁵⁵ We then performed single-point divide-and-conquer (DivCon) self-consistent reaction field (SCRf) calculations on the remaining structures with 6926–6928 atoms, where the reaction field generated by the bulk solvent was accounted for by solving the Poisson-Boltzmann (PB) equation using our own finite-difference PB solver.⁵⁶ Our analysis of the relative stability of the various protonation states was adapted from the theory initially pioneered by Warshel and co-workers⁴² and extended by Bashford and Karplus⁴³ among others.^{44,45} Figure 5 shows the thermodynamic cycle used to analyze the relative stability among the various protonation states, where capped aspartic acid in solution is used as the reference point. In this scheme, the free energy difference between the AspH and Asp⁻ states of either aspartate in the protein environment is computed

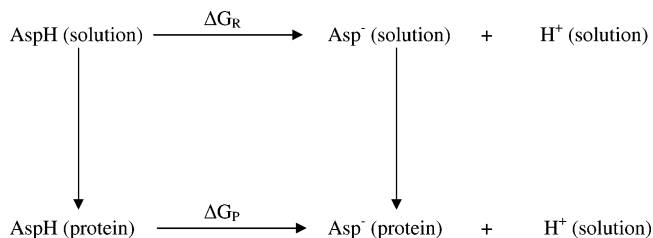


Figure 5. Schematic representation of the thermodynamic cycle used to evaluate the relative stability between protonated and ionized Asp in the protein environment. The reference compound AspH(solution)/Asp⁻(solution) may be substituted with other ionizable groups.

by the following expression

$$\Delta\Delta G = \Delta G_P - \Delta G_S \approx [E(\text{Asp}^-; \text{prot}) - E(\text{AspH}; \text{prot})] - [E(\text{Asp}^-; \text{soln}) - E(\text{AspH}; \text{soln})] \quad (2)$$

where we approximate the differences in free energy between protonation states as differences in potential energy as in previous theoretical work.⁴³ Likewise, this approach is equivalent to the homodesmotic reaction formalism employed by Rajamani et al.,²⁰ with the main difference being their use of propionic acid as the reference compound.

All the calculations were performed on our local Linux PC clusters. Each preprocessing run took 3–5 h, and the wall clock time for each QM/MM refinement was 7–9 h. The DivCon QM/SCRf calculations were carried out with the default dielectric constants of 1 and 80 for the interior and exterior of the protein. The grid box used in our finite difference PB solver was set up in such a way that the solute spanned at most 60% of the box length in each dimension with a grid resolution of 3 points per Å. Since these calculations allow the solute to be polarized by the solvent through perturbations on the Hamiltonian, they are computationally much more demanding than classical Poisson-Boltzmann solvers: on average, 17 h were required to reach SCF convergence for each single-point DivCon QM/SCRf calculation.

Results

Preprocessing of the Crystal Structure. Since the crystal structure was solved with CNS using an incomplete energy function with parameters determined by Engh and Huber, it is expected that the stereochemical details of the structures refined with our method will change slightly. Thus, we minimized the initial structure with an MM energy function at several weights (denoted $w_{X\text{-ray}}$ in eq 1) ranging from 0.01 to 10. It should be noted that a more efficient method exists which obtains a quick estimate of the ideal weight by matching the average energy gradient with the average X-ray gradient in a short molecular dynamics (MD) simulation. Nevertheless this was not undertaken because we are employing an energy function that is different from the one typically used in eq 1, and we wish to understand thoroughly how the weight influences the refinements. As a control, this step of preprocessing was carried out in 4 different protocols: in Protocol 1, the whole protein, the inhibitor, and the solvent molecules were refined; in Protocol 2, only the

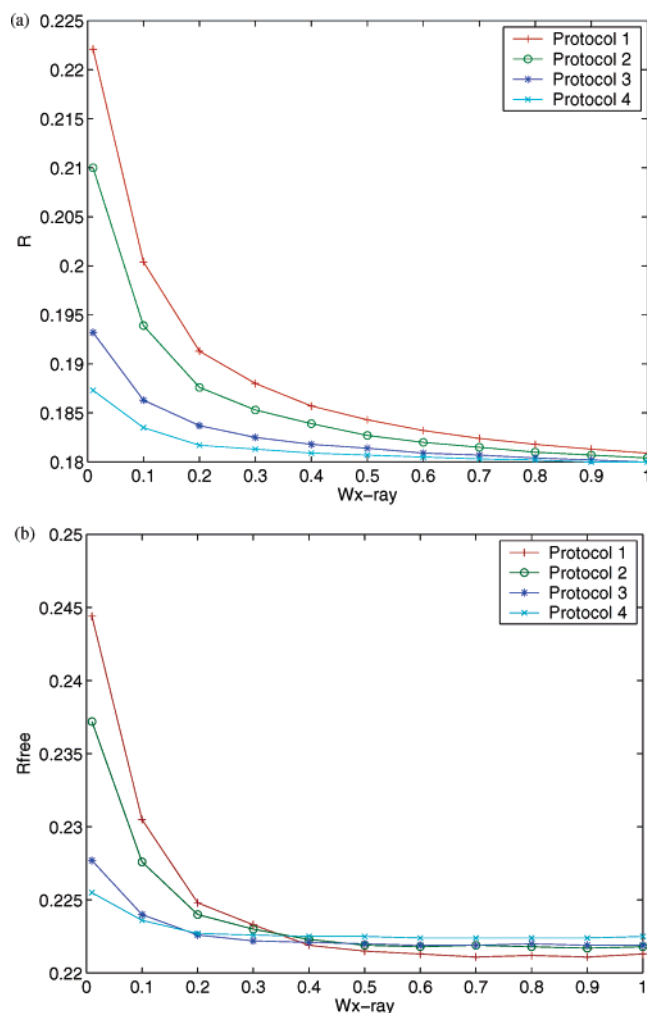


Figure 6. R (a) and R_{free} (b) values as functions of the weight of the X-ray target for 4 preprocessing protocols (see text for definitions).

protein and the inhibitor, but not the solvent, were refined; in Protocol 3 and 4, only the atoms that were within a distance of 20 and 15 Å, respectively, from the center of the active site were refined. In all these Protocols, the coordinates of the hydrogen atoms were allowed to relax.

The R values and R_{free} values of the refined structures are plotted as functions of the weight in Figure 6. Clearly, as $w_{X\text{-ray}}$ decreases the R values of the structures refined using all 4 Protocols increase to between 0.187 and 0.223 as shown in Figure 6(a). In addition, the order of the final R values at the lowest weight suggests that freezing the coordinates of the atoms at a distance away from the active site helps maintain the compatibility of the structure with the observed X-ray signals. Fixing the coordinates of the solvent molecules had similar effects, which was evident from the results of Protocol 2. The explanation for this observation is that when the weight of the X-ray restraint is reduced, the parts of protein structures that are mostly affected are the regions with less well-resolved electron densities, e.g. surface residues and discrete solvent molecules. In these regions, minimizing the structure on a potential energy function containing electrostatic interactions without modeling the solvent can cause significant errors in the structure. In fact,

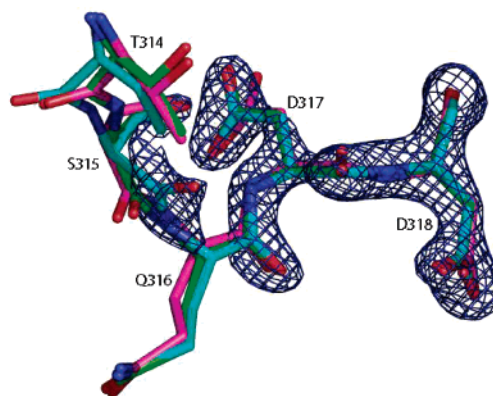


Figure 7. Snapshot of the loop containing residues from Thr314 to Asp318 after preprocessing with Protocol 1 at 3 weights, 10 (cyan), 0.4 (green), and 0.01 (magenta), together with the σ_A -weighted 2Fo-Fc electron density map contoured at 1.8σ level.

a visual inspection of the refined structures resulting from Protocol 1 at the lowest X-ray weight showed that many water molecules were pulled out of their densities and drawn toward charged groups. In examining the plot of the R_{free} values against the weight shown in Figure 6(b), more insight into the effects of the weight can be made. The corresponding R_{free} values at lower weights show a similar trend, but in the intermediate range between 0.4 and 0.9 we observe that the order of agreement is almost reversed from that for the R values at the lowest weight. Here Protocol 1 gives the best R_{free} values among the 4 Protocols. Since the R_{free} value is an unbiased indicator of the agreement with experimental data, the implication is that the complete energy function improves the structure in the all-atom refinement protocol only at intermediate weights. What is particularly encouraging is that all of the R_{free} values for the all-atom refinement protocol in this region are lower than the literature R_{free} value of 0.224, suggesting the potential usefulness of our method in further enhancing the accuracy of X-ray structure refinements. The crossover of the curves between the weights of 0.1 and 0.4 defines these weights as the borderline between the regions where the refinement is dominated by different determinants. Figure 7 displays the preprocessed structures of a fragment consisting of residues Thr314 through Asp318 with the electron density map, where the structures refined with Protocol 1 at 3 different weights, 10 (cyan), 0.4 (green), and 0.01 (magenta), are shown. This example demonstrates how the structure in a relatively flexible loop region is influenced by the weight. The cyan structure stays closest to the crystal structure due to the heavy X-ray constraint, whereas the magenta structure shows significant deviations. The green structure is somewhere in the middle: in regions where the electron density is strong, for example Asp318, it stays close to the crystal structure; in regions where the electron density is weak, for example Thr314, it is more similar to the structure dominated by the energy function.

At this point, we selected the structure refined with Protocol 1 at the weight of 0.4 as the template to construct the various protonation states. It was expected that this structure had undergone substantial adjustments in response

Table 1. Calculated Gas-Phase QM Energies E_{qm} , Free Energies of Solvation ΔG_{Solv} , and Free Energy Changes for the Reference Reaction ΔG_{R} for Capped Asp and Propionic Acid

reference compound	E_{qm} (kcal/mol)	ΔG_{Solv} (kcal/mol)	ΔG_{R} (kcal/mol)
$\text{AspH (solution)} \xrightarrow{\Delta G_{\text{R}}} \text{Asp}^- \text{ (solution)} + \text{H}^+ \text{ (solution)}$			
AspH	-163.4	-20.9	-84.8
Asp ⁻	-180.0	-89.1	
$\text{EtCOOH (solution)} \xrightarrow{\Delta G_{\text{R}}} \text{EtCOO}^- \text{ (solution)} + \text{H}^+ \text{ (solution)}$			
EtCOOH	-108.5	-5.7	-83.5
EtCOO ⁻	-122.0	-75.7	

to the new energy function and yet still fit the experimental signals well. It is also worth noting that the artifacts caused by treating the system in the gas phase may be reduced if the energy function contains a description of solvent in an implicit manner using, for example, the Generalized Born (GB) model, as Moulinier et al. demonstrated in their recent work.⁵⁷ Nonetheless, in our own refinement studies our observation was that introducing the GB/SA terms only resulted in significant differences at very low weights, providing a justification for the choice of the structure refined in the gas phase at $w_{\text{X-ray}} = 0.4$ for further processing.

Relative Stability. Using the thermodynamic cycle shown in Figure 5 we analyzed the relative stability of the 8 possible protonation states. First, we calculated the free energy change, ΔG_{R} , for the reference reaction using QM/SCRF calculations on the AM1-minimized structures of the reference compounds. Initially we considered both capped aspartic acid and propionic acid. From the results shown in Table 1, it can be seen that the difference between the two reference reactions was very small and within the margin of error of the AM1 level of theory.

The differences in the free energy changes, $\Delta\Delta G$, were then calculated and are collected in Table 2. As discussed above, these quantities represent the differences in free energy changes of ionizing Asp32 or Asp228 in the protein environment relative to the reference reactions in solution. In Table 2, we compare the results for 3 different refinement protocols: in Protocol I, we refined only the atoms within 10 Å away from the center of the active site; in Protocol II, we refined the coordinates of all-atoms. In both Protocols I

and II, the refinements were restrained by the X-ray data with a weight of 0.4, whereas in Protocol III all the atomic coordinates were minimized on the QM/MM potential without X-ray constraint. A glance at Table 2 shows that the quantum mechanical energies, E_{qm} , of the 8 protonation states separate into 3 groups: the monoprotonated states, the diprotonated states, and the dideprotonated states. The average gap in E_{qm} between the monoprotonated states and the diprotonated states is about 60–70 kcal/mol; the same gap between the dideprotonated states and the monoprotonated states is about 110–130 kcal/mol. These gaps indicate that the diprotonated states, with a net charge closest to neutral, are most stable in the gas phase. The gaps in E_{qm} are offset by the solvation free energy, ΔG_{Solv} , which gives the most heavily charged dideprotonated states the largest stabilization. Due to cancellations in E_{qm} and ΔG_{Solv} , the resulting $\Delta\Delta G$ between the least stable t228 and the most stable configuration 32i is less than 43 kcal/mol. The differences in the results between Protocols I and II suggest that allowing the all-atom coordinates to be refined produces lower gas-phase energies and larger gaps between the groups of states. The results from Protocol III show that in the absence of the X-ray constraint, the gas-phase energies of the diprotonated and the dideprotonated states, with the exception of 228i32o where the minimization was probably trapped in a local minimum, relax further by varying degrees. However, the lowering in the gas-phase energies are more than offset by the unfavorable free energies of solvation. Overall, these states have higher solution-phase energies, but the gaps between them are reduced.

Final Structures. Since protons are excluded from calculation of structure factors, the initial structures of the 8 protonation configurations have an identical R value of 0.186 and an identical R_{free} value of 0.222. The QM/MM refinements yielded final structures with R and R_{free} values very close to the starting ones, with variations in the R and R_{free} values in the statistically insignificant fourth decimal place. Figure 8 shows a cross-eye stereo picture of the refined active site and the corresponding electron density map in the most favored 32i state, where the inner oxygen atom of Asp32 donates a hydrogen bond to the hydroxyl oxygen of the inhibitor. The good agreement between the observed and calculated structure factors, as the R and R_{free} values suggest, is reflected by the excellent fit of the structure to the electron density map.

Table 2. Calculated Gas-Phase QM Energies E_{qm} , Free Energies of Solvation ΔG_{Solv} , and Differences in Free Energy Change $\Delta\Delta G$ for Ionizing the Asp's in the 8 Protonation Configurations Using Capped Asp as the Reference Compound^a

state	Protocol I			Protocol II			Protocol III		
	E_{qm}	ΔG_{Solv}	$\Delta\Delta G$	E_{qm}	ΔG_{Solv}	$\Delta\Delta G$	E_{qm}	ΔG_{Solv}	$\Delta\Delta G$
32i	-34763.3	-3593.4	0.0	-34783.2	-3587.5	0.0	-34786.7	-3237.3	6.2
228i	-34757.0	-3597.3	2.4	-34766.8	-3595.5	8.3	-34787.3	-3242.9	0.0
32o	-34761.0	-3592.8	2.9	-34768.9	-3594.3	7.4	-34780.7	-3240.0	9.5
228o	-34753.5	-3597.4	5.8	-34768.4	-3589.1	13.1	-34783.0	-3243.2	4.0
228i32o	-34825.6	-3440.2	6.1	-34845.8	-3426.9	13.1	-34836.9	-3094.7	13.8
228o32i	-34821.2	-3439.8	10.9	-34830.1	-3435.0	20.7	-34839.6	-3101.7	4.2
t32	-34640.3	-3766.7	34.5	-34658.1	-3756.7	40.6	-34679.8	-3402.3	32.8
t228	-34637.3	-3766.5	37.7	-34657.2	-3755.7	42.6	-34676.1	-3401.6	37.3

^a All the $\Delta\Delta G$ values are relative to the minimum among the 8 configurations. All the values are in the unit of kcal/mol.

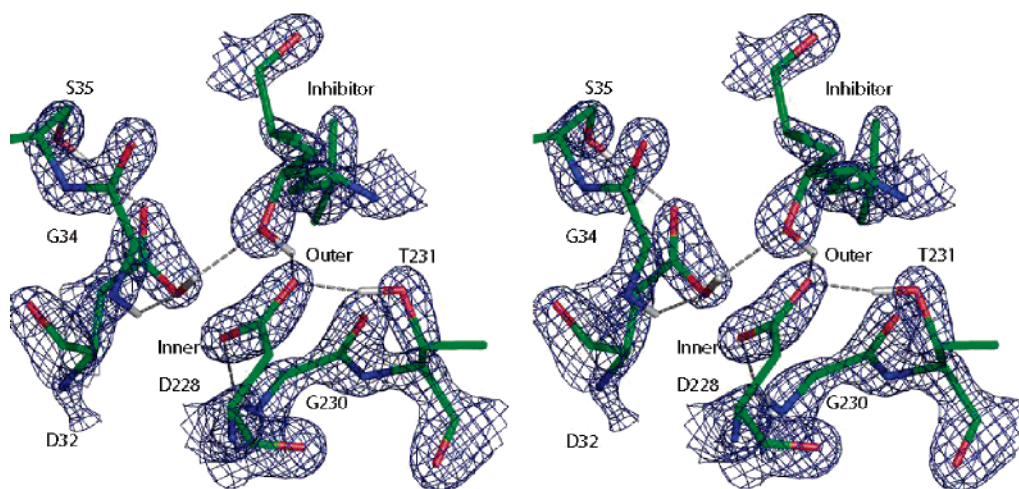


Figure 8. Cross-eye stereoview of the key residues of BACE at the end of the 32i refinement, together with the σ_A -weighted 2Fo-Fc electron density maps contoured at the 2.7σ level.

It can also be seen that the structure of the active site is somewhat symmetric about the hydroxyl of the inhibitor as noted in previous studies:¹⁰ both of the inner oxygen atoms of Asp32 and Asp228 are stabilized by main-chain N–H hydrogen bonds from Gly34 and Gly230, while the outer oxygen atoms accept hydrogen bonds from Ser35 and Thr231. The γ oxygen atoms of Ser35 and Thr231 both accept a hydrogen bond from water molecules Wat2 and Wat101, respectively. Given this symmetry, it is surprising to see a reversal of the order of $\Delta\Delta G$ s for the monoprotonated states in Table 2: when refined with Protocol II, the 32i and 32o states are more stable than their “images”, 228i and 228o, whereas when refined with Protocol III they are destabilized relative to their images by 14.5 and 11.2 kcal/mol, respectively. To rationalize this result, we superimpose the 32i and 228i structures refined with Protocols II and III in Figure 9 together with the electron density map. Both structures minimized with the QM/MM energy function in the absence of the X-ray constraint show significant deviations from the electron density, especially for Wat2 and Wat101. Upon closer examination, it appears that Asp32 refined using Protocol III shifts to the same position relative to Protocol II in both the 32i and 228i states, while Asp228 does not. In fact, the Asp228 structure for Protocol III stays close to the one for Protocol II in the 32i state, whereas it moves away considerably when protonated. This observation is reinforced by the key hydrogen bond lengths shown in Table 3, where the distances are measured between the heavy atoms of donors and acceptors in order to facilitate comparison with the original 1FKN structure. It appears from Table 3 that the distance between the inner oxygen of Asp32 and main-chain nitrogen of Gly34 in the 228i state for Protocol III is much shorter than that in the 1FKN structure, while the distance between the inner oxygen of Asp228 and the main-chain nitrogen of Gly230 is much longer, and these deviations are larger than those found in the 32i state. Altogether, these results suggest that the 32i state is the most consistent with the X-ray diffraction data, while the structural and energy differences between Protocol II and III could be due to a number of factors including crystal packing forces,

the errors introduced by the QM/MM method or the partitioning scheme, and the inherent errors in the semiempirical electronic structure method.

Discussion

Comparisons with Previous Theoretical Work. It is interesting to compare the results of this work to previous theoretical work. Here we will focus on the comparison between the relative energies for the 8 protonation configurations in this work and those calculated by Rajamani et al.²⁰ The relative gas-phase QM energies, solution-phase QM energies, and differences in free energy changes computed by Rajamani et al. are reproduced in Table 4. Qualitatively both their study and this work find the monoprotonated states to be most stable, followed by the diprotonated states. Nevertheless, a few important differences can be observed: first, the ordering of relative solution-phase energies within each group of states (i.e., monoprotonated, diprotonated, and dideprotonated) is different especially among the monoprotonated configurations, for which Rajamani et al. consistently favor the states in which Asp228 is protonated; second, their most stable diprotonated state, 228i32o, is only 3.5 kcal/mol in energy above their most stable monoprotonated state, 32i. Noting the similarity between the $\Delta\Delta G$ s for Protocol III in Table 2 and those in column 3 of Table 4, the second observation may be explained by the fact that Rajamani et al. did not restrain the structure with X-ray data and thus the structures of the diprotonated and dideprotonated states relaxed, resulting in the observed smaller energy gaps. However, the first observation is more subtle and warrants further discussion.

In ref 20 the relative stability of the 8 protonation states were evaluated at the same level of theory as this work, and thus the energy function is unlikely to be responsible for this discrepancy. Nonetheless, to make full QM structure optimization of a very large system feasible in ref 20, the crystal structure had to be truncated. Specifically, the crystallographic waters were selectively retained and the atoms outside a 15 Å cutoff of any atom of the ligand were removed, resulting in a simplified system of about 1477

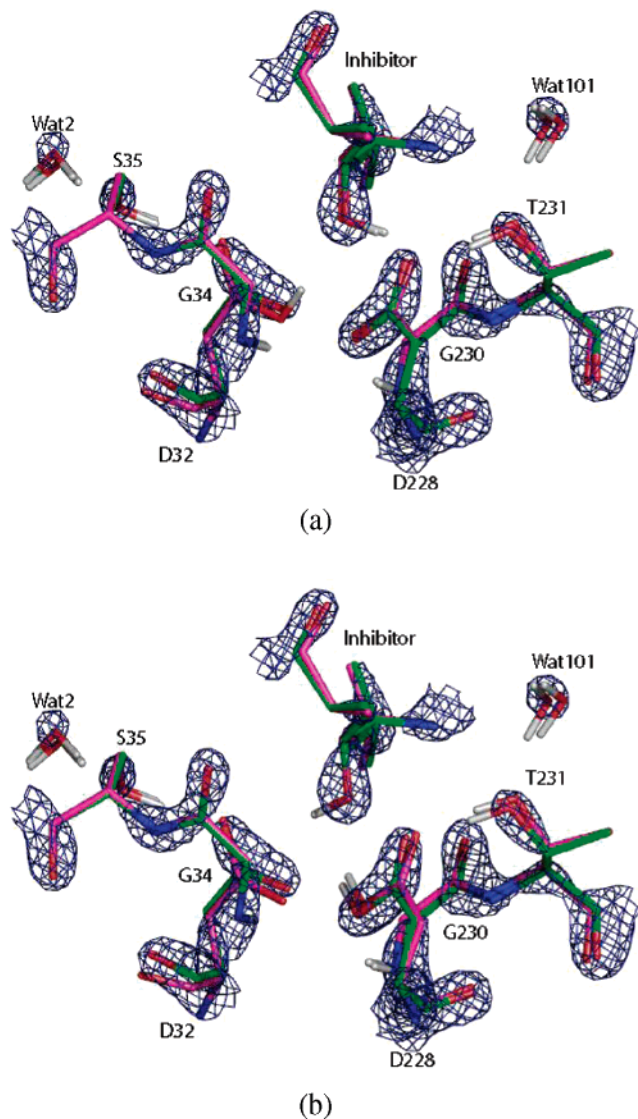


Figure 9. Key residues in the 32i (a) and 228i (b) states at the end of the refinements by Protocol II (green) and Protocol III (magenta), together with the σ_A -weighted 2Fo-Fc electron density map contoured at the 3.5σ level. The waters that deviate more from the densities were refined with Protocol III.

atoms including 4 waters.²⁰ The extensive truncation of the system could have introduced some asymmetry in the protein structure even though the net charge is close to 0, which would artificially favor one arrangement over the other. In addition, the omission of some of the water molecules that are part of the active-site electrostatic network¹⁰ is another major source of difference. Figure 10 shows the active site structure of the 32i configuration computed by Rajamani et al. Comparing Figure 10 to Figures 8 and 9, it appears that the structure of the simplified model fits the density fairly well, if not as well as our refined structure. Despite this, in the binding pocket there are electron densities for a few discrete water molecules, which are in the crystal structure Wat2 and Wat42 on the side of Asp32 and Wat61, Wat101, Wat162, and Wat204 on the side of Asp228. Most of these waters are absent in ref 20 except for Wat2. It is likely that the omission of these solvent molecules might have tilted

Table 3. Key Hydrogen Bond Lengths in the Structures of the 32i and 228i States Refined by Protocol II and Protocol III and in the 1FKN Structure

hydrogen bond		Protocol II		Protocol III		1FKN structure
acceptor	donor	32i	228i	32i	228i	
O _{outer} (Asp32)	O γ (Ser35)	2.66	2.65	2.63	2.61	2.64
O _{outer} (Asp32)	O(Inhibitor)	3.12	3.17	3.00	3.06	3.26
O γ (Ser35)	O(Wat2)	2.70	2.69	2.74	2.75	2.67
O _{inner} (Asp32)	N(Gly34)	3.39	3.35	3.02	2.82	3.56
O _{inner} (Asp32)	O(Inhibitor)	2.65	2.66	2.77	2.89	2.51
O _{outer} (Asp228)	O γ (Thr230)	2.60	2.63	2.61	2.67	2.53
O _{outer} (Asp228)	O(Inhibitor)	2.64	2.58	2.84	2.97	2.54
O γ (Thr230)	O(Wat101)	2.76	2.73	2.76	2.77	2.80
O _{inner} (Asp228)	N(Gly230)	2.81	2.87	2.84	2.99	2.66
O _{inner} (Asp228)	O(Inhibitor)	3.15	3.10	3.20	3.05	3.10
O _{inner} (Asp228)	O _{inner} (Asp32)	2.75	2.75	2.49	2.58	2.89

Table 4. Calculated Relative Gas-Phase QM Energies E_{qm} , Solution-Phase QM Energies, and Differences in Free Energy Changes $\Delta\Delta G$ for Ionizing the Asp's in the 8 Protonation Configurations Using Propionic Acid as the Reference Compound Reproduced from Ref 20

state	E_{qm}	E_{aq}	$\Delta\Delta G$
32i	23.4	30.5	30.5
228i	0	0	0
32o	31.7	40.7	40.7
228o	16.7	13.9	13.9
228i32o	0	0	3.5
228o32i	18.5	16.9	20.4
t32	7.5	0.5	31.2
t228	0	0	30.7

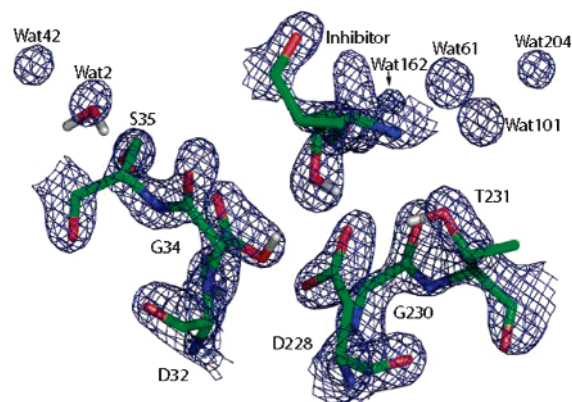


Figure 10. Key residues in the 32i state in the simplified model constructed by Rajamani et al.,²⁰ together with the σ_A -weighted 2Fo-Fc electron density map contoured at 2.7σ level.

the balance between the symmetric configurations because Asp32 accepts a hydrogen bond from Ser35 which in turn accepts a hydrogen bond from Wat2, while on the other side of the active site Wat101, which hydrogen bonds with Thr231, is missing. We suggest that these structural differences provide an explanation to the observed discrepancy, although it is still unclear how much of the gap they would account for. Comparing the 1FKN structure to other crystal structures of BACE such as 1M4H⁷ and 1SGZ,⁸ it appears

Wat2 is a conserved site, whereas Wat101 is not. However, since our results are based on structures that are compatible with experimental data, they are thought to be on a firmer ground than those by Rajamani et al.

Relevance to Structure-Based Design of BACE Inhibitors. The availability of the apo and ligand bound BACE crystal structures^{6,7,9} has provided the basis for structure-based drug design efforts worldwide.^{58–60} Recently, Polgar and Keseru carried out virtual screening of BACE inhibitors and scored their binding affinities.²¹ Their study explored the effects of two factors on the results of docking: (1) by imposing pharmacophore constraints on the docked poses and (2) by taking into account the protonation states of the key aspartates in the docking procedures. The pharmacophore constraints applied by Polgar and Keseru were derived by Miyamoto et al.⁶¹ Two possibilities for the protonation state were considered: the default configuration where both aspartates were ionized and a hypothesized one where Asp32 was protonated and Asp228 ionized. This hypothesis was based on the results of pK_a calculations on the 1FKN structure.²¹ In these calculations all the solvent molecules were removed, and yet the most likely protonation configuration found by Polgar et al. is the same as the one favored by us. Polgar et al. showed that when the hypothesized protonation state was used in the docking experiments, the improvement in the enrichment factor over the results of the default protonation state was very similar to when the pharmacophore constraints were applied. Based on this observation, Polgar et al. proposed that the pharmacophore constraints implicitly encoded information about the protonation state of the receptor. The significance of our results is they suggest the monoprotonated 32i configuration as the appropriate protonation state in docking calculations and de novo inhibitor design.

Relevance to the Reaction Mechanism. The relevance of these results to the understanding of the catalytic mechanism of BACE needs to be assessed with the consideration of the extent to which the crystal structure of the protein–ligand complex resembles the transition-state structure in the reaction. It may be argued that since the hydroethylene inhibitor in the current study does not completely mimic the tetrahedral intermediate in the proposed mechanism, the protonation state of the aspartates may not be the same as that in the reaction. Nevertheless, it is our expectation that if crystal structures of BACE bound to inhibitors containing diol fragments or those of isolated reaction intermediates become available, structure refinements and energetics analysis using our method will be able to provide further insights.

Conclusions

We have presented a novel method, QM/MM X-ray structure refinement, and have applied it to refine atomic-level structures of OM99-2 bound β -secretase including the coordinates for protons. In essence, this method amounts to applying more accurate energy restraints to key regions while retaining the computational efficiency of conventional refinements. The QM treatment is expected to be more accurate than MM for the active site region because it allows an

accurate account of various interactions in a very unusual environment. The divide-and-conquer SCRF calculations provide a means to perform accurate energy evaluations of the structures consisting of BACE, its inhibitor, and ordered water molecules with a continuum representation of the bulk solvent environment. The results of these calculations suggest that the monoprotonated 32i configuration should be favored under the conditions in which the crystallographic experiment was conducted. Indeed, a comprehensive study of all the crystal structures of BACE along the lines of this work would be an interesting future direction and will prove beneficial to a complete understanding of the mechanistic details of the catalytic reaction and to provide better guidance to the structure-based design of BACE inhibitors.

It should be emphasized that QM/MM X-ray refinement is not expected to lower crystallographic R values appreciably, since minute changes in local structures do not affect on a larger scale the fit of structures to observed electron densities. Instead, it is an analysis tool that is useful for generating realistic all-atom models from crystal structures.

The advent of ultrahigh-resolution protein X-ray structures has created a unique challenge as well as opportunity for this method.^{62,63} The observation-to-parameter ratios at these resolutions come close to the typical values for small molecule crystals, and the errors in atomic coordinates are also significantly suppressed. Ultrahigh-resolution X-ray crystallography will serve as a tool with tremendous power in defining protonation states and calibrating the QM/MM X-ray refinement method. That said, QM calculations will still be valuable to atomic-resolution protein crystallography, as even at resolutions higher than 0.85 Å the scattering from hydrogen atoms is still too weak to allow complete determination of the coordinates of all the protons. In a recent paper, Schiffer et al. reviewed the newest advances in simulation techniques that would impact the field of protein crystallography, and the utilization of QM methods was recognized as one of the 3 major forefronts.⁶⁴ Future development of QM-based refinement methodologies that allow first-principles calculations to complement experiments, while challenging, will likely be very rewarding.

Acknowledgment. N. Yu thanks Dr. Martha S. Head at GlaxoSmithKline for critical reading of the manuscript and for helpful suggestions. This research was generously supported by the NSF (MCB-0211639) and the NIH (GM44974).

References

- (1) Selkoe, D. J. *Physiol. Rev.* **2001**, *81*, 741.
- (2) Wolfe, M. S.; Xia, W. M.; Ostaszewski, B. L.; Diehl, T. S.; Kimberly, W. T.; Selkoe, D. J. *Nature* **1999**, *398*, 513.
- (3) Lin, X. L.; Koelsch, C.; Wu, S. L.; Downs, D.; Dashti, A.; Tang, J. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1456.

- (4) Yan, R. Q.; Bienkowski, M. J.; Shuck, M. E.; Miao, H. Y.; Tory, M. C.; Pauley, A. M.; Brashler, J. R.; Stratman, N. C.; Mathews, W. R.; Buhl, A. E.; Carter, D. B.; Tomasselli, A. G.; Parodi, L. A.; Heinrikson, R. L.; Gurney, M. E. *Nature* **1999**, *402*, 533.
- (5) Vassar, R.; Bennett, B. D.; Babu-Khan, S.; Kahn, S.; Mendiaz, E. A.; Denis, P.; Teplow, D. B.; Ross, S.; Amarante, P.; Loeloff, R.; Luo, Y.; Fisher, S.; Fuller, L.; Edenson, S.; Lile, J.; Jarosinski, M. A.; Biere, A. L.; Curran, E.; Burgess, T.; Louis, J. C.; Collins, F.; Treanor, J.; Rogers, G.; Citron, M. *Science* **1999**, *286*, 735.
- (6) Hong, L.; Koelsch, G.; Lin, X. L.; Wu, S. L.; Terzyan, S.; Ghosh, A. K.; Zhang, X. C.; Tang, J. *Science* **2000**, *290*, 150.
- (7) Hong, L.; Turner, R. T.; Koelsch, G.; Shin, D. G.; Ghosh, A. K.; Tang, J. *Biochemistry* **2002**, *41*, 10963.
- (8) Hong, L.; Tang, J. *Biochemistry* **2004**, *43*, 4689.
- (9) Patel, S.; Vuillard, L.; Cleasby, A.; Murray, C. W.; Yon, J. *J. Mol. Biol.* **2004**, *343*, 407.
- (10) Andreeva, N. S.; Rumsh, L. D. *Protein Sci.* **2001**, *10*, 2439.
- (11) Dunn, B. M. *Chem. Rev.* **2002**, *102*, 4431.
- (12) Suguna, K.; Padlan, E. A.; Smith, K. W.; Carlson, W. D.; Davies, D. R. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7009.
- (13) Davies, D. R. *Annu. Rev. Biophys. Chem.* **1990**, *19*, 189.
- (14) James, M. N. G.; Sielecki, A. R.; Hayakawa, K.; Gelb, M. H. *Biochemistry* **1992**, *31*, 3872.
- (15) Parris, K. D.; Hoover, D. J.; Damon, D. B.; Davies, D. R. *Biochemistry* **1992**, *31*, 8125.
- (16) Veerapandian, B.; Cooper, J. B.; Sali, A.; Blundell, T. L.; Rosati, R. L.; Dominy, B. W.; Damon, D. B.; Hoover, D. J. *Protein Sci.* **1992**, *1*, 322.
- (17) Coates, L.; Erskine, P. T.; Wood, S. P.; Myles, D. A. A.; Cooper, J. B. *Biochemistry* **2001**, *40*, 13149.
- (18) Touloukhonova, L.; Metzler, W. J.; Witmer, M. R.; Copeland, R. A.; Marcinkeviciene, J. *J. Biol. Chem.* **2003**, *278*, 4582.
- (19) Park, H.; Lee, S. *J. Am. Chem. Soc.* **2003**, *125*, 16416.
- (20) Rajamani, R.; Reynolds, C. H. *J. Med. Chem.* **2004**, *47*, 5159.
- (21) Polgar, T.; Keseru, G. M. *J. Med. Chem.* **2005**, *48*, 3749.
- (22) Yamazaki, T.; Nicholson, L.; Torchia, D.; Wingfield, P.; Stahl, S.; Kaufman, J.; Eyermann, C.; Hodge, C.; Lam, P.; Ru, Y.; Jadhav, P.; Chang, C.; Weber, P. *J. Am. Chem. Soc.* **1994**, *116*, 10791.
- (23) Hyland, L.; Tomaszek, T.; Roberts, G.; Carr, S.; Magaard, V.; Bryan, H.; Fakhoury, S.; Moore, M.; Mimmich, M.; Culp, J.; Desjarlais, R.; Meek, T. *Biochemistry* **1991**, *30*, 8441.
- (24) Hyland, L.; Tomaszek, T.; Meek, T. *Biochemistry* **1991**, *30*, 8454.
- (25) Piana, S.; Carloni, P. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 26.
- (26) Piana, S.; Sebastiani, D.; Carloni, P.; Parrinello, M. *J. Am. Chem. Soc.* **2001**, *123*, 8730.
- (27) Piana, S.; Bucher, D.; Carloni, P.; Rothlisberger, U. *J. Phys. Chem. B* **2004**, *108*, 11139.
- (28) Jack, A.; Levitt, M. *Acta Crystallogr. A* **1978**, *34*, 931.
- (29) Hendrickson, W. A. *Methods Enzymol.* **1985**, *115*, 252.
- (30) Tronrud, D. E.; Teneyck, L. F.; Matthews, B. W. *Acta Crystallogr. A* **1987**, *43*, 489.
- (31) Brunger, A. T.; Adams, P. D. *Acc. Chem. Res.* **2002**, *35*, 404.
- (32) Brunger, A. T.; Krukowski, A.; Erickson, J. W. *Acta Crystallogr. A* **1990**, *46*, 585.
- (33) Engh, R. A.; Huber, R. *Acta Crystallogr. A* **1991**, *47*, 392.
- (34) Yu, N.; Yennawar, H. P.; Merz, K. M. *Acta Crystallogr. D* **2005**, *61*, 322.
- (35) Nilsson, K.; Hersleth, H. P.; Rod, T. H.; Andersson, K. K.; Ryde, U. *Biophys. J.* **2004**, *87*, 3437.
- (36) Nilsson, K.; Lecerof, D.; Sigfridsson, E.; Ryde, U. *Acta Crystallogr. D* **2003**, *59*, 274.
- (37) Nilsson, K.; Ryde, U. *J. Inorg. Biochem.* **2004**, *98*, 1539.
- (38) Ryde, U.; Nilsson, K. *J. Am. Chem. Soc.* **2003**, *125*, 14232.
- (39) Ryde, U.; Nilsson, K. *J. Inorg. Biochem.* **2003**, *96*, 39.
- (40) Ryde, U.; Nilsson, K. *J. Mol. Struct. (THEOCHEM)* **2003**, *632*, 259.
- (41) Ryde, U.; Olsen, L.; Nilsson, K. *J. Comput. Chem.* **2002**, *23*, 1058.
- (42) Warshel, A.; Sussman, F. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 3806.
- (43) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219.
- (44) Warwicker, J. *Protein Sci.* **1999**, *8*, 418.
- (45) Antosiewicz, J.; Miller, M. D.; Krause, K. L.; McCammon, J. A. *Biopolymers* **1997**, *41*, 443.
- (46) Case, D. A.; Darden, T. A.; T. E.; Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER*; 8 ed.; University of California: San Francisco, CA, 2004.
- (47) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. *Acta Crystallogr. D* **1998**, *54*, 905.
- (48) Lee, T.; York, D.; Yang, W. *J. Chem. Phys.* **1996**, *105*, 2744.
- (49) Dixon, S. L.; Merz, K. M. *J. Chem. Phys.* **1996**, *104*, 6643.
- (50) Dixon, S. L.; Merz, K. M. *J. Chem. Phys.* **1997**, *107*, 879.
- (51) van der Vaart, A.; Suarez, D.; Merz, K. M. *J. Chem. Phys.* **2000**, *113*, 10512.
- (52) Wallocot, A. W.; Merz, K. M. Manuscript in preparation.
- (53) Eurenium, K.; Chatfield, D.; Brooks, B.; Hodosek, M. *Int. J. Quantum Chem.* **1996**, *60*, 1189.
- (54) Reuter, N.; Dejaegere, A.; Maigret, B.; Karplus, M. *J. Phys. Chem. A* **2000**, *104*, 1720.
- (55) Collaborative Computational Project, Number 4 *Acta Crystallogr.* **1994**, *D50*, 760.
- (56) Gogonea, V.; Merz, K. M. *J. Phys. Chem.* **1999**, *103*, 5171.
- (57) Moulinier, L.; Case, D. A.; Simonson, T. *Acta Crystallogr. D* **2003**, *59*, 2094.
- (58) Tounge, B.; Reynolds, C. *J. Med. Chem.* **2003**, *46*, 2074.
- (59) Rajamani, R.; Reynolds, C. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4843.

- (60) Coburn, C.; Stachel, S.; Li, Y.; Rush, D.; Steele, T.; Chen-Dodson, E.; Holloway, M.; Xu, M.; Huang, Q.; Lai, M.; DiMuzio, J.; Crouthamel, M.; Shi, X.; Sardana, V.; Chen, Z.; Munshi, S.; Kuo, L.; Makara, G.; Annis, D.; Tadikonda, P.; Nash, H.; Vacca, J.; Wang, T. *J. Med. Chem.* **2004**, *47*, 6117.
- (61) Miyamoto, M.; Matsui, J.; Fukumoto, H.; Tarui, N. In Patent WO 01/187293, 2001.
- (62) Jelsch, C.; Teeter, M. M.; Lamzin, V.; Pichon-Pesme, V.; Blessing, R. H.; Lecomte, C. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 3171.
- (63) Ko, T.-P.; Robinson, H.; Gao, Y.-G.; Cheng, C.-H. C.; Devries, A. L.; Wang, A. H.-J. *Biophys. J.* **2003**, *84*, 1228.
- (64) Schiffer, C.; Hermans, J. *Methods Enzymol.* **2003**, *374*, 412.
CT0600060

JCTC

Journal of Chemical Theory and Computation

Development of a Parametrized Force Field To Reproduce Semiempirical Geometries

Andrew M. Wollacott[†] and Kenneth M. Merz, Jr.*[‡]

*Department of Chemistry, The Pennsylvania State University,
University Park, Pennsylvania 16802*

Received January 11, 2006

Abstract: Here we describe the development of a classical force field parameter set to reproduce the geometry of proteins minimized at the semiempirical quantum mechanical level. The overall goal of the development of this new force field is to provide an inexpensive, yet reliable, method to arrive at geometries that are more consistent with a semiempirical treatment of protein structures. Since the minimization of a large number of protein structures at the semiempirical level can become cost-prohibitive, a “preminimization” with an appropriately parametrized classical treatment could potentially lead to more computationally efficient methods for studying protein structures through semiempirical means. Here we demonstrate that this force field allows for more rapid and stable geometry optimizations at the semiempirical level and can aid in the adoption of quantum mechanical calculations for large biological systems.

Introduction

Although *ab initio* and Density Functional (DFT) methods have proven reliable in the modeling of chemical systems,^{1,2} they cannot be routinely applied to larger biological systems as they scale poorly with the size of the system. Through various approximations in the quantum mechanical (QM) formulation more computationally feasible methods have been developed. One such approach is the semiempirical QM treatment of chemical systems.^{3–5} These methods were developed in the late 1970s and 1980s to model smaller chemical systems at the quantum mechanical level. To account for the various approximations in the semiempirical QM treatment, these methods have been highly parametrized to reproduce experimental data. In fact, these methods have proven to model chemical systems reliably at accuracies rivaling higher-level treatments. Additionally, with the recent development of linear-scaling semiempirical methods it is

now feasible to apply these methods to the study of protein systems.^{6–14}

Semiempirical QM methods have proven valuable in the study of protein structures, and their usefulness has been demonstrated for related applications such as protein–ligand interactions.¹⁵ While more costly than molecular mechanics-based (MM) methods, semiempirical calculations have been shown to outperform their classical counterparts in the discrimination of native structures from misfolded models (Wollacott and Merz, unpublished results). Semiempirical methods are also more sensitive to changes in protein geometry (Wollacott et al., unpublished results). The choice of starting structure for any modeling exercise is extremely important, and this requirement is exaggerated in the case of QM methods because they are more dependent on the internal geometry. To improve molecular structures before analysis, it is typical to first optimize the structure at the level of theory for which the model is being investigated. To speed up convergence, optimizations can be carried out in multiple steps, starting at lower levels of theory followed by an optimization at the desired level of theory so as to bring the model closer to the local energy minimum.

* Corresponding author e-mail: merz@qtp.ufl.edu.

[†] Current address: Department of Biochemistry, University of Washington, Seattle, WA.

[‡] Current address: Department of Chemistry, Quantum Theory Project, University of Florida, 2328 New Physics Building, P.O. Box 118435, Gainesville, Florida 32611-8435.

This is not usually necessary for classical methods such as employed in AMBER¹⁶ but is frequently used with QM methods.

While a divide-and-conquer approach to semiempirical treatments can be used to minimize the energies of small proteins very rapidly relative to other QM methods,^{6–9,14} it is currently infeasible to apply semiempirical minimizations to large-scale biological problems, such as ab initio protein folding simulations. For these systems, it would be desirable to first minimize structures with a fast classical potential before scoring or optimizing with semiempirical QM methods, thereby potentially reducing the computational expense. The assumption here is that the MM potential results in a structure more consistent with a semiempirical treatment. This procedure has been applied to the identification of native structures from sets of misfolded structures (Wollacott and Merz, unpublished results), with very promising results.

The feasibility of reparametrizing a classical force field, in this case AMBER, to better reproduce structures minimized at the semiempirical level has been investigated. The advantage of such a new parameter set would be 2-fold: (1) it would potentially speed up minimizations by utilizing an MM minimization to arrive at structures that are lower in energy with respect to a semiempirical treatment, and (2) it would reduce the overall strain on the system and potentially remove large instabilities during the minimization process that can lead to bond cleavage. For general protein minimization it is undesirable for amino acid groups to undergo bond rearrangement; in such applications the bonding configuration of residues should remain intact. Several aspects of the AMBER force field have been chosen for reparametrization, starting with the parm94 parameter set,¹⁷ to reproduce the geometries of proteins minimized at the PM3⁵ and AM1⁴ levels. These new parameter sets have been named parmPM3 and parmAM1 for the respective Hamiltonian that they were parametrized against.

Bond lengths, bond angles, atomic charges, and Lennard-Jones parameters from the AMBER parm94 force field were reparametrized. An analysis of the proper and improper torsions in proteins minimized at the semiempirical level indicates that these values may not be optimal for proteins (Wollacott et al., unpublished results). Out of plane bending (controlled by improper torsions) and unfavorable rotameric states of side chains were noted as unfavorable artifacts of semiempirical QM minimizations. Thus, the torsion parameters from AMBER were retained as semiempirical QM methods poorly model the dihedrals, whereas classical methods are better suited to treat these terms.

It should be stressed that the purpose of the parmAM1 and parmPM3 parameter sets is not to yield geometries that are in better agreement with experimentally determined structures. Rather, the parmAM1 and parmPM3 sets have been developed to reproduce protein structures minimized using semiempirical QM methods, regardless of whether these geometries are more or less nativelike. The resulting structures can then be more reliably used in large-scale semiempirical calculations on biological systems.

Table 1. Protein Systems Comprising the Training Set for the Parameterization of ParmAM1 and ParmPM3

PDB ID	description	resolution (Å)	N_{res}
1A80	HIV capsid C-terminal domain	1.70	70
1AIL	N-Ter fragment of Ns1 protein	1.90	70
1B0X	Epha4 receptor tyrosine kinase	2.00	72
1BCG	scorpion toxin Bixtr-Ir	2.10	74
1BMG	bovine beta-2 microglobulin	2.50	98
1CEI	colicin E7 immunity protein	1.80	85
1CQY	starch-binding domain of <i>Bacillus beta-amylase</i>	1.95	99
1CSP	major cold shock protein	2.50	67
1DSL	beta crystallin (C-ter)	1.55	88
1EM7	helix variant of B1 domain from strep protein G	2.00	56
1ENH	engrailed homeodomain	2.10	54
1F0M	ephrin type-B receptor	2.20	71
1FAS	fasciculin 1 (toxin)	1.80	61
1FNA	fibronectin cell-adhesion module	1.80	91
1H75	glutaredoxin-like protein Nrdh	1.70	76
1HPT	human pancreatic secretory trypsin inhibitor	2.30	56
1HYP	hydrophobic protein from soybean	1.80	75
1KW4	polyhomeotic sam domain structure	1.75	70
1LPL	hypothetical 25.4 Kda protein	1.77	95
1MJC	major cold shock protein	2.00	69
1MWP	amyloid A4 protein	1.80	96
1OPS	type III antifreeze protein	2.00	64
1ORC	Cro repressor insertion mutant	1.54	64
1PWT	alpha spectrin SH3	1.77	61
1WHO	allergen Phl P 2	1.90	94
1R69	phage 434 repressor (N-ter)	2.00	63
1SN1	neurotoxin Bmk M1	1.70	64
1UBI	ubiquitin	1.80	76
2CRO	434 Cro protein	2.35	65
2OVO	ovomucoid third domain	1.50	56

Methods

To arrive at a set of parametrized values for the bond lengths, angles, atomic charges, and van der Waals parameters, a set of small proteins from the protein databank was collected, ranging in size from 600 to 1500 atoms (Table 1). Due to the computational expense associated with optimization at the semiempirical level, the training set was limited to 30 small proteins. These proteins were selected to obtain a fairly representative set of topological features, including structures with secondary structural content composed of all α -helices, all β -sheets, a mix of helices and sheets, and random coils. All structures were solved using X-ray crystallography, ranging in resolution from 1.54 Å to 2.5 Å, and contained no cofactors or metal ions. The protein systems used for the training set are listed in Table 1. Hydrogen atoms were added to all proteins using the LEaP module of AMBER (AMBER 8.0). Hydrogen atoms were minimized using the Sander package from AMBER with the parm94 force field for 300 steepest descent steps, followed by 700 conjugate gradient steps. These protein systems were then minimized with either the PM3 or AM1 Hamiltonians using conjugate gradient as the minimization protocol. The resulting optimized structures were chosen as targets for the parametrization. In some cases hydrogen atom transfer reactions occurred between charged

Table 2. Amino Acid Frequencies in the Training Set of Proteins

amino acid	frequency	percentage (%)
ALA	130	6.18
CYS	56	2.66
ASP	122	5.80
GLU	137	6.51
PHE	73	3.47
GLY	160	7.60
HIS	31	1.47
ILE	110	5.23
LYS	152	7.22
LEU	154	7.32
MET	53	2.52
ASN	98	4.66
PRO	96	4.56
GLN	96	4.56
ARG	107	5.09
SER	125	5.94
THR	139	6.61
VAL	155	7.37
TRP	31	1.47
TYR	79	3.75

groups during optimization in vacuo. These were fixed by removing and then rebuilding the transferred hydrogen atoms using AMBER,¹⁶ followed by a short optimization of only the rebuilt hydrogen atom coordinates.

In order for the training set of protein systems to be used as targets for parametrization, there should be an adequate representation of each type of amino acid to obtain reliable statistics. Although only 30 proteins were used, the majority of amino acids were well represented, as shown by the frequency of residue types in the training set in Table 2. Cysteine residues were found as either single residues or as part of disulfide bridges, although the two forms were not distinguished between when parametrizing the bond length and angle values. In general, the frequencies of amino acids in the training set are similar to those found across the protein database.¹⁸

In developing the parmPM3 and parmAM1 parameter sets for AMBER, the atomtype designations used in parm94 were retained. Since parmPM3 and parmAM1 were only parametrized against proteins, only those atomtypes found in protein systems were included in the parametrization. While current versions of these parameter sets are applicable only for proteins, all other parameters were kept unaltered from their parm94 values, retaining the ability to model other biologically relevant molecules.

Reparametrizing Bond Lengths and Angles. The parameters for bond lengths and bond angles were taken as the average found in the training set of minimized structures for bonds between atoms of designated AMBER atomtypes. The force constant for each bond length and angle (K_{eq}) was taken from the AMBER parm94 parameter set, with only the equilibrium value of the lengths and angles (R_{eq}) modified to match the average from the target set. In general the difference in internal geometries between AM1 and PM3 minimized structures is small. The frequency of different

bond types found in the protein systems varied considerably, with a large number of peptide and aliphatic bonds being represented in the database of bond lengths, but only few bond types specific to underrepresented amino acid side chains such as tryptophan and cystine. However, a comparison of the underrepresented bond types to those found in a large set of pentapeptides (10 000 of sequence GGXGG where X is any amino acid) minimized at the semiempirical level revealed very small differences between the two geometries. Furthermore, the deviation in bond lengths across systems was limited. Thus, undersampling of bond lengths and angles does not seem to be a major problem for this data set. For the case of disulfide bonds, there were very few S–S bonds in the training set, so this bond type was not included in the parametrization. The values for bond lengths and angles for the parmPM3 and parmAM1 parameter sets are listed in the Supporting Information in Tables S1 and S2.

Parametrizing Atomic Charges. To parametrize the charges for each amino acid, the average CM2¹⁹ charge on each atom in the protein training set was determined from the semiempirical calculation. These charges were taken from vacuum calculations, although the differences in average atomic charges from in vacuo or solvation calculations is small (Wollacott et al., unpublished results). In comparison, the atomic charges used in the parm94 set were derived by fitting with a restrained ESP-fit (RESP) model.²⁰

Since a semiempirical treatment allows for charge transfer to occur, the average charge on each residue did not sum to an integral value. The charge for each atom in a residue was, therefore, normalized via the scaling of charges such that the residue possessed an integral charge. In addition, charges were averaged for chemically equivalent groups. For example, the three hydrogens in the methyl group of alanine (H^{β} 's) possessed slightly differing charges, so the charge assigned to each atom was taken as the average of the three. Charges for the cysteine residue were not reparametrized because the residue could be found in disulfide bridges, which affected the charge distribution. While the parm94 force field differentiates between cysteine (CYS) and cystine (CYX), the limited number of cysteines in the training set prevented their inclusion in the reparametrization effort. In this case, charges from the parm94 set were used for both forms of cysteine. The charges derived for the parmPM3 and parmAM1 parameter sets are listed in the Supporting Information in Table S3.

Reparametrization of Lennard-Jones Parameters. The shape of the Lennard-Jones potential can be specified by two variables, the R^* value and the ϵ value.¹⁷ The R^* parameters define the closest approach of two atoms, while the ϵ parameters describe the well depth. These parameters were derived for the parm94 parameter set by fitting to Monte Carlo liquid simulations to reproduce the densities and enthalpies of vaporization for hydrocarbons.²¹

For the development of the parmAM1 and parmPM3 parameters, the van der Waals parameters were adjusted such that minimization with AMBER using these parameters would best reproduce protein conformations generated by minimization at the AM1 or PM3 level. To accomplish this

would require a parametrization scheme whereby van der Waals parameters are initially modified from their parm94 values, followed by an AMBER minimization of the crystal structures with the new force field and an evaluation of the RMSD of these minimized models to the target. Parameters that minimized the RMSD of the AMBER minimized structures compared to the semiempirical-minimized structures would then be accepted. On current computer hardware (2.4 GHz AMD Opteron), an AMBER minimization of a small protein can take up to 30 minutes. With 30 proteins in the training set, and 20 van der Waals parameters that must be optimized, this parametrization scheme quickly becomes difficult even on modern hardware.

To parametrize van der Waals parameters for the parmAM1 and parmPM3 parameter sets, the R^* Lennard-Jones parameters were modified such that the AMBER energy and gradients were reduced for the semiempirical-minimized target proteins. This has the advantage that a full AMBER minimization does not have to be performed for each step of the parametrization. However, adjusting the R^* values so as to reduce the AMBER energy of these semiempirical-minimized proteins does not guarantee that a minimization with these parameters will reach the target geometries. With this limitation in mind, the R^* values for the 20 protein-relevant atomtypes were parametrized using a genetic algorithm²² to yield lower energies for structures that have been minimized at the semiempirical level. The ϵ parameters were not, however, modified from their parm94 values. The van der Waals parameters derived for the parmAM1 and parmPM3 parameter sets are listed in the Supporting Information in Table S4.

Results and Discussion

Minimizing with parmAM1 and parmPM3. Optimizing protein geometries with either parmAM1 or parmPM3 results in structures that are similar to those minimized with parm94. The average all-atom RMSD compared to the crystal structure after optimization with parm94 was 0.73 Å, with parmPM3 was 0.83 Å, and with PM3 was 0.95 Å. The all-atom RMSD between parm94 minimized structures and parmPM3 structures was smaller, on average 0.47 Å. Minimizing with parmAM1 lead to structures that had an average RMSD of 0.81 Å compared to the crystal structure. Structures minimized with parmAM1 were very similar in overall topology to those structures minimized with parmPM3 (RMSD 0.3 Å). In general, the parmAM1 and parmPM3 parameter sets are similar, so it is not surprising that when optimized under the same potential energy function they result in comparable structures.

Minimizing with parmAM1 or parmPM3 maintained many of the favorable traits of AMBER models, such as reducing large side-chain motions that form salt bridges. This effect can be most likely attributed to the inclusion of implicit solvation with AMBER minimizations, that were missing with semiempirical optimizations. When minimizing with MM force fields, planar groups retained their planarity, which was problematic when minimizing with semiempirical methods for nitrogen-containing groups such as

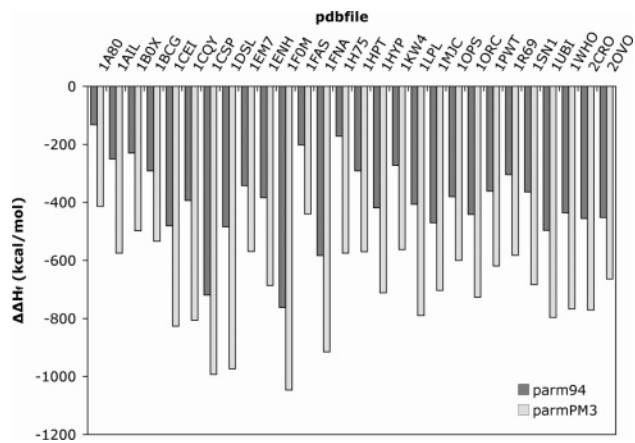


Figure 1. Improvements in the calculated heat of formation for structures minimized with parm94 and parmPM3 relative to their crystal structures for the training set.

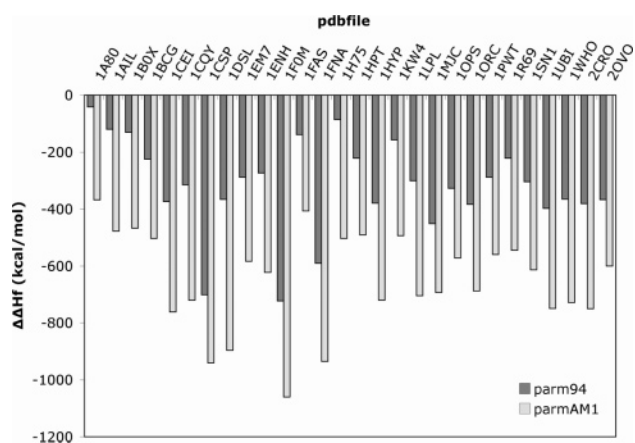


Figure 2. Improvements in the calculated heat of formation for structures minimized with parm94 and parmAM1 relative to their crystal structures for the training set.

the guanidyl group of arginine. The MM treatment of these groups explicitly enforces improper torsions to maintain planarity.

The improvement of preminimizing with MM force fields can be seen by the reduction in the heats of formation for proteins in the training set, as illustrated in Figures 1 and 2. Minimizing a structure with AMBER using the parm94 parameter set results in heats of formation that are lower by an average 397 kcal/mol compared to the crystal structure. This represents a significant improvement over the starting structure, resulting from improved internal geometries and the reduction of atomic clashes. Minimizing using parmPM3 improves the heat of formation by over 692 kcal/mol compared to the starting structure, while parmAM1 improves the heat of formation by approximately 649 kcal/mol. Clearly, a rapid minimization using the parmAM1 or parmPM3 parameter sets improves the structure with respect to semiempirical QM treatments of the protein.

The average force on each atom in a structure is another telling feature of its quality with respect to the potential used. The gradient on each atom is calculated as the first derivative of the energy potential with respect to the Cartesian coordinates. The gradients as evaluated by DivCon are performed numerically, and the average gradient (GAVRG)

Table 3. Summary of Improvements in the Average Heat of Formation ($\Delta\bar{H}_f$) and the Mean of the Average Gradient (GAVRG_{avrg}) for Structures Minimized with parm94, parmAM1, and parmPM3 Compared to the Crystal Structures

	crystal	parm94	parmAM1	parmPM3
$\Delta\bar{H}_f$ (kcal/mol)	0.0	-391.9	-648.6	-692.6
GAVRG _{avrg} (kcal/mol Å)	14.1	13.0	6.5	5.1

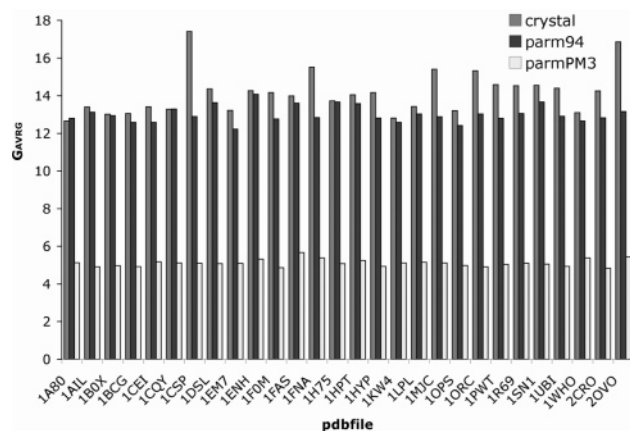


Figure 3. Average atomic gradients for crystal structures, structures minimized with parm94, and structures minimized with parmPM3 for the training set.

is calculated as shown in eq 1

$$\text{GAVRG} = \frac{1}{N} \sum_{i=1}^N \frac{\partial E}{\partial r_i} \quad (1)$$

where N is the total number of atoms, and $\partial E/\partial r_i$ is the derivative of the total energy with respect to the Cartesian coordinates.

There is a marked improvement in the average GAVRG for protein structures after minimization with the parmAM1 and parmPM3 parameters (Table 3). The mean of the initial average gradient for the crystal structure is 14.1 kcal/molÅ, with improvements after optimizing with the MM force fields parm94 (13.0 kcal/molÅ), parmAM1 (6.5 kcal/molÅ), and parmPM3 (5.1 kcal/molÅ). Thus, structures minimized with parmAM1 and parmPM3 are less strained according to the semiempirical calculations, as shown in Figure 3.

The observed improvements in the heats of formation and atomic forces are due primarily to the reparametrization of the bond lengths and angles. Semiempirical treatments were found to be very sensitive to the optimal bond geometry. Despite the improvements, since these minimizations were carried out with a classical force field, the method is still restricted by the limitations in MM modeling such as the use of a fixed point-charge model.

These results illustrate the utility of the parmPM3 parameters to allow proteins to be rapidly minimized with an MM potential before being evaluated at the semiempirical level. By significantly reducing the heat of formation of the protein systems and the atomic forces, preminimizing with parmAM1 or parmPM3 significantly improves the stability of systems before use in QM calculations.

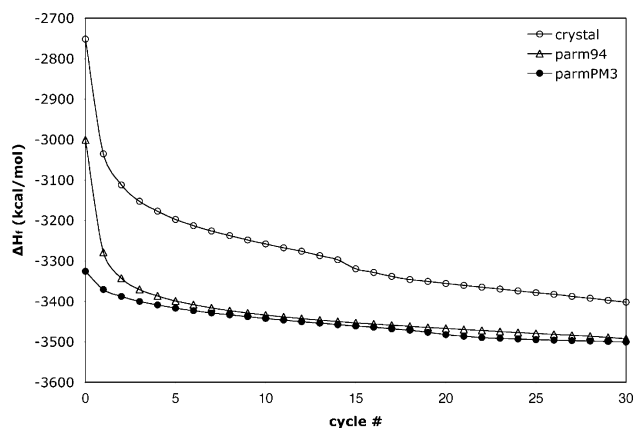


Figure 4. Minimization of 1AIL with PM3 using steepest descent for the crystal structure and the structures minimized with parm94 and parmPM3.

In general, the improvements seen by minimizing with parmAM1 or parmPM3 before scoring with their respective Hamiltonian are comparable. Optimization with parmAM1 leads to structures that are more consistent with the AM1 Hamiltonian, and the same is true for minimizing with parmPM3 in relation to the PM3 Hamiltonian. Since the observed trends and results for both parameter sets mirror each other so closely, we have focused here primarily on the improvements obtained with parmPM3, although the general results and their interpretations also hold true for parmAM1.

Improvements during Minimization. Figure 4 shows a minimization profile for a select protein system in the training set (1AIL). PM3 minimizations were performed for 30 steps using steepest descent starting with either the crystal structure or the structures minimized with parm94 or parmPM3. As shown, there is an absence of a steep drop-off in energy when minimizing the structure that had been preminimized with parmPM3. The energies of the structures preminimized with parm94 and parmPM3 converge to similar values, with structures that are close in overall conformation. A similar trend is seen across the systems in the training set, although in several cases, while the initial heat of formation for structures minimized with parmPM3 is lower than those minimized with parm94, the order of final heats of formation is reversed. This trend is seen for the 1ENH protein system and shown in Figure 5. In general, the parm94 and parmPM3 preminimized structures converge to similar heats of formation upon limited minimization with DivCon.

To investigate the stability of starting structures relative to a semiempirical treatment, structures were minimized using the LBFGS routine instead of steepest descent. The LBFGS minimization scheme reaches a local minimum structure faster than using steepest descent or conjugate gradient techniques, but the starting structure should be close to the local minimum before invoking LBFGS. Figure 6 shows the minimization profile for 1ENH taking starting structures as the crystal structure and structures preminimized with parm94 and parmPM3. As illustrated, in the early stages of minimization of the crystal structure and parm94 preminimized structures, the energy increases before the steep drop-off in energy. The parmPM3 preminimized structure,

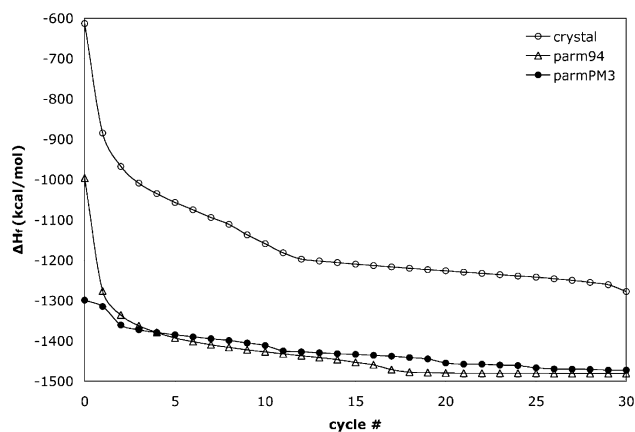


Figure 5. Minimization of 1ENH with PM3 using steepest descent for the crystal structure and the structures minimized with parm94 and parmPM3.

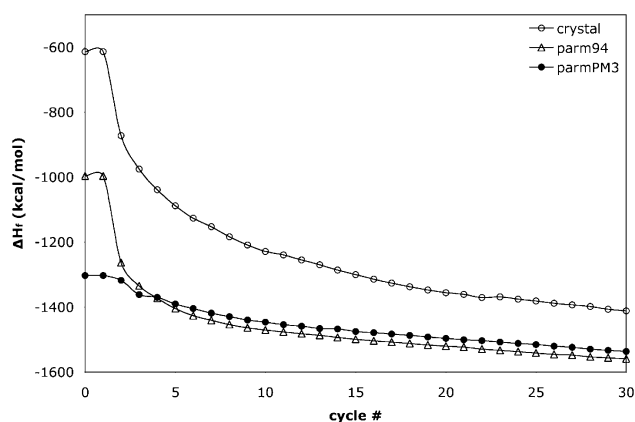


Figure 6. Minimization of 1ENH with PM3 using LBFGS for the crystal structure and the structures minimized with parm94 and parmPM3.

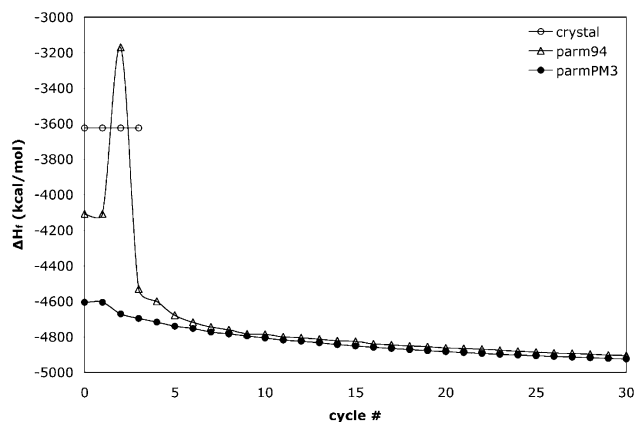


Figure 7. Minimization of 1DSL with PM3 using LBFGS for the crystal structure and the structures minimized with parm94 and parmPM3.

however, exhibits a more stable minimization profile and again demonstrates only a relatively small total decrease in heat of formation upon minimization.

The LBFGS minimization of the 1DSL structure exhibits exaggerated instabilities as seen in Figure 7. In this case, the LBFGS optimization becomes unstable for the crystal structure and is terminated prematurely as the energy continues increasing. The minimization of the parm94

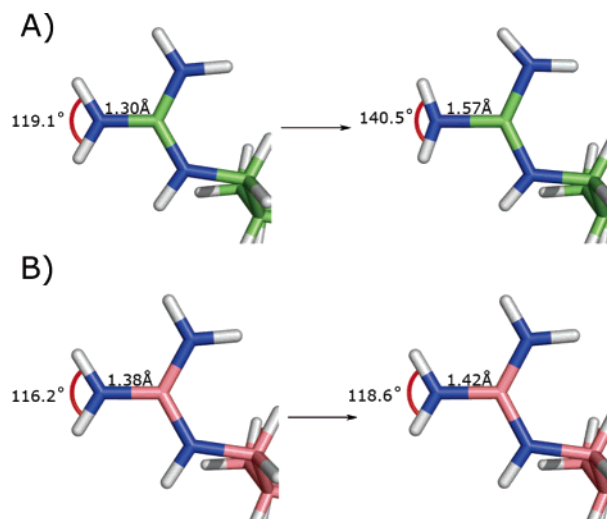


Figure 8. Structure of LYS-13 of 1DSL during LBFGS minimization with parm94 and parmPM3 preminimized structures. (A) Preminimizing with parm94 results in structures with elongated bond lengths and large angles for arginine. (B) Preminimizing with parmPM3 results in more stable bond lengths and angles during minimization.

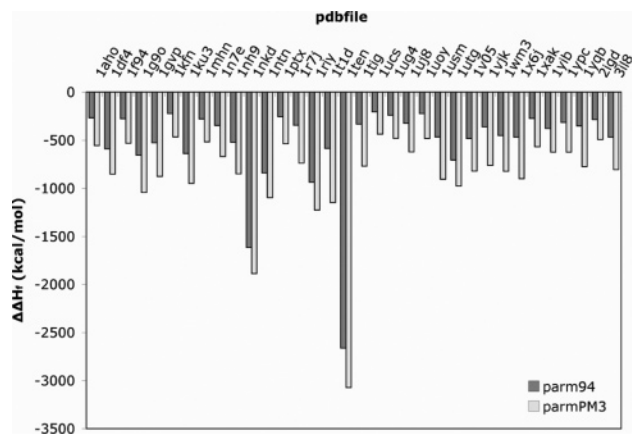


Figure 9. Improvements in the heat of formation for structures minimized with parm94 and parmAM1 relative to their crystal structures for the test set.

preminimized structure exhibits a large increase in energy early in the minimization, while the parmPM3 preminimized structure again shows a stable minimization profile. The energy spike for the parm94 preminimized structure is caused by the large initial forces on the atoms, created by an unfavorable starting geometry. This is illustrated in Figure 8 for the Arg 13 residue of 1DSL. The large forces cause the Cartesian coordinates of the atoms to move by too great a distance, leading to elongated bond lengths and angles in one minimization step, destabilizing the system, and increasing the heat of formation by over 900 kcal/mol. Since the geometry of the parmPM3 preminimized structure is more consistent with a semiempirical approach, the atomic forces for the structure are lower in overall magnitude and so the atomic positions do not vary as much and are more stable during minimizations.

These results highlight the ability of the parmAM1 and parmPM3 force fields to clean up structures for subsequent semiempirical studies. ParmAM1 and parmPM3 premini-

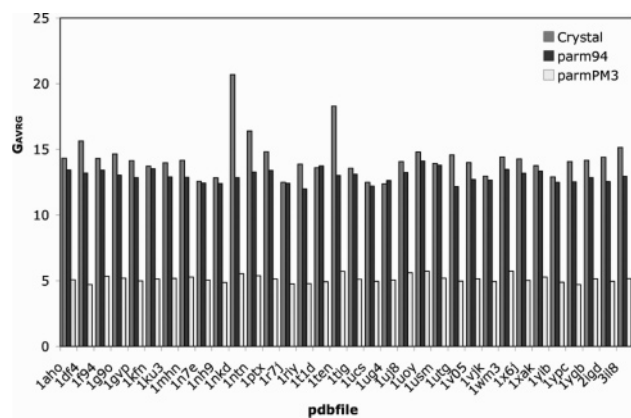


Figure 10. Average atomic gradients for crystal structures, structures minimized with parm94, and structures minimized with parmPM3 for the test set.

minimized structures not only score much lower with respect to heats of formation but also result in less strained structures that are better behaved during the minimization process.

Evaluating the Extensibility of parmAM1 and parmPM3 Parameters. To assess the extensibility of parmAM1 and parmPM3 beyond the training set of proteins, a test set of 34 small proteins structures that had been solved using X-ray crystallography was chosen. As with the training set, the test set listed in Table 4 covers a range of topological features. The reduction in heats of formation and the average gradient for this data set are shown in Figures 9 and 10. As with the test set, the parmPM3 minimized structures have heats of formation over 300 kcal/mol lower than parm94 minimized structures on average, and the average gradient is cut by over 50% for parmPM3 minimized structures. Again, structures preminimized with parmPM3 were more stable during semiempirical minimizations. Overall this illustrates the general applicability of the parmPM3 force field to studying small proteins using semiempirical QM approaches.

Conclusion

Minimizations of proteins at the semiempirical level are time-consuming, even when utilizing linear scaling approaches. In addition to their computational expense, semiempirical calculations are more sensitive to the initial conformation of the starting structure. Since the internal geometries of atoms in X-ray and NMR structures are different from those found in structures minimized at the semiempirical level, many atoms experience large initial forces during minimization. Without the constraint of explicit bonds in the quantum mechanical treatment of the structure, this can lead to undesirable features during the minimization process such as bond cleavage. From these results, it appears that both the parmAM1 and parmPM3 parameter sets are a valuable addition to the semiempirical treatment of proteins. By creating structures that are more geometrically consistent with proteins minimized at the semiempirical level, these structures score and behave better in semiempirical QM calculations. This approach is suitable for use in QM/MM calculations where the QM region is treated at the semiempirical

Table 4. Protein Systems Comprising the Test Set for the Evaluation of parmAM1 and parmPM3

PDB ID	description	resolution (Å)	N_{res}
1AHO	scorpion toxin	0.96	64
1DF4	HIV-1 envelope glycoprotein	1.45	68
1F94	bucandin toxin	0.97	63
1G9O	Pdz domain	1.50	91
1GVP	gene V protein	1.60	87
1KFN	major outer membrane	1.65	56
1KU3	RNA polymerase σ subunit	1.80	73
1MHN	SMN tudor domain	1.80	59
1N7E	sixth PDZ domain of Grip1	1.50	97
1NH9	DNA binding protein Mja10B	2.00	87
1NKD	Cole1 repressor of primer	1.07	65
1NTN	neurotoxin-I	1.90	72
1PTX	scorpion toxin II	1.30	64
1R7J	DNA binding protein Sso10A	1.47	95
1RIY	histone-like DNA binding protein	1.80	90
1TLD	shaker potassium channel	1.51	100
1TEN	fibronectin type III domain	1.80	90
1TIG	translation initiation factor	2.00	94
1UCS	type III antifreeze protein Rd1	0.62	64
1UG4	carditoxin VI	1.60	60
1UJ8	hypothetical protein	1.75	77
1UOY	bubble protein	1.50	64
1USM	protein-binding transcriptional coactivator	1.20	80
1UTG	oxidized uteroglobin	1.34	70
1V05	human filamin	1.43	96
1VJK	molybdopterin converting factor	1.51	98
1W3M	human sumo-2 protein	1.20	72
1X6J	hypothetical protein Yfgy	2.00	91
1XAK	sar-coronavirus Orf7A accessory protein	1.80	83
1YIB	microtubule-associated protein Rp/Eb	1.80	76
1YPC	chymotrypsin inhibitor 2	1.70	64
1YQB	ubiquitin 3	2.00	100
2IGD	protein G IgG-binding domain III	1.10	61
3IL8	interleukin 8	2.00	72

level as well as in QM studies of, for example, protein–ligand interactions.

The parmAM1 and parmPM3 parameter sets represent a fast and effective preminimization step for semiempirical quantum mechanical calculations. By providing a consistent approach to removing strain in the protein, these new parameter sets allow for subsequently more reliable calculations using semiempirical QM methods.

Abbreviations Used. Quantum mechanics, QM; molecular mechanics, MM.

Acknowledgment. We thank Kenneth Ayers for his assistance with managing the computational resources and for valuable discussions. We would also like to thank Duane Williams for his input and proofreading of this manuscript. We thank the NSF (MCB-0211639) and the NIH (GM 44974) for support.

Supporting Information Available: Parameters for parmAM1 and parmPM3; bond length, bond angle, Lennard-Jones, and charge parameters; the parameter sets in formats consistent with the AMBER molecular modeling package (parmAM1.dat and parmPM3.dat) as well as files containing the charge information (all_aminoAM1.in and all_aminoPM3.in). This information is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Kohn, W.; Becke, A. D.; Parr, R. G. Density functional theory of electronic structure. *J. Phys. Chem.* **1996**, *100* (31), 12974–12980.
- (2) Hehre, W. J.; Radom, L.; Pople, J. A.; Schleyer, P. V. R. *Ab Initio Molecular Orbital Theory*; Wiley-Interscience: 1986.
- (3) 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.
- (4) 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.
- (5) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. 1. Method. *J. Comput. Chem.* **1989**, *10* (2), 209–220.
- (6) Dixon, S. L.; Merz, K. M., Jr. Semiempirical Molecular Orbital Calculations with Linear System Size Scaling. *J. Chem. Phys.* **1996**, *104*, 6643–6649.
- (7) Dixon, S. L.; Merz, K. M., Jr. Fast, Accurate Semiempirical Molecular Orbital Calculations for Macromolecules. *J. Chem. Phys.* **1997**, *107*, 879–893.
- (8) 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.
- (9) van der Vaart, A.; Suarez, D.; Merz, K. M. Critical assessment of the performance of the semiempirical divide and conquer method for single point calculations and geometry optimizations of large chemical systems. *J. Chem. Phys.* **2000**, *113* (23), 10512–10523.
- (10) 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.
- (11) 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.
- (12) Lee, T. S.; York, D. M.; Yang, W. Linear-scaling semiempirical quantum calculations for macromolecules. *J. Chem. Phys.* **1996**, *105*, 2744–2750.
- (13) Stewart, J. J. P. *MOPAC2000*; Fujitsu Ltd.: Tokyo, 1999.
- (14) Daniels, A. D.; Milliam, J. M.; Scuseria, G. E. Semiempirical Methods with Conjugate Gradient Density Matrix Search to replace Diagonalization for Molecular Systems Containing Thousands of Atoms. *J. Chem. Phys.* **1997**, *107*, 425–431.
- (15) Raha, K.; Merz, K. M., Jr. Calculating Binding Free Energy in Protein–ligand Interaction. *Ann. Reports Comput. Chem.* **2005**, *1*, 113.
- (16) Case, D. A.; Darden, T. A.; Cheatham, I. T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8.0*; 2004.
- (17) 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.
- (18) Gilis, D.; Massar, S.; Cerf, N. J.; Rooman, M. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol.* **2001**, *2* (11), 1–12.
- (19) Li, J. B.; Zhu, T. H.; Cramer, C. J.; Truhlar, D. G. New Class IV Charge Model for Extracting Accurate Partial Charges from Wave Functions. *J. Phys. Chem.* **1998**, *102*, 2 (10), 1820–1831.
- (20) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges – the Resp Model. *J. Phys. Chem.* **1993**, *97*, 7 (40), 10269–10280.
- (21) Jorgensen, W. L.; Pranata, J. Importance of Secondary Interactions in Triply Hydrogen-Bonded Complexes – Guanine-Cytosine Vs Uracil-2,6-Diaminopyridine. *J. Am. Chem. Soc.* **1990**, *112* (5), 2008–2010.
- (22) Goldberg, D. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: San Mateo, CA, 1989.

CT0600161

JCTC Journal of Chemical Theory and Computation

Theoretical Investigation of Cheletropic Decarbonylation Reactions

Chin-Hung Lai,[†] Elise Y. Li,[†] Kew-Yu Chen,[†] Tahsin J. Chow,^{*,‡} and Pi-Tai Chou^{*,†}

*Department of Chemistry, National Taiwan University, 106, Taipei, Taiwan, R. O. C.,
and Institute of Chemistry, Academia Sinica, Taipei, 115, Taiwan, R.O.C.*

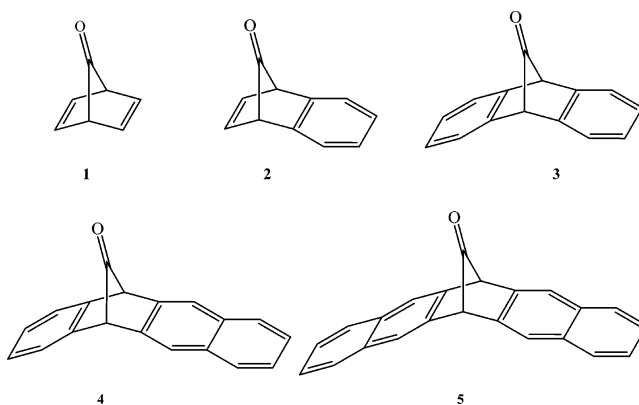
Received January 10, 2006

Abstract: In this study, B3LYP is used to calculate the decarbonylation reactions of the bicyclo-[2.2.1]hepta-2,5-dien-7-one (7-norbornadienone, **1**) and its related extended fused aromatic analogues **2–5**. On the basis of our results, all of the reactions tend to proceed synchronously to expel CO, forming the corresponding aromatic hydrocarbons. It is found that the more exothermic the reaction is, the less of a reaction barrier it needs to overcome. Moreover, upon a decrease of the reaction exothermicity, the structure of the transition state is farther away from the reactant, and the reaction barrier increases. The results agree well with the Hammond postulate as well as the Bell–Evans–Polanyi principle. Studies predict an activation energy of 27.83 kcal/mol for **5**, so that the production of pentacene from compound **5** might proceed at elevated temperatures such as 400 K.

1. Introduction

Since the early 1900s, it has been recognized that bicyclo-[2.2.1]hepta-2,5-dien-7-one (7-norbornadienone, **1**) was remarkably prone to fragmentation, losing carbon monoxide.¹ The previously reported activation energy for its fragmentation to carbon monoxide and benzene was 15 ± 2.5 kcal/mol.^{2,3} Landesberg and Siczkowski⁴ suggested that **1** was destabilized by electron repulsion between the olefinic and carbonyl π orbitals. Woodward and Hoffmann⁵ discussed that the decarbonylation of **1** as an orbital symmetry allowed cheletropic reaction, which was in turn a class of pericyclic reactions. They highlighted it as a prime example of how the availability of an allowed pathway may lower the activation energy of a reaction. In this study, we have performed a systematic approach on a series of analogues of **1** by fusing various numbers of the benzene ring (see Scheme 1) and developed some discussions extracted from the results. An extension of this study that can be made is whether pentacene can be produced from the decarbonylation reaction. Pentacene is a key prototype used in organic single-crystal field effect transistors (FETs). Interest in organic

Scheme 1. Structures of Various Compounds in This Study



devices stems from their mechanical flexibility, their potential for interfacing to biological systems, and their ease of processing over large areas.^{6–9}

2. Theoretical Method

All calculations are done with the Gaussian 03 program.¹⁰ The B3LYP functional is used with the basis set 6-31G* (hereafter designated as B3LYP).^{11,12} The calculated minima and transition states (TSs) have been carefully checked by frequency analyses to examine whether the number of the

* Corresponding author fax: +2-23695208; e-mail: chop@ntu.edu.tw.

[†] National Taiwan University.

[‡] Academia Sinica.

imaginary frequency is zero or one. All mentioned energetic values are corrected for zero-point vibrational energy unless otherwise specified. All rate constants are calculated according to transition-state theory incorporating partition functions.¹³ In transition-state theory, the rate constant could be expressed as

$$k = \frac{k_B T}{h} K^\ddagger \quad (1)$$

where K^\ddagger denotes an equilibrium constant between the reactant and the activated complex. In statistical mechanics, the equilibrium constant K^\ddagger could be represented by molecular partition functions of the activated complex and reactant. In Born–Oppenheimer approximation with a neglect of vibrational coupling, the molecular partition function can be factorized into its translational, rotational, vibrational, electronic, and nuclear parts.¹⁴ The translational and nuclear partition functions of the activated complex and reactant are assumed to be unchanged and can thus be canceled out. Accordingly, K^\ddagger can be expressed as shown in eq 2.

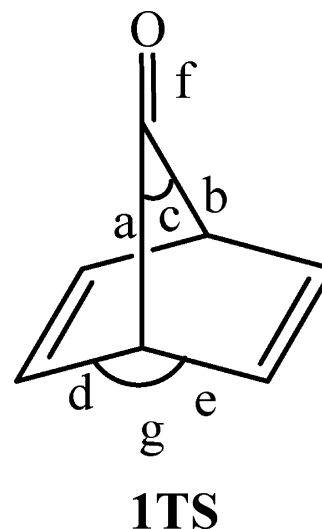
$$K^\ddagger = \frac{q_r^\ddagger q_v^\ddagger q_e^\ddagger}{q_r q_v q_e} \quad (2)$$

where the superscript \ddagger gives an indication of the partition functions of the activated complex and q_r , q_v , and q_e are the rotational, vibrational, and electronic partition functions, respectively. q_v^\ddagger excludes the contribution of the reaction coordinate, and q_e just considers the contribution of the ground state. The critical rotational constants and vibrational frequencies for the rotational and vibrational partition functions are calculated by B3LYP. More detailed information is provided in the Supporting Information.

3. Results and Discussion

3.1. The Reliability of the B3LYP Results. First of all, the reliability of our B3LYP results is investigated. In Table 1, the results of B3LYP and MP2¹⁵ with the same basis set (hereafter designated as MP2) are compared with the previous results for the cases of compounds **1** and **2**.^{2,16} The geometric parameters of interest for further discussion are also defined in Table 1, in which a truncated structure of the transition state is depicted as **1TS**. Similar tendencies can be found between these two results. Both methods predict that upon fusing one benzene ring to the 7-norbornadienone skeleton, forming **2**, the activation energy is raised and the exothermicity of the reaction is decreased in comparison to that of **1**. More importantly, one could find that both MP2 and B3LYP predict similar lengths for geometric parameter a, defined in Table 1. As can be seen in Table 1, our results agree with most of the listed previous experimental and theoretical values.^{2,16} Thus, both results render firm support in that the reaction is synchronous rather than nonsynchronous. On the basis of these benchmark tests, we conclude that the results of B3LYP are trustworthy for dealing with the current system. Furthermore, the results for **1** calculated by the B3LYP method with various basis sets (6-31G*, 6-31+G*, 6-311G*, and 6-311+G*) are compared and listed in Table 2, along with the data in previous literature.^{2,16}

Table 1. B3LYP and MP2 Results for the Decarbonylation Reactions and the Transition State Geometric Structures of **1** and **2** (Energetic Values in kcal/mol, Bond Lengths in Å, and Bond Angles in deg)



	HF	B3LYP	MP2	previous result
1				
E_a^a	21.98	12.95	12.06	7.8 ^b 18 ^c 16 ± 2.5 ^d 15.2 ^e
ΔH^f	-56.10	-39.51	-36.98	-51 ^c -32.5 ^e
a	1.970	1.990	1.986	1.980 ^b 1.986 ^g
c	78.85	78.84	79.21	80.0 ^b
d	1.450	1.450	1.450	1.451 ^g
e	1.450	1.450	1.450	1.451 ^g
f	1.130	1.150	1.150	1.170 ^g
g	115.5	115.6	115.2	
2				
E_a		16.44	15.61	
ΔH		-28.86	-25.72	
a		2.021	2.021	
c		78.81	78.96	
d		1.440	1.430	
e		1.460	1.450	
f		1.150	1.150	
g		116.3	115.8	

^a E_a : the activation energy. ^b The results were calculated by MP2/4-31G, see ref 15. ^c The ΔG^\ddagger and ΔH_f values in ref 2b. ^d The experimental E_a value from NMR, see ref 2a. ^e The values are calculated by MP4(SDTQ)/D95**, see ref 2c. ^f ΔH : the heat of formation. ^g The values are calculated by MP2/6-31G*, see ref 2c.

Because the reaction does not involve hydrogen atoms, the addition of second diffuse and polarization functions, that is, extra functions on hydrogen, is not necessary. This viewpoint can be supported by the results that 6-31G* has similarities with 6-31G** (see Table 2) in this study. Likewise, all basis sets predict similar lengths for geometric parameter a, defined in Table 1, indicating that synchronicity of the reaction is maintained among these basis sets. In advance, we also perform the HF/6-31G* calculation and

Table 2. Results of the Decarbonylation Reaction of **1** Using Different Basis Sets with the Same Method B3LYP (Energetic Values in kcal/mol, Bond Lengths in Å and Bond Angles in deg)

	6-31G*	6-31G**	6-31+G*	6-311G*	6-311+G*	previous data
E_a	12.95	12.90	12.86	11.65	12.03	7.8 ^a 18 ^b 16 ± 2.5 ^c 15.2 ^d
ΔH	-39.51	-39.48	-41.39	-43.23	-42.81	-32.5 ^d
Selected Structural Parameter of Transition State						
a^e	1.990	1.990	1.985	1.970	1.979	1.980 ^a 1.986 ^f
c^e	78.84	78.91	79.23	79.39	79.28	80.0 ^a
d^e	1.450	1.450	1.450	1.450	1.450	1.451 ^f
e^e	1.450	1.450	1.460	1.450	1.450	1.451 ^f
f^e	1.150	1.150	1.150	1.150	1.150	1.170 ^f
g^e	115.6	115.6	115.6	115.4	115.3	

^a The results were obtained by MP2/4-31G, see ref 15. ^b The ΔG^\ddagger and ΔH_r values are taken from ref 2b. ^c The experimental E_a values are taken from NMR, see ref 2a. ^d The values were calculated by MP4(SDTQ)/D95**, see ref 2c. ^e The parameters are defined in Table 1. ^f The values were calculated by MP2/6-31G*, see ref 2c.

compare it with B3LYP, MP2, and the previous results listed in Table 1.^{2,16} As shown in Table 1, B3LYP can handle some electron correlation because of its prediction of similar tendencies in parameters of interest, such as the enthalpy of reaction (ΔH) and the activation energy (E_a), as well as some critical bond angles and distances with respect to MP2. In summary, all of the aforementioned results support our choice of using B3LYP to calculate even larger compounds such as **3–5**.

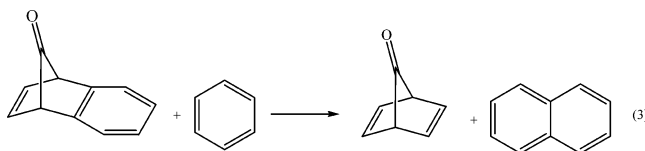
3.2. The Synchronicity of the Reaction. A synchronous reaction is defined as the breakage (or formation) of several chemical bonds simultaneously during the reaction. Whether concerted pericyclic reactions occur via synchronous or nonsynchronous pathways and which theoretical or experimental criteria should be used to differentiate between them appeared to be a core issue in the past.^{5,17–20} In fact, ab initio calculations on pericyclic reactions have generally found synchronous transition states.²¹ Occasionally, nonsynchronous transition states have been found at low levels of theory, but they may disappear at higher correlated levels of theory.^{20a,b} For the decarbonylation of **1**, nonsynchronous transition structures have been located only at the UMNDO and UHF/STO-3G levels. But at higher levels of theory such as (U)HF/4-31G, only the synchronous transition state has been found. In this study, we have made an attempt to locate nonsynchronous transition structures using both UB3LYP and UMP2 methods. In this attempt, different lengths for geometric parameters a and b (defined in Table 1) were set as the initial guess for the transition-state optimization. Even though such methods might be expected to favor nonsynchronous pathways, these two methods only optimized to synchronous transition structures.

3.3. The Location of the Transition State. The properties of the transition state play a critical role in chemical reactions. The calculated reaction heats and activation energies are summarized in Table 3. It was found that the exothermicity

Table 3. Activation Energies (E_a), the Heats of Formation (ΔH), and Marcus and Miller Parameters of Compounds **1–5** for the Decarbonylation Reactions

	E_a	ΔH	Marcus	Miller
1	12.95	-39.51	0.1186	0.1980
2	16.44	-28.86	0.2806	0.2663
3	22.47	-14.29	0.4205	0.3794
4	24.91	-8.731	0.4562	0.4254
5	27.83	-2.502	0.4888	0.4785

decreased as the number of fused benzene rings increased from **1** to **5**. The tendency of the reaction heat could be rationalized by the available resonance energy after reaction. The resonance energy is usually defined as a measure of the extra stability in conjugated systems relative to their corresponding isolated double-bond analogues. The isodesmic reaction²⁴ shown in eq 3 (taking **2** as an example) can be used to estimate the available resonance energy after the decarbonylation reaction of **2**, **3**, **4**, or **5** relative to **1**. The reaction enthalpy of eq 3, that is, ΔH_2 , is equivalent to the difference in resonance stabilization energy between **1** (RE_1) and **2** (or **3**, **4**, or **5**) (RE_2 (or 3, 4, or 5)). $\Delta H_2 < 0$ stands for $(RE)_2$ (or 3, 4, 5) $>$ $(RE)_1$ and vice versa. Accordingly, values of ΔH_2 are calculated to be 10.66, 25.22, 30.78, and 37.01 kcal/mol for **2**, **3**, **4**, and **5**, respectively. Apparently, as the number of fused benzene rings increase from **1** to **5**, the resonance stabilization energy of the corresponding products from benzene to pentacene decreases accordingly.



As shown in Table 3, the more exothermic the reaction is, the less of an activation barrier it has, the results of which agree well with the Hammond postulate,²⁵ which proposes a simple qualitative correlation to relate the position of the transition state with respect to the energies of the reaction. This postulate has been proven to be valid for most chemical reactions, although some exceptions have been reported.^{26,27} In 1986, Birney and Berson found a good correlation between the kinetic and thermodynamic stabilities of decarbonylation of **1** and other orbital-symmetry-allowed cycloreversion reactions by plotting the activation free energy (ΔG^\ddagger) and the enthalpy of reaction (ΔH_r).^{2b} On the basis of a similar plot (see Figure 1), a good linear relationship ($R^2 = 0.9991$) was found between ΔG^\ddagger and ΔH_r among **1–5**, consistent with Birney and Benson's kinetics/thermodynamics correlation.

Several relevant models have been developed to quantitatively characterize the transition-state position, of which two are introduced here. Marcus²⁸ suggested an expression

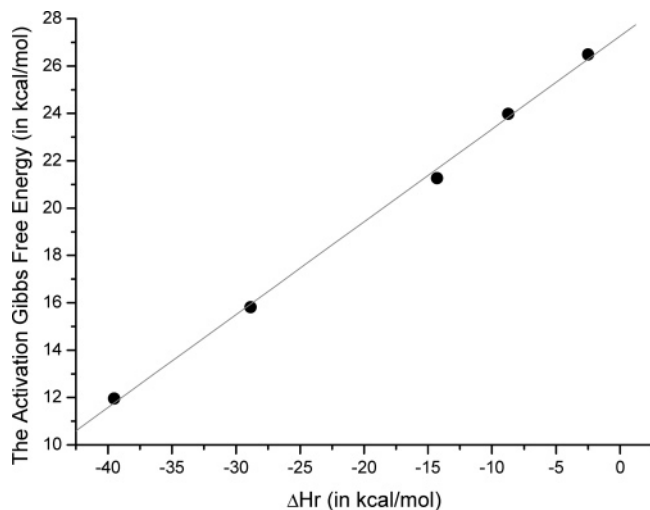


Figure 1. Linear fitting of the activation Gibbs free energy (ΔG^\ddagger) and the reaction heat (ΔH_r). Solid points are theoretical results ($R^2 = 0.9991$)

for the position of the transition state (χ^\ddagger), as given by eq 4.

$$\chi^\ddagger (\text{Marcus}) = 0.5 + \frac{\Delta H_r}{8E_a^0} \quad (4)$$

where ΔH_r is the heat of reaction and E_a^0 is the intrinsic activation energy. This equation was originally proposed to characterize electron-transfer reactions. Chen and Murdoch²⁹ as well as others³⁰ pointed out that eq 4 could also be used for the interpretation of other types of chemical reactions. Miller³¹ devised a similar formulation for the energetic behavior of the chemical reaction, expressed in eq 5

$$\chi^\ddagger (\text{Miller}) = \frac{1}{2 - \Delta H_r/E_a} \quad (5)$$

For both Marcus and Miller parameters, the smaller χ^\ddagger value indicates that the transition state is closer to the reactant. When the value was larger (smaller) than 0.5, the transition state was productlike (reactantlike).

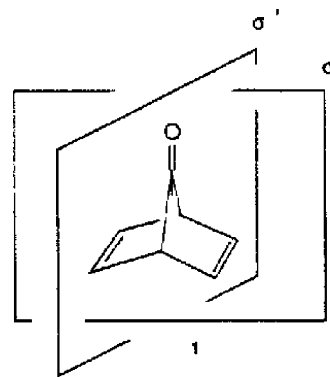
Accordingly, the Marcus and Miller parameters are used to calculate the transition-state position, and the results are also listed in Table 3. Because of the exothermic reaction in the decarboxylation reaction of **1–5**, the resulting χ^\ddagger values of <0.5 are expected. Furthermore, one can see that, as the number of fused benzene rings increases, the Marcus and Miller parameters become larger, the results from which indicate that the transition state shifts to the product side as the decarboxylation reaction takes place in compounds **1–5**. Furthermore, according to the Bell–Evans–Polanyi principle,³² for closely related reactions, there exists a linear relation between E_a and ΔH_r , expressed as

$$E_a = A + B\Delta H_r \quad (6)$$

As depicted in Figure S1 (see the Supporting Information), a good linear relationship was observed ($R^2 = 0.9983$) between E_a and ΔH_r for **1–5**, consistent with the Bell–Evans–Polanyi principle.

Additional firm support is provided by the structural analysis. The selected geometric parameters and the B3LYP

Table 4. Selected Geometric Parameters of **1–5** (Bond Lengths in Å and Bond Angles in deg)



	a ^a	c ^a	d ^a	e ^a	f ^a	g ^a
1	1.584	93.80	1.520	1.520	1.190	110.2
	1.607 ^b	93.4 ^b	1.533 ^b	1.533 ^b	1.226 ^b	
	1.575 ^c		1.528 ^c	1.528 ^c	1.193 ^c	
	1.580 ^d		1.512 ^d	1.512 ^d	1.201 ^d	
2	1.578	94.86	1.520	1.520	1.190	109.1
3	1.574	95.78	1.520	1.520	1.190	108.4
4	1.572	96.05	1.520	1.520	1.190	108.1
5	1.570	96.25	1.520	1.510	1.190	108.1

^a The parameters are defined in Table 1, and the geometry is optimized under a Cs symmetry. The σ planes in **1–5** are defined above (taking **1** as an example). Note that **2** and **4** have no σ' plane. ^b The values were calculated by MP2/4-31G, see ref 15. ^c The values were calculated by MNDO/3, see ref 21. ^d The values were calculated by MP2/6-31G*, see ref 2c.

results of **1–5** are extracted and summarized in Table 4. Some previous results for **1** are also listed in Table 4 for comparison.^{2,16,22} One can promptly see that when the number of fused benzenes on either side of 7-norbornadienone increases, that is, from **1** to **5**, the geometry of the 7-norbornadienone skeleton changes only slightly, even for the unsymmetric analogues **2** and **4**. Geometric parameter a decreases slightly as the number of fused benzenes on either side of 7-norbornadienone increases. Apparently, as the decarbonylative reaction becomes more favorable, geometric parameter a or b increases accordingly. This trend agrees with the previous discussions about the retro Diels–Alder reaction and cheletropic fragmentation.³³ These reactions have been proven in accord with the structure correlation principle of Dunitz et al.³⁴ The structure correlation principle suggests that the structures of molecules can show distortions along a reaction coordinate, but only when the electronic factors that stabilize the transition state are present in an appropriate ground-state geometry. This is indeed a corollary of the Hammond postulate and of the Bell–Evans–Polanyi principle. Table 5 lists some critical bond distances and angles for the transition states of **3–5** (also see Table 1 for **1** and **2**). In comparison, it was found that the change of the O=C...C bonding distance between the reactant and the corresponding transition state increases with an increase in the number of fused benzenes from **1** to **5** (see Table 5). The results clearly show that the structure of the transition state is shifted to the product once the exothermicity of the decarboxylation reaction is decreased from **1** to **5**, accompanied by an increase of the activation energy.

Table 5. B3LYP Results for the Decarbonylation Reactions and the Transition State Geometric Structures of **3–5** (Energetic Values in kcal/mol, Bond Lengths in Å, and Bond Angles in deg)

	3	4	5
E_a	22.47	24.91	27.83
ΔH	-14.29	-8.731	-2.502
a^a	2.098	2.120	2.156
c^a	77.00	76.60	75.73
d^a	1.440	1.440	1.440
e^a	1.440	1.450	1.440
f^a	1.150	1.150	1.150
g^a	117.2	117.6	117.9

^a The parameters are defined in Table 1.

Table 6. Calculated Rate Constants (s^{-1}) of the Cheletropic Decarbonylation Reactions for Compounds **1–5**

	1	2	3	4	5
			$\xrightarrow{k_r}$		
				benzene	
				naphthalene	
				anthracene + CO	
				tetracene	
				pentacene	
	k_r at 200 K	k_r at 250 K	k_r at 300 K	k_r at 400 K	
1	6.014×10^1 $4.62 \times 10^{-4}{}^a$	1.250×10^4	4.196×10^5	3.601×10^7	
2	1.020×10^{-5}	5.682×10^{-2}	1.876×10^1	2.843×10^4	
3	6.429×10^{-12}	7.836×10^{-7}	2.043×10^{-3} $2.39 \times 10^{-5}{}^b$	4.163×10^1	
4	9.012×10^{-15}	3.729×10^{-9}	2.202×10^{-5}	1.253	
5	8.814×10^{-18}	1.649×10^{-11}	2.667×10^{-7}	5.367×10^{-2}	

^a The experimental result at 213 K, see ref 2a. ^b The experimental result at 298 K, see ref 35.

3.4. The Rate Constants. In 1979, Irie and Tanida reported the effect of the electron-withdrawing group, NO_2 , on the decarbonylation rate of compound **3** and found that the reaction rate increased as the number of NO_2 groups on the benzene increased. The decarbonylation rates for compound **3** in dioxane were 2.39×10^{-5} , 6.64×10^{-5} , and $1.45 \times 10^{-4} s^{-1}$ when the number of NO_2 groups was zero, one, and two at 25 °C, respectively.³⁵ In this study, we make an attempt to calculate the decarboxylation reaction rate constant according to the transition-state theory (see the section on the theoretical method).¹³ Details of the derivation of rate constants listed in Table 6 are provided in the Supporting Information. At the ambient temperature of 300 K, the rate of decarboxylation for **1–5** is on the order of 4.2×10^5 , 1.9×10^1 , 2.0×10^{-3} , 2.2×10^{-5} , and $2.7 \times 10^{-7} s^{-1}$, respectively. For **1**, the experimental value of $k = 4.62 \times 10^{-4} s^{-1}$, deduced from a half-life of 25 min at 213 K,^{2a} is somewhat smaller than the $6.014 \times 10^1 s^{-1}$ at 200 K calculated by the theoretical approach. This discrepancy may be due to the fact that transition-state theory neglects the probability of the backward direction of the reaction. Furthermore, solvation effects may play a major role, which is not considered in the current approach.

Nevertheless, the results indicate a minor to negligible degree of decomposition for **4** and **5** at 300 K, whereas **1** is readily decomposed. The results are in qualitative agreement

with the existing experimental evidence,^{1–3} which reports that the optimal temperature needed to observe decarbonylation reactions for **1** and **3** is 195 and 300 K, respectively. If one anticipates the production of pentacene by the cheletropic decarbonylation reaction of **5**, a temperature of 400 K, in which the rate of reaction is calculated to be $5.4 \times 10^{-2} s^{-1}$, may serve as an optimal experimental condition. At this elevated temperature, the disadvantage for **5** seems to be the result of its small exothermic reaction with ΔH of only -2.502 kcal/mol. Nevertheless, because the product CO is in the gas phase at, for example, 1 atm and 400 K, the equilibrium should favor pentacene formation according to Le Châtelier's principle.

4. Conclusion

In conclusion, we have reported a systematic approach on the cheletropic decarbonylation reaction of **1–5**. The results clearly conclude that the thermal cheletropic decarbonylation reaction tends to proceed synchronously for **1–5**. We are particularly interested in the decarbonylation of compound **5** in producing pentacene. In view of the FET application, although thermally generated pentacene could be made by retro Diels–Alder reaction,^{36–39} side products are unavoidable and may exist as the impurity in the film. Alternatively, photolysis of the pentacene precursor, generating pentacene, has been reported. However, the product, that is, pentacene, is also subject to photolysis.^{39,40} Accordingly, **5** may serve as a prototypical pentacene precursor for organic field-effect transistors because its thermally activated process to produce pentacene is feasible under an elevated temperature of, for example, 400 K. Our results also draw a conclusion in that, as the number of fused benzene rings on either side of 7-norbornadienone increases, that is, from **1** to **5**, the corresponding E_a and the endothermicity of the reaction both increase, and the structure of the transition state leans toward the product side. This tendency can be well-described by the Hammond postulate as well as the Bell–Evans–Polanyi principle.

Acknowledgment. We are grateful to the National Center for High-Performance Computing of Taiwan for allowing us generous amounts of computing time. We also thank the National Science Council for financial support and Prof. Hu of National Chung Cheng University for his kind support in the calculation of rate constants.

Supporting Information Available: Figure S1, giving the linear fitting of activation energy (E_a) and the reaction heat (ΔH_r), and the detailed method of deriving rate constants are available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) (a) Dilthey, W.; Schommer, W.; Trosken, O. *Ber. Dtsch. Chem. Ges.* **1933**, *66*, 1627. (b) Ogliaruso, M. A.; Romanelli, M. G.; Becker, E. I. *Chem. Res.* **1965**, *65*, 261. (c) Allen, C. F. H.; Van Allan, J. *J. Am. Chem. Soc.* **1942**, *64*, 1260. (d) Fieser, L. F.; Fieser, M. *Organic Experiments*; D. C. Heath and Co.: Boston, MA, 1964; pp 303–317. (e) Bartlett, P. D.; Giddings, W. P. *J. Am. Chem. Soc.* **1960**, *82*, 1240. (f)

- Story, P. R.; Fahrenholtz, S. R. *J. Am. Chem. Soc.* **1964**, *86*, 1270. (g) Wilt, J. W.; Chenier, P. J. *J. Org. Chem.* **1970**, *35*, 1562. (h) Gassman, P. G.; Aue, D. H.; Patton, D. S. *J. Am. Chem. Soc.* **1964**, *86*, 4211. (i) Hoffmann, R. W.; Hauser, H. *Tetrahedron* **1965**, *21*, 891. (j) Lemal, D. M.; Gosselink, E. P.; Adult, A. *Tetrahedron Lett.* **1964**, 579. (k) Lemal, D. M.; Gosselink, E. P.; McGregor, S. D. *J. Am. Chem. Soc.* **1966**, *88*, 582. (l) Halton, B.; Battiste, M. A.; Rehberg, R.; Deyrup, C. L.; Brennan, M. E. *J. Am. Chem. Soc.* **1967**, *89*, 5964. (m) Zhang, J.; Ho, D. M.; Pascal, R. A., Jr. *J. Am. Chem. Soc.* **2001**, *123*, 10919. (n) Warrener, R. N.; Harrison, P. A. *Molecules* **2001**, *6*, 353. (o) Tobe, Y.; Kubota, K.; Naemura, K. *J. Org. Chem.* **1997**, *62*, 3430. (p) Plummer, B. F.; Currey, J. A.; Russell, S. J.; Steffen, L. K.; Watson, W. H.; Bourne, S. A. *Struct. Chem.* **1995**, *6*, 167. (q) Simpson, C. J. S. M.; Price, J.; Holmes, G.; Adam, W.; Martin, H.-D.; Bish, S. *J. Am. Chem. Soc.* **1990**, *112*, 5089.
- (2) (a) Birney, D. M.; Berson, J. A. *J. Am. Chem. Soc.* **1985**, *107*, 4553. (b) Birney, D. M.; Berson, J. A. *Tetrahedron* **1986**, *42*, 1561. (c) Birney, D. M.; Ham, S.; Unruh, G. R. *J. Am. Chem. Soc.* **1997**, *119*, 4509.
- (3) LeBlanc, B. F.; Sheridan, R. S. *J. Am. Chem. Soc.* **1985**, *107*, 4554.
- (4) (a) Landesberg, J. M.; Sieczkowski, J. *J. Am. Chem. Soc.* **1968**, *90*, 1655. (b) Landesberg, J. M.; Sieczkowski, J. *J. Am. Chem. Soc.* **1969**, *91*, 2120. (c) Landesberg, J. M.; Sieczkowski, J. *J. Am. Chem. Soc.* **1971**, *93*, 972.
- (5) Woodward, R. B.; Hoffmann, R. W. *The Conservation of Orbital Symmetry*; Academic Press: New York, 1970.
- (6) Peumans, P.; Uchida, S.; Forrest, S. R. *Nature* **2003**, *425*, 158.
- (7) Volkel, A. R.; Street, R. A.; Knipp, D. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2002**, *66*, 195336.
- (8) Campbell, I. H.; Smith, D. L. *Solid State Phys.* **2001**, *55*, 1.
- (9) Nelson, S. F.; Lin, Y. Y.; Gundlach, D. J. *Appl. Phys. Lett.* **1998**, *72*, 1854.
- (10) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T., Jr.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Pittsburgh, PA, 2004.
- (11) (a) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.
- (12) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724.
- (13) (a) Gilbert, R. G.; Smith, S. C. *Theory of Unimolecular and Recombination Reactions*; Blackwell Scientific Publications: Oxford, U. K., 1990. (b) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. *J. Phys. Chem.* **1996**, *100*, 12771. (c) Robertson, S. H.; Wagner, A. F.; Wardlaw, D. M. *J. Chem. Phys.* **1995**, *103*, 2917. (d) Holbrook, K. A.; Pilling, M. J.; Robertson, S. H. *Unimolecular Reactions*, 2nd ed.; John Wiley & Sons: Chichester, U. K., 1996.
- (14) (a) Zhang, M.; Huang, Z.; Lin, Z. *J. Chem. Phys.* **2005**, *122*, 134313. (b) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*; John Wiley & Sons: Chichester, U. K., 2002.
- (15) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *Chem. Phys. Lett.* **1988**, *153*, 503.
- (16) Birney, D. M.; Wiberg, K. B.; Berson, J. A. *J. Am. Chem. Soc.* **1988**, *110*, 6631.
- (17) (a) Hoffmann, R. W.; Woodward, R. B. *J. Am. Chem. Soc.* **1965**, *87*, 2046. (b) Fukui, K.; Yonezawa, T.; Nagata, C.; Shingu, H. *J. Chem. Phys.* **1954**, *22*, 1433. (c) Salem, L. *J. Am. Chem. Soc.* **1968**, *90*, 543, 553. (d) Longuet-Higgins, H. C.; Abrahamson, E. W. *J. Am. Chem. Soc.* **1965**, *87*, 2045. (e) Van der Lugt, W. T. A. M.; Oosterhoff, L. J. *J. Am. Chem. Soc.* **1969**, *91*, 6042.
- (18) Dewar, M. J. S. *J. Am. Chem. Soc.* **1984**, *106*, 209.
- (19) (a) Dewar, M. J. S.; Pierini, A. B. *J. Am. Chem. Soc.* **1984**, *106*, 203. (b) Gajewski, J. J.; Peterson, K. B.; Kagel, J. R. *J. Am. Chem. Soc.* **1987**, *109*, 5545. (c) Huisgen, R. J. *J. Org. Chem.* **1976**, *41*, 403. (d) Firestone, R. A. *J. Org. Chem.* **1972**, *37*, 2181. (e) Houk, K. N. *J. Am. Chem. Soc.* **1981**, *103*, 2436. (f) Taagepera, M.; Thornton, E. R. *J. Am. Chem. Soc.* **1972**, *94*, 1168. (g) Tolbert, L. M.; Ali, M. B. *J. Am. Chem. Soc.* **1984**, *106*, 3806.
- (20) (a) McIver, J. W., Jr. *Acc. Chem. Res.* **1974**, *7*, 72. (b) McIver, J. W., Jr. *J. Am. Chem. Soc.* **1972**, *94*, 4782.
- (21) (a) Osamura, Y.; Kato, S.; Morokuma, K.; Feller, D.; Davidson, E. R.; Borden, W. T. *J. Am. Chem. Soc.* **1984**, *106*, 3362. (b) Bernardi, F.; Bottoni, A.; Robb, M. A.; Field, M. J.; Hiller, I. H.; Guest, M. F. *J. Chem. Soc., Chem. Commun.* **1985**, 1051. (c) Houk, K. N.; Lin, Y. T.; Brown, F. K. *J. Am. Chem. Soc.* **1986**, *108*, 554. (d) Ortega, M.; Oliva, A.; Lluch, J. M.; Bertram, J. *Chem. Phys. Lett.* **1983**, *102*, 317. (e) Burke, L. A.; Leroy, G. G.; Sana, M. *Theor. Chim. Acta* **1975**, *40*, 313. (f) Burke, L. A.; Leroy, G. G. *Theor. Chim. Acta* **1977**, *44*, 219. (g) Brown, F. K.; Houk, K. N. *Tetrahedron Lett.* **1984**, 4609. (h) Townshend, R. E.; Ramunni, G.; Segal, G.; Hehre, W. J.; Salem, L. *J. Am. Chem. Soc.* **1976**, *98*, 2190. (i) Dewar, M. J. S.; Ford, G. P.; McKee, M. L.; Rzepa, H. S.; Wade, L. E. *J. Am. Chem. Soc.* **1977**, *99*, 5069.
- (22) Dewar, M. J. S.; Chantranupong, L. *J. Am. Chem. Soc.* **1983**, *105*, 7152.
- (23) (a) Pople, J. A.; Nesbet, R. K. *J. Chem. Phys.* **1954**, *21*, 571. (b) McKee, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 1900.
- (24) Wiberg, K. B.; Hadad, C. M.; Rablen, P. A.; Cioslowski, J. *J. Am. Chem. Soc.* **1992**, *114*, 8644.
- (25) Hammond, G. S. *J. Am. Chem. Soc.* **1955**, *77*, 334.
- (26) Colthurst, M. J.; Williams, A. *J. Chem. Soc., Perkin Trans. 2* **1997**, 1493.
- (27) Lopez, X.; Dejaegere, M.; Karplus, M. *J. Am. Chem. Soc.* **2001**, *123*, 11755.
- (28) Marcus, R. A. *J. Chem. Phys.* **1968**, *72*, 891.

- (29) Chen, M. Y.; Murdoch, J. R. *J. Am. Chem. Soc.* **1989**, *108*, 4735.
- (30) Lee, W. T.; Masel, R. I. *J. Phys. Chem. A* **1998**, *102*, 2332.
- (31) Miller, A. R. *J. Am. Chem. Soc.* **1978**, *100*, 1984.
- (32) Jencks, W. P. *Chem. Rev.* **1985**, *85*, 511.
- (33) (a) Pool, B. R.; White, J. M. *Org. Lett.* **2000**, *2*, 3505. (b) Wei, H.-X.; Zhou, C.; Ham, S.; White, J. M.; Birney, D. M. *Org. Lett.* **2004**, *6*, 4289. (c) Unruh, G. R.; Birney, D. M. *J. Am. Chem. Soc.* **2003**, *125*, 8529.
- (34) (a) Bürgi, H. B.; Dunitz, J. D.; Shefter, E. *J. Am. Chem. Soc.* **1973**, *95*, 5065. (b) Jones, P. G.; Kirby, A. J. *J. Chem. Soc., Chem. Commun.* **1979**, 288. (c) Bürgi, H. B.; Dunitz, J. D. *Acc. Chem. Res.* **1983**, *16*, 153.
- (35) Irie, T.; Tanida, H. *J. Org. Chem.* **1979**, *44*, 1002.
- (36) Herwig, P. T.; Mullen, K. *Adv. Mater.* **1999**, *11*, 480.
- (37) Afzali, A.; Dimitrakopoulos, C. D.; Breen, T. L. *J. Am. Chem. Soc.* **2002**, *124*, 8812.
- (38) Weidkamp, K. P.; Afzali, A.; Tromp, R. M.; Hamers, R. J. *J. Am. Chem. Soc.* **2004**, *126*, 12740.
- (39) Uno, H.; Yamashita, Y.; Kikuchi, M.; Watanabe, H.; Yamada, H.; Okujima, T.; Ogawa, T.; Ono, N. *Tetrahedron Lett.* **2005**, *46*, 1981.
- (40) Yamada, H.; Yamashita, Y.; Kikuchi, M.; Watanabe, H.; Okujima, T.; Uno, H.; Ogawa, T.; Ohara, K.; Ono, N. *Chem.—Eur. J.* **2005**, *11*, 6212.

CT0600130

Hybrid Density Functional Methods Empirically Optimized for the Computation of ^{13}C and ^1H Chemical Shifts in Chloroform Solution

Keith W. Wiitala, Thomas R. Hoye, and Christopher J. Cramer*

*Department of Chemistry and Supercomputer Institute, University of Minnesota,
207 Pleasant Street SE, Minneapolis, Minnesota 55455-0431*

Received March 20, 2006

Abstract: Two hybrid generalized-gradient approximation density functionals, WC04 and WP04, are optimized for the prediction of ^{13}C and ^1H chemical shifts, respectively, using a training set of 43 molecules in chloroform solution. Tests on molecules not included in the training set, namely six stereoisomeric methylcyclohexanols and a β -lactam antibiotic, indicate the models to be robust and moreover to provide results more accurate than those from equivalent B3LYP, PBE1, or *mPW1PW91* calculations, particularly for the prediction of downfield resonances in nuclear magnetic resonance spectra. However, linear regression of the B3LYP, PBE1, and *mPW1PW91* predicted values on the experimental data improves the accuracy of those models so that they are comparable to WC04 and WP04.

Introduction

Nuclear magnetic resonance (NMR) spectroscopy is a powerful technique for the determination of molecular structure.¹ Its utility in the pharmaceutical discovery process has been emphasized,^{2–4} and recent developments in the field include taking advantage of increasingly high-field magnets, novel pulse sequences, and the design of multidimensional spectroscopic experiments.

A fundamental observable in NMR spectroscopy is the nuclear chemical shift, δ . The chemical shift is sensitive to molecular environment, thereby providing insight into local functionality and stereochemistry. Many predictive models have been advanced in order to assist in the interpretation of experimental chemical shifts. The oldest such models are purely empirical and typically adopt a fragment substitution approach,⁵ although more modern variations are increasingly sophisticated in their ability to account for local fields and anisotropies.⁶ However, fragment models are limited in their ability to account for chemical shift differences associated with nonlinear interactions between multiple fragments or with stereoisomerism. Thus, there has been substantial interest in the use of quantum chemical models to predict

chemical shift values from first principles for use in spectral interpretation.^{7–12}

The theory associated with the computation of chemical shifts is well developed,¹³ and it has been demonstrated that highly correlated electronic structure methods using very large basis sets are capable of achieving high accuracy. However, such models are not computationally practical for large molecules or for databases containing a very large number of molecules. For molecules of moderate to large size, density functional theory (DFT) arguably provides the best combination of accuracy and efficiency among quantum chemical models, and the utility of a number of functionals for the prediction of chemical shift values has been evaluated.^{13–23} In general, modern functionals with large basis sets provide results of reasonable quantitative accuracy, but improvements are possible. For example, correcting the energy separation between occupied and virtual Kohn–Sham eigenvalues has been observed to provide improved chemical shift predictions.^{24–26} Model exchange-correlation potentials uncoupled from the self-consistent field procedure have also demonstrated good accuracy.^{27,28} Allen et al.²³ have recently compared results from some of these approaches with those from the KT2 functional,²⁹ which was developed specifically to be a generalized gradient approximation (GGA) functional useful for the direct prediction of chemical shifts.

* Corresponding author phone: (612)624-0859; fax: (612)626-2006; e-mail: cramer@chem.umn.edu.

From a more statistical standpoint, a number of authors have explored linear regression approaches to correct systematic errors associated with smaller basis sets and/or inaccurate functionals.^{30–34} In this work we consider an alternative somewhat along those lines. Many density functionals include one or more parameters whose values are chosen either on the basis of theoretical arguments associated with ideal model systems or by optimization against experimental data. In the latter instance, the data of choice have tended to be dominated by thermodynamic quantities (e.g., atomization energies), structural features (e.g., bond lengths), and in some instances molecular vibrational frequencies. In this paper, we present two new functionals optimized specifically to predict ¹³C and ¹H chemical shifts in chloroform solution, and we demonstrate that they have substantially improved accuracy compared to popular, current “off-the-shelf” functionals. Such optimized functionals should facilitate interpretation of NMR spectra of moderate to highly complex structures containing these two nuclei. Our approach is similar in spirit to prior work by Patchkovskii and Thiel,³⁵ who reparametrized the semiempirical modified neglect of differential overlap (MNDO) model to create a model named MB3, which is designed to give improved accuracy for ¹H, ¹³C, ¹⁵N, and ¹⁷O chemical shifts.

Computational Methods

A total of 160 conformers spanning the 43 molecules in the NMR training set (described below) were fully optimized at the B3LYP level^{36–39} using the 6-31G(d) basis set.⁴⁰ In addition to gas-phase geometries, geometries taking account of chloroform as solvent were optimized using the integral equation formalism⁴¹ of the polarized continuum model⁴² (IEFPCM). The molecular cavity for these calculations was constructed as a sum of atom-centered spheres using the radii of Bondi.⁴³

For each individual geometry, atomic chemical shielding tensors σ were computed¹⁷ using the gauge independent atomic orbital (GIAO) formalism^{44–46} and including the effects of chloroform solvation via the PCM model^{47,48} (this inclusion is at the level of the electronic structure irrespective of whether solvated geometries are employed). Isotropic atomic chemical shifts δ in units of ppm were computed as differences between atomic isotropic shieldings in solutes and corresponding reference atoms in tetramethylsilane (TMS). When more than a single conformer merits consideration, δ values are reported as an average over a Boltzmann-weighted population of conformers according to^{9,49}

$$\delta = \sum_i \left(\frac{\delta_i e^{-G_i/RT}}{\sum_j e^{-G_j/RT}} \right) \quad (1)$$

where i and j run over conformers, G is the free energy of the conformer in solution, R is the universal gas constant, and T is 298 K. The free energy in solution is taken as the sum of the electronic energy and solvation free energy computed at the B3LYP level using the 6-311+G(2d,p) basis set⁴⁰ and IEFPCM chloroform solvation free energies. This

same level of theory was used for the computation of the chemical shifts.

We define the energy E for a general hybrid exchange-correlation (xc) functional as

$$E_{xc}^{\text{B3LYP}} = P_2 E_x^{\text{HF}} + P_3 \Delta E_x^{\text{B}} + P_4 E_x^{\text{LSDA}} + P_5 \Delta E_c^{\text{LYP}} + P_6 E_c^{\text{LSDA}} \quad (2)$$

where P_2 – P_6 are weighting parameters ranging from 0 to 1, and the terms on the right-hand side correspond, respectively, to the Hartree–Fock (HF) exchange energy,⁴⁹ the Becke³⁶ (B) gradient correction to the local spin-density approximation (LSDA) exchange energy, and the Lee, Yang, and Parr³⁷ (LYP) correction to the local spin-density approximation (LSDA) correlation energy of Vosko, Wilk, and Nusair⁵⁰ (VWN). The popular B3LYP functional^{36–39} is defined by $P_2 = 0.20$, $P_3 = 0.72$, $P_4 = 0.80$, $P_5 = 0.81$, and $P_6 = 1.00$.

For the weighted carbon (WC04) and proton (WP04) functionals, the 5 weighting parameters were optimized using a central composite design.⁵¹ Separate optimizations were done for ¹³C chemical shifts and for ¹H chemical shifts of protons bonded to carbon (protons bonded to heteroatoms were not considered because of the extreme sensitivity of their chemical shifts to the purity of experimental NMR solvents, solute concentration, and variations in pH). The design response was the total absolute error over all chemical shifts in each data set defined as

$$|\Delta\delta| = \sum_i \sum_j |\delta_{\text{exp},ij} - \delta_{\text{calc},ij}| \quad (3)$$

where i runs over the 43 molecules in the training set and j runs over the number of carbon or hydrogen atoms in molecule i . Separate ¹³C- and ¹H NMR response surfaces were generated with the parameters permitted to range from 0.0001 to 0.9999. The numbers of unique ¹³C- and ¹H chemical shifts are 141 and 255, respectively.

The 43 molecules in the NMR training set (including Chemical Abstracts Service numbers) were as follows: acetaldehyde (75-07-0), acetamide (60-35-5), acetic acid (64-19-7), acetic anhydride (108-24-7), acetone (67-64-1), acetone oxime (127-06-0), acetonitrile (75-05-8), acetyl chloride (75-36-5), acrolein (107-02-8), 1-bromopropane (106-94-5), 3-buten-2-one (78-94-4), 1-chloropropane (540-54-5), cyclohexane (110-82-7), diacetamide (625-77-4), diethyl ether (60-29-7), dimethyl carbonate (616-38-6), 2,2-dimethyl-1,3-dioxolane (2916-31-6), dimethyl sulfate (77-78-1), dimethyl sulfide (75-18-3), dimethyl sulfite (616-42-2), dimethyl sulfone (67-71-0), dimethyl sulfoxide (67-68-5), 1,3-dimethylurea (96-31-1), ethanethiol (75-08-1), ethylbenzene (100-41-4), ethyl carbamate (51-79-6), ethyl isocyanate (109-90-0), furan (110-00-9), methanol (67-56-1), methyl acetate (79-20-9), methyl acrylate (96-33-3), methyl isothiocyanate (556-61-6), methyl thiocyanate (556-64-9), nitrobenzene (98-95-3), nitromethane (75-52-5), *N*-nitrosodimethylamine (62-75-9), *n*-propylamine (107-10-8), *n*-pentane (109-66-0), 1-pentene (109-67-1), 1-pentyne (627-19-0), phenol (108-95-2), 2-methyloxirane (75-56-9), and pyridine (110-86-1). Experimental reference data in deuteriochloroform solution were taken from the Spectral

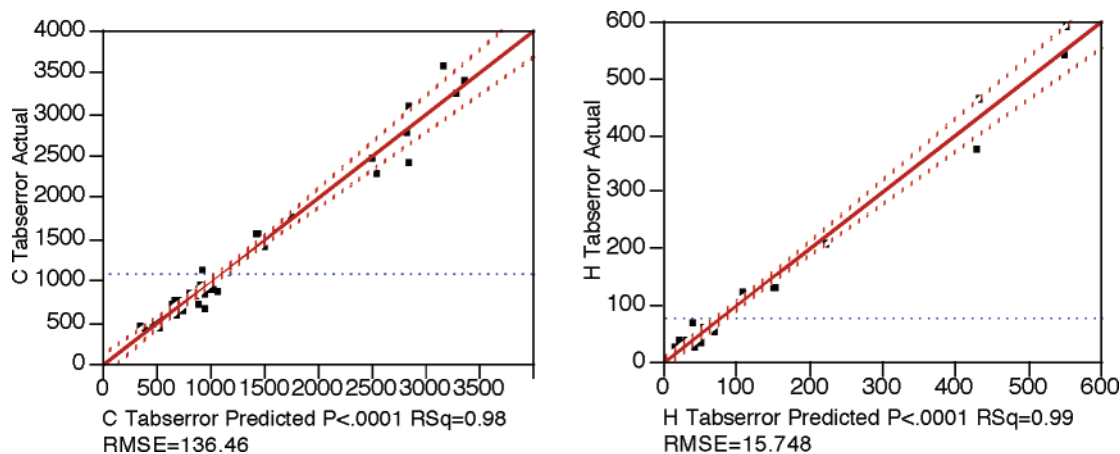


Figure 1. Response-surface predicted vs observed $|\Delta\delta|$ plots (ppm) with 99.99% confidence limits for ^{13}C (left) and ^1H (right) chemical shifts over the 59 parameter set choices.

Database for Organic Compounds, SDBS, organized by the National Institute of Advanced Industrial Science and Technology (AIST) of Japan (<http://www.aist.go.jp/RIODB/SDBS>).

Central composite design optimizations were performed using JMP 5.1.⁵² Electronic structure calculations were carried out using the Gaussian 03 suite of programs.⁵³ Parameter settings (eq 2) were enforced in Gaussian 03 by use of the keywords BLYP, IOp(3/76 = 10000nnnnn), which sets P_2 to nnnnn/10000, IOp(3/77 = mmmmmnnnnn), which sets P_3 to mmmmm/10000 and P_4 to nnnnn/10000, and IOp(3/78 = mmmmmnnnnn), which sets P_5 to mmmmm/10000 and P_6 to nnnnn/10000.

Results and Discussion

Parameter Optimization. An initial selection of 28 points on the parameter response surface was made by JMP 5.1 and corresponding ^{13}C and ^1H $|\Delta\delta|$ values were computed. Subsequently, 31 additional points were selected in order to improve the polynomial fit of the response surfaces. The 59 parameter sets and their associated errors are provided as Supporting Information.

Full second-order polynomial response surfaces (11 degrees of freedom) were fit to the ^{13}C and ^1H data. Terms, coefficients, and term t ratios for the fitted surfaces are provided as Supporting Information. The fits for the ^{13}C and ^1H surfaces provided Pearson correlation coefficient R^2 of 0.9753 and 0.9866 and F ratios of 169 and 256, respectively; these are reasonable levels of statistical significance. Observed vs predicted values for $|\Delta\delta|$ are plotted in Figure 1 with 99.99% confidence limits. Analysis of the term t ratios for the ^{13}C surface indicates only modest sensitivity to coefficients of the correlation functional (absolute t ratios of 1.44 and 0.97 for P_5 and P_6 , respectively). All other terms have absolute t ratios ranging from 3.3 to 31.7 with the exception of P_2 , which has an absolute t ratio of 1.74. Interestingly, the ^1H surface exhibits different sensitivity to the primary terms. In the case of ^1H , the least important terms are the gradient corrections to the exchange and correlation functionals (absolute t ratios of 0.06 and 0.48 for P_3 and P_5 , respectively). All other terms have absolute t ratios ranging from 2.3 to 36.0 with the exception of P_2 , which has an

Table 1. Functionals Optimized for Prediction of ^{13}C and ^1H Chemical Shifts

functional	P_2	P_3	P_4	P_5	P_6
WC04	0.7400	0.9999	0.0001	0.0001	0.9999
WP04	0.1189	0.9614	0.9999	0.0001	0.9999

Table 2. Mean (ME), Mean Unsigned (MUE), and Root-Mean Square (RMSE) Errors (ppm) in Predicted Absolute ^{13}C and ^1H Chemical Shifts^a

theory	^{13}C			^1H		
	ME	MUE	RMSE	ME	MUE	RMSE
WC04	0.7	3.1	3.8	0.06	0.13	0.20
WP04	6.4	6.4	7.6	0.01	0.09	0.13
HF	5.2	5.8	8.3	0.05	0.17	0.27
B3LYP	6.4	6.4	7.7	0.08	0.12	0.19
PBE1	5.5	5.5	6.9	0.07	0.13	0.22
<i>m</i> PW1PW91	5.6	5.6	7.0	0.07	0.13	0.21

^a All calculations used PCM(chloroform)/B3LYP/6-31G(d) geometries and chemical shifts were computed at the PCM(chloroform)/method/6-311+G(2d,p) level.

absolute t ratio of 1.63. The lower sensitivity of the ^1H surface to gradient corrections likely reflects a lack of variation in reduced density gradients at the hydrogen nucleus (which is only a single proton), at least in carbon-bound environments. The relatively modest sensitivity of the two surfaces to the percentage of HF exchange in the functional suggests that this term plays more of a role in affecting the bonding in interatomic regions than it does at the nucleus, since it is certainly well-known that inclusion of HF exchange in hybrid functionals dramatically improves bond energies, for example.⁴⁹

Based on the fitted surfaces, global minimum parameter values were identified (Table 1). The performance of the optimized WC04 and WP04 functionals for chemical shift prediction over the training set is compared to four other methods in Table 2. The other methods are Hartree–Fock theory (which is generally regarded as insufficiently accurate for first-principles calculations),⁴⁹ the one-parameter hybrid generalized-gradient approximation (GGA) functionals PBE1 and *m*PW1PW91, and the three-parameter hybrid GGA functional B3LYP.

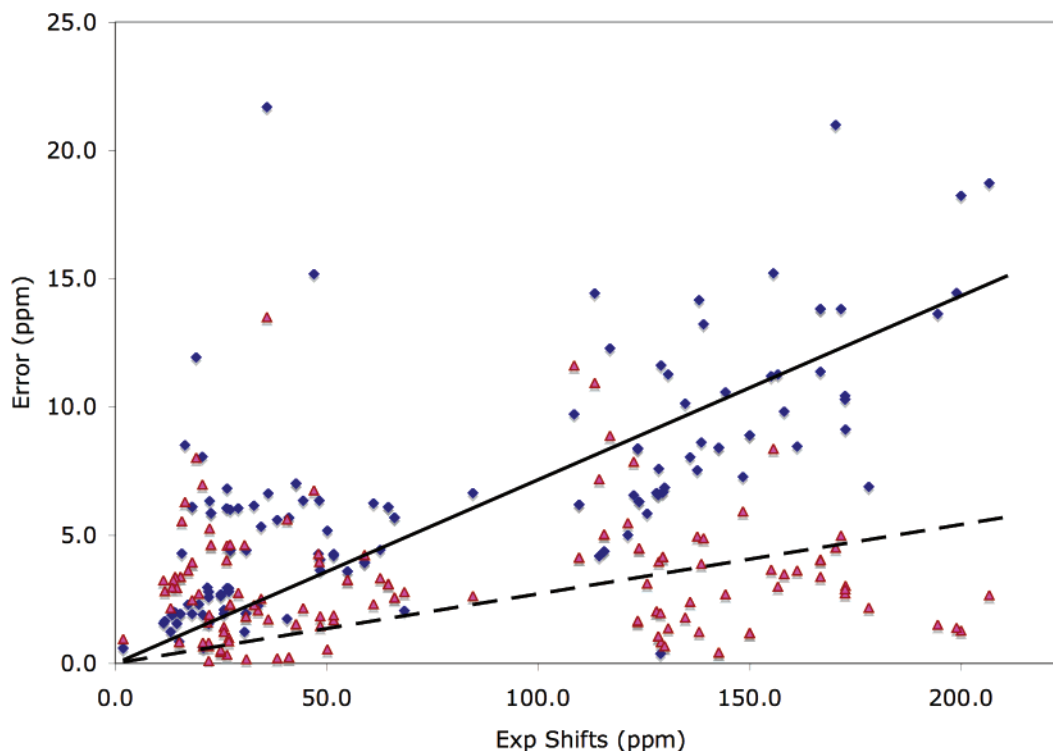


Figure 2. Absolute errors in 141 ^{13}C chemical shift predictions for B3LYP (blue diamonds) and WC04 (magenta triangles) as a function of chemical shift. The trendlines are linear fits to the errors forced to include the origin.

The optimized parameter values in Table 1 are quite different for WC04 and WP04, and both sets of parameters are themselves significantly different from the popular B3LYP functional. On the one hand, it is not particularly desirable to have to perform two separate calculations in order to obtain chemical shifts for the two different magnetically active nuclei. On the other hand, as Table 2 makes clear, the differences in parameter values lead to substantial differences in predictive accuracy. For ^{13}C chemical shifts, WC04 is more than twice as accurate as WP04. It is also substantially more accurate than either HF or any of the 3 hybrid GGA functionals. The nearest competitor is PBE1, which has an error that is 79% larger than WC04.

The variation in model accuracy over the ^1H training set data is smaller compared to the ^{13}C data set but still substantial. The WP04 method shows the highest accuracy, with the next nearest competitor, B3LYP, having an error that is 28% larger. The errors for WC04, PBE1, and *m*PW1PW91 are all within a few tenths of a ppm of one another in magnitude, while that for HF is substantially higher.

An analysis of errors as a function of the experimental chemical shift values indicates that the optimized functionals are more robust than B3LYP especially in the downfield regions of the spectrum (Figures 2 and 3). For ^{13}C data, WC04 remains more accurate in the upfield region by a statistically significant amount, but the magnitude is smaller. For ^1H data, WP04 and B3LYP are both accurate to within 0.1 ppm for most chemical shifts between 0 and 4 ppm.

Improvements from Linear Regression. Errors in chemical shift predictions from standard functionals have previously been shown to be reasonably systematic in various test sets,^{30–34} so that substantial improvements in accuracy may

Table 3. Mean (ME), Mean Unsigned (MUE), and Root-Mean Square (RMSE) Errors (ppm) in Predicted ^{13}C and ^1H Chemical Shifts Following Linear Regression^{a,b}

theory	^{13}C			^1H		
	ME	MUE	RMSE	ME	MUE	RMSE
WC04	0.0	3.0	3.8	0.00	0.11	0.14
WP04	0.0	2.3	3.5	0.00	0.07	0.10
HF	0.0	2.8	3.9	0.00	0.12	0.17
B3LYP	0.0	2.1	3.0	0.00	0.07	0.10
PBE1	0.1	1.8	2.8	0.00	0.08	0.11
<i>m</i> PW1PW91	0.0	1.8	2.8	0.00	0.08	0.11
regression data	<i>m</i>	<i>b</i>	<i>R</i>	<i>m</i>	<i>b</i>	<i>R</i>
WC04	1.0032	−0.9647	0.9958	0.9451	0.1157	0.9943
WP04	0.9601	−3.0273	0.9964	0.9587	0.1127	0.9969
HF	0.9164	1.7078	0.9955	0.9077	0.2318	0.9925
B3LYP	0.9488	−2.1134	0.9973	0.9333	0.1203	0.9974
PBE1	0.9486	−1.257	0.9977	0.9169	0.1895	0.9969
<i>m</i> PW1PW91	0.9487	−1.3423	0.9977	0.9191	0.1834	0.9977

^a All calculations used PCM(chloroform)/B3LYP/6-31G(d) geometries and chemical shifts were computed at the PCM(chloroform)/method/6-311+G(2d,p) level. ^b Regression data are slopes (*m*), intercepts (*b*), and Pearson correlation coefficients (*R*).

be obtained from linear regression of the predicted data on experimental data. We have examined this approach for the various functionals and our training set, and the results are summarized in Table 3. Not surprisingly, since they have been optimized by design, WC04 and WP04 have regression slopes and intercepts most near unity for their respective nuclei. However, the performance of the other approaches improves substantially following regression. While WP04 continues to have the best accuracy for ^1H (albeit by a very

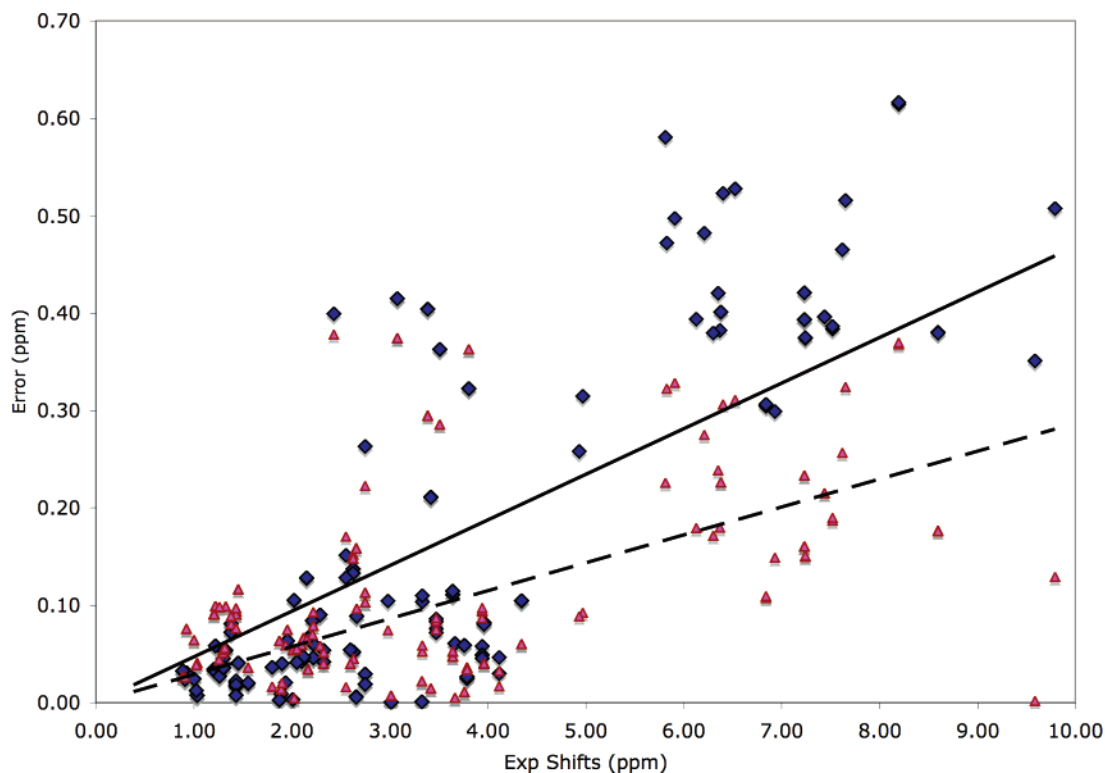


Figure 3. Absolute errors in 255 ^1H chemical shift predictions for B3LYP (blue diamonds) and WC04 (magenta triangles) as a function of chemical shift. The trendlines are linear fits to the errors forced to include the origin.

small margin), WC04 shows decreased accuracy compared to the other density functionals.

Applications to Molecules Not Included in the Training Set. To further assess the range of applicability of the WC04 and WP04 models, we applied them to the six isomeric *n*-methylcyclohexanols ($n = 2, 3,$ or 4) and also to a pharmaceutically relevant β -lactam. Results for the methylcyclohexanols are provided in Table 4, which also summarize predictions using B3LYP. In those stereoisomers where both ring substituents must be either simultaneously axial or equatorial in the chair conformer, only the latter was considered because of its much greater stability. In those stereoisomers where one substituent must be axial, the lower-energy conformer always had the hydroxyl group axial, consistent with its lower A ,⁵⁴ but the population of the other chair conformer was accounted for using a Boltzmann weighting. In every chair that was considered, a Boltzmann weighting over all hydroxyl rotamers was also applied.

For every isomer, the WC04 functional is two to three times more accurate than the B3LYP functional for ^{13}C chemical shifts prior to any correction through linear regression. After linear regression, the two models are about equally accurate. For ^1H chemical shifts, raw B3LYP is roughly twice as accurate as WP04, although both methods are accurate to within 0.1 ppm for these particular molecules. Except for the proton attached to the hydroxyl-substituted carbon, all of the resonances in the methylcyclohexanols are in the far upfield region where the performances of WP04 and B3LYP are generally similar. Linear regression in this case has fairly little effect on the B3LYP predictions but improves the WP04 predictions so that the two models are comparable in accuracy.

Table 4. Mean Unsigned Errors (MUEs, ppm) in Predicted $^{13}\text{C}^a$ and $^1\text{H}^b$ Chemical Shifts for Methylcyclohexanol Stereoisomers^c

compound	^{13}C		^1H	
	WC04	B3LYP	WP04	B3LYP
<i>cis</i> -2-methylcyclohexanol	2.10 ^d (1.65)	4.75 (1.21)	0.056 (0.032)	0.041 (0.050)
<i>trans</i> -2-methylcyclohexanol	2.03 (1.69)	4.82 (1.18)	0.072 (0.059)	0.037 (0.042)
<i>cis</i> -3-methylcyclohexanol	1.60 (1.29)	4.78 (1.14)	0.084 (0.067)	0.051 (0.052)
<i>trans</i> -3-methylcyclohexanol	1.73 (1.23)	4.96 (1.32)	0.080 (0.045)	0.033 (0.052)
<i>cis</i> -4-methylcyclohexanol	1.50 (0.87)	4.73 (1.15)	0.095 (0.055)	0.031 (0.026)
<i>trans</i> -4-methylcyclohexanol	1.57 (1.08)	4.77 (1.04)	0.081 (0.054)	0.041 (0.033)
average MUE	1.76 (1.30)	4.80 (1.17)	0.078 (0.052)	0.039 (0.042)

^a MUEs are $|\Delta\delta|$ as defined in eq 3 divided by 7. ^b MUEs are $|\Delta\delta|$ as defined in eq 3 divided by 13; the hydroxyl proton is not included. ^c All calculations used PCM(chloroform)/B3LYP/6-31G(d) geometries and chemical shifts were computed at the PCM(chloroform)/method/6-311+G(2d,p) level. ^d Raw errors are listed above, and errors after linear regression (using regression equations derived from training set data) are listed below in parentheses.

To more comprehensively examine the utility of the WP04 functional, we next consider the pharmaceutically relevant, heteroatom-rich molecule (+)-(2*S*,5*R*,6*R*)-3,3-dimethyl-7-oxo-6-phthalimido-4-thia-1-azabicyclo(3.2.0)heptane-2-carboxylic acid (**1**, Figure 4). We averaged chemical shifts for **1** over a family of 18 conformers according to eq 1 and compared them to experimental ^{13}C - and ^1H NMR data in

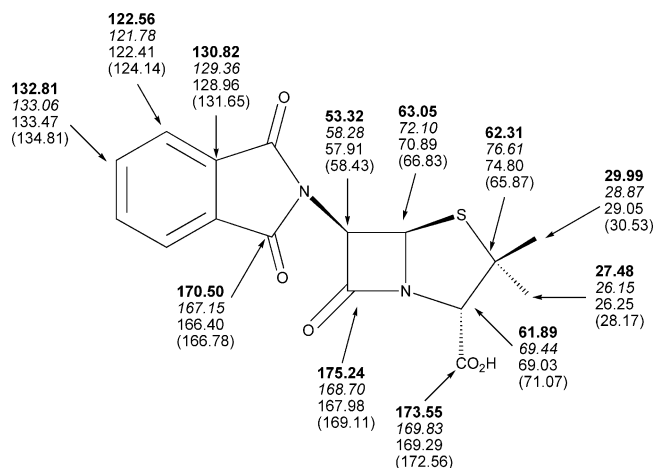


Figure 4. Predicted (after linear regression using regression equations developed on training set data for B3LYP and PBE1) and observed carbon chemical shifts for **1** in chloroform (from top to bottom: WC04 bold; B3LYP italic; PBE1 roman; experiment in parentheses).

Table 5. Mean and Mean Unsigned Errors (ppm) in Predicted Chemical Shifts for **1**^a

level of theory	¹³ C		¹ H	
	ME	MUE	ME	MUE
WC04	-1.1 ^b	2.9	-0.02	0.23
	(-1.7)	(3.3)	(-0.15)	(0.16)
WP04	8.0	8.0	0.01	0.10
	(0.3)	(3.1)	(-0.06)	(0.06)
HF	5.9	7.5	0.10	0.30
	(-4.7)	(4.8)	(-0.09)	(0.14)
B3LYP	7.9	7.9	0.11	0.15
	(-0.3)	(2.4)	(-0.07)	(0.07)
PBE1	6.7	6.7	0.14	0.19
	(-0.6)	(2.2)	(-0.05)	(0.05)
mPW1PW91	6.8	6.8	0.13	0.18
	(-0.6)	(2.2)	(-0.05)	(0.06)

^a MUEs are $|\Delta\delta|$ as defined in eq 3 divided by the number of relevant nuclei. All calculations used PCM(chloroform)/B3LYP/6-31G(d) geometries and chemical shifts were computed at the PCM(chloroform)/method/6-311+G(2d,p) level. ^b Raw errors are listed above, and errors after linear regression (using regression equations derived from training set data) are listed below in parentheses.

CDCl₃.⁵⁵ The performances of the WC04, WP04, B3LYP, PBE1, and mPW1PW91 functionals are summarized in Table 5, as are results from HF theory. Specific ¹³C chemical shifts are also provided in Figure 4 from experiment, WC04, B3LYP, and PBE1. This molecule, substantially decorated with polar functionality (giving rise to more downfield chemical shifts), demonstrates the superior performances of the WC04 and WP04 functionals in terms of mean and mean unsigned errors prior to linear regression. After linear regression, however, the PBE1 functional is overall the most accurate for both nuclei, and it is the only one of those listed in Figure 4 that correctly predicts the relative positions of the ¹³C chemical shifts for the carboxyl carbons of the carboxylic acid and the β -lactam.

Solvation Effects. As a point of technical as well as practical interest, we examined the degree to which chloroform solvation, as implemented via the PCM continuum

Table 6. Mean Unsigned Errors (MUE, ppm) in Predicted WC04 ¹³C and WP04 ¹H Chemical Shifts as a Function of Computational Protocol^a

model	¹³ C	¹ H
gas//gas	3.3	0.14
gas//chloroform	3.2	0.14
chloroform//chloroform	3.1	0.10

^a Errors $|\Delta\delta|$ are defined in eq 3. Calculations used either B3LYP(gas)/6-31G(d) or PCM(chloroform)/B3LYP/6-31G(d) geometries (indicated after the double solidus), and chemical shifts were computed at either the Wx04/6-311+G(2d,p) or PCM(chloroform)/Wx04/6-311+G(2d,p) levels (indicated before the double solidus).

solvation model, influenced the predicted chemical shifts. In particular, for the training set we computed WC04 and WP04 chemical shifts for gas-phase densities at gas-phase geometries, for gas-phase densities at solvated geometries, and for solvated densities at solvated geometries (as already discussed above). The corresponding $|\Delta\delta|$ values are provided in Table 6.

In the case of the ¹³C data set, there is a reduction in error of about 4% when the geometries are relaxed in solution and an additional 4% when solvated densities are used for the NMR calculations. In the ¹H data set, on the other hand, the use of gas-phase densities with relaxed geometries actually increases the total error by a small amount. However, the use of solvated densities leads to a substantial improvement in the predictive accuracy.

In principle, there might be some value in optimizing parameter sets designed to predict chemical shifts in solution from gas-phase densities at gas-phase geometries.³⁴ Gas-phase calculations are efficient and, as long as solvation effects are systematic, the statistical approach might be expected to absorb deviations into the parameter set. However, the cost of including a continuum solvent model into a self-consistent reaction field model is typically no more than 15% or so of the total computational time,^{56,57} so we consider it worthwhile to adopt this approach in order to more accurately capture the physics of solvation when the goal is to predict data for solutes in solution.

Klein et al.⁵⁸ have pointed out that a continuum model alone is generally *not* sufficient for the computation of the ¹⁷O chemical shift of liquid water because of the strong, nonisotropic interactions between water oxygen atoms and the protons of neighboring water molecules. In that case, explicit supermolecular clusters surrounded by a continuum are required to accurately model the polarization of the system. However, interactions of this magnitude are unlikely to be associated with carbon atoms and nonheteroatom bound hydrogen atoms, so a pure continuum approach to account for solvation is within the spirit of the WC04 and WP04 models, which are designed to balance accuracy and efficiency for the interpretation of NMR spectra.

Acknowledgment. This work was supported in part by the National Science Foundation (C.J.C., CHE-0203346). We thank Dr. Vadim Dvornikov and Ziyad Al-Rashid for providing ¹³C- and ¹H NMR chemical shift data for the methylcyclohexanols and discussions. Loren Greenman,

Corey Stotts, and Dr. Jason Thompson are thanked for programming assistance.

Supporting Information Available: Initial and subsequent design worksheets and computed responses, terms, and coefficients for the ^{13}C and ^1H response surfaces, Cartesian coordinates and electronic energies for all conformers of all molecules, and experimental chemical shift data for the methylcyclohexanols. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Claridge, T. *High-resolution NMR Techniques in Organic Chemistry*; Elsevier: London, 1999.
- (2) Lipinski, C. A.; Lombardo, F.; Domin, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.
- (3) Uccello-Barretta, G.; Balzano, F.; Sicoli, G.; Friglola, C.; Aldana, I.; Monge, A.; Paolino, D.; Guccione, S. *Bioorg. Med. Chem.* **2004**, *12*, 447–458.
- (4) Lipinski, C. A. *Pharmacol. Toxicol. Methods* **2001**, *44*, 235–249.
- (5) Pretsch, E.; Bühlmann, P.; Affolter, C. *Structure Determination of Organic Compounds: Tables of Spectral Data*; Springer: Berlin, 2003.
- (6) Abraham, R. J.; Byrne, J. J.; Griffiths, L.; Konioutou, R. *Magn. Reson. Chem.* **2005**, *43*, 611–624.
- (7) Klemp, C.; Bruns, M.; Gauss, J.; Haussermann, U.; Stosser, G.; van Wullen, L.; Jansen, M.; Schnockel, H. *J. Am. Chem. Soc.* **2001**, *123*, 9099–9106.
- (8) Ochsenfeld, C.; Brown, S. P.; Schnell, I.; Gauss, J.; Spiess, H. W. *J. Am. Chem. Soc.* **2001**, *123*, 2597–2606.
- (9) Barone, G.; Duca, D.; Silvestri, A.; Gomez-Paloma, L.; Riccio, R.; Bifulco, G. *Chem.--Eur. J.* **2002**, *8*, 3240–3245.
- (10) Barone, G.; Gomez-Paloma, L.; Duca, D.; Silvestri, A.; Riccio, R.; Bifulco, G. *Chem.--Eur. J.* **2002**, *8*, 3233–3239.
- (11) Price, D. R.; Stanton, J. F. *Org. Lett.* **2002**, *4*, 2809–2811.
- (12) Balandina, A.; Mamedov, V.; Franck, X.; Figadère, B.; Latypov, S. *Tetrahedron Lett.* **2004**, *45*, 4003–4007.
- (13) *Calculation of NMR and EPR Parameters: Theory and Applications*; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; John Wiley & Sons: New York, 2004.
- (14) Wiberg, K. B. *J. Comput. Chem.* **1999**, *20*, 1299–1303.
- (15) Wiberg, K. B.; Hammer, J. D.; Zilm, K. W.; Keith, T. A.; Cheeseman, J. R.; Duchamp, J. C. *J. Org. Chem.* **2004**, *69*, 1086–1096.
- (16) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1998**, *298*, 113–119.
- (17) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J. *J. Chem. Phys.* **1996**, *104*, 5497–5509.
- (18) Malkin, V. G.; Malkina, O. L.; Eriksson, L. A.; Salahub, D. R. In *Modern Density Functional Theory: A Tool for Chemistry*; Politzer, P., Seminario, J., Eds.; Elsevier: Amsterdam, 1995; Vol. 2, p 273–374.
- (19) Bühl, M.; Kaupp, M.; Malkina, O. L.; Malkin, V. G. *J. Comput. Chem.* **1999**, *20*, 91–105.
- (20) Schreckenbach, G.; Ziegler, T. *Theor. Chem. Acc.* **1998**, *99*, 71–82.
- (21) Adamo, C.; Cossi, M.; Barone, V. *J. Mol. Struct. (THEO-CHEM)* **1999**, *493*, 145–147.
- (22) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. *J. Chem. Phys.* **2001**, *115*, 9233–9241.
- (23) Allen, M. J.; Keal, T. W.; Tozer, D. J. *Chem. Phys. Lett.* **2003**, *380*, 70–77.
- (24) Malkin, V. G.; Malkina, O. L.; Casida, M. E.; Salahub, D. R. *J. Am. Chem. Soc.* **1994**, *116*, 5898–5908.
- (25) Fadda, E.; Casida, M. E.; Salahub, D. R. *Int. J. Quantum Chem.* **2003**, *91*, 67–83.
- (26) Wilson, P. J.; Amos, R. D.; Handy, N. C. *Chem. Phys. Lett.* **1999**, *312*, 475–484.
- (27) Patchkovskii, S.; Autschbach, J.; Ziegler, T. *J. Chem. Phys.* **2001**, *115*, 26–42.
- (28) Poater, J.; van Lenthe, E.; Baerends, E. J. *J. Chem. Phys.* **2003**, *118*, 8584–8593.
- (29) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2003**, *119*, 3015–3024.
- (30) Rablen, P. R.; Pearlman, S. A.; Finkbiner, J. *J. Phys. Chem. A* **1999**, *103*, 7357–7363.
- (31) Sebag, A. B.; Forsyth, D. A.; Plante, M. A. *J. Org. Chem.* **2001**, *66*, 7967–7973.
- (32) Giesen, D. J.; Zumbulyadis, N. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5498–5507.
- (33) Wang, B.; Fleischer, U.; Hinton, J. F.; Pulay, P. *J. Comput. Chem.* **2001**, *22*, 1887–1895.
- (34) Wang, B.; Hinton, J. F.; Pulay, P. *J. Comput. Chem.* **2002**, *23*, 492–497.
- (35) Patchkovskii, S.; Thiel, W. *J. Comput. Chem.* **1999**, *20*, 1220–1245.
- (36) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (37) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (38) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (39) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- (40) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.
- (41) Cancès, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032–3041.
- (42) Miertus, S.; Scrocco, E.; Tomasi, J. *J. Chem. Phys.* **1981**, *55*, 117–129.
- (43) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–451.
- (44) London, F. *J. Phys. Radium (Paris)* **1937**, *8*, 397–409.
- (45) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251–8260.
- (46) Gauss, J. *J. Chem. Phys.* **1993**, *99*, 3629–3643.
- (47) Cammi, R. *J. Chem. Phys.* **1998**, *109*, 3185–3196.
- (48) Cammi, R.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1999**, *110*, 7627–7638.
- (49) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*, 2nd ed.; John Wiley & Sons: Chichester, 2004.
- (50) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.

- (51) Khuri, A. I. *Response Surfaces: Designs and Analyses*; Marcel Dekker: New York, 1996.
- (52) *JMP version 5.1*; SAS Institute Inc.: Cary, NC, 2005.
- (53) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03 (Revision B.05)*; Gaussian, Inc.: Pittsburgh, PA, 2003.
- (54) Eliel, E. L.; Wilen, S. H. *Stereochemistry of Organic Compounds*; John Wiley and Sons: New York, 1994.
- (55) Fekner, T.; Baldwin, J. E.; Adlington, R. M.; Jones, T. W.; Prout, C. K.; Schofield, C. J. *Tetrahedron* **2000**, *56*, 6053–6074.
- (56) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.
- (57) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3093.
- (58) Klein, R. A.; Mennucci, B.; Tomasi, J. *J. Phys. Chem. A* **2004**, *108*, 5851–5863.

CT6001016

JCTC

Journal of Chemical Theory and Computation

Adsorption of Benzene on Copper, Silver, and Gold Surfaces

Ante Bilić,[†] Jeffrey R. Reimers,^{*,†} Noel S. Hush,^{†,‡} Rainer C. Hoft,[§] and Michael J. Ford[§]

School of Chemistry and School of Molecular and Microbial Biosciences, The University of Sydney, New South Wales 2006, Australia, and Institute for Nanoscale Technology, University of Technology, Sydney, P.O. Box 123, Broadway, New South Wales 2007, Australia

Received September 21, 2005

Abstract: The adsorption of benzene on the Cu(111), Ag(111), Au(111), and Cu(110) surfaces at low coverage is modeled using density-functional theory (DFT) using periodic-slab models of the surfaces as well as using both DFT and complete-active-space self-consistent field theory with second-order Møller–Plesset perturbation corrections (CASPT2) for the interaction of benzene with a Cu₁₃ cluster model for the Cu(110) surface. For the binding to the (111) surfaces, key qualitative features of the results such as weak physisorption, the relative orientation of the adsorbate on the surface, and surface potential changes are in good agreement with experimental findings. Also, the binding to Cu(110) is predicted to be much stronger than that to Cu(111) and much weaker than that seen in previous calculations for Ni(110), as observed. However, a range of physisorptive-like and chemisorptive-like structures are found for benzene on Cu(110) that are roughly consistent with observed spectroscopic data, with these structures differing dramatically in geometry but trivially in energy. For all systems, the bonding is found to be purely dispersive in nature with minimal covalent character. As dispersive energies are reproduced very poorly by DFT, the calculated binding energies are found to dramatically underestimate the observed ones, while CASPT2 calculations indicate that there is no binding at the Hartree–Fock level and demonstrate that the expected intermolecular correlation (dispersive) energy is of the correct order to explain the experimental binding-energy data. DFT calculations performed for benzene on Cu(110) and for benzene on the model cluster indicate that this cluster is actually too reactive and provides a poor chemical model for the system.

1. Introduction

The nature of the interaction of benzene with metal surfaces is of interest in various fields of applied research such as corrosion protection, lubrication, and dye adhesion as these all involve interfaces between organic matter and metals. This problem has also attracted attention in the area of hetero-

geneous catalysis owing to the role of metals as catalysts in ring-cracking reactions.¹ Being the smallest aromatic molecule, benzene has frequently been employed as a model system for larger hydrocarbons. Recently, new interest has arisen in the interaction of aromatic compounds with metals because of their potential application in the design of devices based on electroactive organic molecules.² For this, the prototype system is a two terminal device formed by two gold electrodes spanned by a single chemisorbed 1,4-benzenedithiol molecule.³ Thiols are known to form strong bonds of order 30 kcal mol⁻¹ to gold^{4–9} and to provide weaker and possibly more flexible means of attachment, the

* Corresponding author e-mail: reimers@chem.usyd.edu.au.

[†] School of Chemistry, The University of Sydney.

[‡] School of Molecular and Microbial Biosciences, The University of Sydney.

[§] University of Technology, Sydney.

binding of aromatic azines to Au(111) such as pyridine and 1,10-phenanthroline have also been investigated.^{10–12} Always, the conformation of the molecule at the surface is critical to function; this involves both site geometry and internal rearrangements of the metal and adsorbate. Given the progress in theoretical chemistry combined with increasing computer power, it can be expected that computational methods will be able to reveal significant detailed information concerning these processes. A prerequisite for this, however, is to establish that computational methods give reliable predictions for each of the properties of interest not only structural properties but also thermodynamic, spectroscopic, and process-related ones.

In previous computational studies, we have investigated the adsorption of pyridine¹² and phenylthiol⁹ on Au(111). Both of these adsorbates have end groups that anchor to Au(111), producing strong binding in the case of the thiol and medium-strength binding in the case of the azine. However, both adsorbates are predicted to bind over a wide range of orientational angles to the surface. For pyridine, the vertical orientation involving interactions between the nitrogen donor and the surface is predicted to be the most stable one, but flat structures dominated by π -stacking are found to be only 5 kcal mol⁻¹ higher in energy. For phenylthiol, little preference for either sp² or sp³ hybridization of the sulfur is predicted, with low-energy configurations occurring for both vertically oriented and near-flat adsorbates. For both chemical systems, the nature of the intrinsic interaction between an aromatic π system and the surface is thus quite important, and to elucidate this more directly we study herein the adsorption of benzene on Au(111). However, as experimental studies of this system are rare, we consider also the related systems of benzene on Ag(111), Cu(111), and Cu(110) for which more information is available to characterize the effectiveness of the available computational procedures.

In an early work by Somorjai^{1,13} it was deduced that benzene does not adsorb on either clean or stepped Au(111), whereas naphthalene interacts strongly with both.¹³ However, Wöll¹⁴ recently studied monolayers of several hydrocarbons adsorbed on various metal surfaces using X-ray absorption spectroscopy and found that benzene does indeed weakly physisorb on Au(111). The spectra indicate a high degree of molecular orientation and preserved adsorbate planarity. For benzene on Ag(111), a (3 × 3) ordered superstructure has been reported.¹⁵ Given the similarity in both atomic and electronic structure between gold and silver crystals, qualitative comparison of our findings for C₆H₆/Au(111) with those on Ag(111) can be made. In general, benzene adsorption on coinage metals takes place only below^{15,16} 280 K, indicative of the relatively weak binding. Extensive experimental data are available for the adsorption of benzene on Cu(111) and Cu(110), with the observed desorption temperature ranges being^{17,18} ~225 K for Cu(111) and ~280 K for the more open Cu(110) surface.¹⁹ Several computational studies^{20–23} have also considered C₆H₆/Cu(110), with variable degrees of success.

Here we report results from calculations on the adsorption of benzene on the (111) surface of Cu, Ag, and Au as well as on the Cu(110) surface. The study is carried out initially

using density-functional theory (DFT), employing both atomic slab and cluster representations of the metal substrate. As significant computational problems arise for interactions such as this involving very shallow potential-energy surfaces supporting very different structures with similar binding energies, a range of computational methods is investigated. In addition, we also perform second-order Møller–Plesset (MP2) perturbation-theory calculations²⁴ based on Hartree–Fock self-consistent field (SCF) wave functions as well as multireference complete active space (CASSCF) perturbation-theory (CASPT2) calculations.²⁵ The results predict that the most significant contribution to the binding comes from the dispersive interaction, an interaction which at present is poorly and inconsistently accounted for by the exchange–correlation functionals used in modern applications of DFT.

2. Methods

DFT computations were carried out using the packages VASP,^{26,27} CASTEP,²⁸ and SIESTA.^{29,30} In the VASP and CASTEP calculations, plane-wave basis sets are employed to expand the electronic wave functions. Electron–ion interactions are accounted for through the use of ultrasoft pseudopotentials,^{31,32} allowing for the use of a low-energy cutoff for the plane-wave basis set. For electron–electron exchange and correlation interactions the functional of Perdew and Wang (PW91),³³ a form of the generalized gradient approximation (GGA), was used in both the VASP and CASTEP calculations with an energy cutoff of the basis set set at 290 eV, as dictated by the pseudopotential for carbon. CASTEP computations were also performed using the GGA functional of Perdew, Burke, and Ernzerhof (PBE)³⁴ with the energy cutoff set to 400 eV. In the SIESTA calculations, norm-conserving pseudopotentials were used, generated according to the scheme of Troullier and Martins,³⁵ with relativistic corrections added for the Cu atoms. The atomic basis set for the valence electron wave function expansion was of double- ζ plus polarization quality. These atomic orbitals have finite range with an excitation energy of 5 mRy arising due to the confinement. Only the PBE functional was used in the SIESTA computations, while the effects of basis-set superposition error (BSSE) associated with the atomic-orbital basis set were examined using the counterpoise method.³⁶

The surfaces of Cu, Ag, and Au were modeled by supercells consisting of several atomic layers and vacuum. The application of periodic boundary conditions in all three Cartesian directions yields an infinite array of periodically repeated slabs separated by regions of vacuum. A single molecule was placed in the vacuum region on the upper side of the slab. Calculations pertinent to gas-phase molecules employed a cell of the same size as the supercell of the complex, an integration using the Γ -point only, and Gaussian smearing. For the VASP calculations, the dipole moment arising from the asymmetric slab was compensated for by the introduction of a dipole sheet of the same strength and opposite direction in the middle of the vacuum.³⁷ This correction can be essential for systems involving strongly dipolar or polarizing adsorbates but has minimal affect for physisorbed benzene.

Only VASP calculations were performed for adsorbates on (111) surfaces. For these, the slabs were four atomic layers thick, while the vacuum was ten. For Au, the interlayer spacing was taken from the previously evaluated³⁸ value of the bulk lattice parameter, 4.20 Å, while for Ag and Cu the corresponding values were 4.170 and 3.655 Å, respectively. The calculations employed (3 × 3) superstructure resulting in nine metal atoms per layer. This represents a 1/9 monolayer (ML) coverage, sufficiently low that the molecules in adjacent cells are well separated. Brillouin-zone integrations were performed using the 3 × 3 × 1 *k*-point Monkhorst-Pack grid, with a Methfessel–Paxton smearing³⁹ of 0.2 eV. In all computations involving the (111) slabs, the top layer and adsorbed species were allowed to relax, with other layers frozen so as to simulate a semi-infinite solid.

To keep the distance between adsorbates in neighboring cells on the Cu(110) lattice close to that for the (111) surface, a (2 × 3) surface supercell of the original unit cell was used. This corresponds to a 1/6 ML benzene coverage. For the VASP calculations on the Cu(110) surface, the slab was six atomic layers thick, while the vacuum was 15 (20 Å), with the lattice parameter of 3.655 Å set to match the appropriate calculated value for bulk copper. The top three layers of Cu(110) and adsorbed species were allowed to relax, with other layers fixed in their bulk positions. Brillouin-zone integrations were performed using the 4 × 3 × 1 *k*-point Monkhorst-Pack grid, with a Methfessel–Paxton smearing³⁹ of 0.2 eV. In the CASTEP and SIESTA computations, the Cu(110) slab was four atomic layers thick with a vacuum region of 20 Å. Test SIESTA calculations indicate a maximum variation in binding energy of 0.35 kcal mol⁻¹ on expansion through to 7 layers. Lattice parameters of 3.636 (CASTEP) and 3.680 (SIESTA) Å for Cu were used consistent with the optimized values for the bulk material obtained using the PBE density functional with the appropriate basis sets. Also, 5 × 5 × 1 and 3 × 3 × 1 Monkhorst-Pack meshes were employed in the SIESTA and CASTEP calculations, respectively. The Cu slabs were fixed at their bulk geometry, however, as test calculations for the Cu(110) slab where the top two layers were allowed to relax, showed that this had negligible effect on binding energies.

The cluster model of Triguero et al.,²¹ sketched elsewhere,²² was also used to study C₆H₆ on Cu(110). It comprises a Cu₁₃ cluster with two atomic layers containing four atoms in the first layer that are bonded to a benzene molecule. The cluster has overall C_{2v} symmetry; strong chemisorptive-type interactions distort the benzene ring, however, giving it an inverted boat shape akin to the quinonoid form of the lowest excited triplet state⁴⁰ of the benzene in the gas phase.²¹

The CASPT2 calculations²⁵ were performed using the MOLCAS package.⁴¹ The Stuttgart basis set ECP10MWB⁴² with its 1s+2s+2p effective core potential was used for Cu in conjunction with the 6-31+G* basis set⁴³ for C and H. The active space was chosen in a way that would comprise all 13 Cu 4s electrons distributed through all 13 Cu 4s orbitals. The chosen orbitals were 5a₁, 2a₂, 3b₁, and 3b₂, while the doubly occupied orbitals were 41a₁, 28a₂, 34b₁, and 35b₂. Orbital rotations distorted this picture, however, with some

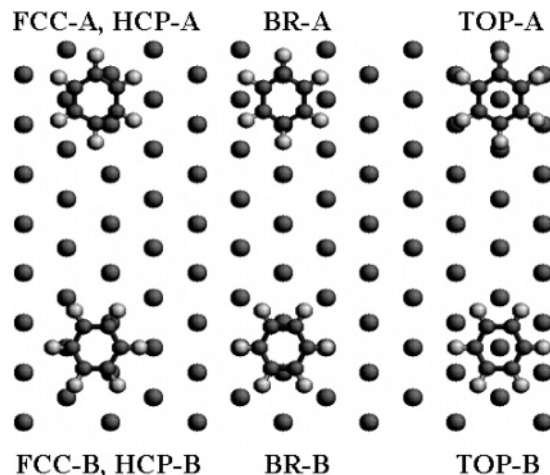


Figure 1. Starting geometries for benzene adsorbed on the (111) faces of Cu, Ag, and Au.

benzene occupied orbitals, benzene virtual orbitals, copper 3d occupied orbitals, and copper 4p virtual orbitals being swapped into the active space instead of some of the copper 4s orbitals. One Cu 3p orbital was also occasionally rotated into the active space. To eliminate the effects of this rotation, the complete-active-space self-consistent-field (CASSCF) part of the CASPT2 calculations was also performed using a frozen core consisting of the C 1s and Cu 3s + 3p orbitals. The quantitative effects of this restriction were insubstantial, however, and the results are not presented. In both cases the Møller–Plesset perturbation aspect of the CASPT2 calculations was performed using frozen C 1s and Cu 3s + 3p orbitals. In addition to the CASPT2 calculations, second-order Møller–Plesset perturbation (MP2)²⁴ calculations were also performed based on a two-determinant restricted open-shell Hartree–Fock (ROHF) wave function using the GAUSSIAN03 program package⁴⁴ with the same basis sets and frozen orbitals. All binding energies were corrected for basis-set superposition error using the counterpoise method.³⁶ Also, some constrained optimizations of the geometry of the adsorbate above the Cu₁₃ cluster were performed by GAUSSIAN03 using the PW91 density functional³³ with the ECP10MWB⁴² and 6-31G* basis sets but without use of BSSE correction.

3. Results and Discussion

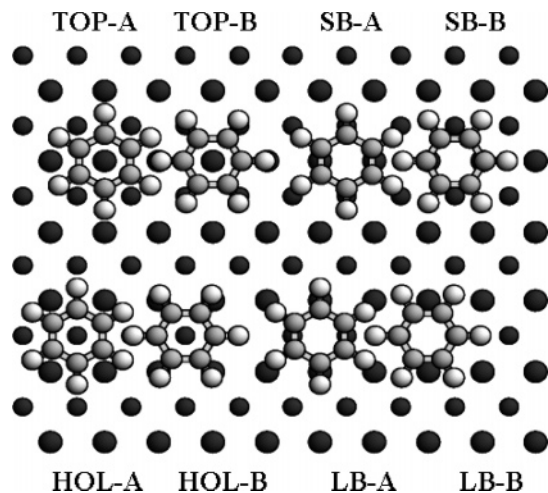
3.1. VASP PW91 Calculations of Adsorption in the Low-Coverage Limit. The adsorption of benzene on the Cu(111), Ag(111), and Au(111) surfaces is considered for flat-lying orientations in which the center of the ring is classified as being either above TOP, bridge (BR), or FCC/HCP 3-fold hollow sites on the surface. Six high-symmetry binding configurations are illustrated in Figure 1 for which the corresponding adsorption energy changes ΔE , evaluated using VASP, are listed in Table 1. These comprise two orientations each, named A and B, for binding at the four sites. All optimized coordinates are provided in the Supporting Information.

For benzene on Cu(111), the two TOP sites are calculated to provide no binding at all, while the BR, FCC, and HCP sites support only very weak binding of $\Delta E \sim -0.5$ kcal

Table 1: Adsorption Energy Changes ΔE as Calculated by VASP for the C_6H_6 -(3 × 3)-M(111) System^a

M	TOP-A	TOP-B	BR-A	BR-B	FCC-A	FCC-B	HCP-A	HCP-B
Cu	0.22	1.02	-0.48	-0.36	-0.39	-0.54	-0.42	-0.60
Ag	0.00	-0.02	-1.02	-0.99	-1.23	-1.09	-1.13	-1.13
Au	-0.50	-0.60	-1.75	-1.32	-1.86	-1.51	-1.92	-1.63

^a M = Cu, Ag, and Au, given in kcal mol⁻¹, for respective molecule conformations. Positive values indicate endothermic reactions and are obtained as the calculations terminate simply when the forces generated are smaller than a preselected limit and the potential-energy surfaces are very flat.

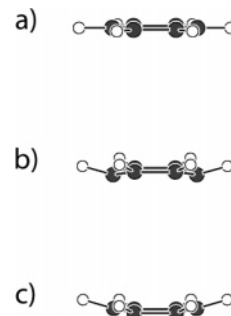
**Figure 2.** Initial geometries for benzene adsorbed on the Cu(110) surfaces.

mol⁻¹. This result is in stark contrast with the observed binding of $\Delta E = -14$ kcal mol⁻¹ (-0.6 eV).^{17,18}

For the case of benzene on Ag(111), the calculated interaction energies for the 8 structures are all more attractive by 0.1–1.0 kcal mol⁻¹ than the corresponding values for benzene on Cu(111). No experimental value for the binding energy is available, but chemical arguments⁴⁵ suggest that instead it should *not* be as strongly bound. However, the strongest interaction of $\Delta E \sim -1.2$ kcal mol⁻¹ on Ag(111) is predicted for the FCC hollow site in the orientation “A” in a pattern that actually corresponds to the experimentally observed (3 × 3) superstructure found after an exposure to 5 L of benzene on Ag(111).¹⁵ An early molecular-orbital calculation also suggested that a 3-fold hollow site is most favored for benzene adsorption on a silver cluster.⁴⁶ The most significant conclusions to be drawn from the calculations, however, is that benzene is predicted to wander freely across the surface with little barrier, even at low temperatures.

The VASP calculations predict also that for each possible binding site the interaction energy is 0.6–0.8 kcal mol⁻¹ more favorable for binding to Au(111) than to Ag(111). Binding to the HCP hollow site A is only 0.1 kcal mol⁻¹ more favorable than the FCC hollow site, however, again indicating no significant preference for any particular binding site.

The adsorption of benzene on Cu(110) is considered at four different binding sites, as illustrated in Figure 2, termed the TOP, hollow (HOL), short bridge (SB), and long bridge (LB) sites. On each site two high-symmetry orientations of the molecule are considered, and these are named A and B in the figure. In addition, three possible adsorbate structures are also considered corresponding to physisorption of flat molecules and possible chemisorption involving quinonoid

**Figure 3.** The structures of benzene considered for the adsorption on Cu(110): (a) planar, (b) quinonoid,²² and (c) H-flipped.²²

and H-flipped configurations of the ring;^{21,22} these structures are illustrated in Figure 3. The chemisorbed structures correspond directly to the local minima identified^{21,22} in calculations of benzene above the HOL site of a Cu₁₃ cluster used as a model for the (110) surface. The computed binding energies are given in Table 2, while key structural properties are given in Table 3 and all optimized coordinates are provided in the Supporting Information. Some of these key properties include the average height of the carbon atoms above the surface, Δz , the maximum difference in CC bond lengths, ΔR_{CC} , and the maximum CCCC and CCCH torsional angles, τ_{CCCC} and τ_{CCCH} , respectively. Based upon them, the optimized structures are classified as being either “flat” ($\Delta z = 2.7$ – 2.9 Å, $\Delta R_{CC} < 0.006$ Å, and torsional angles $< 2.5^\circ$ in magnitude), “quinonoid” ($\Delta z = 2.0$ – 2.4 Å, ΔR_{CC} up to 0.06 Å, large positive τ_{CCCC} , and large positive τ_{CCCH}), and “H-flipped” ($\Delta z = 2.0$ – 2.4 Å, ΔR_{CC} up to 0.02 Å, significant negative τ_{CCCC} , and large positive τ_{CCCH}). In some cases, geometry optimization leads to local-minimum structures with qualitative properties preserved, while for the remainder the structures relaxed to an alternate configuration, as indicated in Table 2.

From the results in Table 2, an important qualitative feature is that increased binding by ca. 5 kcal mol⁻¹ is predicted for benzene binding to Cu(110) compared to Cu(111). While this is consistent with the observed increase of 9 kcal mol⁻¹,^{17,18,23} the absolute magnitude of the binding energies remain in poor agreement, 6 kcal mol⁻¹ calculated compared to 23 kcal mol⁻¹ observed.²³ Physisorbed structures are predicted to be more stable than chemisorbed ones, but the energy difference is only 0.6 kcal mol⁻¹, a value that is most likely less in magnitude than the accuracy of the methodology. As Table 3 shows that these local minima differ dramatically in structure, and as qualitatively we find no significant barriers separating them, it is clear that very large amplitude motions may be sustainable on the surface and hence proper quantum thermal treatment of the vibrational

Table 2: Calculated Adsorption Energy Changes ΔE for C_6H_6 -(2 \times 3) on Cu(110)^d

method	adsorbate	TOP-A	TOP-B	SB-A	SB-B	HOL-A	HOL-B	LB-A	LB-B
VASP-PW91	flat	-1.55	-1.75	-4.33	-4.81	-4.93	-5.81	-5.28	-4.93
	quinonoid	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>	-2.87	<i>b</i>	-4.35	<i>b</i>
	H-flipped	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>	<i>c</i>	-5.19	<i>c</i>	-3.33
SIESTA-PBE start, BSSE	flat	-1.0	-1.2	-1.6	-1.5	-2.1	-2.1	-1.8	-1.7
	quinonoid	1.6	1.4	1.1	0.5	-0.1	0.1	-0.4	-1.2
SIESTA-PBE opt, raw	H-flipped	10.6	10.4	8.7	10.3	1.2	-0.4	4.7	4.1
	flat	-7.0	-7.1	-13.8	<i>c</i>	-8.9	<i>b</i>	<i>c</i>	<i>b</i>
	quinonoid	<i>a</i>	<i>a</i>	<i>a</i>	-13.7	<i>b</i>	<i>b</i>	-14.2	<i>b</i>
SIESTA-PBE opt, BSSE	H-flipped	<i>a</i>	<i>a</i>	<i>a</i>	<i>c</i>	-16.3	-18.5	<i>c</i>	-12.8
	flat	-1.0	-1.2	0.6	<i>c</i>	-1.9	<i>b</i>	<i>c</i>	<i>b</i>
	quinonoid	<i>a</i>	<i>a</i>	<i>a</i>	0.0	<i>b</i>	<i>b</i>	0.4	<i>b</i>
	H-flipped	<i>a</i>	<i>a</i>	<i>a</i>	<i>c</i>	0.8	-2.7	<i>c</i>	1.7

^a Collapses to the flat structure upon optimization. ^b Collapses to the H-flipped structure upon optimization. ^c Collapses to the quinonoid structure upon optimization. ^d Given in kcal mol⁻¹, for the adsorbate orientations illustrated in Figure 2, obtained using VASP and SIESTA (with and without correction for BSSE) with the PW91 and PBE density functionals, at starting intermolecular-only optimized geometries and full optimized ones.

Table 3: Calculated Properties for VASP PW91-Calculated C_6H_6 -(2 \times 3) on Cu(110)^b

adsorbate	structure	Δz	ΔR_{CC}	τ_{CCCC}	τ_{CCCH}	$\epsilon_H - E_F$	$\Delta\epsilon_L^{gas}$	$\Delta\epsilon_L^{111}$	q_{mol}
flat	TOP-A	2.86	.002	.2	-2.1	-3.11	-0.96	-0.57	-0.01
	TOP-B	2.81	.004	.4	-2.4	-3.11	-0.96	-0.57	-0.01
	SB-A	2.74	.005	.7	.7	-3.41	-1.00	-0.61	.01
	SB-B	2.74	.006	.7	1.2	-3.41	-1.00	-0.61	.00
	HOL-A	2.71	.004	-0.5	1.6	-3.48	-1.17	-0.77	-0.02
	HOL-B	2.70	.006	-0.9	-2.4	-3.52	-1.57, -1.17	-1.18, -0.78	.01
	LB-A	2.72	.008	.9	1.6	-3.49	-1.21	-0.82	.00
	LB-B	2.73	.007	-1.0	-2.3	-3.46	-1.10	-0.70	-0.01
quinonoid	HOL-A ^a	2.00	.062	7.4	16.0	-7.27	-2.06	-1.67	.09
	HOL-A	2.20	.002	1.0	8.5	-7.27	-1.69	-1.30	.05
	LB-A	2.41	.011	2.0	4.0	-3.98	-1.52	-1.13	.02
H-flipped	HOL-B ^a	2.09	.011	-4.3	15.9	-7.21	-2.62	-2.23	.12
	HOL-B	2.25	.019	-1.8	9.0	-4.29	-1.50	-1.11	.09
	LB-B	2.38	.020	1.9	6.4	-4.02	-1.37	-0.98	-0.01

^a Embodies the optimized geometry of benzene on a Cu₁₃ cluster.²² ^b Δz is the average height of C above Cu, ΔR_{CC} is the maximum difference in CC bond lengths, τ_{CCCC} and τ_{CCCH} are maximum torsion angles, $\epsilon_H - E_F$ is the shift in the HOMO orbital energy from the Fermi energy, $\Delta\epsilon_L^{gas}$ is the shift in LUMO energy from the gas phase, $\Delta\epsilon_L^{111}$ is the shift in LUMO energy from that for adsorption on Cu(111), and q_{mol} is the adsorbate charge from SIESTA Mulliken orbital analysis.

motion will be essential in any quantitative comparison of computed and experimental properties for the system. However, STM images of benzene on Cu(110) have revealed that the adsorbates stick over both the long-bridge site⁴⁷ and the hollow site,⁴⁸ the two lowest-energy sites revealed in Table 2. Also, it has been observed⁴⁷ that benzene is easily dragged over Cu(110) by an STM tip; this is consistent with the basic qualitative scenario predicted by the calculations of poorly site-specific binding.

Significant differences are found between the optimized structures of benzene on Cu(110) and those reported previously by Triguero et al.^{47,48} for the binding of benzene to a Cu₁₃ cluster. Shown in Table 3 are the geometrical parameters from these calculations for both the cluster-optimized HOL-A (quinonoid) and HOL-B (H-flipped) structures as well as those for our corresponding surface-optimized structures. On the surface, the distortion to the benzene geometry is dramatically reduced, and the molecule floats ca. 0.2 Å higher above the surface. The calculated interaction energies are also very different, with those for the cluster being -18 and -14 kcal mol⁻¹ for the quinonoid and H-flipped structures, respectively, compared to -2.9 kcal

mol⁻¹ and -5.2 kcal mol⁻¹, respectively, on the surface. Further, no flat structures are found above the cluster, whereas a flat structure forms the most stable structure, of interaction energy -5.8 kcal mol⁻¹, on the surface. As there are some computational differences between the original DFT implementation and that used herein, we repeated the previous cluster calculations using VASP and PW91 for the HOL-A quinonoid structure, obtaining $\Delta E = -19$ kcal mol⁻¹ in excellent agreement with the previous value. Hence the differences are due primarily to the differing reactivities of the cluster and the surface. The reasonable agreement found previously between the cluster binding energy and the surface observed adsorption energy is thus found to be due to the near cancellation of two significant effects: the underestimation of the binding due to limitations associated with modern DFT functionals and the enhanced reactivity of the cluster. Note also that variation of the DFT functional used does not lead to qualitative changes in the results and that the problems encountered with the calculation by DFT of the binding of benzene to coinage-metal surfaces is general.

The adsorption of benzene on Cu(111) is known^{14,15,17} to be physisorptive in nature. A series of calculations has been

performed to determine whether VASP predicts stable chemisorbed species for this surface, as it does for Cu(110). Geometry optimizations were performed starting at analogous quinonoid and H-flipped conformations. In all cases, the geometries relaxed to the flat ones, indicating that the surface calculations do not intrinsically overestimate the significance of the chemisorbed structures, and hence they remain as viable alternatives for the actual structure on Cu(110). We return to the question of the experimental determination of whether the interaction is fundamentally chemisorptive or physisorptive in section 3.3.

3.2. Verification of the Major Results Using CASTEP and SIESTA Calculations. The VASP calculations reveal potential-energy surfaces that support 0.6 Å changes in the metal–adsorbate separation Δz and large intramolecular distortions ΔR_{CC} , τ_{CCCC} , and τ_{CCCH} to the adsorbate at the cost of the very small amount of ca. 1 kcal mol⁻¹ in energy. Such energy changes are less than absolute error magnitudes expected for modern density-functionals, pseudopotentials, and basis sets, while the determination of precise results is computationally challenging in terms of the algorithms used for geometry optimization, etc. To verify that the major conclusions reached from the VASP calculations are robust to these effects, some analogous calculations have been performed using CASTEP and SIESTA for the binding of benzene to the surface of greatest contention, the Cu(110) surface.

The CASTEP calculations were performed for the LB-A and LB-B structures involving translational scans of the potential-energy surface for frozen metal and adsorbate components, using the flat, quinonoid, and H-flipped adsorbate structures. In all cases, the same qualitative conclusions were reached as from the VASP optimizations. For one structure a full optimization was performed, and this yielded a binding energy within 1 kcal mol⁻¹ of the corresponding VASP one. For most problems such quantitative agreement would be considered excellent, but for this system this amounts to 20% of the binding energy. The significant factor, however, is that the primary qualitative conclusions remain invariant. CASTEP was also used to compare results from the PW91 and PBE density functionals; good agreement was found, with the PBE binding strengths being slightly less than the PW91 values by just 0.1–0.3 kcal mol⁻¹.

SIESTA calculations were performed for all adsorbate structures and binding locations, and the results are provided along with the VASP ones in Table 2. These calculations were performed by first adjusting the height of the adsorbate above the surface at fixed metal and adsorbate geometry so as to provide a best-estimate starting structure, and then these structures were fully relaxed. Both the energy of the z -optimized structure and the fully relaxed one are given in the table. The resultant quinonoid and H-flipped structures show even less variations in bond lengths and torsional angles than those from the VASP calculations reported in Table 3, with in particular the HOL-A structure being very flat; significant differences in the height above the surface are still found between the physisorbed and chemisorbed structures, however.

Direct comparison of the binding energies from the SIESTA and VASP calculations is difficult owing to the

significant BSSE that arises from the use of atomic basis sets in the SIESTA calculations. While atomic basis sets are much more conducive to mechanistic analyses than are plane-wave ones, an advantage exploited in the next subsection, the presence of BSSE arising from the incompleteness of the atomic basis set used provides a significant disadvantage. In Table 2, the energy changes due to binding are shown both with and without the use of BSSE corrections. The energies of binding without BSSE correction fall in the range of $\Delta E = -7$ to -20 kcal mol⁻¹, but after correction the binding is or is very nearly lost altogether. The corresponding values obtained using VASP and CASTEP fall mid-way between the BSSE corrected and uncorrected values. When large atomic basis sets are used, the BSSE correction is usually small and typically of the wrong sign, and so BSSE corrections should not be applied.⁴⁹ However, for small atomic basis sets, the BSSE correction is large and of the correct sign, and its application is essential. The double- ζ plus polarization basis set used in these SIESTA calculations does not have the augmented functions that are crucial to BSSE reduction, and hence its application appears essential. However, for intermediate-sized basis sets such as this, a technique of fractional BSSE correction is often used⁵⁰ involving the addition of some set fraction of the full correction. This technique may be applicable here, with fractional corrections of 0.6–0.9 being required to bring the SIESTA and VASP results into quantitative agreement.

A significant difficulty with the atomic basis set approach, however, is that all geometry optimizations are performed on the raw, uncorrected energies. As the BSSE is of order 6 kcal mol⁻¹ for the distant physisorbed structures and of order 14 kcal mol⁻¹ for the close-lying chemisorbed ones, the method used to treat it induces significant changes to the shape of the potential-energy surfaces. As a result, e.g., the raw SIESTA energies for the HOL-A structure strongly favor the H-flipped structure, while after correction they favor the flat one. As the reduction of BSSE to the level required for realistic geometry optimization in these systems requires Gaussian basis sets that are at least an order of magnitude larger than those used herein,⁴⁹ any previous or foreseeable calculation of this type is likely to be unreliable. Such calculations will artificially favor closely interacting, highly distorted chemisorbed structures over physisorbed ones. While this effect cannot account for the perceived high reactivity of the C₁₃ cluster to benzene, it could account for the high degree of distortion found in the cluster-optimized structures.

3.3. The Electronic Structure of the Adsorbate Layer.

Adsorption-induced changes to the electronic structure of the surface and adsorbate provide important indicators of the nature of the surface–molecule interaction. Atomic-basis-set programs such as SIESTA provide insight into this process through the ready application of simple methods such as Mulliken analysis of the charge flow to the adsorbate, methods not available for use with plane-wave basis sets. SIESTA results for benzene on the (111) surfaces of Cu, Ag, and Au indicate negligible charge transfer to benzene, q_{mol} , of magnitude less than 0.01 e , where e is the magnitude of the charge on the electron. For benzene in various

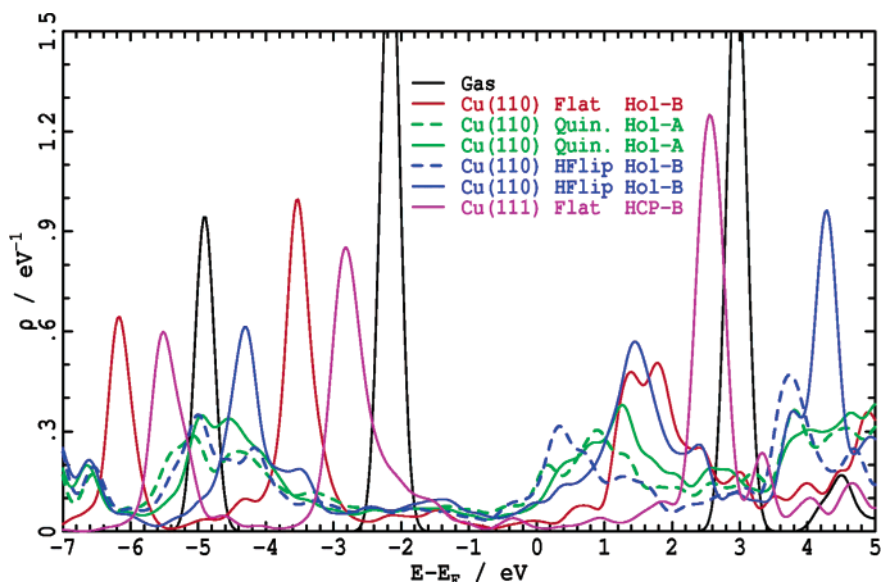


Figure 4. VASP PW91 calculated average carbon p_z (π) density of states ρ as a function of the orbital energy difference from the Fermi energy, $E - E_F$, for gas-phase benzene, benzene on Cu(111), and benzene on Cu(110) in the flat, quinonoid, and H-flipped structures; (solid)- fully optimized structures on the surface, (dashed)- starting structures based on cluster-optimized geometries.²²

structures above Cu(110), SIESTA results are given in Table 3; again, negligible charge transfer is found for the flat physisorbed structures, but some charge flow up to 0.12 e is predicted for the chemisorbed ones, especially those at the cluster-optimized geometries.

A commonly used method to estimate charge flow using plane-wave based calculations is Helmholtz analysis of the change in the surface work function. A reduction in the work function of the surface of -0.44 eV is predicted by the VASP calculations for the C_6H_6 -(3×3)-Ag(111) surface, in reasonable agreement with the measured value of -0.3 eV.^{15,51} On Au(111) the computed value is -0.42 eV, while on Cu it is somewhat bigger: -0.55 eV (obs.⁵² -0.3 eV) and -0.84 eV on the (3×3)-(111) and (2×3)-(110) surfaces, respectively. Using the Helmholtz equation¹ applied in the low coverage limit, charge transfer to the adsorbate may be estimated from these changes in surface potential, yielding 0.71, 0.74, 0.81, and 1.26 D per adsorbate molecule for the calculated lowest-energy structures on Au(111), Ag(111), Cu(111), and Cu(110), respectively. Dipole-moment changes arise from polarization of the metal surface, polarization of the molecule, and from charge-transfer between the molecule and surface. Neglecting polarization effects completely leads to estimated charge transfers of 0.04, 0.04, 0.05, and 0.10 e for these flat adsorbates, much larger than the values of <0.01 e deduced from the SIESTA Mulliken analysis. This discrepancy could arise as the molecular and surface polarization terms are also naively expected to be of this order but of opposite sign to each other. Also, non-Helmholtz terms do contribute to changes in the surface potential,⁵³⁻⁵⁵ and such effects could dominate the process especially for weakly bound adsorbates.

The predicted and observed changes in the work function are much smaller than those predicted and observed for benzene chemisorbed on reactive transition metals such as Ni, Pd, and Pt of ca. 1.4 eV.¹ There is thus a significant

qualitative difference found between the results of the present calculations and those for a system in which full organometallic bonds are implied. Analysis of the work function changes calculated from the plane-wave-based methods thus corroborates the conclusions reached from Mulliken analysis of the SIESTA results that DFT predicts only weak to very weak interactions between benzene and the various surfaces.

To gain further insight from the VASP calculations into the electronic variations that arise because of adsorption, projected densities of states (PDOS) have been evaluated. Results are presented in Figures 4 and 5 for benzene and Cu(110) well separated, at the optimized flat HOL-B lowest-energy structure, and for the starting (i.e., Cu_{13} cluster optimized^{21,22}) and surface-optimized HOL-A quinonoid and HOL-B H-flipped structures. Figure 4 shows the average C p_z (π) orbital density, while Figure 5 shows the density for the surface copper d_z^2 orbitals. As it has been shown that the computational methods predict much more realistic changes in binding energies between the Cu(111) and Cu(110) surfaces than absolute binding energies, results for benzene on Cu(111) at the lowest-energy optimized HCP-hollow B are also included in these figures. In addition, results from the quantitative analysis of the calculated PDOS for all cluster and surface optimized structures are provided in Table 3. These include the highest-occupied molecular orbital (HOMO) energy with respect to the Fermi level $\epsilon_H - E_F$, the shifts of the lowest-unoccupied molecular orbital (LUMO) from the gas phase, ϵ_L^{gas} , as well as from its calculated value on Cu(111), ϵ_L^{111} .

The gas-phase densities show the benzene HOMO at 2.15 eV below the Fermi energy E_F , whereas this band is observed⁵⁶ to be at 4.5 eV below the Cu(110) Fermi energy (this Fermi energy is at -4.8 eV with respect to the vacuum level); alternatively, the calculated LUMO appears at 2.96 eV above E_F compared to 5.9 eV observed. These discrepancies are due to the asymptotic potential error and band-

gap error, respectively, that are inherent in modern density functionals;^{9,57} the calculated band gap is 5.1 eV, while the observed one is 10.3 eV. The HOMO orbital error is somewhat compensated for in calculations of surface adsorbates by charge transfer processes that act to align the energy level systems. As a result, DFT calculations tend to give qualitatively reasonable occupied electronic structures of adsorbates but fail to quantitatively reproduce charge transfer.⁹

For benzene on Cu(111), Figure 4 shows that the calculated C π PDOS is broadened slightly due to the weak interaction with the metal surface and shifted downward by 0.66 eV (HOMO) and 0.40 eV (LUMO). These shifts reflect the net effects of orbital-specific molecule-surface interactions and charge transfer. More significant interactions are evident for benzene on Cu(110), however. For the deduced lowest-energy physisorbed flat structure HOL-B, the calculated C π PDOS shown in Figure 4 is broadened significantly, shifted downward, and the LUMO is split into two peaks at 1.36 and 1.77 eV, changes $\Delta\epsilon_L^{\text{gas}}$ given in Table 3 of -1.57 and -1.17 eV from the gas-phase values. The corresponding HOMO level shifts down by 1.37 eV. For the optimized chemisorbed structures, the broadening of the LUMO (and HOMO) is further increased, while the orbitals shift to lower energy as the separation Δz decreases and interactions become significantly larger. However, the PDOS evaluated for the quinonoid and H-flipped geometries optimized in previous cluster calculations^{21,22} show even greater broadening and shifts, with the LUMOs shifted until they cross the Fermi energy. The SIESTA Mulliken charge analysis results shown in Table 3 also indicate the appearance of detectable charge transfer in these chemisorbed structures, up to 0.12 e at the cluster-optimized geometries. Note that a possible consequence of the DFT band-gap error is that the charge-transfer process associated with the donation of electrons from the benzene π orbitals to the metal is artificially curtailed by the apparent back-bonding that is enforced when the LUMO prematurely crosses the metal Fermi energy.

Qualitatively, the interaction of benzene with Cu(110) is known to be much weaker than that with Ni(100),⁵⁸ a surface on which it is clearly chemisorbed. Also, the benzene–Cu(110) interaction is much weaker than that of acetylene with Cu(110),⁵⁹ another chemisorptive interaction, but it is significantly stronger than the interaction of benzene and Cu(111), a clearly physisorptive interaction.^{14,15,17} The adsorption of benzene on Ag(111) is also unambiguously physisorptive.^{15,51} While some experimental results⁵⁸ have been interpreted in terms of weak physisorption of benzene on Cu(110), others^{22,21,60} have been interpreted as indicating that σ – π mixing does occur and hence some degree of chemisorption is implicated. Indeed, the DOS for *all* of the possible optimized structures shown in Figure 4 or summarized in Table 3 depict significant interactions between benzene and Cu(110), interactions that are much stronger than those with Cu(111), in agreement with the general scenario depicted experimentally. PW91 calculations⁶¹ for benzene on Ni(110) predict a binding energy of 41 kcal mol⁻¹ and torsional angles up to 35°, clearly depicting strong chemisorption as

apposed to the much weaker binding on Cu(110), also in agreement with experimental findings.

More quantitative experimental information is available, however, that could in principle discriminate between the various calculated binding possibilities. X-ray emission spectroscopy observes the nature of the occupied orbitals. For benzene on Ni(110), the HOMO orbital is observed at an energy of -4.3 to -4.6 eV with respect to the Fermi level and calculated in good agreement by PW91 to be at -4.5 eV.⁶¹ For benzene on Cu(110), the observed value⁵⁸ is very similar, -4.4 eV. Figure 4 shows that the calculated HOMO levels are of this order but vary considerably depending on site and structure. In Table 3, the calculated HOMO energies $\epsilon_H - E_F$ are listed. Most calculated structures predict $\Delta\epsilon_H$ within 0.4 eV of the observed value, the exceptions being the cluster-optimized structures, at ca. -7 eV, and the high-energy TOP-A structures at > -3 eV. The best results are obtained for the optimized quinonoid LB-A and H-flipped HOL-B and LB-B structures (-4.0 to -4.3 eV), while the low-energy flat structures all more significantly removed (-3.4 to -3.5 eV).

The energy difference between the LUMO orbital and the Fermi energy cannot be reliably determined using modern DFT owing to the DFT band-gap error, but changes in this quantity between different structures should be better described. The calculated changes $\Delta\epsilon_L^{111}$ between adsorbates on Cu(110) and Cu(111) given in Table 3 and are -1.2 and -0.8 eV for the two peaks associated with the lowest-energy surface optimized (HOL-B) structure. These are somewhat less for the other physisorbed structures, -1.7 eV and -2.2 eV for the cluster-optimized quinonoid (HOL-A) and H-flipped (HOL-B) structures, respectively, and -1.0 to -1.3 eV for surface optimized quinonoid and H-flipped structures. Experimentally the LUMO energy for benzene on Cu(111) has been determined from inverse photoemission spectroscopy.⁶² Figure 6 shows the original results⁶² fitted to Gaussian-shaped peaks on a piecewise-linear background. The inverse photoemission spectrum for clean Cu(111) is also shown; it contains two peaks that are lost in the adsorbate but more significantly a very similar underlying background. For benzene on Cu(111), the fitted Gaussian has a center of 4.4 eV and standard deviation of 0.8 eV. The LUMO energy for benzene on Cu(110) has been measured by scanning-tunneling spectroscopy,⁴⁸ and the original current–voltage ($I(V)$) curve is shown in Figure 7. There the curve is fitted to the sum of three *arctan* functions depicting molecular resonances⁶³ at 3.0 eV (LUMO), 3.7 eV, and 4.5 eV. Hence the observed value of $\Delta\epsilon_L^{111}$ is -1.4 eV, in best agreement with the calculated results for the optimized chemisorption structures, again, although the flat HOL-B structure is quite close and actually has the two-peaked structure found experimentally.

Hence, from consideration of the PDOS, it is clear that the cluster-optimized structures depict unreasonable possibilities, while the optimized chemisorption structures are most favored and the flat HOL-B structure is not implausible. Authoritative conclusions cannot be made, however, due to the complex nature of the calculated PDOS structures and the lack of treatment of quantum-mechanical and thermal

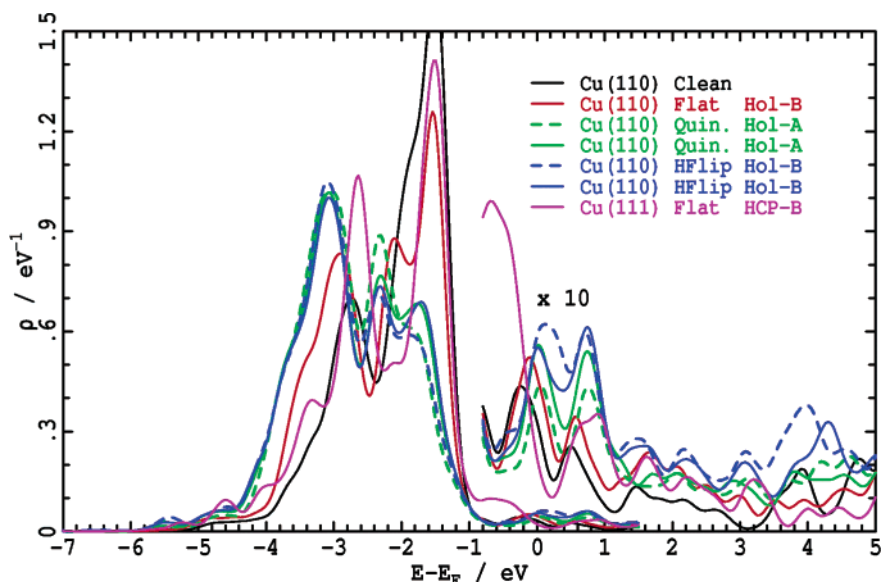


Figure 5. VASP PW91 calculated average copper d_z density of states ρ as a function of the orbital energy difference from the Fermi energy, $E - E_F$, for a clean surface, benzene on Cu(111), and benzene on Cu(110) in the flat, quinonoid, and H-flipped structures; (solid)- fully optimized structures on the surface, (dashed)- starting structures based on cluster-optimized geometries.²²

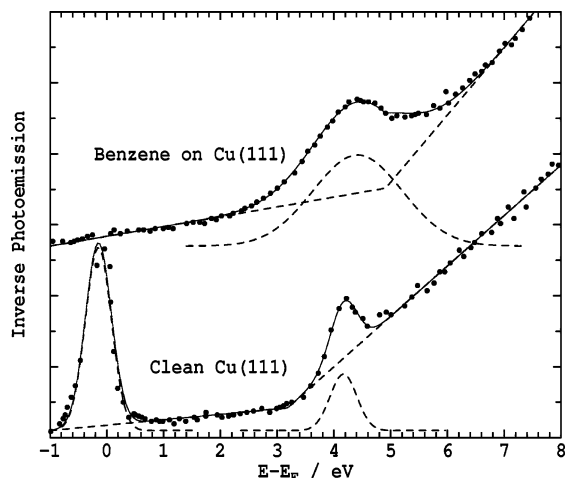


Figure 6. Deconvolution of the observed⁶² inverse photoemission spectra of a clean Cu(111) surface and surface with benzene adsorbed into Gaussian-shaped bands on a piecewise-linear background.

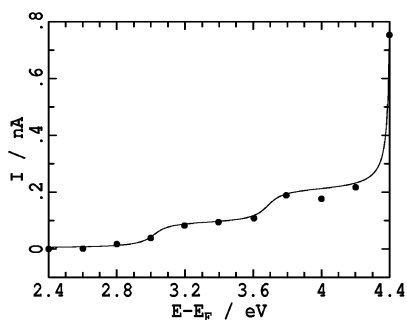


Figure 7. Fitting of the observed⁴⁸ STM current as a function of voltage (the energy above the Fermi energy) for benzene on Cu(110), revealing molecular resonances centered at 3.0, 3.7, and 4.5 eV.

vibrational effects. Given the small energy differences predicted between the most probable chemisorptive and

physisorptive structures, it could be that zero-point or thermal vibrational effects are sufficient to mix all of these structures, resulting in an observed average structure that could be quite different in appearance to any of the local-energy minima found on the potential-energy surface.

In general, only small perturbations to the DOS of the copper atoms are found on adsorption of benzene. The copper orbital that interacts most significantly is the d_z^2 orbital whose PDOS is shown in Figure 5. While a weak tail to this distribution above the Fermi energy is found indicating bonding interactions with the unoccupied molecular orbitals, the effect is clearly quite weak. Instead, large downward shifts of the orbital energies are found, especially for the cluster-optimized chemisorptive structures, indicating a strong interaction with the occupied orbitals. It is indicative of a strong dispersive interaction between the copper and benzene, an interaction for which DFT does not correctly include the resultant attractive energy contribution, especially for complexes with coinage metals.^{49,38} It is hence reasonable to hypothesize that the binding is dispersive in nature and that this is the cause of the poor agreement between calculated and observed absolute binding energies.

3.4. Lack of Involvement of the Triplet States of Benzene in the Binding. The chemisorptive interactions observed between some alkenes and reactive metal surfaces cannot be accounted for assuming that the surface interacts with the ground state of the alkene.²¹ Instead, strong interactions with excited states of alkenes have been invoked. For benzene on Cu(110), the local structure of the benzene is reminiscent of the equilibrium geometry⁴⁰ of the lowest-triplet excited state, and hence it has been postulated that it is this state that interacts with the metal. Assuming that two covalent bonds form in this chemisorptive process,²¹ the DFT-calculated absorption energy of 18 kcal mol^{-1} for benzene on the Cu_{13} cluster has been interpreted as indicating a Cu-C bond strength of 58 kcal mol^{-1} , opposed by the energy of 98 kcal mol^{-1} required to form the triplet state.

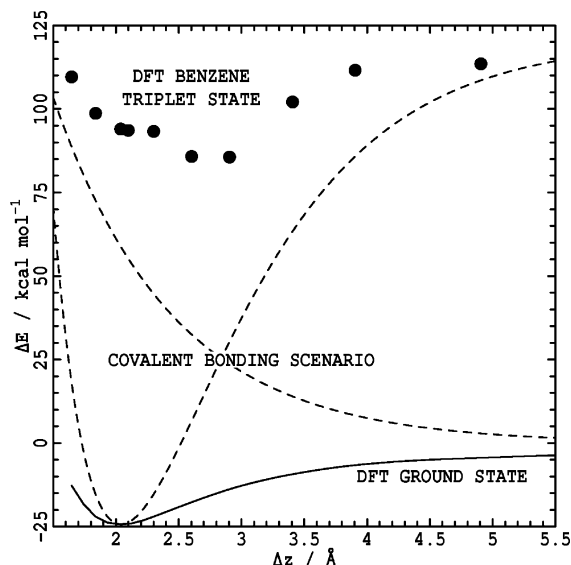


Figure 8. The DFT calculated ground-state potential-energy surface (—) for benzene approaching the Cu_{13} cluster at minimum carbon to copper surface-plane separation Δz , along with the calculated surface for the lowest triplet state (●) of benzene, compared to the generic form²¹ of these two surfaces (---) expected in the scenario that a covalent bond is formed between the two species.

Triguero et al.²¹ have depicted the generic form of the potential-energy surfaces expected in this situation, and these are sketched in Figure 8: the ground-state of the adsorbate correlates to the asymptotic benzene triplet state, while the asymptotic ground state becomes an excited state of the adsorbate. Note that both depicted states are actually doublets, the upper one being a triplet-coupled state known as a *tripdoublet* state.^{64,65} In this figure is also shown the actual potential-energy surfaces for these two states calculated using DFT at the PW91/6-31G* level using GAUSS-IAN03. These surfaces were obtained by freezing the cluster geometry and the height of the nearest carbon atoms above the surface, allowing all other coordinates of the benzene molecule to relax on the ground state. The ground state of the adsorbate is clearly seen to correlate to the asymptotic ground state, in contrast to the predictions of the triplet-interaction model. Note that the calculated binding energy of 24 kcal mol⁻¹ is somewhat greater than the value of 19 kcal mol⁻¹ obtained in this work previously using VASP; this is due largely to the neglect of BSSE corrections in the current calculations.

Accurate calculations of the energy of the benzene tripdoublet state are not feasible as there are many hundreds of excited states of lower energy involving the cluster; at the equilibrium geometry, we determined the nature of the lowest 100 excited states using time-dependent DFT, finding none to contain benzene triplet character. However, a crude estimate of the tripdoublet energy is obtained as the difference in the benzene LUMO and HOMO orbital energies, and this is shown in the figure. Though only approximate and not optimized for the electronic state of the cluster, these results support the emphatic results obtained for the ground-state surface that the adhesion of benzene to the C_{13} cluster does not involve covalent bonding to the triplet state. Instead,

the DFT results depict a weak intermolecular interaction typical of hydrogen bonding, dispersive interactions, or possibly dative covalent bonding involving the benzene ground-state only.

Inspection of the form of the molecular orbitals reflects the same scenarios discussed in the previous section for the surface-benzene interaction: all benzene π orbitals of the adsorbate are significantly depressed in energy, with the dominant mixings being between different occupied levels (or between different virtual levels) arising from dispersive intermolecular interactions. However, there are also some significant interactions evident between the benzene HOMO and unoccupied cluster orbitals. This results in significant charge transfer, the net effect of which is, after BSSE correction, a Mulliken-charge transfer of $q_{\text{mol}} = 0.3 e$; this could indicate the action of dative covalent bonding. The depression of the benzene LUMO orbital appears to be dominated by interactions with copper 4p orbitals. If this interaction was slightly stronger, the LUMO could become significantly occupied, and hence the benzene would appear to take on triplet character. Hence the triplet interaction model, while being shown to be inappropriate for benzene on Cu(110), may be quite apt for systems with slightly stronger interactions.

3.5. Quantification of the Contribution of Dispersion to the Binding Energy. While DFT calculations indicate that the Cu_{13} cluster introduced by Triguero et al.^{47,48} is too reactive for quantitative modeling of the reactivity of Cu(110) surfaces, it can provide a useful guide as to the significance of dative bonding and dispersive force as it facilitates the application of high-end ab initio approaches designed for discrete molecular systems. We have performed CASPT2 calculations for the Cu_{13} -benzene interaction for the quinonoid structure HOL-A at the previous cluster-optimized geometry²¹ and compared them to DFT calculations for the same system. CASPT2 is a Møller-Plesset perturbation method similar to MP2 but generalized to treat systems with open-shell bands such as the s band of the copper cluster and is, in principle, the simplest ab initio method that is appropriate for problems of this type. It can provide an a priori estimate of the magnitude of the dispersive interaction that acts in parallel to, and independent of, covalent-bonding forces; it has been shown to be reliable for the study of related problems involving a small number of metal atoms,^{49,66} but it becomes much more difficult to apply to large metal clusters such as Cu_{13} .

At the cluster optimized binding geometry of Triguero et al.,²¹ the raw interaction energy of the two fragments is calculated using CASPT2 to be -77 kcal mol⁻¹, reducing to -51 kcal mol⁻¹ after BSSE correction. The distortion energy of the benzene molecule required to produce the quinonoid geometry is calculated at this level to be 18 kcal mol⁻¹, so the total calculated interaction energy is -33 kcal mol⁻¹. The CASSCF calculations used as a starting point for the perturbation calculations in CASPT2 predict that the cluster is highly unbound, however: the raw interaction energy is $+17$ kcal mol⁻¹, becomes $+20$ kcal mol⁻¹ after BSSE correction, and gives the total interaction energy as $+38$ kcal mol⁻¹. As covalent bonding is described at a

usefully realistic level at the CASSCF level and no binding is predicted, it is clear that covalent bonding plays an insignificant role in the interaction. This includes both simple dative bonding in which say benzene acts as an electron donor to fill partially occupied copper orbitals as well as more sophisticated scenarios such as the interaction of the surface with benzene excited states. The correlation-energy correction is thus a massive $-71 \text{ kcal mol}^{-1}$, a correction that is noncovalent in origin.

The correlation energy of two interacting species can be separated into contributions from the changes to the fluctuations on each species as modified by the presence of the other as well as the dispersive contribution that arises from correlated fluctuations on both species. If a covalent bond forms between the two species, then the bond formation alters the valence electrons on each, and these electrons then interact with the local cores. This gives rise to the nondispersive *core - valence* correlation⁶⁷ that can act to significantly deepen covalent wells,⁶⁷ being especially significant for interactions with transition metals.⁶⁸ Core-valence correlation acts in response to bonding interactions but does not constitute a bonding mechanism. In the present application, there is no intrinsic covalent bond for core-valence correlation to enhance, but there is clear evidence of many dispersive bonding interactions that mix say the copper 3d orbitals with the benzene occupied orbitals. It is thus clear that the primary source of the binding is dispersive in nature and that the effect of any core-valence correlation is thus to enhance the significance of the dispersive interactions.

The 13-orbital active space optimized in the CASSCF calculations excluded some of the Cu 4s orbitals that initially constituted it, including in their stead some occupied Cu 3d and benzene orbitals as well as some unoccupied Cu 4p and benzene orbitals. Like results obtained using DFT, the orbital coefficients reflect strong mixing of the occupied benzene and Cu 3d orbitals and strong mixing of the virtual benzene and Cu 4p orbitals. However, much stronger benzene $\sigma-\pi$ mixing is perceived at the CASSCF level, owing, most likely to the neglect of dynamical electron correlation in CASSCF. It appears that some of the possible metal to benzene-triplet bonding interactions are directly included as some of the excitations available within the active space. All possible interactions may have been included, if the ground-state energy could have been reduced in this process, however, and all interactions are indeed included at the CASPT2 level. Because of the high level of $\sigma-\pi$ mixing, quantification of the significance of the benzene triplet states at either the CASSCF or CASPT2 levels is difficult, although it is clear that they do not dominate the binding.

The interaction energy of benzene and the cluster calculated by VASP DFT is $-19 \text{ kcal mol}^{-1}$, some 14 kcal mol^{-1} less than the CASPT2 value; the correlation energy from the DFT calculation is thus perceived to be ca. $-57 \text{ kcal mol}^{-1}$ or only 80 of that determined by CASPT2. While core-valence correlation energies are well represented by DFT,⁶⁸ a major cause of the underestimation of the binding energies is the inadequate treatment of dispersion forces offered by all currently available density functions.⁵⁷ These forces dominate weakly interacting systems such as physisorbed

adsorbates on solid surfaces and, whether by design or otherwise, are accounted for as part of a perceived covalent bonding interaction.³⁸ Consequently, the quality of DFT predictions for weakly bound systems varies dramatically, possibly either underestimating or overestimating binding energies by an order of magnitude. Dispersion contributions to strong interactions involving covalent binding are usually of the same magnitude as those to physisorbed interactions, but as the covalent forces are much larger, improper treatment of dispersion does not present a critical problem. While there have been attempts to incorporate realistic descriptions of dispersive interactions empirically within DFT,^{66,69,70} such refinements are not yet applicable or well characterized for practical purposes.

While the CASPT2 calculations clearly indicate the major qualitative features controlling the binding, accurate quantitative calculations at this level are difficult to perform. Basis sets of the size used herein are generally considered to give results of accuracy of ca. 5 kcal mol^{-1} for second-row complexes; however, the number of electrons retained in the calculation, and the extent of electron correlation within the metal bands, can have profound effects on the accuracy of the calculation.

A simple test that verifies that the active space used is not unrealistic is provided by an MP2 calculation using only a single spin-adapted reference determinant. The calculated MP2 interaction energy is $-26 \text{ kcal mol}^{-1}$, quite close to the CASPT2 value of $-33 \text{ kcal mol}^{-1}$. Previous MP2 calculations on this system²⁰ have predicted either no binding or weak binding, in contrast to this result. In the current calculations a 10-electron effective core potential is used, explicitly including 19 electrons per copper atom in the calculations, compared to 1–11 electrons included previously. The dispersion and core-valence correlation energies are proportional to the number of nearby electrons, with the effect of reducing the number of electrons per atom to 11 being enhanced by the inclusion of only 6 copper atoms in the earlier calculations. It is thus anticipated that the inclusion of 13 atoms containing 19 electrons in the present calculations could approximate the asymptotic limit. However, the core-valence correlation energy is accounted for in the present calculations at the CASPT2 level, and, as these contributions to the binding may be significant,⁶⁸ enhanced quantitative accuracy would be expected if the 3d orbitals were included in the active space.

A simple test for the adequacy of the basis set is the magnitude of the BSSE correction. At 26 kcal mol^{-1} , this correction is quite large. Even the adequacy of the application of the BSSE correction on systems of this type using sufficiently large basis sets, of the sized used herein, has been questioned⁴⁹ as it may change calculated binding energies in the *wrong* direction or double the actual effect.⁵⁰ Clearly, much larger basis sets⁴⁹ are required in quantitative calculations.

4. Conclusions

The typically strong interaction of benzene with surfaces of transition metals has previously been extensively studied owing to its technological relevance. Here, the nature of the

adsorption of benzene to the coinage metals copper, silver, and gold is shown to be significantly different with the adsorbate only weakly interacting with the surfaces. For benzene on Cu(111), Ag(111), and Au(111) the binding is clearly identified as being weak and physisorptive, with all major qualitative features of the available experimental results being reproduced by the calculations. A major quantitative feature not properly predicted is the magnitude of the binding energy, a quantity that is dramatically underestimated, however. For benzene on Cu(110), a variety of feasible physisorption-like and chemisorption-like structures are predicted. Owing to the underestimation of the binding energies, authoritative discrimination between these possibilities based solely on calculated energies is not feasible. However, the calculated PDOS of the various structures are shown to differ significantly, and the chemisorbed-like ones actually appear to give the best agreement with experimental results. Nevertheless, neither the possibility that the physisorbed structures prevail nor the possibility that zero-point and thermal fluctuations dominate by mixing the structures can be eliminated. Previously, the prominence of highly distorted chemisorbed structures for benzene on Cu(110) had been anticipated through model DFT calculations of benzene on a Cu₁₃ cluster,^{21,22} but our analogous calculations for that system and for benzene on a periodic surface indicate that the model cluster is too reactive for use in quantitative studies of adsorption structure and energetics. The predicted PDOS for these cluster-optimized distorted structures are inconsistent with the available experimental information and are hence excluded from contention for the structure of benzene on Cu(110).

A significant feature of the calculated binding between benzene and the (110) and (111) surfaces is that covalent bonding contributions are insignificant. Instead, the interactions are dominated by dispersive forces. Calculations of the interaction of benzene with the Cu₁₃ cluster used to model the Cu(110) surface provide quantitative support to this conclusion: no binding at all is predicted between benzene and the cluster at the CASSCF level, while CASPT2 calculations reveal a massive intermolecular correlation energy. In addition, spin-uncoupling models²¹ that anticipate strong covalent interactions between Cu(110) and the lowest triplet excited state of benzene are shown to be inappropriate by both the DFT calculations, which indicate that the ground state of the adsorbate correlates to the ground states of the separated species, and by the CASPT2 calculations, calculations that explicitly include all possible states of bond preparedness of the cluster. The DFT calculations hint that spin-uncoupling may become quite significant for other systems with stronger metal–alkene interactions, however. It is the preeminence of dispersive forces in the benzene–Cu(110) interaction that leads to the previously noted very poor quantitative predictions of binding strengths by DFT methods.

The authoritative prediction of the structure and properties of aromatic molecules interacting through π -stacking interactions with surfaces of copper, silver, and gold is thus shown to be a very difficult task that is not currently feasible. Such calculations must include proper treatment of the periodic

metallic surfaces, accurate treatment of the dispersive forces between the surface and molecule, and adequate treatment of quantum molecular motion. DFT-based methods recognize that dispersive forces modulate the DOS of the system but fail to include the contribution of that modulation to the total energy. Hence these methods may be of greater use in determining electronic properties than in predicting structural equilibria, but such use of DFT remains limited by entrenched problems such as band-lineup error and band-gap error.

Acknowledgment. The work was supported by the Australian Research Council. The use of computer facilities at the Australian Partnership for Advanced Computing (APAC) and Australian Centre for Advanced Computing and Communications (AC3) is gratefully acknowledged.

Supporting Information Available: Optimized coordinates. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Somorjai, G. A. *Introduction to Surface Chemistry and Catalysis*; Wiley: New York, 1994.
- (2) Tour, J. M.; Jones, L., II.; Pearson, D. L.; Lamba, J. J. S.; Burgin, T. P.; Whitesides, G. M.; Allara, D. L.; Parikh, A. N.; Atre, S. V. *J. Am. Chem. Soc.* **1995**, *117*, 9529–9534.
- (3) Reed, M. A.; Zhou, C.; Muller, C. J.; Burgin, T. P.; Tour, J. M. *Science* **1997**, *278*, 252–254.
- (4) Ulman, A. *Chem. Rev.* **1996**, *96*, 1533–1554.
- (5) Lavrich, D. J.; Wetterer, S. M.; Bernasek, S. L.; Scoles, G. *J. Phys. Chem. B* **1998**, *102*, 3456–3465.
- (6) Grönbeck, H.; Curioni, A.; Andreoni, W. *J. Am. Chem. Soc.* **2000**, *122*, 3839–3842.
- (7) Andreoni, W.; Curioni, A.; Grönbeck, H. *Int. J. Quantum Chem.* **2000**, *80*, 598–542.
- (8) Gottschalck, J.; Hammer, B. *J. Chem. Phys.* **2002**, *116*, 784–790.
- (9) Bilić, A.; Reimers, J. R.; Hush, N. S. *J. Chem. Phys.* **2005**, *122*, 094708–1–15.
- (10) Crossley, M. J.; Prashar, J. K. *Tetrahedron Lett.* **1997**, *38*, 6751–6754.
- (11) Reimers, J. R.; Hall, L. E.; Crossley, M. J.; Hush, N. S. *J. Phys. Chem. A* **1999**, *103*, 4385–4397.
- (12) Bilić, A.; Reimers, J. R.; Hush, N. S. *J. Phys. Chem. B* **2002**, *106*, 6740–6747.
- (13) Chesters, M. A.; Somorjai, G. A. *Surf. Sci.* **1975**, *52*, 21–28.
- (14) Wöll, C. *J. Synchrotron Radiat.* **2001**, *8*, 129–135.
- (15) Dudde, R.; Frank, K.-H.; Koch, E.-E. *Surf. Sci.* **1990**, *225*, 267–272.
- (16) Netzer, F. P. *Langmuir* **1991**, *7*, 2544–2547.
- (17) Xi, M.; Yang, M. X.; Jo, S. K.; Bent, B. E.; Stevens, P. *J. Chem. Phys.* **1994**, *101*, 9122–9131.
- (18) Lukes, S.; Vollmex, S.; Witte, G.; Wöll, C. *J. Chem. Phys.* **2001**, *114*, 10123–10130.
- (19) Lomas, J. R.; Baddeley, C. J.; Tikhov, M. S.; Lambert, R. M. *Langmuir* **1995**, *11*, 3048–3053.

- (20) Lomas, J. R.; Pacchioni, G. *Surf. Sci.* **1996**, *365*, 297–309.
- (21) Triguero, L.; Pettersson, L. G. M.; Minaev, B.; Ågren, H. *J. Chem. Phys.* **1998**, *108*, 1194–1205.
- (22) Pettersson, L. G. M.; Ågren, H.; Luo, Y.; Triguero, L. *Surf. Sci.* **1998**, *408*, 1–20.
- (23) Rogers, B. L.; Shapter, J. G.; Ford, M. J. *Surf. Sci.* **2004**, *548*, 29–40.
- (24) Møller, C.; Plesset, M. S. *Phys. Rev. A* **1934**, *46*, 618–622.
- (25) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218–1226.
- (26) Kresse, G.; Hafner, J. *Phys. Rev. B* **1993**, *47*, 558–561.
- (27) Kresse, G.; Furthmüller, J. *Comput. Mater. Sci.* **1996**, *6*, 15–30.
- (28) Segall, M. D.; Lindan, P. J. D.; Probert, M. J.; Pickard, C. J.; Hasnip, P. J.; Clark, S. J.; Payne, M. C. *J. Phys.: Condens. Matter* **2002**, *14*, 2717–2744.
- (29) Ordejón, P.; Artacho, E.; Soler, J. M. *Phys. Rev. B* **1996**, *53*, R10441–R10444.
- (30) Soler, J. M.; Artacho, E.; Gale, J. D.; García, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. *J. Phys.: Condens. Matter* **2002**, *14*, 2745–2779.
- (31) Vanderbilt, D. *Phys. Rev. B* **1990**, *41*, 7892–7895.
- (32) Kresse, G.; Hafner, J. *J. Phys.: Condens. Matter* **1994**, *6*, 8245–8257.
- (33) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249.
- (34) Perdew, J. P.; Burke, W.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3965–3968.
- (35) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
- (36) Boys, S. F.; Benardi, F. *Mol. Phys.* **1970**, *19*, 553–557.
- (37) Neugebauer, J.; Scheffler, M. *Phys. Rev. B* **1992**, *46*, 16067–16080.
- (38) Bilić, A.; Reimers, J. R.; Hush, N. S.; Hafner, J. *J. Chem. Phys.* **2002**, *116*, 8981–8987.
- (39) Methfessel, A.; Paxton, A. T. *Phys. Rev. B* **1989**, *40*, 3616–3621.
- (40) Burns, W. J.; van der Waals, J. H.; van Herner, M. C. *J. Am. Chem. Soc.* **1989**, *111*, 86–87.
- (41) Andersson, K.; Blomberg, M. R. A.; Fülcher, M. P.; Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Neogrády, P.; Olsen, J.; Roos, B. O.; Sadlej, A. J.; Seijo, L.; Serrano-Andrés, L.; Siegbahn, P. E. M.; Widmark, P.-O. *Molcas Version 4*; University of Lund: Lund, Sweden, 1997.
- (42) Andrae, D.; Häussermann, U.; Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chim. Acta* **1990**, *77*, 123–141.
- (43) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257–2261.
- (44) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; B2, et al. *GAUSSIAN 03 Rev.*; Gaussian Inc.: Pittsburgh, PA, 2003.
- (45) Chatterjee, R.; Postawa, Z.; Winograd, N.; Garrison, B. J. *J. Phys. Chem. B* **1999**, *103*, 151–163.
- (46) Anderson, A. B.; McDevitt, M. R.; Urbach, F. L. *Surf. Sci.* **1984**, *146*, 80–92.
- (47) Doering, M.; Rust, H.-P.; Briner, B. G.; Bradshaw, A. M. *Surf. Sci.* **1988**, *410*, L736–L740.
- (48) Komeda, T.; Kim, Y.; Fujita, Y.; Sainoo, Y.; Kawai, M. *J. Chem. Phys.* **2004**, *120*, 5347–5352.
- (49) Lambropoulos, N. A.; Reimers, J. R.; Hush, N. S. *J. Chem. Phys.* **2002**, *116*, 10277–10286.
- (50) Cai, Z.-L.; Reimers, J. R. *J. Phys. Chem. A* **2002**, *106*, 8769–8778.
- (51) Zhou, X.-L.; Castro, M. E.; White, J. M. *Surf. Sci.* **1990**, *238*, 215–225.
- (52) Munakata, T.; Shudo, K. *Surf. Sci.* **1999**, *433*, 184–187.
- (53) Pettersson, L. G. M.; Bagus, P. S. *Phys. Rev. Lett.* **1986**, *56*, 500–503.
- (54) Michaelides, A.; Hu, P.; Lee, M.-H.; Alavi, A.; King, D. A. *Phys. Rev. Lett.* **2003**, *90*, 246103-1-4.
- (55) Migani, A.; Sousa, C.; Illas, F. *Surf. Sci.* **2005**, *574*, 297–305.
- (56) Becker, R. S.; Wentworth, W. E. *J. Am. Chem. Soc.* **1963**, *85*, 4.
- (57) Reimers, J. R.; Cai, Z.-L.; Bilić, A.; Hush, N. S. *Ann. N. Y. Acad. Sci.* **2003**, *1006*, 235–251.
- (58) Weinelt, M.; Wassdahl, N.; Weill, T.; Karis, O.; Hasselström, J.; Bennich, P.; Nilsson, A. *Phys. Rev. B* **1998**, *7351*–7360.
- (59) Triguero, L.; Föhlisch, A.; Väterlein, P.; Hasselström, J.; Weinelt, M.; Pettersson, L. G. M.; Luo, Y.; Ågren, H.; Nilsson, A. *J. Am. Chem. Soc.* **2000**, *122*, 12310–12316.
- (60) Triguero, L.; Luo, Y.; Pettersson, L. G. M.; Ågren, H. *Phys. Rev. B* **1999**, *5189*–5200.
- (61) Mittendorfer, F.; Hafner, J. *Surf. Sci.* **2001**, *472*, 133–135.
- (62) Frank, K. H.; Dudde, R.; Koch, E. E. *Chem. Phys. Lett.* **1986**, *132*, 83–87.
- (63) Reimers, J. R.; Hall, L. E.; Hush, N. S.; Silverbrook, K. *Ann. N. Y. Acad. Sci.* **1998**, *852*, 38–53.
- (64) Reimers, J. R.; Hush, N. S. *Inorg. Chim. Acta* **1994**, *226*, 33–42.
- (65) Reimers, J. R.; Shapley, W. A.; Hush, N. S. *J. Chem. Phys.* **2003**, *119*, 3240–3248.
- (66) Zhechkov, L.; Heine, T.; Patchkovskii, S.; Seifert, G.; Duarte, H. A. *J. Chem. Theory Comput.* **2005**, *1*, 841–847.
- (67) Müller, M.; Meyer, W. *J. Chem. Phys.* **1984**, *80*, 3311–3320.
- (68) Triguero, L.; Wahlgren, U.; Pettersson, L. G. M.; Siegbahn, P.
- (69) Wu, X.; Vargas, M. C.; Nayak, S.; Lotrich, V.; Scoles, G. *J. Chem. Phys.* **2001**, *115*, 8748–8757.
- (70) Xu, X.; Goddard, W. A., III *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673–2677.

JCTC

Journal of Chemical Theory and Computation

Effect of the f-Orbital Delocalization on the Ligand-Field Splitting Energies in Lanthanide-Containing Elpasolites

Mohamed Zbiri,[†] Claude A. Daul,[†] and Tomasz A. Wesolowski^{*,‡}

Département de Chimie, Université de Fribourg - 9, Chemin du Musée, Pérolles, CH-1700 Fribourg, Switzerland, and Département de Chimie, Université de Genève - 30, quai Ernest-Ansermet, CH-1211 Genève 4, Switzerland

Received February 1, 2006

Abstract: The ligand-field induced splitting energies of f-levels in lanthanide-containing elpasolites are derived using the first-principles universal orbital-free embedding formalism [Wesolowski and Warshel, *J. Phys. Chem.* **1993**, *97*, 8050]. In our previous work concerning chloroelpasolite lattice (Cs₂NaLnCl₆), embedded orbitals and their energies were obtained using an additional assumption concerning the localization of embedded orbitals on preselected atoms leading to rather good ligand-field parameters. In this work, the validity of the localization assumption is examined by lifting it. In variational calculations, each component of the total electron density (this of the cation and that of the ligands) spreads over the whole system. It is found that the corresponding electron densities remain localized around the cation and the ligands, respectively. The calculated splitting energies of f-orbitals in chloroelpasolites are not affected noticeably in the whole lanthanide series. The same computational procedure is used also for other elpasolite lattices (Cs₂NaLnX₆, where X=F, Br, and I)—materials which have not been fabricated or for which the ligand-field splitting parameters are not available.

1. Introduction

Lanthanide complexes offer potential applications in chemistry, physics, and other related areas.^{1–13} Theoretical modeling of such complexes involves high-cost methods because of the role of electron correlation and the necessity of taking into account the effects of the environment of the f-elements.^{14–29} Density-functional-theory methods based on the Kohn–Sham equations (KS-DFT) became standard tools in modeling large polyatomic systems.^{30–32} In practice, KS-DFT calculations apply approximations to the exchange-correlation functional and the associated potential which are usually rather adequate. Typically, they lead to results of reasonable accuracy at computational cost which is significantly lower than that of traditional wave function-based methods. For some systems and/or properties, however, standard approximations face difficulties. As far as the f-elements are concerned, they lead to rather satisfactory

results concerning structure, energetics, and vibrational properties^{33–35} but lead sometimes to qualitatively wrong results as far as the details of the electronic structure are concerned.^{36–41} Alternatively, following the spirit of the ligand-field theory, the orbitals of key interest can be obtained using the embedding strategy, in which only the lanthanide is described at the orbital level, whereas its environment is represented by some “effective embedding potential”.^{42–46} In this work, we apply the nonempirical embedding formalism⁴⁸ in which the embedded subsystem is described at the orbital-level, whereas its environment is characterized by the electron density (ρ_{II}). For a given ρ_{II} , the embedded orbitals ($\phi_{(I)i}$) used to construct the electron density of the subsystem under investigation ($\rho_I = \sum_{i=1}^{N_I} n_i^I |\phi_{(I)i}|^2$) are obtained from one-electron Kohn–Sham-like equations:⁴⁸

$$\left[-\frac{1}{2}\nabla^2 + V_{\text{eff}}^{\text{KS CED}}[\vec{r}, \rho_I, \rho_{II}] \right] \phi_{(I)i} = \epsilon_{(I)i} \phi_{(I)i} \quad (1)$$

The superscript KSCED (Kohn–Sham Equations with Constrained Electron Density) is used to indicate the difference between the effective potential in eq 1 and that in the

* Corresponding author e-mail: Tomasz.Wesolowski@chiphy.unige.ch.

[†] Université de Fribourg - 9.

[‡] Université de Genève - 30.

Kohn–Sham formalism.³¹ Fully variational variant of the above scheme, where instead of assuming some ρ_{II} it is obtained from a complementary embedding equation in which ρ_I and ρ_{II} exchange their roles, represents one of the possible practical realizations of the subsystem formulation of density functional theory by Cortona.⁴⁷ The total effective potential $V_{\text{eff}}^{\text{KSCED}}[\rho_I, \rho_{II}, \vec{r}]$ can be conveniently split into two components: the Kohn–Sham effective potential for the isolated subsystem ($V^{\text{KS}}[\vec{r}, \rho_I]$) and the remaining part representing the environment ($V_{\text{emb}}^{\text{eff}}[\vec{r}, \rho_I, \rho_{II}]$) which reads

$$V_{\text{emb}}^{\text{eff}}[\vec{r}, \rho_I, \rho_{II}] = \sum_{A_{II}} - \frac{Z_{A_{II}}}{|\vec{r} - \vec{R}_{A_{II}}|} + \int \frac{\rho_{II}(\vec{r}')}{|\vec{r}' - \vec{r}|} d\vec{r}' + \frac{\delta E_{\text{xc}}[\rho_I + \rho_{II}]}{\delta \rho_I} - \frac{\delta E_{\text{xc}}[\rho_I]}{\delta \rho_I} + \frac{\delta T_s^{\text{nad}}[\rho_I, \rho_{II}]}{\delta \rho_I} \quad (2)$$

where $T_s^{\text{nad}}[\rho_I, \rho_{II}] = T_s[\rho_I + \rho_{II}] - T_s[\rho_I] - T_s[\rho_{II}]$, and the functionals $E_{\text{xc}}[\rho]$ and $T_s[\rho]$ are defined in the Kohn–Sham formalism.³¹ Neither $V^{\text{KS}}[\vec{r}, \rho_I]$ nor $V_{\text{emb}}^{\text{eff}}[\vec{r}, \rho_I, \rho_{II}]$ depend on the orbitals but only on the electron densities of the two subsystems.

The numerical solution of eq 1 proceeds by representing embedded orbitals as a linear combination of atom-centered basis functions ($\{\chi_i^I\}$ and $\{\chi_i^{II}\}$). In such a case, two types of expansion are of great practical relevance: the approximated one, in which only selected atom-centered functions are used in the construction of embedded orbitals, and another one, in which all available atom-centered functions are used. The first type of expansion is an approximation, and such calculations are labeled here by KSCED(m) following the convention of ref 49. It was used in our previously reported work on the ligand-field parameters of the f-levels of lanthanide cations in chloroelpasolites. It is referred to also as “monomolecular expansion”. This type of expansion is obviously attractive computationally. Its drawback is, however, the absence of the terms of the $\chi_k^I(\mathbf{r})^* \chi_l^{II}(\mathbf{r})$ type in the expansion of the total electron density ($\rho_{\text{total}} = \rho_I + \rho_{II}$). This makes the cases with possible intersubsystem charge-transfer and/or covalency computationally unattractive because of the very slow convergence of the KSCED(m) results with the basis set.⁵⁰

Our previous studies showed that the differences between ligand-field splitting energies derived from KSCED(m) calculations and deduced from experiment⁵² were rather small (relative errors within 10–20%).⁵³ Such errors are qualitatively smaller than the ones corresponding to calculations applying conventional Kohn–Sham calculations or electrostatic-only embedding.⁵³ Several factors contribute to the deviations from experimental data: the intrinsic errors in the applied approximation for the exchange-correlation effective potential, the use of the average-of-configuration reference state, errors in the applied approximation to the nonadditive kinetic energy effective potential, and the absence of the $\chi_k^I(\mathbf{r})^* \chi_l^{II}(\mathbf{r})$ terms.

In the present work, one among possible sources of deviations between the calculated and experimental parameters reported previously is investigated in detail. The effect of charge transfer and covalency is quantified by comparing

the ligand-field splitting energies derived from the two types of KSCED embedding calculations which use either monomer or supermolecular expansion of both components of the total electron density (ρ_I and ρ_{II}). Following the convention of ref 49, the calculations using the supermolecular expansion are labeled by KSCED(s) in this work.

It is worthwhile to notice that the possibility for a complete delocalization of f-orbitals and charge transfer might either improve or worsen the calculated splitting energy. The worsening of the results would indicate that the applied approximate functionals in the embedding potential given in eq 2 are not adequate, and their flaws are exposed by adding more flexibility to the embedded orbitals. One of the key issues of this work is, therefore, the determination whether the good quality of the obtained previously KSCED(m) results is due to the localization assumption. This assumption is no longer made in the present work. The possibility of the intersystem charge-flow exposes the possible flaws of the approximations used in the embedding potential given in eq 2 such as an artificial charge-leak from ligands to the cation.⁵¹ Due to the variational character of the applied method, the use of more centers in the orbital expansion leads to the results which are closer to the basis set limit. It is especially important in view of the possible extension of the present studies toward modeling the complete spectra of lanthanide centers in solids. Such a task hinges, however, not only on a reliable description of the effect of the environment—the main issue of this work—but also on a proper representation of the electronic structure of the isolated cation.

2. Computational Details

Applications of eqs 1 and 2 in computer modeling rely on the approximations to the relevant functionals: $T_s^{\text{nad}}[\rho_I, \rho_{II}]$ and $E_{\text{xc}}[\rho]$. The used functionals approximate reasonably well the exact embedding potential of eq 2 in the case of small overlap between the electron densities ρ_I and ρ_{II} . The applied gradient-dependent approximation for $T_s^{\text{nad}}[\rho_I, \rho_{II}]$ was chosen based on dedicated numerical tests in the case of such pairs of ρ_I and ρ_{II} ,⁴⁹ which do not overlap significantly—a case relevant for the present studies.

The exchange-correlation component of the effective embedding potential given in eq 2 was approximated by means of the functional of Perdew and Wang (PW91).⁵⁴ The van Leeuwen-Baerends (LB94) exchange-correlation potential⁵⁵ was used to approximate the exchange-correlation component of $V^{\text{KS}}[\vec{r}, \rho_I]$ in eq 1. This choice was motivated by the fact that one component of the system (ligands) is negatively charged, and such systems are not well described by means of the Kohn–Sham equations applying semilocal functionals. The orbital-free embedding potential given in eq 2 depends not only on the choice of the approximations used to evaluate its exchange-correlation- and kinetic-energy dependent components but also on the choice of the electron density ρ_{II} . All the reported numerical values were obtained from fully variational calculations in which both ρ_{II} and ρ_I are derived from the minimization of the total-energy bifunctional $E[\rho_I, \rho_{II}]$ in eq 2. Such a minimization is

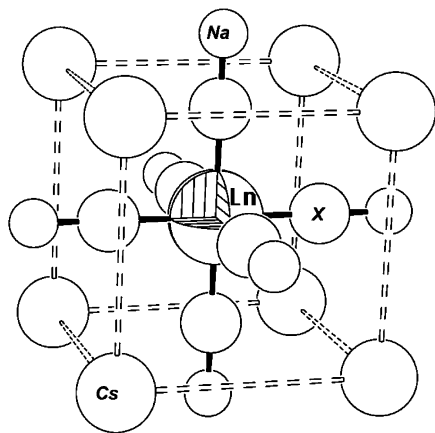


Figure 1. Schematic view on the environment of studied lanthanide cations. Each Ln^{3+} is hexacoordinated to six X^- ions (halides). The second coordination sphere comprises eight Cs^+ ions at the corners of the cube. The third coordination sphere comprises six Na^+ ions occupying the vertices of the octahedron.

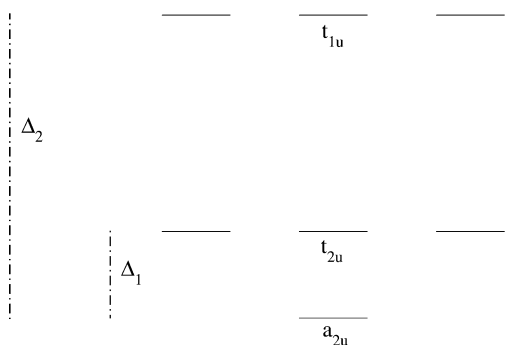


Figure 2. The f-orbital levels of Ln^{3+} in the octahedral environment.

performed by means of the “freeze-and-thaw” cycle of iterations described in ref 56.

The orbital-levels of an embedded lanthanide cation (Ln^{3+}) were obtained from eq 1 in which ρ_I corresponds to Ln^{3+} and ρ_{II} to the environment. The numerical implementation of eqs 1 and 2 into the Amsterdam Density Functional (ADF) package^{60,61} was used in all calculations. Relativistic scalar ZORA,^{57,58} all electron calculations were performed using the ZORA triple- ζ STO set plus one polarization function (ZORA/TZP).⁵⁹ Figure 1 shows the investigated system comprising the octahedral arrangement of the lanthanide cation Ln^{3+} and its ligands. O_h symmetry was assumed in all calculations.

Figure 2 shows the expected order of f-levels (a_{2u} , t_{2u} , and t_{1u}) and defines the two ligand-field splitting parameters Δ_1 and Δ_2 . The energy levels were calculated for average-of-configuration, in which each f-orbital was partially occupied (occupation number $(n/7)$) for a given f^n configuration. The occupations of orbitals used to express the electron density of the ligands (ρ_{II}) were chosen in such a way that the corresponding single-determinantal wave function possesses the full symmetry of the system. In some cases ($\text{Ln}=\text{Ce}$, Pr , Nd , and Sm in $\text{Cs}_2\text{NaLnX}_6$), the orbitals of the ligands were maximally filled (occupations given in Table. 1). The $N^{\text{orb}}A_{1,g}$

Table 1: Electronic Occupation Numbers of the Hexahalide Anions for Each Irreducible Representation of the O_h Symmetry

irreps/halides	$(\text{F}^-)_6$	$(\text{Cl}^-)_6$	$(\text{Br}^-)_6$	$(\text{I}^-)_6$
$A_{1,g}$	6	10	16	22
$A_{2,g}$	0	0	2	4
E_g	12	20	36	52
$T_{1,g}$	6	12	24	36
$T_{2,g}$	6	12	30	48
$A_{2,u}$	0	0	2	4
E_u	0	0	4	8
$T_{1,u}$	24	42	72	102
$T_{2,u}$	6	12	30	48
$N_{\text{electrons}}$	60	108	216	324

orbitals ($N^{\text{orb}} = 2, 4, 7$ and 8 corresponding to $\text{X}=\text{F}$, Cl , Br , and I in $\text{Cs}_2\text{NaLnX}_6$, respectively.) were, therefore, emptied.

3. Results and Discussion

This section comprises two parts. In the first one, the results of KSCED(s) and KSCED(m) calculations are compared in order to show the role of f-orbital delocalization on the calculated ligand-field splitting energies. The following section concerns the ligand-field splitting energies for a number of other elpasolites, for which either experimental ligand-field splitting were not accurately measured yet, or for materials which do not exist.

Table 2 collects the ligand-field splitting parameters Δ_1 and Δ_2 in lanthanide-containing chloroelpasolites $\text{Cs}_2\text{-NaLnCl}_6$ derived from KSCED(m) and KSCED(s) calculations (see also Figure 3). Experimental results are also given for the sake of comparison. Note that the Δ_1 and Δ_2 values given in refs 40 and 53 for Yb (220 and 799 cm^{-1} , respectively) are erroneous, and we use the correct ones (301 and 747 cm^{-1} , respectively) here. In the whole lanthanide series, lifting the localization assumption for embedded orbitals does not affect significantly the calculated values of Δ_1 . Both KSCED(m) and KSCED(s) results are very similar and agree very well with experiment. The experimental values of Δ_1 decrease almost monotonically in the whole series from 390 cm^{-1} (Ce) to 301 cm^{-1} (Yb). However, the dependence of the calculated values of Δ_1 on the number of f-electrons n_f is smoother than that deduced from experiment. The average and the maximal deviation from experimental data amount to 28 and 100 cm^{-1} (Sm) using the KSCED(m) scheme and 50 and 201 cm^{-1} (Ce) using KSCED(s), respectively. The corresponding mean absolute errors amount to 52 and 83 cm^{-1} .

Compared to Δ_1 , the effect of lifting the localization assumption on Δ_2 is different. For cations with $n_f > 7$, the KSCED(m) and KSCED(s) values are almost identical. For $n_f < 7$ cations, the possibility of delocalization increases the calculated Δ_2 parameter by about 100 cm^{-1} bringing the calculated values closer to the experimental data.

The average and the maximal deviation from the experimental data amount to 178 and 320 cm^{-1} (Eu) for KSCED(m) and 141 and 315 cm^{-1} (Ho) for KSCED(s), respectively. The corresponding mean absolute errors amount to 100 and 168 cm^{-1} .

Table 2: Experimental and Calculated Ligand-Field Splitting Parameters Δ_1 and Δ_2 (in cm^{-1}) Derived from KSCED(s) and KSCED(m) Calculations^a

		Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb
experiment	Δ_1	390	462	343		250	341		349	345	358	300	299	301
	Δ_2	1072	1172	988		803	973		840	808	865	764	790	747
eq 2	Δ_1	591	536	480	445	434	410	351	341	319	315	303	294	290
KSCED(s)	Δ_2	1122	1006	896	830	825	716	641	621	570	550	530	519	503
eq 2	Δ_1	478	435	392	365	350	329	312	312	291	291	279	266	273
KSCED(m)	Δ_2	929	855	773	725	696	653	620	629	576	577	553	528	537

^a Calculations were made at the ab initio optimized⁶³ ion–ligand distances.

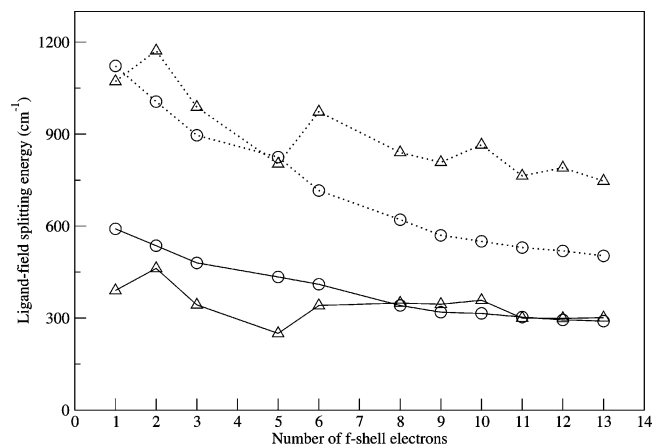


Figure 3. Ligand-field splitting parameters (Δ_1 and Δ_2) in the octahedrally coordinated lanthanide ions in $\text{Cs}_2\text{NaLnCl}_6$ elpasolites: the splitting energies calculated using effective embedding potential of eq 2 and the observed splitting energies. Calculations were made at the ab initio optimized cation–ligand distances taken from the literature.⁶³ Solid and dotted lines are used to indicate Δ_1 and Δ_2 parameters, respectively. Triangles and circles are used to guide the eye for experimental⁵² and calculated values using KSCED(s) schemes, respectively. The estimated error bars of experimental parameters are not shown because they are of the size of the applied symbols.

The small differences between the KSCED(m) and KSCED(s) results (Δ_1 and Δ_2) for the whole series of embedded lanthanide cations, indicate clearly that lifting the localization assumption does not affect significantly the orbital levels. In some cases, the agreement between the calculated and experimental ligand-field splitting parameters slightly improves. As measured by mean absolute errors in the whole lanthanide series, lifting the localization assumption leads only to a slight deterioration of the calculated splitting energies. It is worthwhile to stress at this point that the intersystem charge-flow possible in KSCED(s) calculations makes the KSCED embedding potential prone to possible flaws of the applied approximations in the relevant functionals.⁵¹ Moreover, the KSCED(s) results approach better the basis set limit for the applied method which is based on the variational principle. The remaining deviations between the KSCED calculated and experimental parameters should be attributed to other assumptions/approximations used in the applied computational scheme: the use of average-of-configurations and approximations for the exchange-correlation- and nonadditive-kinetic-energy potentials.

In the following part, the results were obtained for a number of other elpasolites for which either experimental splitting parameters were not accurately measured yet or do not exist.

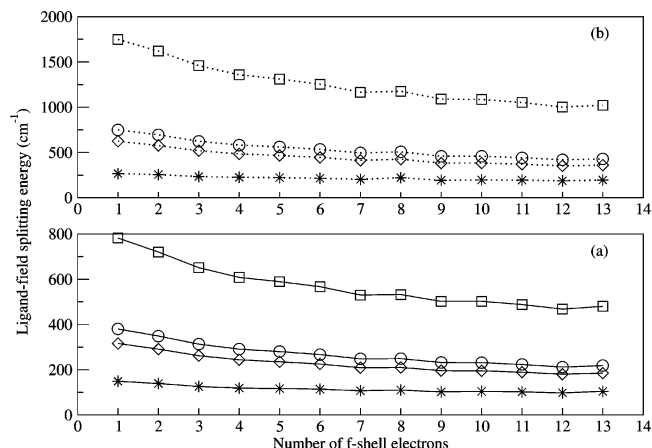


Figure 4. Ligand-field splitting parameters (Δ_1 and Δ_2) in the octahedrally coordinated lanthanide ions for the whole $\text{Cs}_2\text{NaLnX}_6$ elpasolites series ($X=\text{F}, \text{Cl}, \text{Br}, \text{I}$) from KSCED(m) calculations using the sum of ionic radii cation–ligand distances.⁶⁴ Solid and dotted lines are used to indicate (a) Δ_1 and (b) Δ_2 parameters, respectively. Squares, circles, diamonds, and stars are used to guide the eye for calculated values corresponding to LnF_6^{3-} , LnCl_6^{3-} , LnBr_6^{3-} , and LnI_6^{3-} , respectively.

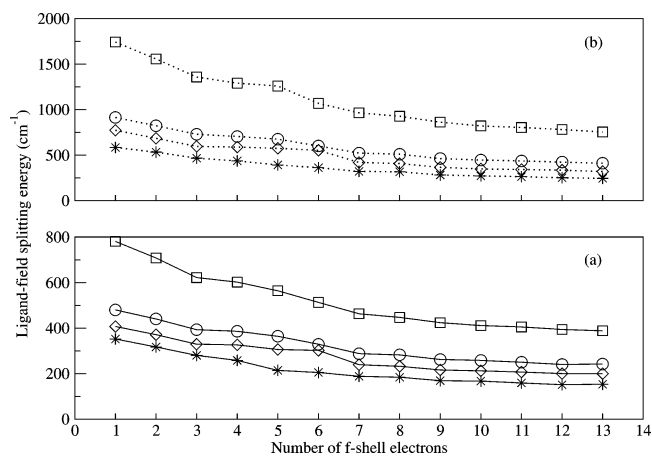


Figure 5. Ligand-field splitting parameters (Δ_1 and Δ_2) in the octahedrally coordinated lanthanide ions for the whole $\text{Cs}_2\text{NaLnX}_6$ elpasolites series ($X=\text{F}, \text{Cl}, \text{Br}, \text{I}$) from KSCED(s) calculations using the sum of ionic radii cation–ligand distances.⁶⁴ Solid and dotted lines are used to indicate (a) Δ_1 and (b) Δ_2 parameters, respectively. Squares, circles, diamonds, and stars are used to guide the eye for calculated values corresponding to LnF_6^{3-} , LnCl_6^{3-} , LnBr_6^{3-} , and LnI_6^{3-} , respectively.

Figures 4 and 5 show the calculated values of Δ_1 and Δ_2 for the whole series of $\text{Cs}_2\text{NaLnX}_6$ (Tables 3 and 4 collect the corresponding numerical values) derived from either KSCED(m) or KSCED(s) calculations). The ligand-field splitting energies calculated using both techniques increase

Table 3: Ligand-Field Splitting Parameters Δ_1 and Δ_2 (in cm^{-1}) from KSCED(m) Calculations Using the Sum of Ionic Radii Cation–Ligand Distances⁶⁴

	Ce ³⁺	Pr ³⁺	Nd ³⁺	Pm ³⁺	Sm ³⁺	Eu ³⁺	Gd ³⁺	Tb ³⁺	Dy ³⁺	Ho ³⁺	Er ³⁺	Tm ³⁺	Yb ³⁺
						Δ_1							
F ₆ ⁻	782	720	651	608	589	567	530	532	502	502	488	468	480
Cl ₆ ⁻	380	349	313	291	280	267	248	249	232	231	223	212	218
Br ₆ ⁻	316	291	262	244	235	225	209	210	196	195	189	180	185
I ₆ ⁻	149	139	126	119	117	114	107	110	102	104	102	98	104
						Δ_2							
F ₆ ⁻	1749	1620	1459	1358	1312	1254	1165	1176	1090	1086	1052	1001	1022
Cl ₆ ⁻	750	695	624	582	561	534	495	507	459	458	442	420	428
Br ₆ ⁻	624	577	519	484	467	445	413	426	383	383	370	352	360
I ₆ ⁻	267	256	234	225	222	215	203	221	193	197	194	187	195

Table 4: Ligand-Field Splitting Parameters Δ_1 and Δ_2 (in cm^{-1}) from KSCED(s) Calculations Using the Sum of Ionic Radii Cation–Ligand Distances⁶⁴

	Ce ³⁺	Pr ³⁺	Nd ³⁺	Pm ³⁺	Sm ³⁺	Eu ³⁺	Gd ³⁺	Tb ³⁺	Dy ³⁺	Ho ³⁺	Er ³⁺	Tm ³⁺	Yb ³⁺
						Δ_1							
F ₆ ⁻	781	708	622	602	564	513	463	447	424	411	405	394	388
Cl ₆ ⁻	480	440	393	386	364	329	288	282	262	258	250	240	242
Br ₆ ⁻	407	371	329	326	306	302	239	232	216	212	207	200	200
I ₆ ⁻	352	317	280	258	214	205	188	184	169	167	159	151	153
						Δ_2							
F ₆ ⁻	1742	1556	1358	1291	1259	1069	965	927	862	821	804	781	754
Cl ₆ ⁻	914	823	729	705	676	600	523	510	462	446	436	423	411
Br ₆ ⁻	773	685	596	589	575	553	420	407	366	348	342	335	320
I ₆ ⁻	584	533	466	437	392	362	321	317	282	272	263	252	245

in the expected order⁶² along the series $\text{I}^- < \text{Br}^- < \text{Cl}^- < \text{F}^-$. It is worthwhile to note that the ligand–cation ($\text{X}^- - \text{Ln}^{3+}$) distance increases along the series F, Cl, Br, and I. Except for iodide elpasolites $\text{Cs}_2\text{NaLnI}_6$, the numerical values derived from KSCED(s) and KSCED(m) calculations are very similar. This exceptional behavior of iodide elpasolites $\text{Cs}_2\text{NaLnI}_6$ results probably from the fact that iodine has the smallest electron affinity among the considered ligands. In view of the analysis concerning chloroelpasolites, the numerical values derived from KSCED(s) calculations are probably more accurate.

4. Conclusions

In this study, the ligand-field splitting parameters Δ_1 and Δ_2 obtained from orbital-free embedding calculations are reported. To take into account the f-orbital delocalization and the possibility of the ligand \leftrightarrow metal charge transfer, supermolecular expansion of basis sets functions was used for each subsystem. The results obtained previously⁵³ using selected atom-centered functions in the linear combination of atomic orbitals expansion of embedded orbitals (monomolecular expansions for ρ_I and ρ_{II}) are not affected for heavier lanthanides ($f_n > 7$) and are slightly improved for lighter ones ($f_n < 7$) in chloroelpasolites. Our calculations confirm that localizing the cation and ligand orbitals in different regions in space, an intuitive approximation applied in our previous work, is adequate because lifting this assumption does not affect the calculated parameters significantly. Nevertheless, the calculated difference between the t_{1u} and a_{2u} levels (Δ_2 parameter) is underestimated by about 200 cm^{-1} for cations with the f-shell more than half-filled. This underestimation is probably the result of the use of the “average-of-configuration” Ansatz or the inherent errors of the applied approximations for the effective potential in KSCED. The present analysis does not justify a more precise determination of the relative significance of

these two effects. Another possible source of deviations between the ligand-field parameters deduced from experiment and the calculated ones might be the result of their strong dependence ($r^{-5} - r^{-6}$) on the metal–ligand distances. In fact, the actual geometry in the crystal lattice might be different from the standard geometries applied in this work. The current study provides also predictions of the ligand-field splitting parameters for homologous materials: fluoroelpasolites $\text{Cs}_2\text{NaLnF}_6$, bromoelpasolites $\text{Cs}_2\text{NaLnBr}_6$, and iodoelpasolites $\text{Cs}_2\text{NaLnI}_6$. The KSCED(s) results are recommended because the additional atom-centered basis functions approach better the complete basis set, whereas their use was found to be numerically stable despite possible flaws in the used approximations for the orbital-free embedding potential given in eq 2.

Acknowledgment. This work is supported by the Swiss National Science Foundation. It is also part of the COST D26 Action.

References

- (1) Meyer, G. *Prog. Solid State Chem.* **1982**, *14* (3), 141–219.
- (2) Molander, G. A.; Romero, J. A. C. *Chem. Rev.* **2002**, *102*, 2161–2186.
- (3) Morss, L. R.; Fuger, J. *Inorg. Chem.* **1969**, *8*, 1433–1439.
- (4) Gudel, H. U.; Furrer, A.; Blank, H. *Inorg. Chem.* **1990**, *29*, 4081–4084.
- (5) Eldelmann, F. T.; Freckmann, D. M. M.; Schumann, H. *Chem. Rev.* **2002**, *102*, 1851–1896.
- (6) Eisenstein, O.; Hitchcock, P. B.; Khvostov, A. V.; Lappert, M. F.; Maron, L.; Perrin, L.; Protchenko, A. V. *J. Am. Chem. Soc.* **2003**, *125*, 10790–10791.
- (7) Aparna, K.; Ferguson, M.; Cavell, R. G. *J. Am. Chem. Soc.* **2000**, *122*, 726–727.
- (8) Tanner, P. A. *Mol. Phys.* **1985**, *58*, 317–328.

- (9) Schwartz, R. W. *Inorg. Chem.* **1977**, *16*, 1694–1697.
- (10) Case, D. A.; Lopez, J. P. *J. Chem. Phys.* **1983**, *80*, 3270–3277.
- (11) Roser, M. R.; Xu, J.; White, S. J.; Corruccini, L. R. *Phys. Rev. B* **1992**, *45*, 12337–12342.
- (12) Eisenstein, O.; Maron, L. *J. Organomet. Chem.* **2002**, *647*, 190–197.
- (13) Zhao, C. Y.; Wang, D.; Phillips, D. L. *J. Am. Chem. Soc.* **2003**, *125*, 15200–15209.
- (14) Gordon, J. C.; Giesbrecht, G. R.; Clark, D. L.; Hay, P. J.; Koegh, D. W.; Poli, R.; Scott, B. L.; Watkin, J. G. *Organometallics* **2002**, *21*, 4726–4734.
- (15) Clark, D. L.; Gordon, J. C.; Hay, P. J.; Martin, R. L.; Poli, R. *Organometallics* **2002**, *21*, 5000–5006.
- (16) Cao, X.; Dolg, M. *Mol. Phys.* **2003**, *101*, 2427–2435.
- (17) Luo, Y.; Selvam, P.; Ito, Y.; Endou, A.; Kubo, M.; Miyamoto, A. *J. Organomet. Chem.* **2003**, *679*, 84–92.
- (18) Jayasankar, C. K.; Richardson, F. S.; Tanner, P. A.; Reid, M. F. *Mol. Phys.* **1987**, *61*, 635–644.
- (19) Reid, M. F.; Richardson, F. S.; Tanner, P. A. *Mol. Phys.* **1986**, *60*, 881–886.
- (20) Falin, M. L.; Latypov, V. A.; Kazakov, B. N.; Leushin, A. M.; Bill, H.; Lovy, D. *Phys. Rev. B* **2000**, *61*, 9441–9448.
- (21) Foster, D. R.; Reid, M. F.; Richardson, F. S. *J. Chem. Phys.* **1985**, *83*, 3225–3233.
- (22) Tanner, P. A.; Yulong, L.; Edelstein, N. M.; Murdoch, K. M.; Khaidukov, N. M. *J. Phys.* **1997**, *9*, 7817–7836.
- (23) Berry, A. J.; McCaw, C. S.; Morisson, I. D.; Denning, R. G. *J. Lumin.* **1996**, *66*, 272–277.
- (24) McCaw, C. S.; Murdoch, K. M.; Denning, R. G. *Mol. Phys.* **2002**, *101*, 427–438.
- (25) Denning, R. G.; Berry, A. J.; McCaw, C. S. *Phys. Rev. B* **1997**, *57*, 2021–2024.
- (26) Tanner, P. A.; Kumar, V. V. R. K.; Jayasanka, C. K.; Reid, M. F. *J. Alloys Compd.* **1994**, *215*, 349–370.
- (27) Tanner, P. A.; Mak, C. S. K.; Edelstein, N. M.; Murdoch, K. M.; Liu, G.; Huang, J.; Seijo, L.; Barandiaran, Z. *J. Am. Chem. Soc.* **2003**, *125*, 13225–13233.
- (28) Tanner, P. A.; Mak, C. S. K.; Faucher, M. D. *J. Chem. Phys.* **2001**, *114*, 10860–10871.
- (29) Tanner, P. A.; Chua, M.; Reid, M. F. *J. Alloys Compd.* **1995**, *225*, 20–23.
- (30) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864–B871.
- (31) Kohn, W.; Sham, L. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- (32) Jones, R. O.; Gunnarsson, O. *Rev. Mod. Phys.* **1989**, *61*, 689–746.
- (33) Jiang, L.; Xu, Q. *J. Phys. Chem. A* **2006**, *110*, 5636–5641.
- (34) Yakuphanoglu, F.; Atalay, Y.; Erol, I. *Mol. Phys.* **2005**, *103*, 3309–3314.
- (35) Otani, M.; Okada, S.; Oshiyama, A. *Phys. Rev. B* **2003**, *68*, 125424.
- (36) Wang, S. G.; Pan, D. K.; Schwarz, W. H. E. *J. Chem. Phys.* **1995**, *102*, 9296–9307.
- (37) Forstreuter, J.; Steinbeck, L.; Richter, M.; Eschrig, H. *Phys. Rev. B* **1997**, *55*, 9415–9421.
- (38) Said, M.; Zid, F. B.; Bertoni, C. M.; Ossicini, S. *Eur. Phys. J. B* **2001**, *23*, 191–199.
- (39) Gutierrez, F.; Rabbe, C.; Poteau, R.; Daudey, J. P. *J. Phys. Chem. A* **2005**, *109*, 4325–4330.
- (40) Atanasov, M.; Daul, C.; Gudel, H. U.; Wesolowski, T. A.; Zbiri, M. *Inorg. Chem.* **2005**, *44*, 2954–2963.
- (41) Liu, W.; Hong, G.; Dai, D.; Li, L.; Dolg, M. *Theor. Chem. Acc.* **1997**, *96*, 75–83.
- (42) Sommerfeld, A.; Welker, H. *Ann. Phys.* **1938**, *32*, 56–65.
- (43) Schäffer, C. E. *Mol. Phys.* **1965**, *9*, 401–412.
- (44) Umland, W. *Chem. Phys.* **1976**, *14*, 393–401.
- (45) Yang, W. *Phys. Rev. Lett.* **1991**, *66*, 1438–1441.
- (46) Yang, W. *Phys. Rev. A* **1991**, *44*, 7823–7826.
- (47) Cortona, P. *Phys. Rev. B* **1991**, *44*, 8454–8458.
- (48) Wesolowski, T. A.; Warshel, A. *J. Chem. Phys.* **1993**, *97*, 8050–8053.
- (49) Wesolowski, T. A. *J. Chem. Phys.* **1997**, *106*, 8516–8526.
- (50) Kevorkiants, R.; Dulak, M.; Wesolowski, T. A. *J. Chem. Phys.* **2006**, *124*, 024104.
- (51) Dulak, M.; Wesolowski, T. A. *J. Chem. Phys.* **2006**, *124*, 164101.
- (52) Foster, D. R.; Reid, M. F.; Richardson, F. S. *J. Chem. Phys.* **1985**, *83*, 3813–3830.
- (53) Zbiri, M.; Atanasov, M.; Daul, C.; Lastra, J.; Wesolowski, T. A. *Chem. Phys. Lett.* **2004**, *397*, 441–446.
- (54) Perdew, J.; Chevary, J.; Vosko, S.; Jackson, K.; Pederson, M.; Singh, D.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–6687.
- (55) van Leeuwen, R.; Baerends, E. *Phys. Rev. A* **1994**, *49*, 2421–2431.
- (56) Wesolowski, T. A.; Weber, J. *Chem. Phys. Lett.* **1996**, *248*, 71–76.
- (57) van Lenthe, E.; Snijders, J. G.; Baerends, E. J. *J. Chem. Phys.* **1996**, *105*, 6505–6516.
- (58) van Lenthe, E.; van Leeuwen, R.; Baerends, E. J.; Snijders, J. G. *Int. J. Quantum Chem.* **1996**, *57*, 281–293.
- (59) Lenthe, E. V.; Baerends, E. J. *J. Comput. Chem.* **2003**, *24*, 1142–1156.
- (60) ADF, A. d. f. p. *Theoretical Chemistry*; Vrije Universiteit: Amsterdam, 2005; URL: <http://www.scm.com>.
- (61) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Guerra, C. F.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967.
- (62) Denning, R. G.; Berry, A. J.; McCaw, C. S. *Phys. Rev. B* **1998**, *57* (4), 2021–2024.
- (63) Ordejón, B.; Seijo, L.; Barandiarán, Z. *J. Chem. Phys.* **2003**, *119*, 6143–6149.
- (64) Shanon, R. D. *Acta Crystallogr. A* **1976**, *32*, 751–767.

Simulation of Actuation by Polymeric Polyelectrolyte Helicenes

Pawel Rempala and Benjamin T. King*

University of Nevada, Reno, Department of Chemistry/216, Reno, Nevada 89557

Received March 20, 2006

Abstract: The potential of several peripherally substituted [6.3.1] helicenes to serve as linear actuators was investigated using molecular dynamics calculations. Reversible extension upon ionization of pendant functionality was observed in three of four cases. The largest extensions were obtained for molecules with amino groups or ionized phosphate groups attached directly to the helical backbone (extensions of $176 \pm 4\%$ and $184 \pm 4\%$, respectively). Electrostatic forces and swelling drive the actuation.

Introduction

Motion is a fundamental physical phenomenon. Indeed, the management and utilization of motion at the molecular level is an emerging theme in nanotechnology. Internal rotational motion in molecules and its potential use in molecular machines attracts considerable attention, as evidenced by a recent review article.¹ Interest in translational motion is also apparent, for example, research on rotaxanes as molecular shuttles.² Rotation may induce translation, as occurs in biological muscles. Of course, the cumulative function of molecular objects can produce macroscopic motion, as in artificial muscles.³ Chemically driven artificial muscles have been demonstrated. For example, a film of triblock copolymer with hydrophobic ends [poly(methyl methacrylate)] and a midblock of poly(methacrylic acid) exhibited reversible actuation driven by changing pH.⁴

Biological molecular motors, which transform chemical energy into motion (e.g., the myosin–actin system⁵), are complex, and their synthetic imitation is daunting. We propose a synthetically feasible molecular actuator based on a springlike [6.3.1] helicene (we use Balaban's nomenclature⁶ because IUPAC nomenclature is inadequate for this family of helicenes) with peripheral functionality (Figure 1). Ionization, as the result of a chemical reaction of peripheral functionality (Figure 2), could induce actuation.

The proposed systems possess the unusual combination of properties necessary for effective actuation: a high aspect ratio, elasticity, shape persistence, and the ability to change

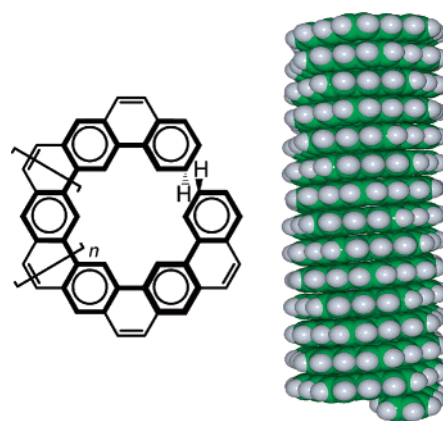


Figure 1. [6.3.1] helicene investigated as prospective actuator backbone.

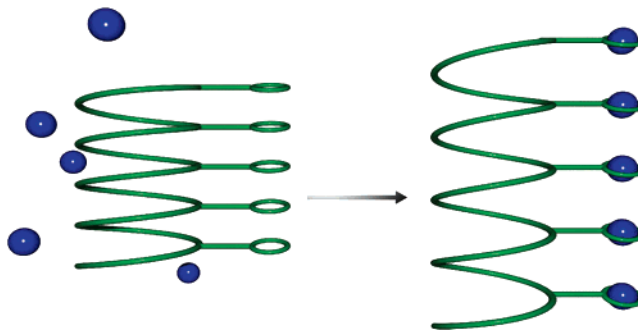


Figure 2. General concept of a chemically driven helical molecular actuator. Ions are represented by blue spheres.

chemical state. Most shape-persistent, high aspect ratio molecules, for example, poly(*p*-phenylene) and carbon nano-

* Corresponding author fax: (775) 784-6804; e-mail: king@chem.unr.edu.

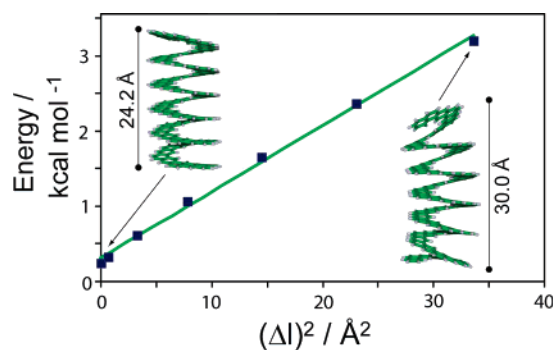


Figure 3. Energy of [6.3.1] helicene as a function of the square of extension from the equilibrium length.

tubes, are not easily stretched. Most polyelectrolytes and dopable polymers are not shape-persistent.

Elasticity, which is fundamental to actuation, is commonly expressed using Young's modulus (E). This intrinsic measure of elasticity is defined by eq 1, where A is the cross section area, l_0 is the unperturbed length, and l is the length when force F is applied.

$$E = Fl_0/A(l - l_0) = Fl_0/A\Delta l \quad (1)$$

We treat the helicene as a solid rod with a constant cross section. Energy as a function of length was obtained using the semiempirical PM3 Hamiltonian. For a spring obeying Hooke's law, the elastic potential energy, V , is described by eq 2, where k is the force constant. Hence, the total energy of the system (H_{tot}) should be a linear function of $(\Delta l)^2$, with a slope equal to $k/2$ (eq 3). Substituting $k\Delta l$ for F in eq 1 relates Young's modulus (E), the force constant, and the geometry (eq 4).

$$V = k(l - l_0)^2/2 = k(\Delta l)^2/2 \quad (2)$$

$$H_{\text{tot}} = H_0 + V = H_0 + k(\Delta l)^2/2 \quad (3)$$

$$E = (k\Delta l)l_0/A\Delta l = kl_0/A \quad (4)$$

Application of this protocol provided a Young's modulus of 0.16 GPa, compared to 7.5 GPa estimated for a conventional [6.2.1] helicene⁷ (a [n]helicene in IUPAC nomenclature). To put these values into context, the Young's modulus of single-walled carbon nanotubes is ~ 1000 GPa,⁸ steel is ~ 200 GPa, rubber is 0.01–0.1 GPa, and an α -helix peptide (poly-L-glutamic acid) is ~ 3 GPa.⁹ The use of a single end-to-end distance constraint in these optimizations resulted in slightly bent geometries (Figure 3), but the expected linear relationship of energy versus $(\Delta l)^2$ held.

Our initial calculations in vacuo (PM3 and molecular mechanics) on charged helicenes indicated, not surprisingly, severalfold expansion as compared to that of neutral molecules. This vacuum treatment was unrealistic for many reasons. First, it was physically unreasonable—the real systems will operate in the condensed phase and will be electrically neutral. Second, if the system is treated as a set of collinear point charges fixed at even intervals, the electrostatic repulsion energy *per charge (monomer)* in-

creases with the chain size without an upper bound. However, in the opposite limit, a disordered distribution of charge in an electroneutral system does not produce large potentials or fields.¹⁰ For these reasons, the modeling of an intermediate case, in which charges are organized around a shape-persistent yet elastic backbone and counterions and a solvent are present, should be physically reasonable and could demonstrate actuation.

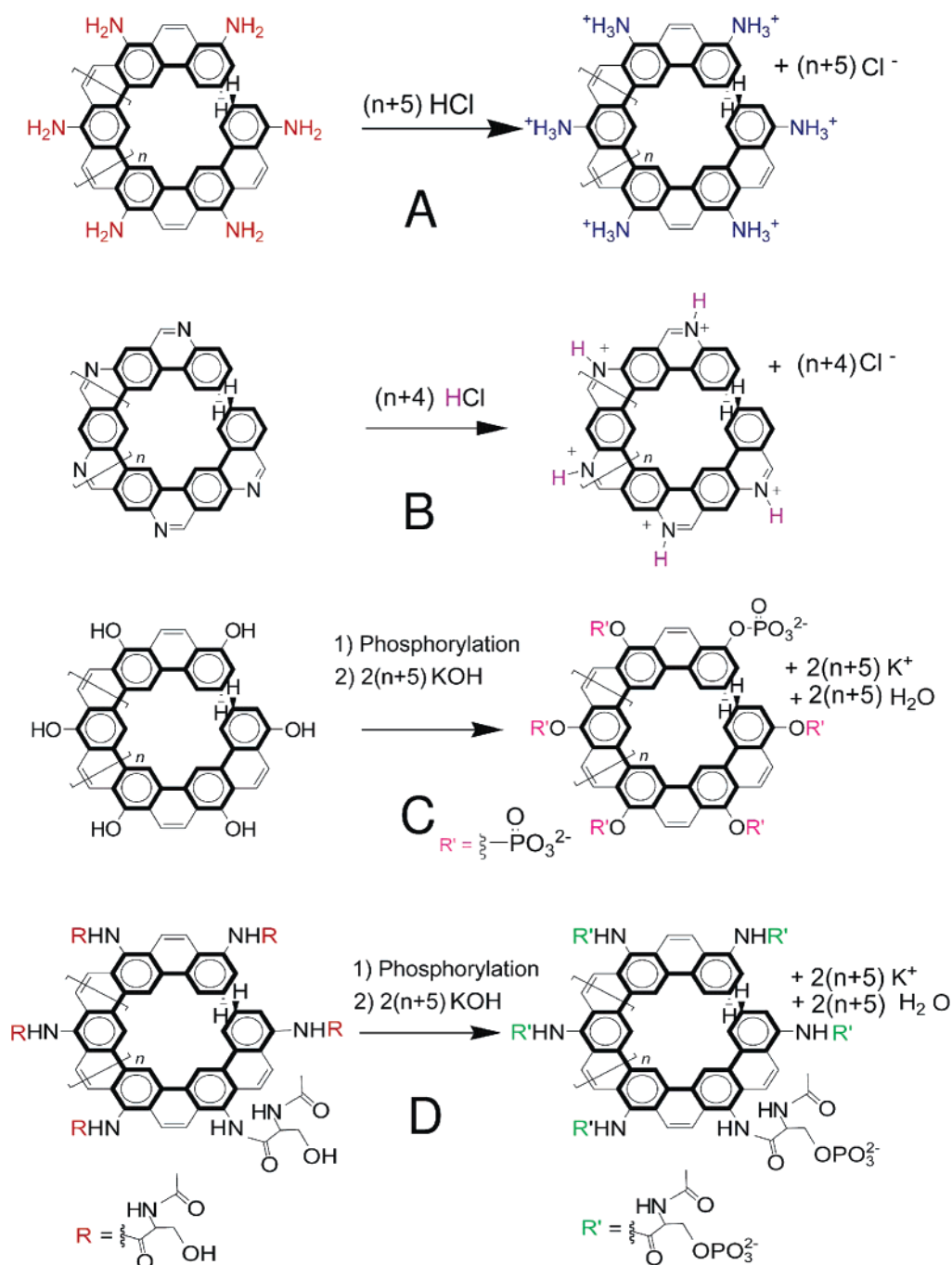
Methods

Molecular dynamics calculations were performed and analyzed using the Amber 7 suite of programs¹¹ (Leap, Sander, Carnal, and Antechamber¹²). The molecules studied are shown in Scheme 1 ($n = 31$). Both neutral and charged forms of the molecules were prepared in compact and extended conformations, and then atomic charges were assigned (see the Supporting Information for details). The charged forms were neutralized with Cl^- or K^+ ions during input preparation. Rectangular periodic boxes of TIP3P water¹³ molecules were added next (approximately 5000–7000 water molecules). AMBER¹⁴ and GAFF (General Amber Force Field)¹⁵ force fields were applied to helicenes in the calculations. All calculations with TIP3P water were run at isobaric and isothermal conditions (target pressure of 1 bar, compressibility of water was assumed, target temperature 298 K) with a 1 fs integration step. To test whether swelling induced by the finite size of water molecules is important, some simulations were run using a generalized Born solvation model.^{16,17} In this model, hydrophobic effects are represented using a surface energy term (gbsa=1 option, AMBER atom type). The screening effect of counterions was included by setting the monovalent salt concentration to 0.1 M (saltcon = 0.1, and also saltcon = 0.0 for comparison). Force field modifications, where a few missing torsional parameters were assumed to be identical to the available parameters based on chemical similarity, are given in the Supporting Information.

Results

All of the molecules investigated share a common backbone design. Their side chains were selected on the basis of synthetic accessibility and the ability to ionize (acid–base chemistry for molecules **A** and **B** and phosphorylation followed by deprotonation in the cases of **C** and **D**). In cases **A** and **B**, basic groups close to the backbone should result in a concentration of positive charge at the edge of the oligomer helix. In the oligoacid derived from the phosphorylation of oligophenol **C**, deprotonation will yield doubly charged negative groups, so even greater repulsion and expansion might be expected. System **D** has long and flexible *N*-acetylphosphoserine side chains, which further separate doubly charged phosphate groups from the backbone.

Phosphorylation, the first step in the biomimetic actuation mechanism of **C** and **D**, is ubiquitous in metabolism and is responsible for energy transformation and storage, enzymatic regulation, and signaling.¹⁸ Kinases (phosphotransferases) catalyze the transfer of a phosphoryl group (terminal phosphoryl group of ATP) to acceptors such as hydroxyl, carboxy, or other phosphate groups.¹⁸ Kinases capable of an indis-

Scheme 1. Potential Actuators Studied

criminate phosphorylation of proteins are available.¹⁹ Phosphorylation of the phenolic OH group in **C** and of the serine residues in **D** might be feasible under enzymatic catalysis, using ATP as a donor of the phosphate and an energy source.

The results for system **A** with explicit water are shown in Figure 4. Starting from a compact geometry, the neutral amino form of **A** remained compact over the time span of simulation (300 ps). Starting from a highly extended geometry, the neutral amino form of **A** contracts within 60 ps to the compact geometry. The attainment of the compact geometry from both compact and extended forms demonstrates that the equilibrium geometry of the neutral amino form of **A** is compact, with an end-to-end length of 20 Å. A similar set of calculations was performed on the protonated form of **A**. Starting from a compact geometry, the protonated

form of **A** extends within 60 ps to an extended geometry. Starting from a highly extended geometry, the protonated form of **A** partially contracts within 60 ps to an extended geometry. This demonstrates that the equilibrium geometry of protonated **A** is extended, with an end-to-end length of 36 Å. The protonated form of **A** is $(35.7 \text{ Å}/20.3 \text{ Å}) = 176\%$ longer than the unprotonated form.

In the neutral molecule, favorable van der Waals attraction between hydrophobic hydrocarbon surfaces of the helix and minimization of the dihedral strain result in a compact structure. This compact structure, in which the tiers of atoms buttress one another, exhibits only small thermal variations in length. In the extended protonated form of **A**, which lacks the buttressing of the compact form, the end-to-end distance fluctuates somewhat.

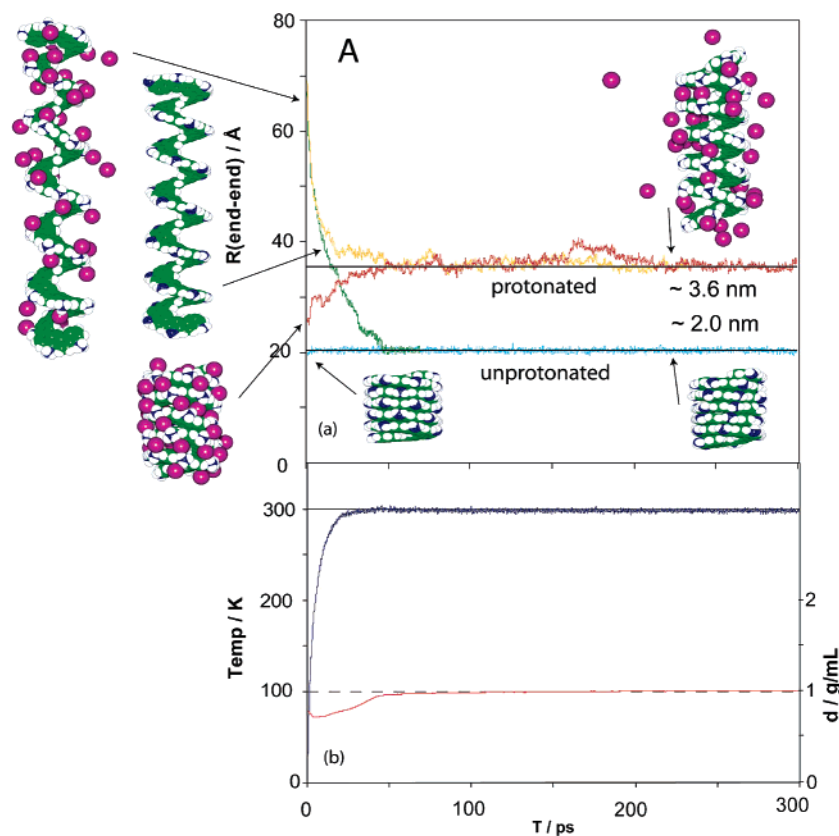


Figure 4. Summary of molecular dynamic simulation for system **A**, TIP3P water, periodic boundary conditions; chloride counterions are represented as purple spheres: (a) end–end distance as function of time and (b) evolution of temperature (blue) and density (red) for contraction of extended neutral form.

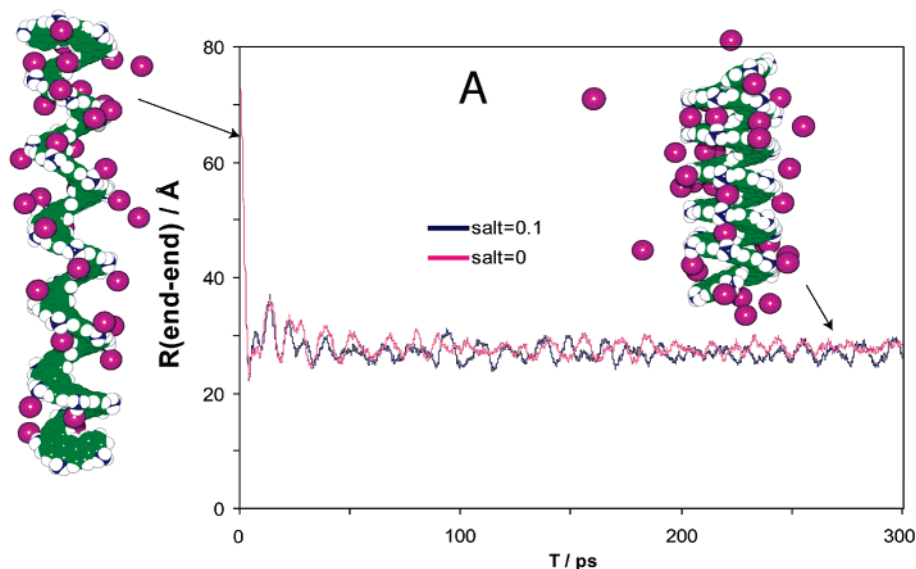


Figure 5. Results of molecular dynamic simulation for system **A** using a generalized Born solvation model. Chloride counterions are represented as purple spheres.

It is tempting to speculate, on the basis of trajectory snapshots, that the intercalation of chloride counterions stabilizes the extended geometry of protonated **A**. Inductive charge delocalization may also contribute to the extension by decreasing the hydrophobic forces on the helicene.

The chemical reactions that trigger actuation are not included in our simulations. The extended basic structure could arise from the rapid deprotonation of an extended oligo

ammonium salt. Evidence exists that the inclusion of proton exchange dynamics can affect the molecular dynamic simulations of peptides.²⁰

Two underlying mechanisms are responsible for the extension: electrostatic repulsion of the charged pendant groups and swelling induced by ions and their solvation. Simulations using a generalized Born solvation model, where the solvent was represented implicitly as a zero-viscosity

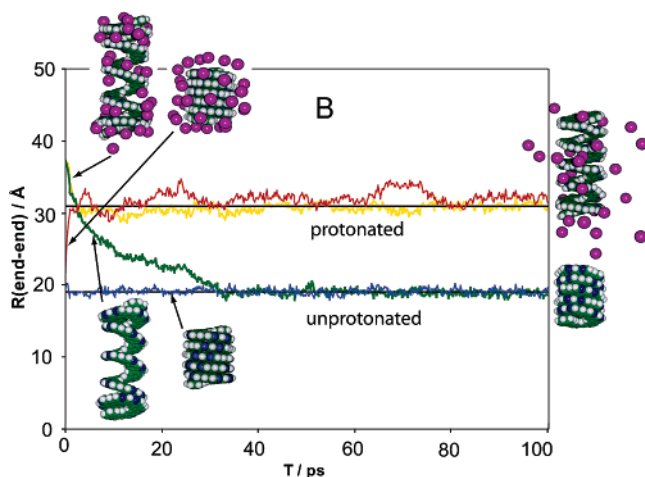


Figure 6. Summary of molecular dynamic simulation for system **B**. Chloride counterions are represented as purple spheres.

continuum with dielectric properties of water and an instantaneous dielectric response, revealed the relative contributions of these underlying phenomena. This model represents an aqueous electrolyte solution, but without the molecular bulk. If full extension occurs in this model, the mechanism of extension must be predominantly electrostatic repulsion, not swelling. If extension does not occur, the mechanism of extension is swelling. Two simulations (0 and 0.1 M salt concentrations) starting from the extended form of protonated **A** were performed. Chloride counterions provided electrical neutrality (Figure 5). The resulting equilibrium length (27.7 Å, 0.98 Å RMSD, 0 salt concentration, averaged over 100–300 ps interval) is between the equilibrium lengths of the simulation with explicit water molecules (35.7 Å) and those using the uncharged helicene (20.3 Å). Both mechanisms operate.

The comparison of the two solvation models requires addressing a few technical points. The generalized Born

solvation model does not use periodic boundary conditions, which results in more accessible volume in the simulation. This permits the chloride ions to drift far from the helicene. For both 0 and 0.1 M salt solutions, the length oscillated on the ~ 10 ps time scale, which is likely due to limited degrees of freedom in this rigid system lacking solvent molecules.

System **B** is similar to system **A**, except that the basic nitrogen sites are incorporated into the backbone (i.e., pyridine versus aniline). The results are summarized in Figure 6.

The behavior of **B** is similar to that of **A**, with an actuation of $31.0 \text{ \AA} / 19.0 \text{ \AA} = 163\%$. This system is also a viable synthetic target.

System **C** relies on a bioinspired two-stage actuation process. A parent helicene with pendant phenolic hydroxyl groups is phosphorylated, and then the resulting polyacid is deprotonated. The simulations were organized in the same manner as for systems **A** and **B**—approaching equilibrium from both directions for both the neutral and charged forms. The results are summarized in Figure 7. At equilibrium, actuator **C** exhibits similar behavior as actuator **A**—the neutral form is compact and the charged form is extended. Indeed, the extent of actuation, $37.1 \text{ \AA} / 20.2 \text{ \AA} = 184\%$, is comparable.

The charged form of **C** shows greater variations in length than charged **A** and **B**. It is also important to note that this two-stage process, which increases molecular bulk [OP(O)(OH)₂ vs OH] and charge (0 vs -2 per unit), does not increase the extension substantially.

System **D** is another bioinspired example which uses the same two-stage process, except, instead of simple hydroxyl groups, *N*-acetylserine side chains are used, which are more amenable to enzymatic phosphorylation. The simulations were performed as before: the equilibrium geometry of the charged and uncharged systems are approached from compact and extended forms. The results are summarized in Figure 8.

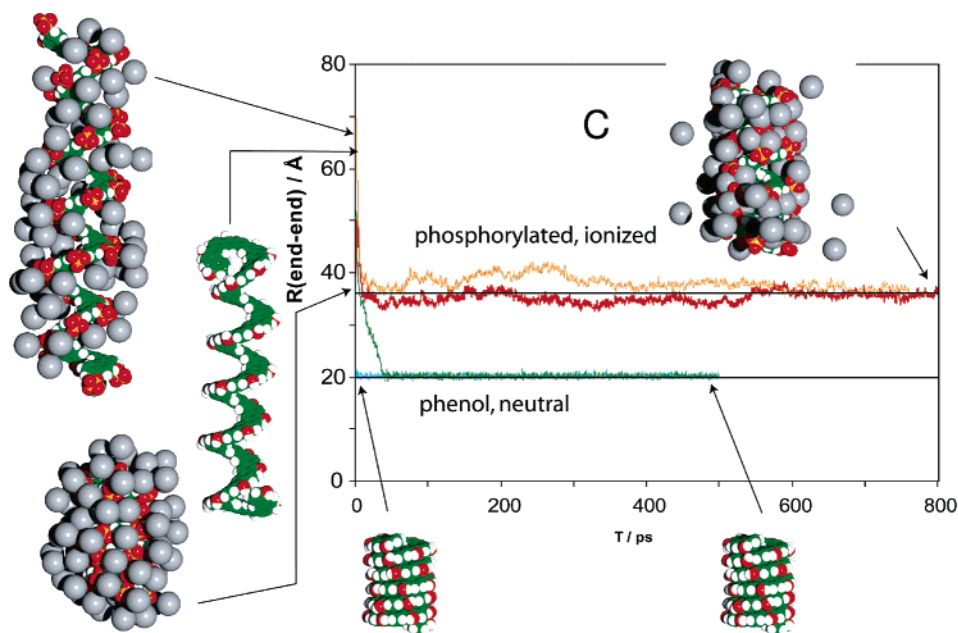


Figure 7. Summary of molecular dynamic simulation for system **C**. Potassium counterions are represented as grey spheres.

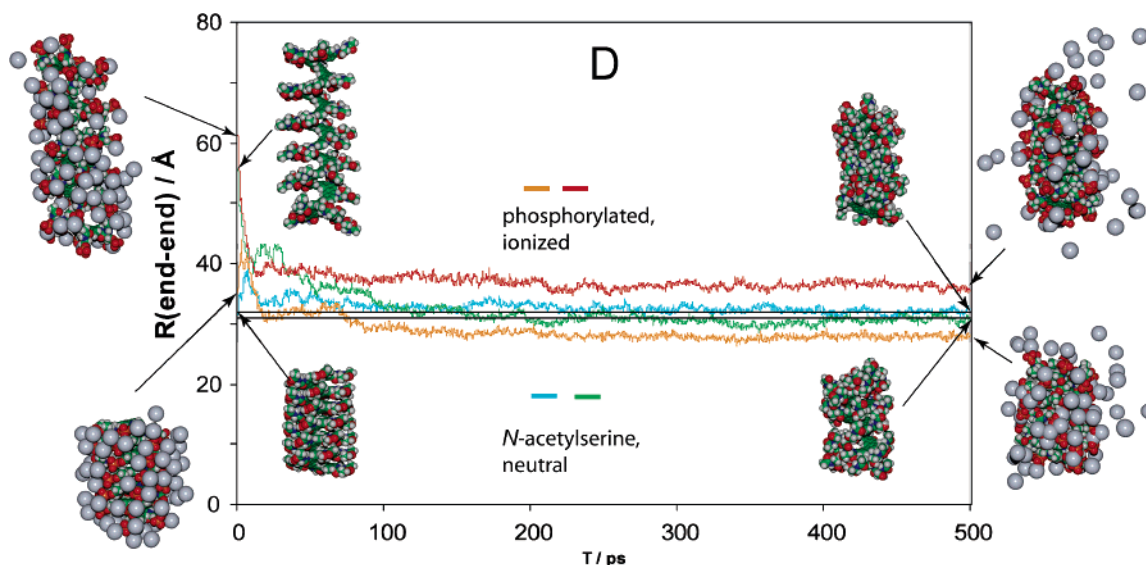


Figure 8. Summary of molecular dynamic simulation for system **D**.

Table 1. Results of MD Simulations for the Charged and Neutral Systems

system	neutral form		charged form		expansion, %	uncertainty, ^e %
	end–end average/Å	RMSD/Å	end–end average/Å	RMSD/Å		
A ^a	20.32	0.31	35.66	0.63	175.5	4.1
B ^b	18.98	0.35	31.03	0.50	163.5	4.0
C ^c	20.19	0.31	37.06	0.72	183.6	4.5
D ^d	31.27	1.03	32.05	4.31	102	14

^a Neutral (base): average over 200–300 ps interval, compact starting geometry. Charged (hydrochloride): average over 200–300 ps, extended starting geometry. ^b Neutral (base): average over 200–300 ps interval, extended starting geometry. Charged (hydrochloride): average over 200–300 ps, extended starting geometry. ^c Neutral (phenol): average over 300–500 ps interval, extended starting geometry. Charged (phosphate potassium salt): average over 550–750 ps, extended starting geometry. ^d Neutral (amino acid): average over 300–500 ps for both trajectories (starting from compact and extended geometries) treated as one data population. Charged (phosphorylated, potassium salt) averaging same as for neutral. The trajectories in system **D** did not converge. ^e RMSDs for end–end distances treated as standard deviations to estimate uncertainty of length ratios.

System **D** does not exhibit actuation within 500 ps. Indeed, equilibrium is not achieved after 500 ps for both the charged and neutral forms, as the lengths depend on the initial geometry. In any case, actuation is not observed after 500 ps, whereas systems **A**, **B**, and **C** exhibit pronounced actuation within 60 ps. It is unlikely that system **D** could serve as a fast response actuator.

System **D** cannot actuate because it cannot contract—contraction is prevented by the steric bulk of the *N*-acetylserine tethers. Indeed, the equilibrium length of the neutral form of system **D** (31.3 Å) is similar to the equilibrium length of the charged forms of systems **A**, **B**, and **C** (35.7, 31.0, and 37.0 Å, respectively).

Conclusion

Our concept of a [6.3.1]-helicene-based, chemically driven molecular actuator seems viable on the basis of molecular dynamics calculations. A summary of results as ratios of lengths at the ionized and neutral states is given in Table 1. The steric bulk of the side chains reduces efficiency dramatically, and doubly charged groups do not improve efficiency compared to singly charged groups. The simple systems **A** and **B** and the biomimetic system **C** are good candidates for molecular actuators. Because the final states of the systems **A**, **B**, and **C** do not depend on the initial state, the actuation is reversible. Both electrostatic repulsion and swelling contribute to actuation.

Acknowledgment. We are grateful for generous support from the Department of Energy, the National Science Foundation (CHE- 0449740), the University of Nevada, and the Office of Naval Research.

Supporting Information Available: Trajectory animations for the contraction of uncharged system **A** and the extension of charged system **A**, details of modulus calculations, details of dynamics calculations, and modifications to the force field. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Kottas, G. S.; Clarke, L. I.; Horinek, D.; Michl, J. *Chem. Rev.* **2005**, *105*, 1281–1376.
- (2) Jang, S. S.; Jang, Y. H.; Kim, Y.-H.; Goddard, W. A., III; Choi, J. W.; Heath, J. R.; Laursen, B. W.; Flood, A. H.; Stoddart, J. F.; Nørgaard, K.; Bjørnholm, T. *J. Am. Chem. Soc.* **2005**, *127*, 14804–14816.
- (3) Lin, K.-J.; Fu, S.-J.; Cheng, C.-Y.; Chen, W.-H.; Kao, H.-M. *Angew. Chem., Int. Ed.* **2004**, *43*, 4186–4189.
- (4) Howse, J. R.; Topham, P.; Crook, C. J.; Gleeson, A. J.; Bras, W.; Jones, R. A. L.; Ryan, A. J. *Nano Lett.* **2006**, *6*, 73–77.
- (5) Piazzesi, G.; Reconditi, M.; Linari, M.; Lucii, L.; Sun, Y.-B.; Narayanan, T.; Boesecke, P.; Lombardi, V.; Irving, M. *Nature* **2002**, *415*, 659–662.

- (6) Balaban, A. T. *Polycyclic Aromat. Compd.* **2003**, *23*, 277–296.
- (7) Jalaie, M.; Weatherhead, S.; Lipkowitz, K. B.; Robertson, D. *Electron. J. Theor. Chem.* **1997**, *2*, 268–272.
- (8) Salvétat, J.-P.; Briggs, G. A. D.; Bonard, J.-M.; Bacsá, R. R.; Kulik, A. J.; Stöckli, T.; Burnham, N. A.; Forró, L. *Phys. Rev. Lett.* **1999**, *82*, 944–947.
- (9) Idiris, A.; Alam, M. T.; Ikai, A. *Protein Eng., Des. Sel.* **2000**, *13*, 763–770.
- (10) Singer, A.; Schuss, Z.; Eisenberg, R. S. *J. Stat. Phys.* **2005**, *119*, 1397–1418.
- (11) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Wang, J.; Ross, W. S.; Simmerling, C.; Darden, T.; Merz, K. M.; Stanton, R. V.; Cheng, A.; Vincent, J. J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P.; Kollman, P. A. *Amber 7*; University of California: San Francisco, CA, 2002.
- (12) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, in press. <http://amber.scripps.edu/antechamber/antechamber.html> (accessed September 8, 2004).
- (13) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (14) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (15) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (16) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122–129.
- (17) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.
- (18) Horton, H. R.; Moran, L. A.; Ochs, R. S.; Rawn, J. D.; Scrimgeour, K. G. *Principles of Biochemistry*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, 1996.
- (19) Hubbard, S. R.; Till, J. H. *Annu. Rev. Biochem.* **2000**, *69*, 373–398.
- (20) Długosz, M.; Antosiewicz, J. M. *J. Phys. Chem. B* **2005**, *109*, 13777–13784.

CT600102R

QM/MM Models of the O₂-Evolving Complex of Photosystem II

Eduardo M. Sproviero, José A. Gascón, James P. McEvoy, Gary W. Brudvig, and Victor S. Batista*

Department of Chemistry, Yale University, P.O. Box 208107,
New Haven, Connecticut 06520-8107

Received January 11, 2006

Abstract: This paper introduces structural models of the oxygen-evolving complex of photosystem II (PSII) in the dark-stable S₁ state, as well as in the reduced S₀ and oxidized S₂ states, with complete ligation of the metal–oxo cluster by amino acid residues, water, hydroxide, and chloride. The models are developed according to state-of-the-art quantum mechanics/molecular mechanics (QM/MM) hybrid methods, applied in conjunction with the X-ray crystal structure of PSII from the cyanobacterium *Thermosynechococcus elongatus*, recently reported at 3.5 Å resolution. Manganese and calcium ions are ligated consistently with standard coordination chemistry assumptions, supported by biochemical and spectroscopic data. Furthermore, the calcium-bound chloride ligand is found to be bound in a position consistent with pulsed electron paramagnetic resonance data obtained from acetate-substituted PSII. The ligation of protein ligands includes monodentate coordination of D1-D342, CP43-E354, and D1-D170 to Mn(1), Mn(3), and Mn(4), respectively; η² coordination of D1-E333 to both Mn(3) and Mn(2); and ligation of D1-E189 and D1-H332 to Mn(2). The resulting QM/MM structural models are consistent with available mechanistic data and also are compatible with X-ray diffraction models and extended X-ray absorption fine structure measurements of PSII. It is, therefore, conjectured that the proposed QM/MM models are particularly relevant to the development and validation of catalytic water-oxidation intermediates.

1. Introduction

The photosynthetic water-oxidation reaction in the thylakoid membranes of cyanobacteria and green-plant chloroplasts releases O₂(g) into the atmosphere according to the four-electron water-splitting reaction



The water-oxidation reaction, given by eq 1, is catalyzed by the so-called oxygen-evolving-complex (OEC) of photosystem II (PSII). This paper develops chemically sensible models of the OEC of PSII in which the Mn₃CaO₄Mn cluster is completely ligated by amino acid residues, water, hydroxide, and chloride ions. State-of-the-art quantum mechanics/

molecular mechanics (QM/MM) hybrid methods^{1–3} are applied in conjunction with the X-ray crystal structure of PSII from the cyanobacterium *Thermosynechococcus elongatus*,⁴ explicitly addressing the perturbational influence of the surrounding protein environment on the structural and electronic properties of the OEC. The resulting structural models are analyzed by comparison to a large body of experimental data and mechanistic hypotheses of photosynthetic oxygen evolution.⁵

In contrast to chemical and electrochemical water oxidation reactions, which are thermodynamically highly demanding, the OEC-catalyzed water-splitting mechanism proceeds with very little driving force and requires only moderate activation energies.^{6–9} Moreover, PSII turns over very rapidly, producing up to 50 dioxygen molecules per second. The high efficiency of the reaction has motivated extensive spectro-

* Corresponding author fax: (203) 432-6144; e-mail: victor.batista@yale.edu.

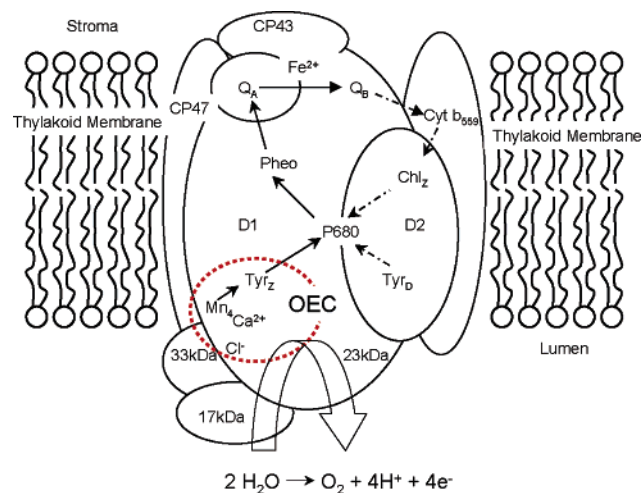


Figure 1. PSII complex and its antenna system, consisting of more than 20 protein subunits, either embedded in the thylakoid membrane or associated with its luminal surface. Light energy is trapped predominantly by the outer antenna and transferred to the photochemically active reaction center, via light-harvesting proteins CP47 and CP43, where it is used to drive the water-splitting reaction at the OEC. The electrons extracted from water are passed from the lumenally located Ca/Mn cluster to P680⁺ via D1-Y161 (Tyr_Z or Y_Z), a process that is coupled to ET from P680 to pheophytin (Pheo), to quinone electron acceptor Q_A, and onto quinone electron acceptor Q_B, near a nonheme iron group, defining the ET pathway marked by the solid arrows. Broken arrows indicate secondary ET pathways, which may play a photoprotective role. The protons and molecular oxygen produced during the water-splitting reaction are released into the lumen.

scopic and biochemical studies of PSII.^{5,6,10,11} However, the complexity of the system and the lack of a complete and unambiguous structure of the OEC have so far hindered the development of rigorous theoretical studies, limiting calculations to QM descriptions of inorganic complexes isolated from the influence of the actual protein environment (see ref 12 and references therein), or more complete OEC models built according to classical MM methods.^{13,14} Therefore, the development of computational studies in which both the intrinsic properties of the cluster and the influence of the protein environment are explicitly considered has yet to be reported and is the subject of this paper.

A complete functional model of the OEC remains elusive, although extensive work over many years of study has provided considerable insight into the OEC functionality and the underlying catalytic mechanism of photosynthetic water oxidation. It is, nowadays, established that photoabsorption by the specialized chlorophyll *a* species, P680, triggers a chain of electron transfer (ET) reactions (see Figure 1). The excited singlet state of P680 decays to the oxidized state P680⁺ by ET to a nearby pheophytin (Pheo) in about 2 ps after photoexcitation of P680. The charge-separated state is stabilized by a subsequent ET to a primary quinone electron acceptor (Q_A), which functions as a one-electron carrier, and subsequently to a secondary quinone electron acceptor (Q_B), which functions as a two-electron carrier, exchanging with free quinone upon two-electron reduction. The photo-

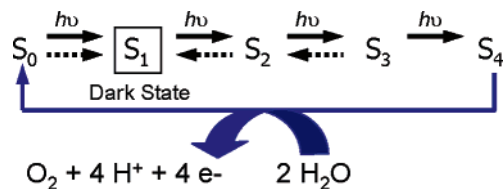


Figure 2. Kok cycle describing photosynthetic water oxidation by the reduction of the OEC from the S₄ to the S₀ state. Dotted arrows indicate reactions that relax the system back to the dark stable state S₁ within minutes. For simplicity, deprotonation reactions during the S₀ → S₁, S₂ → S₃, and S₃ → S₄ oxidation steps are omitted.

oxidized chlorophyll *a* species P680⁺ is reduced by a redox-active tyrosine (Y_Z), which is, in turn, reduced by the oxidation of water, catalyzed by the OEC.

The catalytic cycle of Joliet et al. and Kok et al.^{15,16} (see Figure 2) constitutes the basis of our current understanding of photosynthetic water oxidation as well as the foundation for further studies on the chemical nature of the reaction intermediates. The Kok cycle includes five oxidation states of the OEC, which are called storage states or simply “S states”. Each photoinduced ET from P680 to Q_A oxidizes the OEC to a higher S state. The most oxidized state (S₄) is quickly reduced to the S₀ state by the four-electron water-oxidation reaction, given in eq 1. In the dark, the S₀, S₂, and S₃ states are meta-stable and transform into the S₁ state within minutes (Figure 2, dashed lines). Hence, extensively dark-adapted samples contain only the S₁ state. Because this is the most easily characterized S state, it is our starting point for structural studies of the OEC of PSII. A number of structural models of the OEC^{6–9,11,12,17,18} with mechanistic implications^{5,13,14,18–23} have been proposed in attempts to rationalize the catalytic cycle at the detailed molecular level. However, many fundamental aspects of the proposed mechanisms and structure intermediates are the subject of current debate.^{14,18,19} In fact, unequivocal functional models of the OEC S states are yet to be established.

Until recently, all structural information regarding the OEC and its local environment has been derived from a variety of spectroscopic and biochemical techniques,^{11,24–27} including electron paramagnetic resonance (EPR) spectroscopy,^{23,28–32} X-ray absorption spectroscopy (XAS),^{33–38} optical spectroscopies,³⁹ Fourier transform infrared spectroscopy,^{40–44} and site-directed mutagenesis.^{45–48} In recent years, however, several groups have published X-ray diffraction structures of PSII from the cyanobacteria *Thermosynechococcus elongatus* and *Thermosynechococcus vulcanus*, yielding structures at 3.0–3.8 Å resolution.^{4,49–52} In particular, the recently published X-ray crystal structure of cyanobacterial PSII (PDB access code 1S5L)⁴ resolves most of the amino acid residues in the protein and nearly all cofactors at 3.5 Å resolution and suggests an atomic model of the OEC metal center (see Figure 3), providing a great opportunity for rigorous theoretical studies.

There are many aspects of the X-ray diffraction structure that have met with criticism, including both the geometric features of the Mn cluster^{36,38} and the proposed ligation scheme.^{13,14,19,42–44} In addition to the moderate resolution,

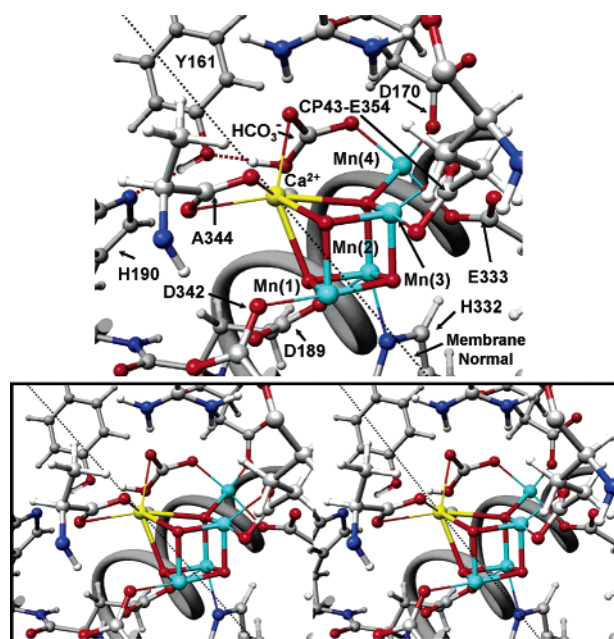


Figure 3. X-ray structure of the OEC of PSII.⁴ Upper and lower panels show the OEC and surrounding residues in mono- and stereoviews, respectively. Note that all amino acid residues correspond to the D1 protein subunit, unless otherwise indicated.

the X-ray diffraction data might correspond to a photo-reduced Mn cluster due to the high doses of X-rays employed.^{4,53} Therefore, the proposed X-ray diffraction models of the OEC remain rather controversial.

In fact, to date, the precise positions of the individual Mn ions could not be resolved in any X-ray diffraction model because the coordinate error in the resulting density maps is usually as high as 1 Å⁵⁰ and the resolution of bridging ligands is typically out of reach.³⁶ Because of these limitations, the model of the OEC metal center in the 1S5L structure, including the “3 + 1 Mn tetramer” proposed by EPR spectroscopic studies,^{23,35} has been based both on the overall electronic density maps and on the Mn–Mn distances reported by previous XAS studies.^{20,54,55} The proposed Mn tetramer (see Figure 3) takes the form of a Mn₃CaO₄ cuboidal cluster, including three closely associated manganese ions linked to a single μ_4 -oxo-ligated Mn ion, often called the “dangling manganese ion”. The proposed cuboidal model is still the subject of current debate³⁶ and disagrees with previously proposed structures in which three Mn ions were placed roughly at three corners of an isosceles triangle, with the fourth Mn ion at the center of the triangle either protruding toward the luminal surface of the membrane^{49,50} or parallel to it.⁵² The 1S5L crystallographic model also assigns potential protein ligands to the cluster (see Figure 3), including several amino acid residues already thought to be ligands on the basis of site-directed mutagenesis and spectroscopic studies.^{47,48} However, the number of protein ligands is surprisingly small, especially considering that Mn ions in high-oxidation states are usually coordinated by five or six ligands. The X-ray structure makes up for part of the ligand deficit by suggesting the presence of a bicarbonate anion bridging Ca²⁺ and Mn(4). Furthermore, a number of

small, nonprotein ligands, such as substrate and nonsubstrate water molecules, as well as hydroxide and chloride ions, are not visible at the current resolution.

Considering that QM/MM studies have played an essential role in revealing structure–function relations in a variety of other biological systems,^{3,56–71} it is expected that many of the controversial aspects of PSII could be resolved by combining the analysis and interpretation of experiments with rigorous QM/MM studies. Despite the incomplete and somewhat provisional nature of the 1S5L crystallographic structure, the proposed cluster architecture constitutes the most valuable point of departure for developing complete functional models of the OEC. The structural models developed in this paper, therefore, build upon the 1S5L structure. Density functional theory (DFT) QM/MM hybrid methods are applied to obtain completely ligated model structures of the OEC in the S₁ state, in an effort to determine whether the proposed 1S5L architecture can lead to a chemically sensible molecular structure of the hydrated OEC in the S₁ state with a complete coordination by water, protein ligands, hydroxide, and Cl[−] ions. The resulting QM/MM structural models are analyzed in terms of the intrinsic structure of the proposed Mn₃CaO₄Mn unit, embedded in the protein environment, as compared to structural XAS data. Some of the important questions addressed by the structural analysis are as follows: How many Mn–Mn vectors of ~ 2.7 Å are predicted by QM/MM models in the S₁ state? What is the most likely binding site for chloride? What proteinaceous ligating motifs give rise to the observed Mn–Mn distances? Finally, considering that water is the substrate for the catalytic reaction and that the positions of water molecules are unlikely to be resolved even by considerably higher-resolution X-ray structures, what are the implied locations of water molecules? Are those compatible with catalysis? The analysis of these fundamental aspects suggests that QM/MM hybrid methods, applied in conjunction with the X-ray crystallographic data, can considerably extend the description of the OEC into chemically sensible models with complete coordination of the metal cluster.

The paper is organized as follows. Section 2 describes the methodology, including the preparation of QM/MM structural models, the description of the QM/MM methodology, and the theoretical methods applied for simulations of extended X-ray absorption fine structure (EXAFS) spectra. Section 3 presents the results and a discussion with emphasis on mechanistic implications. Section 4 summarizes and concludes.

2. Methodology

2.1. Molecular Models. Molecular models are based on the 1S5L X-ray crystal structure of PSII (see Figure 3).⁴ The models build upon previous work,^{13,14} explicitly considering 1987 atoms of PSII, including the proposed Mn₃CaO₄Mn unit and all amino acid residues with α -carbons within 15 Å from any atom in the OEC metal ion cluster, with the addition of a buffer shell of amino acid residues with α -carbons within 15–20 Å from any atom in the OEC ion cluster. The coordination of the Mn ions was completed by

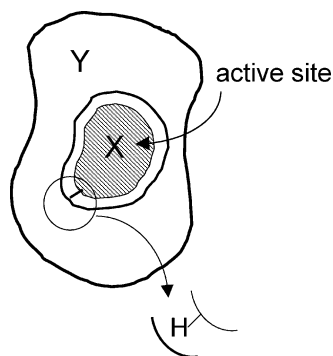


Figure 4. Partition of the biomacromolecular system into a reduced system (region X) and the surrounding molecular environment (region Y).

hydration, assuming a minimum displacement of the ligating residues from their crystallographic positions and the usual coordination of five or six ligands to Mn ions with oxidation states III and IV, respectively. The variable coordination of calcium, typically with six to eight ligands, was satisfied by the coordination resulting from geometry optimization of the water molecules and negative counterions.

The 1S5L X-ray crystal structure of the OEC is consistent with several possible binding sites for water molecules,⁴ including Ca^{2+} as suggested by ^{18}O isotope exchange measurements.^{26,72} Hydrated models were constructed by “soaking” the molecular structures in a large box containing a thermal distribution of water molecules and keeping those water molecules that did not sterically interfere with the protein residues or with existing water molecules in the model.¹⁴ The complete structures were subsequently relaxed. Because geometric optimization often creates new cavities, a series of soaking and relaxation procedures was applied until the number of water molecules converged. Such a computational protocol usually resulted in the addition of ~ 85 water molecules, with a few of them (up to six molecules) attached to calcium and manganese ions in the cuboidal $\text{Mn}_3\text{CaO}_4\text{Mn}$ cluster. Two of the ligated waters bound to Ca^{2+} and $\text{Mn}(4)$ are probable substrate water molecules, responsible for O–O bond formation in the $S_4 \rightarrow S_0$ transition. The resulting hydration of the cluster is, thus, roughly consistent with pulsed EPR experiments, which reveals the presence of several exchangeable deuterons near the Mn cluster in the S_0 , S_1 , and S_2 states.²³

2.2. QM/MM Hybrid Approach. QM/MM computations are based on the two-layer ONIOM electronic-embedding (EE) link-hydrogen atom approach² as implemented in Gaussian 03.⁷³ The ONIOM QM/MM methodology could only be efficiently applied to studies of the OEC of PSII after obtaining high-quality initial-guess states for the ligated cluster of Mn ions (i.e., the reduced system) according to ligand field theory⁷⁴ as implemented in Jaguar 5.5.⁷⁵ The resulting combined approach allowed us to exploit important capabilities of ONIOM, including both the link-hydrogen atom scheme for efficient and flexible definitions of QM layers and the possibility of modeling open-shell systems by performing unrestricted DFT (e.g., UB3LYP) calculations.

The ONIOM-EE method is applied by partitioning the system, as described in Figure 4, according to a reduced

molecular domain (region X) that includes the $\text{Mn}_3\text{CaO}_4\text{Mn}$ complex and the directly ligating proteinaceous carboxylate groups of D1-D189, CP43-E354, D1-A344, D1-E333, D1-D170, D1-D342, and the imidazole ring of D1-H332,⁷⁶ as well as bound water molecules, hydroxide, and chloride ions. The rest of the system defines region Y. The QM/MM boundaries are defined for the corresponding amino acid residues (i.e., D1-D189, CP43-E354, D1-A344, D1-E333, D1-D170, D1-D342, and D1-H332), by completing the covalency of frontier atoms according to the standard link-hydrogen atom scheme depicted in Figure 4.

The total energy E of the system is obtained at the ONIOM-EE level from three independent calculations as follows

$$E = E^{\text{MM},\text{X}+\text{Y}} + E^{\text{QM},\text{X}} - E^{\text{MM},\text{X}} \quad (2)$$

where $E^{\text{MM},\text{X}+\text{Y}}$ is the energy of the complete system computed at the molecular-mechanics level of theory, while $E^{\text{QM},\text{X}}$ and $E^{\text{MM},\text{X}}$ correspond to the energy of the reduced system computed at the QM and MM levels of theory, respectively. Electrostatic interactions between layers X and Y are included in the calculation of both $E^{\text{QM},\text{red}}$ and $E^{\text{MM},\text{red}}$, at the quantum mechanical and molecular mechanical levels, respectively. Therefore, the electrostatic interactions computed at the MM level in $E^{\text{MM},\text{red}}$ and $E^{\text{MM},\text{full}}$ cancel. Thus, the resulting QM/MM evaluation of the total energy at the ONIOM-EE level includes a quantum mechanical description of polarization of the reduced system due to the electrostatic influence of the surrounding protein environment. The analogous QM/MM method where the polarization of the reduced system is neglected is called ONIOM molecular embedding (ME). The self-consistent polarization of the protein environment is modeled according to the “moving domain–QM/MM” (MoD-QM/MM) approach,³ outlined in section 2.3.

The efficiency of the QM/MM calculations is optimized by using a combination of basis sets for the QM layer, including the lacvp basis set for Mn ions in order to consider nonrelativistic electron core potentials, the 6-31G(2df) basis set for bridging O^{2-} ions in order to include polarization functions on μ -oxo bridging oxides, and the 6-31G basis set for the rest of the atoms in the QM layer. Such a choice of basis set has been validated through extensive benchmark calculations on high-valent manganese complexes.^{12,77} The molecular structure beyond the QM layer is described by the Amber MM force field. Fully relaxed QM/MM molecular structures are obtained at the ONIOM-EE (UHF B3LYP/lacvp,6-31G(2df),6-31G:AMBER) level of theory by geometry optimization of the complete structural models in the presence of a buffer shell of amino acid residues with α -carbons within 15–20 Å from any atom in the OEC ion cluster. These are subject to harmonic constraints in order to preserve the natural shape of the system.

The electronic states of the structural models, fully relaxed at the ONIOM QM/MM level of theory, involve antiferromagnetic couplings between manganese centers. These couplings define broken-symmetry states, providing multi-configurational character to the singlet state S_1 .^{78–81} A typical optimization procedure involves the preparation of the QM

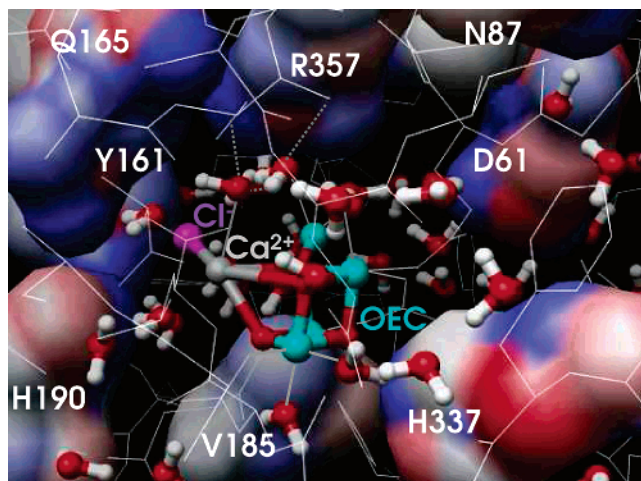


Figure 5. Quantitative analysis of the effect of protein polarization on the charge distribution of amino acid residues surrounding the Mn₃CaO₄Mn cuboidal cluster of the OEC. Blue (red) colors indicate an increase (decrease) in electronic density due to polarization effects (maximum differences, indicated by bright coloring, correspond to changes of atomic charges of about ± 15 – 20%). Note that all amino acid residues correspond to the D1 protein subunit, unless otherwise indicated.

layer in various possible initial spin states, stabilized by specific arrangements of ligands. The subsequent geometry relaxation, carried out at the ONIOM DFT–QM/MM level of theory, locally minimizes the energy of the system by finding the optimized geometry and spin-electronic state. The purity of the state is preserved throughout the geometry optimization process, in the event that the initial-guess electronic state is compatible with the geometry of the Mn₃CaO₄Mn cluster and the specific arrangement of ligands, and there are no other spin states of similar energy found along the optimization process. Otherwise, the optimization process changes the electronic spin state to the ground electronic-state symmetry of the corresponding nuclear configuration. The resulting optimized structures are analyzed and evaluated not only on the basis of the total energy of the system but also as compared to structural, electronic, and mechanistic features that should be consistent with experimental data.

2.3. Polarization of the Protein Environment. Modeling the electrostatic interactions between the QM and MM layers of the OEC of PSII is a challenging task because the high-valent multinuclear oxomanganese cluster Mn₃CaO₄Mn is embedded in a polarizable protein environment, ligated by protein ligands, water, hydroxide, and chloride ions. To describe the resulting self-consistent polarization of the system, the ONIOM-EE method has been applied in conjunction with the MoD-QM/MM computational protocol.³

The protocol MoD-QM/MM involves a simple space domain decomposition scheme where electrostatic potential (ESP) atomic charges of the constituent molecular domains are computed, to account for mutual polarization effects, and iterated until obtaining a self-consistent point-charge model of the electrostatic potential. This is particularly relevant for systems where polarization effects make inappropriate the use of standard molecular mechanics force fields.

Figure 5 shows a color map of the OEC residues, displaying differences in atomic charges obtained by considering, or neglecting, the mutual electrostatic influence at the ONIOM-EE and ONIOM-ME levels of theory, respectively. The residues that are more significantly polarized by the oxomanganese cluster are CP43-R357, D1-H337, D1-Q165, D1-Y161, D1-N87, D1-H190, D1-D61, and D1-V185, in addition to the residues directly ligated to Mn ions. Furthermore, it is shown that protein polarization, induced by the high-valent multinuclear oxomanganese cluster ions, usually introduces small corrections (~ 7 – 20) to the values of atomic charges of surrounding amino acid residues. However, summing these corrections over the whole QM/MM interface typically corrects the total QM/MM energy by 10–15 kcal/mol. The overall energy correction is, thus, significant (e.g., comparable to the energy-level splitting between high-spin and low-spin states of the Mn tetramer) and, therefore, necessary for accurate descriptions of the structure of the OEC of PSII.

2.4. EXAFS Simulations. Simulations of EXAFS spectra of the proposed structural models allow one to make direct comparisons with experimental data. Simulations of EXAFS spectra consider that a monochromatic X-ray beam is directed at a sample and that the photon energy of the X-rays is gradually increased such that it traverses one of the absorption edges of the elements contained within the sample. When the energy is below the absorption edge, the photons cannot excite the electrons of the relevant atomic level, and thus, absorption is low. On the other hand, when the photon energy is sufficiently high, a deep core electron is excited into a state above the Fermi energy. The resulting increase in absorption is known as the absorption edge. The ejected photoelectrons usually have low kinetic energy and can be backscattered by the atoms surrounding the emitting atom source. The interference of these outgoing photoelectrons with the scattered waves from atoms surrounding the central atom causes EXAFS. The regions of constructive and destructive interference are seen as local maxima and minima giving rise to oscillations in EXAFS intensities. These oscillations can be used to determine the atomic number, the distance and coordination number of the atoms surrounding the element whose absorption edge is being examined, the nature of neighboring atoms (their approximate atomic number), and changes in central-atom coordination with changes in experimental conditions.

The theory of the oscillatory structure, due to scattering of the photoelectron (emitted upon absorption of the X-ray) by atoms surrounding the emitting atom, was originally proposed by Kronig^{82,83} and worked out in detail by Sayers et al.,⁸⁴ Stern,⁸⁵ Lee and Pendry,^{86,87} and Ashley and Doniach.⁸⁸ Here, we outline only briefly the calculation of a typical EXAFS experiment, where a monochromatic X-ray beam passes through a homogeneous sample of uniform thickness x .

The absorption coefficient $\mu(E)$ is related to the transmitted (I) and incident (I_0) fluxes by $I = I_0 \exp[-\mu(E)x]$. In the weak-field limit, it is assumed that the main contribution to XAS comes from a dipole-mediated transition. In particular, when an electron in a deep core state i is excited into an

unoccupied state f , the absorption probability is given by time-dependent perturbation theory and is proportional to the square of the transition matrix element:

$$\mu(E) \sim \sum_f^{E_f < E_F} |\langle f | \hat{\epsilon} \cdot \vec{r} \exp(i\vec{k} \cdot \vec{r}) | i \rangle|^2 \delta(E_f) \quad (3)$$

where E_F is the Fermi energy and $\hat{\epsilon}$ and \vec{k} are the X-ray electric polarization and the wave vector, respectively. In the dipole approximation, the exponential is neglected and the absorption probability is independent of the direction of the sample axes with respect to $\hat{\epsilon}$.

There are two ways to solve eq 3. One method involves finding an adequate representation of the i and f states and then evaluating the integral directly. The other approach involves multiple scattering theory, where eq 3 can be written as follows:

$$\mu(E) \propto -\frac{1}{\pi} \text{Im} \langle i | \hat{\epsilon} \cdot \vec{r} G_{r,r'}(E) \hat{\epsilon} \cdot \vec{r}' | i \rangle \Theta(E - E_F) \quad (4)$$

with G and Θ representing the Green and Heaviside functions, respectively.

The results reported in this paper are based on the real space Green's function (RSGF) approach,⁸⁹ which has several advantages over traditional electronic-structure methods, especially for complex systems. The RSGF approach is essential for processes such as X-ray absorption, where symmetry-breaking effects (e.g., the photoelectron mean free path damping due to core-hole and inelastic losses) must be taken into account.

The central quantity in RSGF calculations is the matrix form of the propagator $G_{L'R',LR}(E)$ in a representation $|LR\rangle = i^l j_l(kr_R) Y_{lm}(\hat{r}_R)$, for site R and angular momentum $L = (l, m)$, where $\vec{r}_R = \vec{r} - \vec{R}$ and $L = (l, m)$. The matrix elements represent the transition amplitudes for an electron to propagate between states $|LR\rangle$ and $|L'R'\rangle$, satisfying the multiple-scattering equations⁹⁰ for a cluster with N_R sites,

$$G = G^C + G^{SC} \\ G^{SC} = e^{i\delta} [\mathbf{1} - G^0 T]^{-1} G^0 e^{i\delta'} \quad (5)$$

Here, the matrix indices are suppressed for simplicity, G^C represents the central atom contribution, G^{SC} is the scattering part from the surroundings, G^0 represents the damped free propagators, and T is the dimensionless scattering matrix which incorporates the spherical scattering potentials in terms of partial phase shifts for individual sites.

Once the propagator is obtained by solving eq 5, many physical quantities can be calculated. For example, the contribution to the X-ray absorption spectra from a given site and final state angular momentum L (with a relaxed core hole) is given by the golden rule expression

$$\mu(E) \sim -\frac{1}{\pi} \text{Im} \sum_{L,L'} M_{L'}^*(E) G_{L',L0}(E) M_L(E) \quad (6)$$

where $M_L(E) = \langle L, 0 | \hat{\epsilon} \cdot \vec{r} | c \rangle$ is a transition dipole matrix element between the atomic core state and a local final state $|L, 0\rangle$, with $\hat{\epsilon}$ being the X-ray polarization vector.

The total absorption coefficient $\mu(E)$ can be conveniently described as the isolated atom absorption $\mu_0(E)$ times a correction factor: $\mu = \mu_0(1 + \chi)$, where χ is the fractional change in absorption coefficient induced by neighboring atoms. Within the context of the single scattering approximation, a simple expression for χ is known as 'the standard EXAFS equation' for K-edge excitation.⁸⁴ According to such an equation, the contribution to EXAFS of an atom (index i) is given by

$$\chi(k) = \text{Im} \sum_i \left(\frac{N_i S_0^2 F_i(k)}{k R_i^2} \exp\{i[2\kappa R_i + \Phi_i(k)]\} \right. \\ \left. \exp(-2\sigma_i^2 k^2) \exp[-2R_i/\lambda(k)] \right) \quad (7)$$

where k is the wave vector modulus for the photoelectron; N_i is the number of atoms of type i at distance R_i from the absorber; the Debye–Waller factor $\exp(-2\sigma_i^2 k^2)$ takes account of fluctuations of distances due to a structural or thermal disorder, under the assumption of small displacements and Gaussian distributions of distances; the exponential term $\exp[-2R_i/\lambda(k)]$ takes account of finite elastic mean free paths of photoelectrons $\lambda(k)$ (between 5 and 10 Å for photoelectron energies from 30 to 1000 eV); S_0^2 is an average amplitude reduction factor (its value, usually 0.8–0.9, is the percent weight of the main excitation channel with respect to all possible excitation channels); $F_i(k)$ is a scattering amplitude function characteristic of the i th atom; $\Phi_i(k)$ is a phase function that takes account of the varying potential field along which the photoelectron moves. Equation 7 is valid in the case of nonoriented samples.

In this paper, the Fourier transform (FT) of $k^3 \chi(k)$ is performed with a Kaiser–Bessel-type window. The FT amplitude is normalized so that the maximum amplitude of the simulated spectrum coincides with the maximum amplitude of the experimental spectrum. To model the total number of electrons, the Fermi energy is taken as a free parameter to fit the relative peaks of the simulated spectrum to the experiment. EXAFS simulations on benchmark model compounds, for which high-resolution X-ray structures are known, tend to overestimate the apparent distances by about 0.15 Å. Thus, this shift was also applied to the EXAFS calculations reported here. The QM/MM structural models of the OEC of PSII are analyzed and partially validated by performing simulations of EXAFS spectra, explicitly considering $N_R \sim 10^3$ atomic sites with s, p, and d electrons. Simulations are carried out by using the program FEFF8 (version 8.2),⁹¹ and the resulting simulated spectra are directly compared to readily available experimental data.^{36,38}

3. Results and Discussion

The results are presented in eight subsections. Section 3.1 describes QM/MM models of the OEC of PSII in the S_1 state that are consistent with a broad range of experimental data. The electronic and structural properties of the models, introduced in section 3.1, are analyzed in sections 3.2 and 3.3, respectively. Section 3.4 analyzes the coordination of

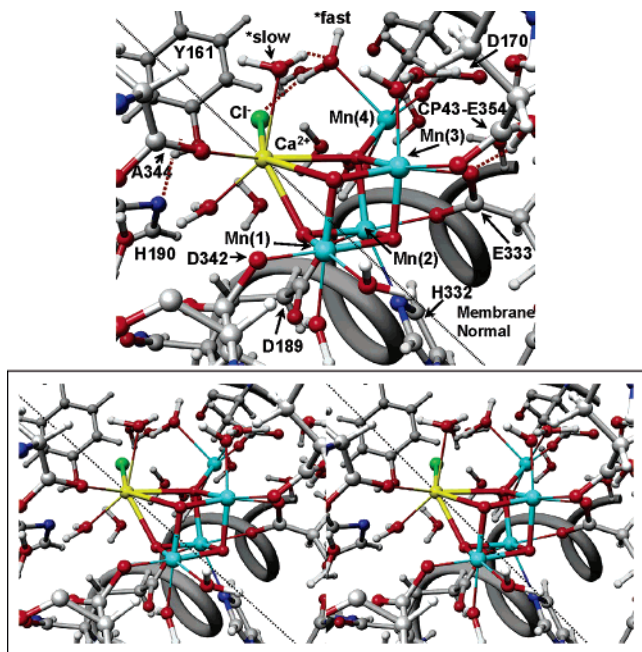


Figure 6. DFT QM/MM minimum energy geometry of the OEC of PSII, obtained at the ONIOM-EE (UHF B3LYP/lacvp,6-31G(2df),6-31G:AMBER) level of theory. Upper and lower panels show the OEC and surrounding residues in mono- and stereoviews, respectively. Putative substrate waters are labeled *slow and *fast (see text for explanation). Note that all amino acid residues correspond to the D1 protein subunit, unless otherwise indicated.

the oxomanganese complex by substrate water molecules in the presence of Ca²⁺ and Cl⁻ ions. The positioning of amino acid residues D1-Y161 and CP43-R357, relative to the oxomanganese complex, is discussed in section 3.6. The effect of oxidation and reduction of the OEC of PSII is analyzed in section 3.7 in terms of the resulting structural and electronic rearrangements as compared to readily available experimental data. Finally, section 3.8 describes a family of molecular structures, closely related to the QM/MM models discussed in sections 3.1–3.7, that are found to be also largely consistent with a wide range of experiments.

3.1. QM/MM Structural Models. Several QM/MM structural models have good agreement with the X-ray structure of Ferreira et al.,⁴ differing only in the protonation states, or number of ligated water molecules, or the coordination of labile ligands. However, only two combinations of spin states were found for the S₁ resting state. These include model **A**, with Mn₄(IV,IV,III,III) or Mn(1) = IV, Mn(2) = IV, Mn(3) = III, Mn(4) = III, in which the dangling manganese is pentacoordinated, and model **B**, with Mn₄(IV,III,III,IV), in which the dangling manganese has an additional ligated water molecule completing the six-coordination shell.

Figure 6 shows model **A**, a fully relaxed QM/MM structural model of the OEC of PSII in the S₁ resting state Mn₄(IV,IV,III,III), obtained at the ONIOM-EE (UHF B3LYP/lacvp,6-31G(2df),6-31G:AMBER) level of theory, including complete ligation of the Mn tetramer. Model **B** includes an additional water molecule ligated to the dangling Mn(4) that completes the hexacoordination of the dangling manganese

in the stabilization of a high-valent oxidation state IV. Because of the slightly strained coordination of D1-H332 to Mn(2) in the Mn cluster, the hexacoordination of Mn(2) becomes less favorable when the coordination sphere of Mn(4) is complete, and this stabilizes an oxidation state III for Mn(2), with a Jahn–Teller elongation along the Mn–N(D1-H332) axis. Therefore, the resulting state is Mn₄(IV,III,III,IV).

Note that, in contrast to the 1S5L structure (see Figure 3), Mn ions with oxidation states III and IV have the usual number of coordinated ligands (i.e., five or six, respectively) and Ca²⁺ has seven ligands. The QM/MM ligation of amino acid residues, however, is slightly different from the ligation scheme suggested by the X-ray diffraction structure.⁴ The proteinaceous ligation in the QM/MM models includes η² coordination of D1-E333 to both Mn(3) and Mn(2) and hydrogen bonding to the protonated (neutral) state of CP43-E354; monodentate coordination of D1-D342, CP43-E354, and D1-D170 to Mn(1), Mn(3), and Mn(4), respectively; and ligation of D1-E189 and D1-H332 to Mn(2).

Table 1 presents a comparative analysis of interatomic bond lengths and bond orientation angles relative to the membrane normal, including models **A** and **B**, obtained at the ONIOM-EE (UHF B3LYP/lacvp,6-31G(2df),6-31G:AMBER) level of theory, the X-ray diffraction structure 1S5L, solved at 3.5 Å resolution,⁴ EXAFS data,⁵⁴ and a reduced model system in the absence of the surrounding protein environment.⁷⁷ It is shown that the configuration of the cuboidal Mn₃CaO₄Mn complex in both QM/MM hybrid models is very similar to the structural model proposed by Ferreira et al.⁴ In fact, for both models, the root-mean-squared displacement of the QM/MM structural models, relative to the X-ray diffraction structure, is 0.6 Å. Therefore, it is difficult to judge whether the oxomanganese complex in the QM/MM models and 3.5 Å resolution X-ray structure are truly identical or whether there are any significant differences. In addition, as discussed in sections 3.2–3.8, the underlying structural and electronic properties of the QM/MM model are found to be in very good agreement with a wide range of experimental data of PSII. Furthermore, the comparison presented in Table 1 also indicates that there are only minor structural rearrangements in the oxomanganese complex when the configuration of the system is relaxed after substituting the surrounding protein environment by a reduced model with ligands that mimic the proteinaceous chelation scheme introduced by the QM/MM hybrid model.⁷⁷

These results suggest that the proposed cuboidal model of the inorganic core of the OEC of PSII, completely ligated with water, OH⁻, and Cl⁻ and proteinaceous ligands, is a stable molecular structure not only in two possible states (models **A** and **B**) associated with a slightly different coordination of the dangling Mn(4) but also in the absence of the surrounding protein environment.⁷⁷ Therefore, it is expected that significant insight could be provided by reduced model systems once the proteinaceous chelation scheme is elucidated by applying QM/MM hybrid methods in conjunction with moderate-resolution X-ray diffraction structures.

Considering the structural similarities between the cuboidal Mn₃CaO₄Mn complexes in the reduced model and those in the QM/MM hybrid structures, it is natural to conjecture that

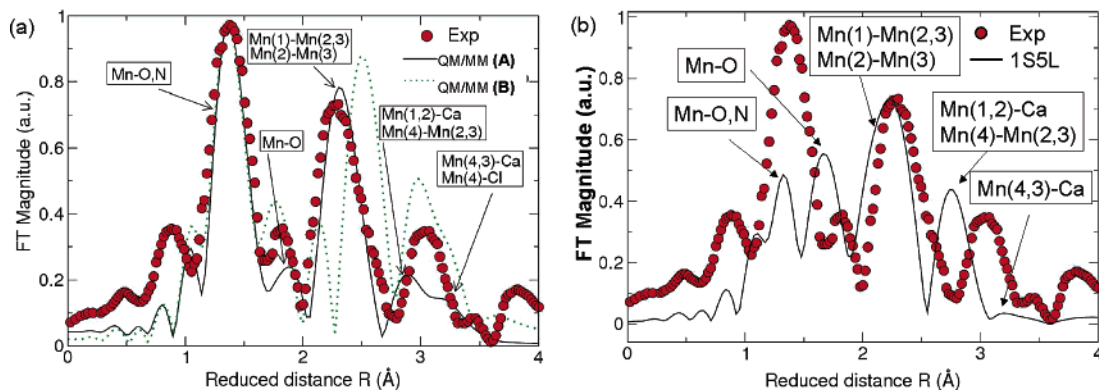


Figure 7. Comparison between the experimental EXAFS spectrum of the OEC of PSII in the S_1 state (red dots)^{36,38} and the calculated EXAFS spectra for (a) the DFT QM/MM models and (b) the X-ray diffraction structure.

Table 1. Interatomic Bond Lengths and Bond Orientation Angles (Relative to the Membrane Normal) in DFT-QM/MM Structural Models **A** and **B** of the OEC of PSII in the S_1 State (Described in the Text), Including Comparisons to the X-ray Diffraction Structure,⁴ EXAFS Data,⁵⁴ and the Configuration of the Reduced Quantum Mechanical Model in the Absence of the Protein Environment⁷⁷

bond vector	bond lengths					bond angles		
	A	B	X-ray	EXAFS ^a	red. model ^b	A	B	X-ray
Mn(1)–Mn(2)	2.76 Å	2.71 Å	2.65 Å	2.7 Å	2.77 Å	57°	59°	59°
Mn(1)–Mn(3)	2.76 Å	2.75 Å	2.67 Å	2.7 Å	2.76 Å	85°	82°	79°
Mn(2)–Mn(3)	2.82 Å	2.78 Å	2.72 Å		2.87 Å	63°	65°	71°
Mn(2)–Mn(4)	3.34 Å	3.79 Å	3.25 Å	3.3 Å	3.42 Å	54°	49°	58°
Mn(3)–Mn(4)	3.72 Å	3.61 Å	3.26 Å		3.74 Å	29°	24°	38°
Ca–Mn(2)	3.31 Å	3.41 Å	3.40 Å	3.4 Å	3.42 Å	53°	55°	59°
Ca–Mn(3)	3.95 Å	3.61 Å	3.38 Å		3.51 Å	35°	38°	39°

^a EXAFS data⁵⁴ include only two Mn–Mn distances of 2.7 Å, one Mn–Mn distance of 3.3 Å, and one Ca–Mn distance of 3.4 Å. ^b Interatomic distances in a reduced quantum mechanical model of the OEC computed at the UB3LYP level of theory.⁷⁷

the biomolecular environment must conform to the intrinsic properties of the ligated inorganic oxomanganese core, achieving catalytic functionality simply by positioning suitable sources and sinks of electrons and protons. The extent to which these results are significant is associated with the intrinsic limitations of moderate resolution X-ray diffraction models, obtained under conditions of unavoidable photo-reduction of Mn ions and the rapid exchange of labile substrate ligands.

3.2. Mn–Mn Distances: EXAFS Spectra. In contrast to the symmetric configuration of the X-ray diffraction model of the OEC of PSII (PDB access code 1S5L),⁴ with three Mn–Mn distances of about 2.7 Å, the DFT QM/MM hybrid models described in section 3.1 suggest that the S_1 state of the OEC has two short Mn–Mn distances of 2.71–2.76 Å per Mn tetramer (one of which is oriented at about 60° relative to the membrane normal), one slightly longer Mn–Mn distance of 2.78–2.82 Å, one ~3.3 Å Mn–Mn distance, and one ~3.3–3.4 Å Mn–Ca distance.⁹² These results are roughly consistent with most XAS data, often interpreted in terms of two 2.7 Å vectors per Mn₄ complex oriented at 60°^{33,93,94} (79°).⁵⁵ This important structural feature, however, remains controversial.³⁶ In fact, it has been proposed that there might be three Mn–Mn vectors of 2.7–2.8 Å per Mn₄ complex already in the S_1 state.³⁷

To make direct comparisons with readily available XAS experimental data, Figure 7a compares the simulated spectra of the DFT QM/MM models **A** and **B** and the experimental

EXAFS spectrum of the OEC of PSII in the S_1 state (red dots).^{36,38} The simulations are based on the real space Green's function methodology described in Section 2.4. It is shown that the simulated EXAFS spectrum of **A** is in very good agreement with experimental data, including the description of the widths and positions of multiscattering peaks associated with Mn–ligand distances of ~1.8 Å (reduced distance of ~1.6 Å) and Mn–Mn distances of ~2.7 Å (reduced distance of ~2.5 Å). In contrast, the simulated EXAFS spectrum based on the X-ray diffraction model (see Figure 7b) is in much worse agreement with the experimental EXAFS spectrum. This disagreement is partly due to the different proteinaceous ligation scheme and the incomplete coordination of metal ions in the cluster. The discrepancies in model **B** are due to inequivalent Mn–Mn distances, splitting a single peak at a reduced distance of ~2.5 Å into a bimodal structure.

3.3. Electronic Structure of the S_1 State. Table 2 shows that the DFT–QM/MM hybrid models predict high-valent configurations of the S_1 state of the OEC of PSII, with oxidation numbers Mn₄(IV,IV,III,III) and Mn₄(IV,III,III,IV) for model structures **A** and **B**, respectively. These results are consistent with EPR and X-ray spectroscopic evidence^{55,94–99} but disagree with low-valent Mn₄(III,III,III,III) proposals.^{100,101}

Table 2 indicates that both models **A** and **B** involve antiferromagnetic coupling between Mn(1) and Mn(2), between Mn(2) and Mn(3), and between Mn(3) and Mn(4)

Table 2. Formal Oxidation Numbers, Mulliken Spin-Population Analysis, and ESP Atomic Charges in the QM/MM Models of the OEC of PSII in the S₁ State

ion center	oxidation #		spin population		ESP charge	
	A	B	A	B	A	B
Mn(1)	+4	+4	+2.80	+2.81	+1.11	+1.16
Mn(2)	+4	+3	-2.75	-3.84	+1.08	+1.43
Mn(3)	+3	+3	+3.82	+3.82	+1.26	+1.30
Mn(4)	+3	+4	-3.80	-2.80	+1.35	+1.41
Ca	+2	+2	-0.01	+0.01	+1.77	+1.60
O(5),O(6)	-2, -2	-2, -2	-0.00, -0.02	+0.02, -0.01	-0.60, -0.80	-0.71, -0.89
O(7),O(8)	-2, -2	-2, -2	+0.08, -0.07	-0.04, +0.04	-0.67, -0.98	-0.64, -1.14

but frustrated coupling between Mn(1) and Mn(3). It is important to note, however, that predicting the correct relative stability of low-lying spin states in multinuclear oxomanganese complexes might be beyond the capabilities of the implemented DFT/B3LYP methodology.⁷⁷ For completeness, Table 2 compares the formal oxidation numbers (columns 2 and 3) as determined by the spin-population analysis (columns 4 and 5) to the actual ESP atomic charges (columns 6 and 7) of the corresponding ions. It is shown that the relation between oxidation numbers and atomic charges is complicated by the fact that there is charge transfer between bridging oxygen and manganese ions, similar to charge delocalization mechanisms observed in synthetic oxomanganese complexes.⁷⁷ Therefore, it is natural to expect that rationalizing certain properties of the oxomanganese complex of the OEC of PSII, such as ligand-exchange rates and the effect of changes in oxidation states on the vibrational spectroscopy of specific ligands,^{42–44} might require a complete electronic analysis in which both atomic charges and Mulliken spin populations are considered. Otherwise, these properties might be difficult to interpret by using an analysis based solely on formal oxidation numbers. As an example, Table 2 shows that the atomic charge of Ca²⁺ of model A (+1.77), with a formal oxidation number of II, is higher than the atomic charges of Mn³⁺ and Mn⁴⁺ (1.08–1.35; with oxidation numbers III and IV, respectively), suggesting that Ca²⁺ with its smaller oxidation number may actually bind the *slowly* exchanging substrate water molecule, in agreement with experimental observations.⁷²

3.4. Substrate Water Binding. The QM/MM hybrid models rationalize the 1S5L electronic density, initially assigned to bicarbonate (see Figure 3), to substrate water molecules bound to Mn(4) and Ca²⁺ (see Figure 6). This arrangement is consistent with mechanistic proposals,^{4,5,14,19,21} because the respective substrate oxygen atoms are 2.72 Å apart and may be brought yet closer together in the S₄ state (following deprotonation of the Mn-bound water) to achieve O–O bond formation in the S₄ → S₀ transition. Further, hydration of the cluster is broadly in line with pulsed EPR experiments which reveal the presence of several exchangeable deuterons near the Mn cluster in the S₀, S₁, and S₂ states.²³ However, the number of water molecules ligated to Mn ions is larger than would be expected from the analysis of exchangeable deuterons, as observed in pulsed EPR studies.²³ Therefore, the substitution of ligated water molecules by bicarbonate elsewhere in the OEC^{19,102} cannot be discounted. In the absence of bicarbonate, however, models

constructed by completing the Mn coordination numbers according to the principle of *minimum number of additional water molecules* have been dismissed, because such structures require unrealistic displacements of the ligating amino acid residues relative to their corresponding positions in the crystallographic 1S5L structure.^{13,14}

Possible binding positions for bicarbonate have been analyzed in the DFT–QM/MM hybrid models, including (i) bidentate coordination to Mn(1), (ii) chelation between Mn(3) and Mn(4), and (iii) coordination to Mn(4) and calcium (analogous to Ferreira's X-ray structure). In each case, bicarbonate replaces two water molecules (or a water molecule and a hydroxide ligand) bound to the OEC model, as shown in Figure 6. It is found that, in the first ligation scheme (i.e., case i), bicarbonate ligates by splitting into CO₂ and OH⁻ (the C–O distance is 0.4 Å longer than the equilibrium value in vacuo). In the second and third schemes (i.e., cases ii and iii), the resulting structure is stable. However, higher-resolution crystal structures^{51,52} have not corroborated the presence of bicarbonate. Further, it has been recently shown that bicarbonate is not the substrate.¹⁰³ Therefore, bicarbonate has not been included in the proposed QM/MM structural models A and B.

3.5. Chloride Binding. The QM/MM hybrid models include a calcium-bound chloride ion (see Figure 6): Cl⁻ is 3.1 Å from Ca²⁺ and 3.2 Å from the phenoxy oxygen of D1-Y161. Such an arrangement of ligands completes a coordination sphere of seven ligands for Ca²⁺ (often chelated by up to eight ligands), including Cl⁻, two water molecules, the monodentate carboxylate terminus of D1-A344, and the three bridging oxides of the cuboidal structure.

The presence of the chloride ion has not been resolved by X-ray diffraction experiments. However, the binding site suggested by the QM/MM hybrid models is consistent with the experimental observation that acetate binds competitively with chloride¹⁰⁴ and blocks catalysis at the S₂ state.¹⁰⁵ It has also been suggested that chloride is required for transitions beyond the S₂ state.¹⁰⁵ Furthermore, the direct binding of chloride to calcium is consistent with the proposal that chloride is part of a proton relay network.¹⁰⁶ In addition, the QM/MM models are consistent with pulsed EPR experiments of the OEC in which chloride has been replaced by acetate, revealing a distance of 3.1 Å from the methyl deuterons of the bound acetate to the phenoxy oxygen of D1-Y161.¹⁰⁷

The chloride/acetate substitution has been analyzed in the QM/MM computational models, to partially validate the proposed Cl⁻ binding site. The chloride/acetate substitution

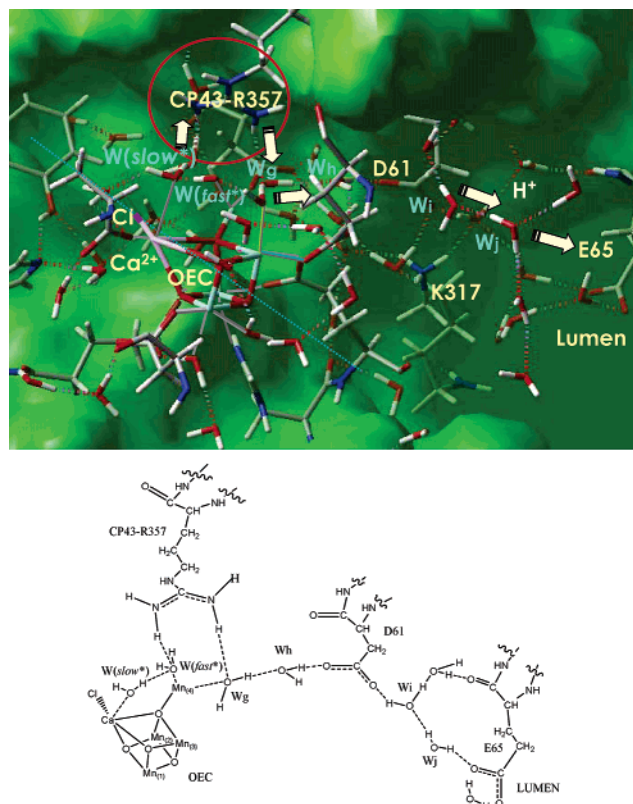


Figure 8. Proton exit channel toward the luminal surface of the membrane, suggested by the DFT QM/MM structural models, involving an extended network of hydrogen bonds from the substrate water molecules W_s and W_f to CP43-R357 and from CP43-R357 to D1-D61, which is the first residue of the putative proton-transfer channel leading to the luminal surface of PSII via hydrogen-bonded water molecules W_g , W_h , W_i , and W_j . Note that all amino acid residues correspond to the D1 protein subunit, unless otherwise indicated.

required the removal of two water molecules (originally coordinated to Ca^{2+}) in order to avoid strong repulsive interactions. The removal of bound water molecules is consistent with the experimental evidence that adding acetate to the OEC displaces several water molecules from the protein cavity.¹⁰⁸ After geometry optimization, the carboxylate group of acetate replaces Cl^- and the C–C bond becomes collinear with the previously modeled Cl^- – Ca^{2+} bond. Furthermore, the phenoxy oxygen of D1-Y161 is found to be 3.2 Å from the averaged position of the acetate methyl hydrogens (3.1 Å from the methyl carbon), in excellent agreement with experimental observations.

3.6. Function of D1-Y161 (Y_Z) and CP43-R357. Tyrosine D1-Y161 has long been viewed as an electron-transport cofactor. As mentioned in section 1, the oxidized state P680^+ is thought to be reduced by the redox-active tyrosine D1-Y161 (Y_Z), which is in turn reduced by an electron from the OEC.^{109,110} The QM/MM structural models of PSII seem to be consistent with such a postulated redox mechanism, especially judging by the proximity of Y_Z to the Mn cluster, although this remains to be demonstrated by rigorous calculations of redox potentials, which we will present elsewhere. Simple inspection of the QM/MM structural models (see Figure 8), however, reveals that the

phenoxy oxygen of Y_Z is close to the chloride ligand (3.4 Å) and that the Ca^{2+} – Cl^- bond length is 3.14 Å. Furthermore, the QM/MM structure shown in Figure 8 indicates that the Y_Z phenol group is hydrogen-bonded to the imidazole ϵ -N of the D1-H190 side chain (see also Figure 6). This hydrogen-bonding partnership is consistent with mutational and spectroscopic studies^{45,111,112} as well as with earlier studies based on MM models.^{13,14}

The possibility that the oxidized Y_Z radical might simultaneously oxidize and deprotonate the hydrated OEC¹¹³ would require a mechanism in which Y_Z abstracts hydrogen atoms and delivers protons to the protein surface via D1-H190. However, on the basis of the lack of a H-bonding pathway leading from D1-H190 to the lumen, it is more likely that D1-H190 accepts a proton from Y_Z during the oxidation of Y_Z and returns the proton to Y_Z upon its reduction. Consistently, the QM/MM structural models suggest that other amino acid residues (e.g., CP43-R357) might be more favorably placed for proton abstraction,^{13,14,19} because substrate water molecules are not directly exposed to Y_Z .

The QM/MM hybrid models show that a network of hydrogen bonds is formed around the catalytically active face of the OEC cluster (see Figure 8), including both substrate water molecules, the side chain of CP43-R357, and the calcium-bound chloride ion. Nearby, two hydrogen-bonded nonligating water molecules are found to fit easily into the structure between Mn(4) and D61, the first residue of the putative proton-transfer channel leading to the luminal surface of PSII. The proximity of CP43-R357 to the Mn cluster in the QM/MM hybrid models suggests that CP43-R357 might play the role of the redox-coupled catalytic base in the latter half of the S-state cycle.¹⁹ A recent computational study indicates that the $\text{p}K_a$ of D1-R357 is indeed particularly sensitive to an increase in the charge of the Mn/Ca cluster.¹¹⁴ We will present our calculations of $\text{p}K_a$ values of acid/base groups along the proton exit channel elsewhere. It is expected that a $\text{p}K_a$ gradient along the channel should facilitate proton transfer into an entropically favored (i.e., irreversible) state in the luminal bulk solution.

3.7. S_0 and S_2 States. The DFT–QM/MM hybrid models of the OEC of PSII in the S_1 state, reported in previous sections, allow for the investigation of structural changes induced by oxidation/reduction of the OEC and the effect of such electronic changes on the underlying ligation scheme. The $S_1 \rightarrow S_2$ transition involves oxidation^{115–117} without deprotonation. This can be achieved by the oxidation of one of the two manganese ions with oxidation state III [i.e., Mn(3) or Mn(4), in model **A**, and Mn(2) or Mn(3), in model **B**].

Minimum-energy structures, obtained by optimizations initialized with both possible spin configurations for each model, indicate that the more likely configurations involve the oxidation of Mn(3) in model **A** and the oxidation of Mn(2) in model **B** (see Table 3 and Figure 9).

Therefore, DFT QM/MM models **A** and **B** predict that the S_2 state, obtained at the ONIOM-EE (UHF B3LYP/lacvp,6-31G(2df),6-31G:AMBER) level of theory, involves the high-valent configuration $\text{Mn}_4(\text{IV},\text{IV},\text{IV},\text{III})$ or

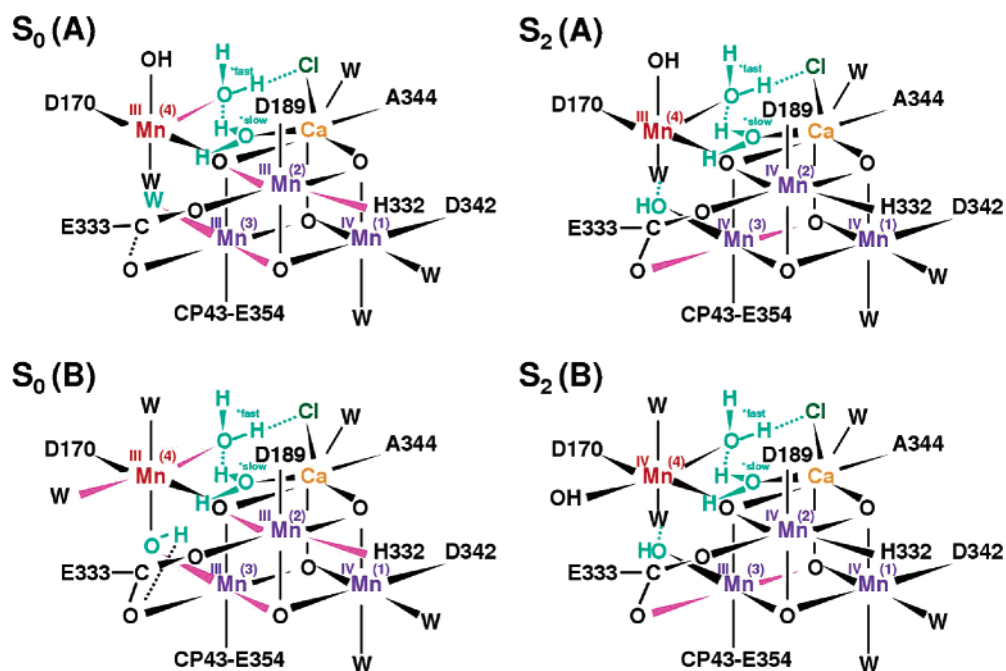


Figure 9. Schematic representation of the proteinaceous ligation scheme in the DFT QM/MM models **A** and **B** of the OEC of PSII in the S_0 (left panel) and S_2 (right panel) states. Elongated bonds due to Jahn–Teller distortion are represented in magenta. Note that all amino acid residues correspond to the D1 protein subunit, unless otherwise indicated.

Table 3. Mulliken Spin Population Analysis and ESP Atomic Charges in the DFT QM/MM Models of the OEC of PSII in the S_0 and S_2 States

ion center	S_0						S_2					
	spin population		oxidation #		ESP charge		spin population		oxidation #		ESP charge	
	A	B	A	B	A	B	A	B	A	B	A	B
Mn(1)	-2.75	-2.86	+4	+4	+1.41	+1.33	-2.79	-2.76	+4	+4	+1.14	+1.25
Mn(2)	+3.76	+3.78	+3	+3	+1.36	+1.35	+2.92	+3.15	+4	+4	+1.02	+1.52
Mn(3)	-3.81	-3.85	+3	+3	+1.43	+1.33	-2.74	-3.82	+4	+3	+1.59	+1.11
Mn(4)	+3.76	+3.82	+3	+3	+1.39	+1.47	+3.79	+3.17	+3	+4	+1.49	+1.50
O(5)	0.00	-0.03	-2	-2	-0.75	-0.60	+0.09	-0.07	-2	-2	-0.53	-0.78
O(6)	-0.04	-0.07	-2	-2	-0.99	-0.86	+0.02	-0.01	-2	-2	-0.81	-0.80
O(7)	+0.08	+0.03	-2	-2	-0.75	-0.73	-0.03	-0.03	-2	-2	-0.78	-0.67
O(8)	-0.04	+0.01	-2	-2	-1.14	-1.14	-0.09	-0.05	-2	-2	-0.86	-1.22
Ca	-0.00	-0.00	+2	+2	+1.62	+1.62	-0.00	-0.02	+2	+2	+1.56	+1.66
Cl	-0.00	-0.00	-1	-1	-0.77	-0.75	+0.00	+0.28	-1	-1	-0.67	-0.41

Table 4. Interionic Distances and Bond Angles, Relative to the Membrane Normal, in the DFT QM/MM Structural Models of the OEC of PSII in the S_0 and S_2 States

bond vector	S_0				S_2			
	bond length		bond angle		bond length		bond angle	
	A	B	A	B	A	B	A	B
Mn(1)–Mn(2)	2.70 Å	2.70 Å	58°	58°	2.78 Å	2.71 Å	58°	58°
Mn(1)–Mn(3)	2.78 Å	2.91 Å	82°	82°	2.76 Å	2.73 Å	81°	81°
Mn(2)–Mn(3)	2.78 Å	2.92 Å	68°	68°	2.86 Å	2.79 Å	65°	65°
Mn(2)–Mn(4)	3.52 Å	3.59 Å	58°	58°	3.31 Å	3.79 Å	59°	59°
Mn(3)–Mn(4)	3.43 Å	2.94 Å	32°	32°	3.55 Å	3.67 Å	35°	35°
Ca–Mn(2)	3.40 Å	3.51 Å	57°	57°	3.78 Å	3.36 Å	57°	57°
Ca–Mn(3)	3.71 Å	3.52 Å	38°	38°	3.98 Å	3.77 Å	36°	36°

Mn₄(IV,IV,III,IV). The analysis of the configurations of the relaxed S_1 and S_2 QM/MM hybrid models (see Tables 1 and 4) indicates that the $S_1 \rightarrow S_2$ oxidation is not expected to involve any significant rearrangement of ligands, or structural

changes in the Mn cluster (Table 4). These results are, thus, in agreement with recent findings of EXAFS studies.³⁸

The $S_0 \rightarrow S_1$ transition involves the oxidation of a manganese ion and the deprotonation of a ligand (probably

a water molecule^{5,14} or bridging oxide³⁸), a process that induces structural rearrangements in the oxomanganese cluster, shortening a Mn–Mn distance by approximately 0.15 Å.³⁷

To elucidate the nature of the S_0 state and the specific electronic and structural changes induced upon oxidation of the system, several DFT QM/MM hybrid models have been investigated. The analysis of QM/MM structures indicates that Mn(2) is oxidized in model **A**, and Mn(4) is oxidized in model **B**, during the $S_0 \rightarrow S_1$ step (see Tables 2 and 3). Therefore, both DFT QM/MM models, obtained at the ONIOM-EE (UHF B3LYP/lacvp,6-31G(2df),6-31G:AMBER) level of theory, predict that the S_0 state has the high-valent configuration Mn₄(IV,III,III,III).

Considering that the proposed DFT QM/MM hybrid model does not involve the protonation of bridging oxides, the potential ligands responsible for deprotonation are water molecules ligated to Mn(3) or Mn(4), because Mn(2) has no water ligands and changes in protonation states of water molecules ligated to Mn(1) would not involve changes in the active site of the cluster. In particular, model **A** suggests that a water molecule ligated to Mn(3) is the best candidate because the ligated molecule is deprotonated (HO^-) in the S_1 state, while model **B** points to deprotonation of a water molecule ligated to Mn(4).

Table 4 summarizes the configurations of models **A** and **B** in the S_0 state, indicating that the DFT QM/MM hybrid models predict a single 2.7 Å Mn–Mn distance per Mn tetramer (oriented at about 58° relative to the membrane normal), which is 0.06 Å shorter than that in the S_1 state. Furthermore, the QM/MM models of the S_0 state indicate that the Jahn–Teller effect in Mn(3) elongates the distance Mn(1)–Mn(3), making it slightly longer than the distance Mn(1)–Mn(2). This effect is more pronounced in model **B** than in model **A**, because in model **B** the elongation is along the direction of the μ -oxo bridge linking Mn(3)–Mn(1) and Mn(3)–Mn(2). In contrast, model **A** involves elongation along the coordination axis with D1-E333. Such a distortion is also responsible for moving the D1-E333 ligand away from Mn(3), partially forming an oxo-bridge with Mn(4) (see Figure 9).

3.8. Ligation of D1-E333. The 1S5L crystal structure indicates that the amino acid residue D1-E333 is ligated at an intermediate position between Mn(2) and Mn(4), with its carboxylate group in close contact with the carboxylate side-chain of CP43-E354. The QM/MM analysis of the OEC of PSII in the S_1 state indicates that the intrinsic stability of the pair of amino acid residues D1-E333 and CP43-E354, in close contact with each other, is not only due to coordination to the oxomanganese complex but also due to hydrogen bonding between the protonated (neutral) CP43-E354 and the carboxylate group of D1-E333.

Several other ligation schemes for the amino acid residue D1-E333 have been analyzed in an effort to investigate the influence of chelation on the geometry of the OEC cluster in the S_1 state. Relevant geometrical parameters for fully optimized QM/MM models, with ligation schemes depicted in Figure 10, are reported in Table 5.

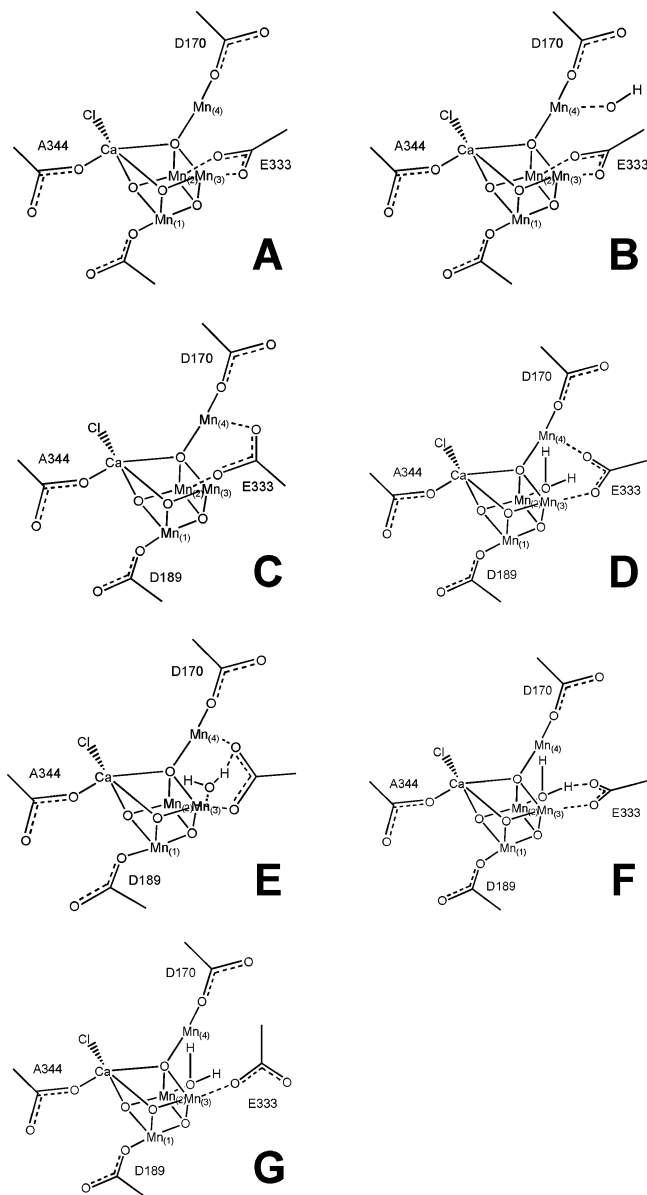


Figure 10. Possible ligation schemes for the carboxylate terminus of amino acid residue D1-E333 and a water molecule ligated to the OEC of PSII. Note that all amino acid residues correspond to the D1 protein subunit, unless otherwise indicated.

Schemes A and B correspond to the QM/MM structures **A** and **B**, introduced in section 3.1. These structures involve η^2 coordination of D1-E333 to Mn(2) and Mn(3). Scheme C involves D1-E333 chelation to both Mn(4) and Mn(2), with a minimum displacement of the side chain of D1-E333 relative to the X-ray configuration. Schemes D–F correspond to D1-E333 coordination to Mn(2) or Mn(3), competing with a water molecule for the other metal center. It is found that all of the ligation schemes described in Figure 10 are within the resolution limits of the X-ray structure because the displacement of each carboxylate oxygen is smaller than 1 Å, relative to their configuration in the X-ray structure.

Direct comparisons between the resulting structural models obtained with each of these possible ligation schemes and readily available EXAFS data are not straightforward. On the basis of EXAFS studies, it was initially concluded that

Table 5. Interionic Distances in the OEC of PSII in the S₁ State for Different Ligation Schemes of the Amino Acid Residue D1-E333

ref. ^a	scheme ^b	1–2 ^c	1–3 ^c	2–3 ^c	2–4 ^c	3–4 ^c	Ca-2 ^c	Ca-3 ^c
A(1)	2–3	2.76 Å	2.76 Å	2.82 Å	3.34 Å	3.72 Å	3.31 Å	3.95 Å
A(2)	2–3 + OH(4) ^d	2.70 Å	2.73 Å	2.80 Å	3.79 Å	3.66 Å	3.34 Å	3.44 Å
B	2–4	2.76 Å	2.77 Å	3.00 Å	3.27 Å	3.68 Å	3.49 Å	3.42 Å
C	3–4 + w(2) ^e	2.79 Å	2.73 Å	2.86 Å	3.61 Å	3.59 Å	3.10 Å	4.20 Å
D	w(3)–2–4 ^f	2.75 Å	2.79 Å	3.02 Å	3.40 Å	3.89 Å	3.37 Å	3.68 Å
E	w(2)–3	2.77 Å	2.76 Å	3.01 Å	3.14 Å	3.64 Å	3.19 Å	4.37 Å
F	3 + w(2)	2.79 Å	2.78 Å	2.93 Å	2.95 Å	3.49 Å	3.12 Å	4.44 Å

^a Schemes are labeled according to Figure 10. ^b Atom numbers correspond to the Mn center to which Glu333 is ligated; w(#), or OH(#), indicate a water molecule or OH[−], respectively, ligated to Mn(#), see Figure 10. ^c For simplicity, Mn symbols in table headers are omitted, e.g., 1–2 stands for Mn(1)–Mn(2), etc. ^d The oxidation state is Mn₄(IV,III,IV,III). ^e The optimization of the structure without a water ligated to Mn(2) (i.e., 3–4) converged to 2–3.

the S₁ state has two 2.7 Å Mn–Mn distances, one 3.3–3.4 Å Mn–Ca distance, and one 3.3 Å Mn–Mn distance.¹¹⁸ However, it was recently reported that a third 2.7 Å Mn–Mn distance may be present.^{37,54} Dau et al.³⁸ reported that a third, longer Mn–Mn distance of about 2.8 Å might also be present, but this was disfavored because its inclusion lowered the fit quality of the EXAFS spectrum simulations.³⁶ Furthermore, there also seems to be no agreement on the number of 3.3–3.4 Å Mn–Ca distances, reported as one or two 3.4 Å distances by the Berkeley group⁵⁴ and two or three 3.3 Å distances by the Berlin group.³⁸

All structures in Table 1 show two short Mn–Mn distances and a third Mn–Mn distance that is longer by 0.06–0.17 Å. The calculations also show one Mn–Ca distance and one Mn–Mn distance of ca. 3.3 Å, consistent with both the Berkeley and Berlin groups' EXAFS analyses. Accordingly, all ligation schemes show qualitative accordance with the number and relative magnitudes of Mn–Mn and Mn–Ca distances proposed by both experimental groups. In particular, the coordination scheme that includes the ligation of D1-E333 to Mn(2) and Mn(3) has the smallest “long” Mn–Mn distance (2–3 in Table 1), making the three Mn–Mn distances lie within the uncertainty of the DFT QM/MM method. In this way, the models introduced in sections 3.1–3.7 are consistent with a whole set of measurements, both of the Berkeley group and of the Berlin group. The other possibilities discussed in this section, however, cannot be ruled out on either experimental or current computational grounds.

Other possible QM/MM structures have been analyzed, differing in the protonation state of ligated water molecules or in the coordination of labile ligands. For simplicity, however, the presentation has been limited to the fully optimized QM/MM models whose structural features are most consistent with EXAFS measurements.

4. Conclusions

We have developed chemically sensible structural models of the OEC of PSII with complete ligation of the metal-oxo cluster by amino acid residues, water, hydroxide, and chloride. The models were developed at the DFT QM/MM ONIOM-EE (UHF B3LYP/lacvp,6-31G(2df),6-31G:AMBER) level of theory. Manganese and calcium ions are ligated consistently with standard coordination chemistry assumptions, supported by much biochemical and spectro-

scopic data, including a calcium-bound chloride ligand which is docked consistently with pulsed EPR data obtained from acetate-substituted PSII. Proteinaceous ligation includes the monodentate coordination of D1-D342, CP43-E354, and D1-D170 to Mn(1), Mn(3), and Mn(4), respectively; the coordination of D1-E333 to both Mn(3) and Mn(2) and hydrogen bonding of D1-E333 to CP43-E354, which is in the protonated (neutral) form; and the ligation of D1-D189 and D1-H332 to Mn(2). The proposed models are found to be stable and entirely consistent with available mechanistic data as well as compatible with EXAFS measurements and X-ray diffraction models of PSII (i.e., with root-mean-squared displacement smaller than 1 Å relative to the X-ray structure). Therefore, it is concluded that the proposed QM/MM structures are particularly relevant to the investigation and validation of reaction intermediates of photosynthetic water oxidation.

We have found a family of closely related QM/MM structural models which are partially consistent with a wide range of experiments. Most of these structures differ only in the protonation state of water molecules ligated to the Mn cluster, or in the coordination of a proteinaceous ligand (D1-E333). It is, therefore, concluded that the intrinsic degeneracy of protonation states and coordination patterns (as well as low-lying spin states) might be necessary to ensure the robustness of the functionality in the presence of thermal fluctuations.

We have found that the DFT QM/MM level of theory predicts high-valent electronic configurations with oxidation numbers Mn₄(III,III,III,IV) for the S₀ state, Mn₄(III,IV,IV,IV) for the S₂ state, and Mn₄(III,III,IV,IV) or Mn₄(IV,III,III,IV) for the S₁ state, consistent with EPR and X-ray spectroscopic evidence.^{55,94–99} However, we caution that further studies exploring the relative stability of different spin states are required, because predicting the correct relative stability of low-lying spin states in multinuclear oxomanganese complexes might be beyond the current capabilities of the DFT B3LYP hybrid functional.⁷⁷ This problem adds one more example to the list of high-valent transition-metal complexes in which DFT might provide an unreliable description of the energetics of the low-lying spin-electronic states.^{119–125}

In agreement with experiments,³⁸ we found that the S₁ → S₂ oxidation does not involve any significant rearrangement of ligands, or structural changes in the Mn cluster. In contrast, the S₀ → S₁ oxidation step deprotonates a water molecule

ligated to Mn(3) and oxidizes Mn(2) from III to IV in model **A** or oxidizes Mn(4) and deprotonates a water molecule ligated to Mn(4) in model **B**. The resulting Jahn–Teller effect in the S_0 state elongates the Mn(1)–Mn(3) distance relative to Mn(1)–Mn(2). In agreement with EXAFS experiments, the proposed DFT–QM/MM structures predict that the S_0 state involves a single Mn–Mn vector close to 2.7 Å.

We conclude that the relation between oxidation numbers and atomic charges is complicated by charge transfer between μ -O and Mn ions, similar to charge delocalization mechanisms observed in synthetic oxomanganese complexes.⁷⁷ Therefore, we found that joint charge- and spin-population analysis might be necessary in order to rationalize certain mechanistic and structural properties of the system, including water exchange rates and the vibrational spectroscopy of ligated residues. In fact, in agreement with experimental measurements of exchange rates, the charge-population analysis indicates that Ca^{2+} carries the highest positive charge and, therefore, might bind the slow exchanging substrate water molecule, even though its formal oxidation number is smaller than that of the dangling manganese.

We have found that the proximity of D1-Y161 (Y_Z) to the Mn cluster in the QM/MM structural model of PSII is consistent with the electron-transfer role of D1-Y161 (Y_Z). However, the actual calculation of redox potentials will be necessary to address this fundamental aspect. These calculations involve work in progress in our group and will be presented elsewhere. In particular, the synergistic modulation of protonation and redox states will be addressed in terms of continuum electrostatic calculations based on the DFT QM/MM molecular structures reported herein. Furthermore, we have found that the substrate water molecules are directly exposed to CP43-R357 in the QM/MM structural models. We found an extended network of hydrogen bonds linking CP43-R357 with D1-D61, suggesting a proton exit channel toward the luminal surface of the membrane.

We have found only minor structural rearrangements in the oxomanganese complex after substituting the surrounding protein environment by a reduced model with ligands that mimic the proposed QM/MM proteinaceous ligation scheme. These results suggest that the cuboidal model of the inorganic core of the OEC of PSII, completely ligated with water, OH^- , Cl^- and proteinaceous ligands, is a stable molecular structure even in the absence of the surrounding protein environment. Therefore, it is natural to conjecture that the biomolecular environment must conform to the intrinsic properties of the ligated inorganic oxomanganese complex, achieving catalytic functionality simply by positioning suitable sources and sinks of electrons and protons.

Acknowledgment. V.S.B. acknowledges a generous allocation of supercomputer time from the National Energy Research Scientific Computing (NERSC) center and financial support from Research Corporation, Research Innovation Award # RI0702, a Petroleum Research Fund Award from the American Chemical Society PRF # 37789-G6, a junior faculty award from the F. Warren Hellman Family, the National Science Foundation (NSF) Career Program Award CHE # 0345984, the NSF Nanoscale Exploratory Research (NER) Award ECS # 0404191, the Alfred P. Sloan Fellow-

ship (2005–2006), a Camille Dreyfus Teacher-Scholar Award for 2005, and a Yale Junior Faculty Fellowship in the Natural Sciences (2005–2006). G.W.B. acknowledges support from the National Institutes of Health Grant GM32715.

References

- (1) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.
- (2) Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *THEOCHEM* **1999**, *461*, 1–21.
- (3) Gascón, J. A.; Leung, S. S. F.; Batista, E. R.; Batista, V. S. *J. Chem. Theory Comput.* **2006**, *2*, 175–186.
- (4) Ferreira, K. N.; Iverson, T. M.; Maghlaoui, K.; Barber, J.; Iwata, S. *Science* **2004**, *303*, 1831–1838.
- (5) Vrettos, J. S.; Limburg, J.; Brudvig, G. W. *Biochim. Biophys. Acta* **2001**, *1503*, 229–245.
- (6) Diner, B. A.; Babcock, G. T. In *Oxygenic Photosynthesis: The Light Reactions*; Ort, D. R., Yocum, C. F., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1996; pp 213–247.
- (7) Debus, R. J. *Biochim. Biophys. Acta* **1992**, *1102*, 269–352.
- (8) Witt, H. T. *Phys. Chem. Chem. Phys.* **1996**, *100*, 1923–1942.
- (9) Britt, R. D. In *Oxygenic Photosynthesis: The Light Reactions*; Ort, D. R., Yocum, C. F., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1996; pp 137–159.
- (10) Barber, J. *Quart. Rev. Biophys.* **2003**, *36*, 71–89.
- (11) Yachandra, V. K.; Sauer, K.; Klein, M. P. *Chem. Rev.* **1996**, *96*, 2927–2950.
- (12) Lundberg, M.; Siegbahn, P. E. M. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4772–4780.
- (13) McEvoy, J. P.; Gascón, J. A.; Sproviero, E. M.; Batista, V. S.; Brudvig, G. W. In *Photosynthesis: Fundamental Aspects to Global Perspectives*; Bruce, D., van der Est, A., Eds.; Allen Press Inc.: Lawrence, Kansas, 2005; Vol. 1, pp 278–280.
- (14) McEvoy, J. P.; Gascón, J. A.; Batista, V. S.; Brudvig, G. W. *Photochem. Photobiol. Sci.* **2005**, *4*, 940–949.
- (15) Joliot, P.; Barbieri, G.; Chabaud, R. *Photochem. Photobiol.* **1969**, *10*, 309–329.
- (16) Kok, B.; Forbush, B.; McGloin, M. *Photochem. Photobiol.* **1970**, *11*, 457–475.
- (17) Messinger, J.; Badger, M.; Wydrzynski, T. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3209–3213.
- (18) Messinger, J. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4764–4771.
- (19) McEvoy, J. P.; Brudvig, G. W. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4754–4763.
- (20) Robblee, J. H.; Cinco, R. M.; Yachandra, V. K. *Biochim. Biophys. Acta* **2001**, *1503*, 7–23.
- (21) Pecoraro, V. L.; Baldwin, M. J.; Caudle, M. T.; Hsieh, W. Y.; Law, N. A. *Pure Appl. Chem.* **1998**, *70*, 925–929.
- (22) Barber, J.; Ferreira, K. N.; Maghlaoui, K.; Iwata, S. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4737–4742.
- (23) Britt, R. D.; Peloquin, J. M.; Gilchrist, M. L.; Aznar, C. P.; Dicus, M. M.; Robblee, J.; Messinger, J. *Biochim. Biophys. Acta* **2004**, *1655*, 158–171.
- (24) Yachandra, V. K.; Klein, K. S. M. P. *Chem. Rev.* **1996**, *96*, 2927–2950.

- (25) Hillier, W.; Wydrzynski, T. *Biochemistry* **2000**, *39*, 4399–4405.
- (26) Hillier, W.; Wydrzynski, T. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4882–4889.
- (27) Mukhopadhyay, S.; Mandal, S. K.; Bhaduri, S.; Armstrong, W. H. *Chem. Rev.* **2004**, *104*, 3981–4026.
- (28) Miller, A. F.; Brudvig, G. W. *Biochim. Biophys. Acta* **1991**, *1056*, 1–18.
- (29) Britt, R. D.; Peloquin, J. M.; Campbell, K. A. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 463–495.
- (30) Mino, H.; Kawamori, A. *Biochim. Biophys. Acta* **2001**, *1503*, 112–122.
- (31) Boussac, A.; Un, S.; Horner, O.; Rutherford, A. W. *Biochemistry* **1998**, *37*, 4001–4007.
- (32) Kulik, L.; Epel, B.; Messinger, J.; Lubitz, W. *Photosynth. Res.* **2005**, *84*, 347–353.
- (33) Dau, H.; Liebisch, P.; Haumann, M. *Anal. Bioanal. Chem.* **2003**, *376*, 562–583.
- (34) Liang, W. C.; Roelofs, T. A.; Cinco, R. M.; Rompel, A.; Latimer, M. J.; Yu, W. O.; Sauer, K.; Klein, M. P.; Yachandra, V. K. *J. Am. Chem. Soc.* **2000**, *122*, 3399–3412.
- (35) Yachandra, V. K. *Philos. Trans. R. Soc. London, Ser. B* **2002**, *357*, 1347–1357.
- (36) Dau, H.; Liebisch, P.; Haumann, M. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4781–4792.
- (37) Robblee, J. H.; Messinger, J.; Cinco, R. M.; McFarlane, K. L.; Fernandez, C.; Pizarro, S. A.; Sauer, K.; Yachandra, V. K. *J. Am. Chem. Soc.* **2002**, *124*, 7459–7471.
- (38) Haumann, M.; Muller, C.; Liebisch, P.; Iuzzolino, L.; Dittmer, J.; Grabolle, M.; Neisius, T.; Meyer-Klaucke, W.; Dau, H. *Biochemistry* **2005**, *4*, 1894–1908.
- (39) Ke, B. *Photosynthesis: Photobiochemistry and Photobiophysics*; Academic Publishers: Dordrecht, The Netherlands, 2001.
- (40) Berthomieu, C.; Hienerwadel, R.; Boussac, A.; Breton, J.; Diner, B. A. *Biochemistry* **1998**, *37*, 10547–10554.
- (41) Chu, H.; Hillier, W.; Debus, R. J. *Biochemistry* **2004**, *43*, 3152–3166.
- (42) Debus, R. J.; Strickler, M. A.; Walker, L. M.; Hillier, W. *Biochemistry* **2005**, *44*, 1367–1374.
- (43) Strickler, M. A.; Walker, L. M.; Hillier, W.; Debus, R. J. *Biochemistry* **2005**, *44*, 8571–8577.
- (44) Kimura, Y.; Mizusawa, N.; Yamanari, T.; Ishii, A.; Ono, T. *J. Biol. Chem.* **2005**, *280*, 2078–2083.
- (45) Roffey, R. A.; Kramer, D. M.; Govindjee; Sayre, R. T. *Biochim. Biophys. Acta* **1994**, *1185*, 257–270.
- (46) Kramer, D. M.; Roffey, R. A.; Govindjee; Sayre, R. T. *Biochim. Biophys. Acta* **1994**, *1185*, 228–237.
- (47) Debus, R. J. *Biochim. Biophys. Acta* **2001**, *1503*, 164–186.
- (48) Diner, B. A. *Biochim. Biophys. Acta* **2001**, *1503*, 147–163.
- (49) Zouni, A.; Witt, H. T.; Kern, J.; Fromme, P.; Krauss, N.; Saenger, W.; Orth, P. *Nature* **2001**, *409*, 739–743.
- (50) Kamiya, N.; Shen, J. R. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 98–103.
- (51) Biesiadka, J.; Loll, B.; Kern, J.; Irrgang, K. D.; Zouni, A. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4733–4736.
- (52) Loll, B.; Kern, J.; Saenger, W.; Zouni, A.; Biesiadka, J. *Nature* **2005**, *438*, 1040–1044.
- (53) Yano, J.; Kern, J.; Irrgang, K.; Latimer, M. J.; Bergmann, U.; Glatzel, P.; Pushkar, Y.; Biesiadka, J.; Loll, B.; Sauer, K.; Messinger, J.; Zouni, A.; Yachandra, V. K. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 12047–12052.
- (54) Sauer, K.; Yachandra, V. K. *Biochim. Biophys. Acta* **2004**, *1655*, 140–148.
- (55) Dau, H.; Iuzzolino, L.; Dittmer, J. *Biochim. Biophys. Acta* **2001**, *1503*, 24–39.
- (56) Crespo, A.; Scherlis, D.; Martí, M. A.; Ordejon, P.; Roitberg, A. E.; Estrin, D. A. *J. Phys. Chem. B* **2003**, *107*, 13728–13736.
- (57) Martí, M. A.; Crespo, A.; Bari, S. E.; Doctorovich, F. A.; Estrin, D. A. *J. Phys. Chem. B* **2004**, *108*, 18073–18080.
- (58) Fernández, M. L.; Martí, M. A.; Crespo, A.; Estrin, D. A. *J. Biol. Inorg. Chem.* **2005**, *10*, 595–604.
- (59) Crespo, A.; Martí, M. A.; Estrin, D. A.; Roitberg, A. E. *J. Am. Chem. Soc.* **2005**, *127*, 6940–6941.
- (60) Friesner, R.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389–427.
- (61) Guallar, V.; Jacobson, M.; McDermott, A.; Friesner, R. A. *J. Mol. Biol.* **2004**, *337*, 227–239.
- (62) Friesner, R. A.; Baik, M.; Gherman, B. F.; Guallar, V.; Wirstam, M.; Murphy, R. B.; Lippard, S. J. *Coord. Chem. Rev.* **2003**, *238–239*, 267–290.
- (63) Gherman, B.; Goldberg, S.; Cornish, V. W.; Friesner, R. J. *J. Am. Chem. Soc.* **2004**, *126*, 7652–7664.
- (64) Derat, E.; Cohen, S.; Shaik, S.; Altun, A.; Thiel, W. *J. Am. Chem. Soc.* **2005**, *127*, 13611–13621.
- (65) Lin, H.; Schoneboom, J.; Cohen, S.; Shaik, S.; Thiel, W. *J. Phys. Chem. B* **2004**, *108*, 10083–10088.
- (66) Schoneboom, J.; Cohen, S.; Lin, H.; Shaik, S.; Thiel, W. *J. Am. Chem. Soc.* **2004**, *126*, 4017–4034.
- (67) Lundberg, M.; Blomberg, M.; Siegbahn, P. *Top. Curr. Chem.* **2004**, *238*, 79–112.
- (68) Siegbahn, P. E. M. *J. Biol. Inorg. Chem.* **2003**, *8*, 567–576.
- (69) Gascón, J. A.; Batista, V. S. *Biophys. J.* **2004**, *87*, 2931–2941.
- (70) Gascón, J. A.; Sproviero, E. M.; Batista, V. S. *J. Chem. Theory Comput.* **2005**, *1*, 674–685.
- (71) Gascón, J. A.; Sproviero, E. M.; Batista, V. S. *Acc. Chem. Res.* **2006**, *39*, 184–193.
- (72) Hendry, G.; Wydrzynski, T. *Biochemistry* **2003**, *42*, 6209–6217.
- (73) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.;

- Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision A.1; Gaussian, Inc.: Wallingford, CT, 2003.
- (74) Vacek, G.; Perry, J. K.; Langlois, J. M. *Chem. Phys. Lett.* **1999**, *310*, 189–194.
- (75) Jaguar 5.5; Schroedinger, LLC: Portland, OR, 2003.
- (76) Note that all amino acid residues are labeled according to the one-letter standard nomenclature and correspond to the D1 protein subunit unless otherwise indicated.
- (77) Sproviero, E. M.; Gascón, J. A.; McEvoy, J. P.; Brudvig, G. W.; Batista, V. S. *J. Inorg. Biochem.* **2005**, *100*, 786–800.
- (78) Noodleman, L. *J. Chem. Phys.* **1981**, *74*, 5737–5743.
- (79) Noodleman, L.; Davidson, E. R. *Chem. Phys.* **1986**, *109*, 131–143.
- (80) Noodleman, L.; Case, D. A. *Adv. Inor. Chem.* **1992**, *38*, 423–470.
- (81) Noodleman, L.; Peng, C. Y.; Case, D. A.; Mouesca, J. M. *Coord. Chem. Rev.* **1995**, *144*, 199–244.
- (82) Kronig, R. Z. *Phys.* **1931**, *70*, 317–323.
- (83) Kronig, R. Z. *Phys.* **1932**, *75*, 190–210.
- (84) Sayers, D. E.; Stern, E. A.; Lytle, F. W. *Phys. Rev. Lett.* **1971**, *27*, 1204–1207.
- (85) Stern, E. *Phys. Rev. B* **1974**, *10*, 3027–3027.
- (86) Pendry, J. In *Low Energy Electron Diffraction*; Academic Press: New York, 1974; pp 20–35.
- (87) Lee, P.; Pendry, J. *Phys. Rev. B* **1975**, *11*, 2795–2811.
- (88) Ashley, C.; Doniach, S. *Phys. Rev. B* **1975**, *11*, 1279–1288.
- (89) Gonis, A. *Green Functions for Ordered and Disordered Systems*; Elsevier: Amsterdam, 1992.
- (90) Ankudinov, A. L.; Ravel, B.; Rehr, J. J.; Conradson, S. D. *Phys. Rev. B* **1998**, *58*, 7565–7576.
- (91) Ankudinov, A. L.; Bouldin, C.; Rehr, J. J.; Sims, J.; Hung, H. *Phys. Rev. B* **2002**, *65*, 104107–104118.
- (92) The capabilities of the DFT B3LYP functional for predicting Mn–Mn distances in biomimetic oxomanganese complexes have been recently investigated.⁷⁷
- (93) Pospisil, P.; Haumann, M.; Dittmer, J.; Sole, V. A.; Dau, H. *Biophys. J.* **2003**, *84*, 1370–1386.
- (94) Yachandra, V. K.; DeRose, V. J.; Latimer, M. J.; Mukerji, L.; Sauer, K.; Klein, M. P. *Science* **1993**, *260*, 675–679.
- (95) Ono, T. A.; Noguchi, T.; Inoue, Y.; Kusunoki, M.; Matsushita, T.; Oyanagi, H. *Science* **1992**, *258*, 1335–1337.
- (96) Roelofs, T. A.; Liang, W. C.; Latimer, M. J.; Cinco, R. M.; Rompel, A.; Andrews, J. C.; Sauer, K.; Yachandra, V. K.; Klein, M. P. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 3335–3340.
- (97) Bergmann, U.; Grush, M. M.; Horne, C. R.; DeMarois, P.; Penner-Hahn, J. E.; Yocum, C. F.; Wright, D. W.; Dube, C. E.; Armstrong, W. H.; Christou, G.; Eppley, H. J.; Cramer, S. P. *J. Phys. Chem. B* **1998**, *102*, 8350–8352.
- (98) Iuzzolino, L.; Dittmer, J.; Dau, H. *Biochemistry* **1998**, *37*, 17112–17119.
- (99) Messinger, J.; Robblee, J. H.; Bergmann, U.; Fernandez, C.; Glatzel, P.; Visser, H.; Cinco, R. M.; McFarlane, K. L.; Bellacchio, E.; Pizarro, S. A.; Cramer, S. P.; Sauer, K.; Klein, M. P.; Yachandra, V. K. *J. Am. Chem. Soc.* **2001**, *123*, 7804–7820.
- (100) Zheng, M.; Dismukes, G. C. *Inorg. Chem.* **1996**, *35*, 3307–3319.
- (101) Kuzek, D.; Pace, R. J. *Biochim. Biophys. Acta* **2001**, *1503*, 123–137.
- (102) Dasgupta, J.; van Willigen, R. T.; Dismukes, G. C. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4793–4802.
- (103) Clausen, J.; Beckman, K.; Junge, W.; Messinger, J. *Plant Physiol.* **2005**, *139*, 1444–1450.
- (104) Kühne, H.; Szalai, V. A.; Brudvig, G. W. *Biochemistry* **1999**, *38*, 6604–6613.
- (105) Wincencjusz, H.; van Gorkom, H. J.; Yocum, C. F. *Biochemistry* **1997**, *36*, 3663–3670.
- (106) Olesen, K.; Andreasson, L. E. *Biochemistry* **2003**, *42*, 2025–2035.
- (107) Force, D. A.; Randall, D. W.; Britt, R. D. *Biochemistry* **1997**, *36*, 12062–12070.
- (108) Clemens, K. L.; Force, D. A.; Britt, R. D. *J. Am. Chem. Soc.* **2002**, *124*, 10921–10933.
- (109) Debus, R. J.; Barry, B. A.; Sithole, I.; Babcock, G. T.; McIntosh, L. *Biochemistry* **1988**, *27*, 9071–9074.
- (110) Metz, J. G.; Nixon, P. J.; Rogner, M.; Brudvig, G. W.; Diner, B. A. *Biochemistry* **1989**, *28*, 6960–6969.
- (111) Hays, A. M. A.; Vassiliev, I. R.; Golbeck, J. H.; Debus, R. J. *Biochemistry* **1998**, *37*, 11352–11365.
- (112) Hays, A. M. A.; Vassiliev, I. R.; Golbeck, J. H.; Debus, R. J. *Biochemistry* **1998**, *38*, 11851–11865.
- (113) Hoganson, C. W.; Babcock, G. T. *Science* **1997**, *277*, 1953–1956.
- (114) Ishikita, H.; Saenger, W.; Loll, B.; Biesiadka, J.; Knapp, E. W. *Biochemistry* **2006**, *45*, 2063–2071.
- (115) Schlodder, E.; Witt, H. T. *J. Biol. Chem.* **1999**, *274*, 30877–30392.
- (116) Junge, W.; Haumann, M.; Ahbrink, R.; Mulikidjanian, A.; Clausen, J. *Philos. Trans. R. Soc. London, Ser. B.* **2002**, *357*, 1407–1418.
- (117) Rappaport, F.; Lavergne, J. *Biochim. Biophys. Acta* **2001**, *1503*, 246–259.
- (118) Mukerji, I.; Andrews, J. C.; DeRose, V. J.; Latimer, M.; Yachandra, V. K.; Sauer, K.; Klein, M. P. *Biochemistry* **1994**, *33*, 9712–9721.
- (119) Ghosh, A.; Taylor, P. R. *Curr. Opin. Chem. Biol.* **2003**, *7*, 113–124.
- (120) Ghosh, A.; Taylor, P. R. *J. Chem. Theory Comput.* **2005**, *1*, 597–600.
- (121) Ghosh, A.; Steene, E. *J. Biol. Inorg. Chem.* **2001**, *6*, 739–752.
- (122) Weiss, R.; Bulach, V.; Gold, A.; Ternner, J.; Trautwein, A. X. *J. Biol. Inorg. Chem.* **2001**, *6*, 831–845.
- (123) Reiher, M.; Salomon, O.; Hess, B. A. *Theor. Chem. Acc.* **2001**, *107*, 48–55.
- (124) Salomon, O.; Reiher, M.; Hess, B. A. *J. Chem. Phys.* **2002**, *117*, 4729–4737.
- (125) Holthausen, M. C. *J. Comput. Chem.* **2005**, *26*, 1505–1518.

Minimalist Explicit Solvation Models for Surface Loops in Proteins

Ronald P. White and Hagai Meirovitch*

*Department of Computational Biology, University of Pittsburgh School of Medicine,
3064 Biomedical Science Tower 3, Pittsburgh, Pennsylvania 15260*

Received December 13, 2005

Abstract: We have performed molecular dynamics simulations of protein surface loops solvated by explicit water, where a prime focus of the study is the small numbers (e.g., ~ 100) of explicit water molecules employed. The models include only part of the protein (typically 500–1000 atoms), and the water molecules are restricted to a region surrounding the loop. In this study, the number of water molecules (N_w) is systematically varied, and convergence with a large N_w is monitored to reveal $N_w(\text{min})$, the minimum number required for the loop to exhibit realistic (fully hydrated) behavior. We have also studied protein surface coverage, as well as diffusion and residence times for water molecules as a function of N_w . A number of other modeling parameters are also tested. These include the number of environmental protein atoms explicitly considered in the model as well as two ways to constrain the water molecules to the vicinity of the loop (where we find one of these methods to perform better when N_w is small). The results (for the root-mean-square deviation and its fluctuations for four loops) are further compared to much larger, fully solvated systems (using $\sim 10\,000$ water molecules under periodic boundary conditions and Ewald electrostatics) and to results for the generalized Born surface area (GBSA) implicit solvation model. We find that the loop backbone can stabilize with a surprisingly small number of water molecules (as low as five molecules per amino acid residue). The side chains of the loop require a somewhat larger N_w , where the atomic fluctuations become too small if N_w is further reduced. Thus, in general, we find adequate hydration to occur at roughly 12 water molecules per residue. This is an important result because, at this hydration level, computational times are comparable to those required for GBSA. Therefore, these “minimalist explicit models” can provide a viable and potentially more accurate alternative. The importance of protein loop modeling is discussed in the context of these, and other, loop models, along with other challenges including the relevance of an appropriate free-energy simulation methodology for the assessment of conformational stability.

I. Introduction

A great amount of work has been devoted in the past 20 years to understanding the function and determining the structure (or structures) of protein loops. The latter is particularly important in homology modeling where one generates initially a partial structure (a template) of unconnected chain segments of a target protein on the basis of the

known X-ray structure of a homologous protein (or proteins); however, it still remains to determine the structure of the connecting (missing) loops. This endeavor, which is carried out by conformational search techniques or comparative modeling, is not a trivial task and is an unsolved problem for large loops;^{1–3} the structure prediction of loops constitutes a challenge also in protein engineering.

Of special interest are surface loops that take part in protein–protein and protein–ligand interactions; such loops can form “lids” over active sites of proteins, and mutagenesis

* Corresponding author. Phone: 412-648-3338. Fax: 412-648-3163. E-mail: hagaim@pitt.edu.

experiments show that residues within these loops are crucial for substrate binding or enzymatic catalysis.⁴ Typically, these loops are flexible, and their flexibility is essential for protein function. Two general recognition mechanisms related to flexibility have been defined, *induced* and *selected fit*. Thus, the conformational change between a free and a bound antibody demonstrates the flexibility of the antibody combining site, which typically includes hypervariable loops; this provides an example of induced fit as a mechanism for antibody–antigen recognition (see, for example, refs 5 and 6). Alternatively, the *selected-fit* mechanism has been suggested, where a free loop interconverts among different microstates in thermodynamic equilibrium, and one of them is selected upon binding⁷ (a microstate is a limited region in conformational space such as the helical region of a peptide). While loop flexibility can be detected by multidimensional nuclear magnetic resonance (NMR) and X-ray crystallography (in terms of elevated B factors in the latter method), using these methods to map the most stable microstates of an unbound loop (i.e., those with the lowest free energy) is problematic, and one has, therefore, to resort to molecular modeling techniques.

The interest in surface loops has yielded extensive theoretical work, where one avenue of research has been the classification of loop structures.^{7,8–15} However, to understand various recognition mechanisms such as those mentioned above, it is mandatory to be able to predict the structure of a loop by theoretical/computational procedures. The commonly used methodologies in this category are comparative modeling based on known loop structures from the Protein Data Bank (PDB),^{16,17} an energetic modeling (based on a force field), and methods that are hybrids of these two approaches. However, mapping the most stable microstates can only be achieved with the energetic approach that consists of calculating the loop–loop and loop–protein interaction energies. To be able to apply such calculations to a large number of loops, the entire protein structure has typically been kept fixed in its X-ray structure (and sometimes only part of it has been considered). Because of the exposure of surface loops to the solvent, the development of adequate modeling of solvation is mandatory. The most stable microstates can then be generated by a combination of conformational search techniques (simulated annealing, the bond relaxation algorithm, the local torsional deformation method, etc.); thermodynamic sampling methods, such as molecular dynamics (MD) simulation or Monte Carlo; and methods for calculating the free energy.^{18–31}

Modeling of the solvent is of special importance. In some of the earlier studies, the solvation problem was not addressed at all, while others only use a distance-dependent dielectric function (i.e., $\epsilon = r_{ij}$ is substituted in the Coulomb potential, $E = q_i q_j / [r_{ij} \epsilon]$, making the interactions decay more rapidly as r_{ij}^{-2}). Better treatments of solvation were applied by Moulton and James²³ and Mas et al.³² A systematic comparison of solvation models was first carried out by Smith and Honig,³³ who tested the $\epsilon = r$ model against results obtained by the finite difference Poisson Boltzmann calculation including a hydrophobic term; the implicit solvation model of Wesson and Eisenberg³⁴ with $\epsilon = r$ was also studied by them. Later,

the generalized Born surface area (GBSA) model³⁵ was applied to loops of ribonuclease A³⁶ and has been found by Blundell's group to discriminate better than other models between the native loop structures and close-to-native “decoy” structures.^{37,38} Very recently, an extensive study of loops was carried out by Jacobson et al.,³⁹ who used the surface GB⁴⁰ and a nonpolar solvation model⁴¹ (SGB–NP) with the OPLS force field.⁴² Zhang et al.⁴³ have tested their knowledge-based statistical potential, DFIRE (distance-scaled, finite ideal gas reference state), by applying it to the loop sets studied in refs 37–39. Another interesting loop prediction algorithm has been suggested by Xiang et al.,⁴⁴ and finally, we mention our loop studies, using a simplified implicit model.^{30,31}

The popularity of implicit solvent models for loops stems from their relative simplicity and the fact that the loops are applicable to a wide range of conformational search techniques, in particular, those that are based on energy minimization. At least in principle, an energy-minimized implicit model can be used as a gauge of loop stability (i.e., the free energy), because the solvent coordinates have been “averaged out”. (Note, however, that this still does not account for the very important free-energy contribution associated with the movement of the loop atoms within a microstate.) On the other hand, explicit solvation—the more accurate modeling—is computationally expensive and allows application of limited types of search techniques. Therefore, systematic studies of loop structure prediction with explicit water have not been carried out; however, certain problems involving loops have been studied with explicit water.⁴⁵

While the quality of these implicit models for loops has not been compared, most of them were found to be adequate for predicting the backbone structure of loops (in the known protein framework) of up to nine residues [i.e., a prediction within 1 Å root-mean-squared deviation (RMSD) from the X-ray crystal structure²³]. However, the correlation between low free energy and low RMSD of structures generated by conformational search were found to be unsatisfactory (in particular, for highly charged loops), meaning that implicit modeling, in most cases, is not suitable for mapping the most stable microstates, and for that, one will have to resort to explicit solvation models. We have a special interest in such problems, as discussed in refs 30, 31, and 46 and in the Conclusions section.

Therefore, the objective of this article is to examine the validity (and efficiency) of explicit solvation models defined within the framework of the limited model mentioned above, where the loop moves in the presence of a fixed protein structure. Here, the loop is “capped” with a number of water molecules (N_w), and our aim is to determine the minimal N_w which still leads to reliable results. More specifically, we use the TIP3P model of water⁴⁷ and simulate the protein–loop–water system by MD,^{48,49} where only the loop atoms and the surrounding waters are allowed to move while the rest of the protein atoms are kept in their X-ray coordinates; moreover, to further save computer time, we retain in the model only the part of the protein that is close to the loop. To gauge performance, the RMSDs of the heavy backbone

and side-chain atoms from the X-ray structure are calculated together with the RMSD fluctuations and other quantities.

For the test cases studied here, the X-ray backbone loop structure is well-determined; that is, its atoms are defined with relatively low B factors; therefore, if the simulation starts from the X-ray structure, for a large enough N_w , one would expect the simulated backbone to demonstrate stability, that is, to remain close to this structure for long simulation times, while for a small N_w , the backbone might escape to another microstate. On the other hand, some of the coordinates of the side-chain atoms are typically poorly resolved (high B factors), and in general, the side-chain environment in the simulation could be expected to mimic the experimental solution environment better than that of the crystal; therefore, for the side chains, one would not expect the simulation to always reproduce the crystallographic data. However, as N_w is increased, the structure of the simulated side chains is expected to stabilize at some microstate. These are some of the criteria according to which the results are analyzed. (It should be emphasized, however, that during a long enough simulation the loop will change microstates, and therefore, such an analysis should be carried out with caution.) Finally, as further criteria to test the validity of the restricted (or “minimalist”) loop models studied here, we solvate the corresponding (entire) proteins with water under periodic boundary conditions, simulate them by MD, and compare the RMSD and fluctuations of the loops to those obtained from the restricted solvation models.

In this work, extensive MD simulation studies are carried out for four loops ranging in size from 8 to 10 residues; the loops are taken from the three proteins ribonuclease A (RNase A), ser-proteinase, and proteinase. (We also report results from less extensive tests conducted on several other loops.) As mentioned above, different N_w 's are tested for each loop (and other modeling parameters discussed below), and the minimal N_w [$N_w(\text{min})$] which reproduces the large N_w behavior is determined. While $N_w(\text{min})$ depends on various properties of the loop and its associated nearby protein environment, for the four primary loops studied, we find $N_w(\text{min}) \sim 12$ per residue, which (using the AMBER96 force field⁵⁰ programmed in the package TINKER⁵¹) requires comparable computer time to running MD based on the implicit solvent, GBSA.³⁵ It is also shown that for two loops the GBSA results deviate significantly from those obtained with the explicit solvent. While these results are expected to be typical, they should be validated for each loop studied.

It should be pointed out that approximate explicit solvation models, where only part of the protein (around the active site) is considered (and solvated), have been suggested before. One of the first was the stochastic boundary model of Karplus' group, where the region of interest (including the protein and the solvent) is divided into subregions of decreasing importance;⁵² we have used this model for calculating the backbone entropy of loops in the protein ras.⁵³ In many other studies of ligands in active sites, caps of water molecules were built around these sites, with the number of water molecules typically increasing as computers have become more powerful. For example, in 1986, Bash et al.⁵⁴ used only 168 waters to cover the active site of thermolysine

in their calculations of the relative free energy of binding of two inhibitors, whereas in 1991, Merz used 300 waters for calculating the binding of CO₂ to human carbonic anhydrase II.⁵⁵ In 1993, Miyamoto and Kollman used 205 waters to solvate the active site of streptavidin in their calculation of the absolute free energy of binding of biotin and other similar ligands to this protein.⁵⁶ In 1997, Jorgensen's group capped 482 waters around the active site of trypsin and calculated the binding affinities of trypsin–benzamidine complexes;⁵⁷ however, in later publications of this group, caps including up to 1600 waters were used.⁵⁸ In most of these works, a systematic investigation of the effect of the number of water molecules has not been carried out. Our present study has been largely motivated by the work of Steinbach and Brooks,⁵⁹ who studied, by MD, the change in the RMSD of protein structures from their X-ray structures with an increasing number of water molecules; they found that a relatively small number of waters led to the behavior of the fully solvated system.

II. Methods

II.1. Models. Our investigations are focused on the solvation of protein surface loops with small numbers of explicit water molecules, N_w . The protein portion of these models is further limited to just the loop atoms, and only the protein atoms belonging to residues that are close to the loop. We will refer to this as the “partial-protein model”. To test the approximations inherent in this model (which are chiefly limited solvation, a reduced protein environment, and lack of flexibility in the template), we also model the entire protein, solvated under periodic boundary conditions with particle mesh Ewald electrostatics. This model is referred to as the “full-protein model”. Both models will be described in detail in the following sections.

All computational work associated with the partial-protein modeling (i.e., structure preparation and simulations) was performed using the TINKER software package (version 4.2),⁵¹ which was modified to suit our specific needs. The computational work for the full-protein models (structure preparation, simulations, and analysis) was performed using a variety of programs in the AMBER software package (version 8). For both models, we used the AMBER96 force field,⁵⁰ where His is in the doubly protonated state (charge = +1) and four other residues are also modeled in their respective neutral pH charged states, Lys (+1), Arg (+1), Asp (−1), and Glu (−1). The water molecules are modeled with the three-site TIP3P potential.⁴⁷

II.2. Construction of the Partial-Protein Model. The starting coordinates for the partial-protein model are taken from the PDB X-ray structure (where hydrogen atoms and disulfide bonds are added in the usual manner). As stated above, the loop atoms, and only the protein atoms that are close to the loop, are included in the model. The nonloop atoms which are retained in the model are collectively referred to as the “template”. To construct the template (see also Figure 1), the center of mass of the loop backbone atoms is calculated as a reference point. We denote the coordinates of this point as \mathbf{x}_{cmb} . A distance (R_{temp}) is chosen such that residues that are greater than R_{temp} from \mathbf{x}_{cmb} are not included

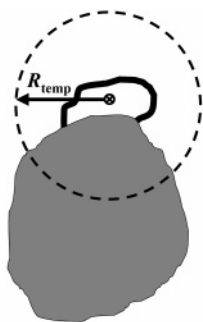


Figure 1. Diagram showing the region of the protein that is retained (the “template”) in the partial-protein model. The loop is represented as the heavy black curve. The remainder of the protein is shown as a gray blob. The center of mass of the loop backbone, \mathbf{x}_{cmb} , is located at the position marked as \otimes . The protein template is “cut out” at the dashed circle (a sphere in three dimensions), which is defined by the distance R_{temp} measured from \mathbf{x}_{cmb} . All protein residues that are inside this region are considered in the model, thus defining the nearby protein environment for the loop.

Table 1. Diffusion Properties of Water Molecules Calculated for the Partial-Protein Model of RNase A [64–71]^a

N_w	R_{cap} (Å)	$\langle N_{\text{surf}} \rangle$	$\langle N_{\text{surf}}/N_w \rangle$	D_{all}	D_{surf}	τ_{all} (ps)	τ_{surf} (ps)
300	20	103.2	0.344	4.96	2.61	6.8	12.9
200	19	96.8	0.484	4.03	2.53	8.4	13.3
120	18	79.1	0.659	2.73	1.98	12.4	17.0
70	17	57.8	0.825	1.71	1.43	19.8	23.6
50	17	44.4	0.888	1.28	1.12	26.5	30.2

^a N_w is the number of water molecules. R_{cap} is the radius of the spherical solvent restraining region (SPH restraint). The same protein template ($R_{\text{temp}} = 15$ Å) was used in all cases. $\langle N_{\text{surf}}/N_w \rangle$ is the (average) fraction of water molecules observed at the surface of the protein. D_{all} is the diffusion constant calculated for all N_w water molecules. D_{surf} is the diffusion constant calculated for just the water molecules at the protein surface. Units for D_{all} and D_{surf} are 10^{-5} cm²/s. τ_{all} and τ_{surf} are estimated residence times defined by the time for a water molecule to diffuse a distance of 4.5 Å. τ_{all} is calculated for all N_w water molecules, and τ_{surf} is for the protein surface water only. Statistical uncertainties in $\langle N_{\text{surf}}/N_w \rangle$, D_{all} (D_{surf}), and τ_{all} (τ_{surf}) are typically less than 0.003, 0.05×10^{-5} cm²/s, and 0.5 ps, respectively. Other details and definitions are given in the text.

in the template. More specifically, if any atom in a protein residue is less than the distance R_{temp} , from \mathbf{x}_{cmb} , the entire residue is included in the template. Otherwise, the residue is eliminated. Obviously, the choice of R_{temp} will determine the number of environmental protein atoms to be included in the model. Atom numbers for various R_{temp} values are given in Table 1 for each of the loops studied.

The starting (PDB) coordinates for the loop and template atoms are relaxed to a nearby geometry. This minimization is carried out using additional harmonic positional restraints ($k = 5$ kcal mol⁻¹ Å⁻²), which are applied to all heavy atoms. This eliminates bad atomic overlaps and strains in the original structure, while keeping the atoms still reasonably close to the PDB coordinates. These resulting relaxed coordinates are referred to as the “X-ray reference coordinates” and are denoted as \mathbf{X}_{ref} . [Note that \mathbf{x}_{cmb} (above) is a single point in 3D space, whereas \mathbf{X}_{ref} specifies the whole coordinate set

for a group of atoms.] The loop coordinates from this configuration are used in the RMSD calculations, described below.

As outlined in the Introduction, MD simulations of the loop are carried out in the presence of the nearby template atoms, along with the N_w water molecules. Specifically, the coordinates of the loop atoms evolve in time under the influence of interactions with the template atoms, the water molecules, and each other. The water molecules are also mobile; they interact with each other, the protein atoms (in both the loop and template), and the boundary of a containment region (described below). The template atoms, however, are fixed in these simulations at their respective coordinates in \mathbf{X}_{ref} (where the purpose of this approximation is to increase the computational efficiency, as it is then unnecessary to calculate template-atom–template-atom interactions).

II.3. Solvation of the Partial-Protein Model. To make best use of (the solvating effects of) the limited number of water molecules, they are restricted to a region that is close to the loop. This also prevents evaporation. The situation is similar to “capping” an active site, where one wishes to keep water molecules near the most critical region of the model investigation. Unlike many active sites, however, which tend to be concave, a solvation region around a surface loop tends to be more convex and, thus, can present more of a challenge. We have implemented two methods to restrain the water molecules to the vicinity of the loop. One involves a (semi-) spherical restraining region, which we call the SPH restraint. The other is a nearest-loop-atom-based restraint, which we call the NLA restraint. Both will be described in detail below.

II.3.1. Spherical Restraining Region. In the SPH restraint, water molecules are restrained with a flat-welled half-harmonic potential (force constant, $k = 5$ kcal mol⁻¹ Å⁻²), based on the distance from the “center” of the loop region. That is, the distance of each water molecule (in practice, the oxygen atom) is measured from a restraining center (\mathbf{x}_{sph}). If this distance is greater than a prescribed distance, R_{cap} , a harmonic restoring force is applied; otherwise, the restraining force is zero.

A reasonable restraining center could be, for example, the center of mass of the loop backbone atoms (i.e., $\mathbf{x}_{\text{sph}} = \mathbf{x}_{\text{cmb}}$). The choice of R_{cap} , on the other hand, should be roughly based on the number (N_w) of water molecules used. It is important to note, however, that a range of reasonable R_{cap} values can be found; but obviously, for large N_w 's, small values of R_{cap} would be undesirable. (This scenario would be evidenced, for example, by a large average value for the restraining potential.) Some examples of R_{cap} at various values of N_w are available in Tables 2–7, where, in general, R_{cap} increases with N_w . In our modeling, the restraining volume is typically quite large for the given number of water molecules (i.e., much of the available volume is empty). A more detailed discussion describing the nature of the solvation within the partial-protein model will be given in section III.

In most cases, we have taken values of R_{cap} where $R_{\text{cap}} \geq R_{\text{temp}}$. (As described above, R_{temp} is the distance value used to determine the size of the template.) However, for large-enough values of R_{cap} , water molecules can migrate away

Table 2. Description of Loops and Modeling Parameters^a

protein	number of atoms: protein	N_w^{pbc}	loop residues	sequence	R	number of atoms: loop	R_{temp}	number of atoms: loop + template
RNase A (1rat)	1860	6808	64–71 (8)	ACKNGQTN	3.2	107	14, 15, 16	526, 572, 590
ser-proteinase (2ptn)	3223	9320	143–151 (9)	NTKSSGTSY	4.9	117	13, 14, 15	498, 578, 738
proteinase (2apr)	4714	12393	128–137 loop1 (10)	DTITTVRGVK	4.3	158	11, 13, 15	497, 731, 1035
proteinase (2apr)	4714	12393	188–196 loop2 (9)	IDNSRGWWG	4.5	143	11, 13, 15	569, 775, 1034

^a Atom numbers are provided for different portions of the system: the entire protein, the loop atoms only, and the loop together with the template. The number of atoms in the latter depends on the template radius parameter R_{temp} (in Å), where the values separated by commas give rise to the corresponding (comma separated) atom numbers. N_w^{pbc} is the number of water molecules used in the full-protein simulations. Loop sequences are given with the charged residues as bold-faced letters. R is the ratio between the length of the stretched loop and the distance between the C $^{\alpha}$ of the first and last residues of the loop.

Table 3. Partial-Protein Model Results for RNase A [64–71]^a

N_w	R_{temp}	water restraint	R_{cap} (or R_{nla})	RMSD(BB)	RMSD(SC)	σ (BB)	σ (SC)	σ^w (BB)	σ^w (SC)
300	15	SPH	20	0.57 (5)	1.31 (4)	0.19 (6)	0.48 (3)	0.14 (1)	0.28 (1)
200	15	SPH	19	0.54 (2)	1.13 (11)	0.17 (1)	0.38 (8)	0.15 (1)	0.24 (2)
200	15	SPH	16	0.55 (2)	1.23 (8)	0.17 (2)	0.44 (5)	0.15 (1)	0.26 (2)
200	16	SPH	19	0.56 (3)	1.22 (12)	0.17 (1)	0.37 (5)	0.15 (1)	0.25 (2)
120	14	SPH	18	0.64 (23)	1.21 (19)	0.18 (4)	0.31 (6)	0.15 (2)	0.21 (5)
120	15	SPH	18	0.54 (4)	1.02 (6)	0.17 (2)	0.29 (3)	0.15 (1)	0.20 (3)
120	15	NLA	8.5	0.67 (19)	1.09 (8)	0.24 (10)	0.36 (6)	0.15 (1)	0.23 (3)
120	16	SPH	18	0.61 (16)	1.35 (4)	0.21 (7)	0.34 (2)	0.14 (1)	0.23 (2)
100	15	SPH	16	0.52 (1)	1.00 (11)	0.15 (1)	0.27 (8)	0.14 (1)	0.20 (3)
70	14	SPH	14	0.50 (2)	1.27 (14)	0.15 (1)	0.30 (4)	0.13 (1)	0.21 (2)
70	15	SPH	17	0.50 (4)	1.02 (14)	0.17 (3)	0.26 (8)	0.14 (1)	0.17 (3)
70	15	NLA	7	0.58 (14)	1.00 (8)	0.19 (9)	0.27 (5)	0.14 (0)	0.19 (2)
70	16	SPH	16	0.52 (4)	1.40 (8)	0.18 (3)	0.33 (7)	0.15 (2)	0.21 (3)
50	15	SPH	17	0.50 (3)	0.99 (9)	0.18 (3)	0.22 (3)	0.15 (1)	0.15 (1)
50	15	NLA	7	0.49 (1)	1.10 (10)	0.14 (1)	0.28 (3)	0.12 (1)	0.18 (2)
50	15	SPH	16	0.57 (13)	1.09 (38)	0.21 (6)	0.29 (20)	0.15 (2)	0.16 (2)
40	15	SPH	16	0.55 (10)	1.10 (7)	0.20 (9)	0.27 (5)	0.14 (3)	0.16 (1)
30	15	SPH	16	0.55 (8)	1.07 (6)	0.21 (7)	0.21 (5)	0.15 (4)	0.15 (3)
20	15	SPH	16	0.51 (5)	0.99 (5)	0.19 (4)	0.17 (2)	0.14 (2)	0.13 (1)
10	15	SPH	16	0.67 (10)	1.21 (5)	0.22 (5)	0.18 (3)	0.18 (3)	0.15 (2)
5	15	SPH	16	1.03 (39)	1.59 (41)	0.27 (6)	0.24 (10)	0.21 (3)	0.16 (2)
0	15			2.30 (85)	2.60 (82)	0.48 (36)	0.44 (37)	0.18 (2)	0.13 (1)
GBSA	15			1.93 (48)	2.71 (62)	0.54 (11)	0.70 (16)	0.29 (5)	0.32 (4)

^a N_w is the number of water molecules. R_{temp} (in Å) is a radius parameter defining the size of the template. The water restraint method is either “SPH” (spherical restraining region) or “NLA” (nearest-loop-atom-based restraint), which are described (respectively) by the parameters R_{cap} or R_{nla} (Å). The RMSD values (eq 2, averaged over all five trajectories) for the loop backbone (BB) and side-chain (SC) atoms are denoted by RMSD(BB) and RMSD(SC), respectively. The corresponding RMSD fluctuations (eqs 3 and 6, averaged over all five trajectories) are denoted σ (BB) and σ (SC), while the window-averaged RMSD fluctuations are denoted as σ^w (BB) and σ^w (SC). The numbers in parentheses are the standard deviations of the individual results from the five trajectories. For example, 1.31 (4) means that the standard deviation is 0.04, and 1.09 (38) implies a standard deviation of 0.38. All RMSD values and their fluctuations, σ , are reported in Å.

from the loop, around to the “back side” of the template, where their solvation effect is wasted. For this reason, we actually choose a restraining center such that

$$\mathbf{x}_{sph} = \mathbf{x}_{cmb} + (R_{cap} - R_{temp})(\mathbf{x}_{cmb} - \mathbf{x}_{cm})/|\mathbf{x}_{cmb} - \mathbf{x}_{cm}| \quad (1)$$

where \mathbf{x}_{cm} is the overall center of mass of the loop-template system. Here, the effect is to shift the center of the restraining sphere (\mathbf{x}_{sph}) toward the “loop side” of the loop-template system (see Figure 2). This serves to keep the water molecules away from the back of the template, because the van der Waals radii of the “back side” template atoms will now be closer to the wall of the restraining sphere. At the same time, there will be sufficient room for water on the “loop side” of the template.

II.3.2. Nearest-Loop-Atom-based Restraint. A slightly more elaborate restraint option for the water molecules is to employ a flat-welled half-harmonic potential ($k = 5 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$) that is based on the distance to the nearest loop atom (for an example, see ref 60). Specifically, for each water molecule, the distance to the nearest loop atom is calculated and then compared to a prescribed distance value, R_{nla} . If this distance is less than R_{nla} , then there are no restraining forces. If the distance is greater than R_{nla} , then a harmonic restoring force is applied to (the oxygen of) the water molecule and is directed along the vector between the water and the nearest loop atom (see Figure 3).

The NLA restraint is arguably advantageous compared to the SPH restraint, because the implementation can be

Table 4. Partial-Protein Model Results for Ser-Proteinase [143–151]^a

N_w	R_{temp}	water restraint	R_{cap} (or R_{nla})	RMSD(BB)	RMSD(SC)	$\sigma(BB)$	$\sigma(SC)$	$\sigma^w(BB)$	$\sigma^w(SC)$
300	13	SPH	20	0.69 (1)	1.51 (3)	0.14 (1)	0.26 (2)	0.13 (1)	0.20 (1)
200	13	SPH	19	0.69 (1)	1.53 (2)	0.14 (1)	0.25 (2)	0.13 (0)	0.20 (1)
200	15	SPH	19	0.69 (1)	1.44 (3)	0.12 (0)	0.26 (2)	0.12 (0)	0.20 (1)
120	13	SPH	18	0.68 (1)	1.51 (7)	0.14 (1)	0.29 (3)	0.12 (1)	0.20 (1)
120	13	NLA	8.5	0.67 (1)	1.50 (7)	0.13 (1)	0.26 (3)	0.12 (1)	0.19 (1)
120	14	SPH	18	0.67 (3)	1.54 (11)	0.14 (1)	0.29 (3)	0.12 (1)	0.20 (1)
120	15	SPH	18	0.64 (1)	1.39 (8)	0.12 (1)	0.27 (2)	0.11 (1)	0.19 (1)
70	13	SPH	17	0.80 (11)	1.49 (13)	0.22 (6)	0.32 (4)	0.15 (2)	0.18 (1)
70	13	NLA	7	0.69 (3)	1.46 (3)	0.17 (3)	0.28 (3)	0.13 (1)	0.20 (1)
70	14	SPH	17	0.83 (10)	1.57 (15)	0.25 (5)	0.37 (13)	0.16 (1)	0.19 (1)
70	14	NLA	7	0.65 (1)	1.40 (4)	0.13 (1)	0.27 (3)	0.12 (1)	0.19 (1)
50	13	SPH	17	1.28 (40)	1.88 (32)	0.30 (7)	0.35 (7)	0.17 (4)	0.17 (3)
50	13	NLA	7	0.75 (5)	1.55 (5)	0.21 (6)	0.32 (1)	0.15 (1)	0.19 (1)
GBSA	13			0.71 (3)	1.52 (9)	0.18 (2)	0.31 (2)	0.16 (1)	0.24 (2)

^a The various parameters are defined in the captions of Table 3.

Table 5. Partial-Protein Model Results for Proteinase [128–137] (Loop 1)^a

N_w	R_{temp}	water restraint	R_{cap} (or R_{nla})	RMSD(BB)	RMSD(SC)	$\sigma(BB)$	$\sigma(SC)$	$\sigma^w(BB)$	$\sigma^w(SC)$
300	13	SPH	20	0.74 (3)	2.19 (12)	0.12 (1)	0.37 (11)	0.09 (1)	0.17 (4)
200	13	SPH	19	0.73 (2)	2.16 (19)	0.12 (0)	0.40 (7)	0.10 (1)	0.21 (5)
120	13	SPH	18	0.66 (2)	2.31 (14)	0.12 (2)	0.30 (7)	0.10 (1)	0.14 (3)
120	15	SPH	18	0.71 (5)	2.25 (13)	0.12 (1)	0.18 (3)	0.10 (1)	0.12 (2)
70	11	SPH	17	0.72 (2)	2.19 (2)	0.10 (0)	0.22 (3)	0.09 (0)	0.11 (2)
70	11	NLA	7	0.70 (3)	2.29 (13)	0.09 (1)	0.27 (5)	0.08 (0)	0.13 (2)
70	13	SPH	17	0.67 (3)	2.41 (8)	0.11 (1)	0.12 (3)	0.09 (1)	0.08 (1)
70	13	NLA	7	0.67 (3)	2.33 (6)	0.12 (1)	0.20 (3)	0.10 (1)	0.11 (2)
70	15	SPH	17	0.74 (3)	2.18 (12)	0.08 (1)	0.12 (4)	0.07 (1)	0.08 (1)
70	15	NLA	7	0.72 (4)	2.21 (11)	0.10 (1)	0.15 (5)	0.08 (1)	0.10 (3)
50	13	SPH	17	0.66 (2)	2.41 (3)	0.10 (1)	0.13 (1)	0.09 (1)	0.08 (1)
50	13	NLA	7	0.63 (1)	2.43 (9)	0.10 (1)	0.12 (6)	0.09 (1)	0.07 (1)
40	13	SPH	16	0.68 (4)	2.34 (16)	0.09 (1)	0.11 (3)	0.08 (1)	0.08 (1)
30	13	SPH	16	0.70 (4)	2.43 (4)	0.09 (2)	0.11 (2)	0.07 (1)	0.07 (0)
20	13	SPH	16	0.72 (5)	2.46 (6)	0.21 (6)	0.32 (1)	0.15 (1)	0.19 (1)
10	13	SPH	16	0.89 (11)	2.63 (9)	0.11 (4)	0.13 (4)	0.07 (1)	0.08 (1)
5	13	SPH	16	0.84 (17)	2.56 (12)	0.07 (1)	0.11 (6)	0.06 (1)	0.07 (1)
0	13			1.08 (13)	2.65 (24)	0.07 (4)	0.11 (2)	0.05 (1)	0.09 (2)
GBSA	13			0.79 (8)	2.88 (14)	0.18 (2)	0.31 (2)	0.16 (1)	0.24 (2)

^a The various parameters are defined in the captions of Table 3.

somewhat less-dependent on the loop-template geometry, as it is able to effect a “glovelike” fit to the loop regardless of the conformation. Again, the choice of R_{nla} should be based roughly on the number of water molecules used in the model (and noting again, however, that acceptable performance can be obtained over a range of reasonable values). For very small N_w 's, we often choose R_{nla} 's to be roughly two water-molecule diameters, plus a little fluctuation room (e.g., 7 Å). For larger N_w 's, R_{nla} is increased somewhat. In general, the restraining volume is typically still large for the given number of water molecules.

II.4. Details of the Partial-Protein Simulations. Above, we described the initial preparation (from PDB coordinates) of the loop-template system, thus resulting in the coordinates \mathbf{X}_{ref} . A cluster of N_w water molecules is then added to this system. The center of mass of the water cluster is initially positioned, away from the protein atoms (such that there are no van der Waals overlaps), in the direction of $(\mathbf{x}_{cmb} - \mathbf{x}_{cm})/$

$|\mathbf{x}_{cmb} - \mathbf{x}_{cm}|$ (i.e., on the “loop side” of the loop-template system). The positions of the water molecules are then energy minimized, keeping all protein atoms fixed at \mathbf{X}_{ref} (and subject to the water restraints described above). Following this minimization, 300 ps of MD simulation is performed to equilibrate the water molecules, keeping the protein atoms fixed at \mathbf{X}_{ref} . The first 50 ps is run at 600 K, followed by 50 ps at 450 K. These higher temperatures allow the water molecules to spread out and explore the entire protein surface (within the allowable restraining volume). The remaining 200 ps is run at 300 K.

As mentioned above, the main (production) MD simulations consist of the moveable loop atoms and water molecules (subject to the SPH or NLA restraints), in the presence of the fixed template. Therefore, following the above equilibration, the protein loop atoms are allowed to move (along with the water) and are equilibrated (at 300 K) for 30 ps. The production MD simulations are performed at 300 K and are

Table 6. Partial-Protein Model Results for Proteinase [188–196] (Loop 2)^a

N_w	R_{temp}	water restraint	R_{cap} (or $R_{\text{nl}a}$)	RMSD(BB)	RMSD(SC)	σ (BB)	σ (SC)	σ^w (BB)	σ^w (SC)
300	13	SPH	20	0.63 (44)	1.50 (50)	0.12 (1)	0.37 (11)	0.09 (1)	0.17 (4)
200	13	SPH	19	0.49 (3)	1.45 (19)	0.18 (5)	0.43 (11)	0.12 (0)	0.23 (1)
120	13	SPH	18	0.46 (4)	1.42 (18)	0.15 (2)	0.32 (10)	0.11 (1)	0.17 (2)
120	15	SPH	18	0.54 (6)	1.69 (36)	0.16 (1)	0.43 (15)	0.11 (1)	0.17 (2)
120	15	NLA	8.5	0.52 (7)	1.67 (41)	0.16 (2)	0.36 (12)	0.12 (1)	0.19 (4)
70	11	SPH	17	0.45 (9)	1.63 (6)	0.18 (12)	0.24 (6)	0.11 (1)	0.16 (2)
70	11	NLA	7	0.44 (2)	1.57 (10)	0.15 (2)	0.29 (2)	0.11 (1)	0.17 (2)
70	13	SPH	17	0.65 (42)	1.95 (36)	0.19 (8)	0.24 (6)	0.12 (2)	0.15 (2)
70	13	NLA	7	0.52 (5)	1.54 (17)	0.13 (1)	0.25 (7)	0.12 (1)	0.17 (5)
70	15	SPH	17	0.46 (5)	2.25 (8)	0.12 (1)	0.23 (10)	0.10 (0)	0.15 (1)
70	15	NLA	7	0.53 (6)	1.73 (39)	0.18 (8)	0.39 (15)	0.12 (2)	0.18 (4)
50	13	SPH	17	0.45 (4)	1.84 (19)	0.14 (2)	0.25 (5)	0.11 (1)	0.13 (2)
50	13	NLA	7	0.57 (13)	1.46 (20)	0.17 (5)	0.22 (4)	0.11 (1)	0.13 (3)
GBSA	13			1.16 (50)	2.86 (70)	0.32 (7)	0.56 (26)	0.17 (2)	0.26 (3)

^a The various parameters are defined in the captions of Table 3.

Table 7. Comparison of the Partial-Protein and Full-Protein Model Results^a

N_w (or N_w^{pbc})	protein model	superpose	RMSD(BB)	RMSD(SC)	σ (BB)	σ (SC)	σ^w (BB)	σ^w (SC)
RNase A [64–71]								
6808	full-protein	yes	0.61 (13)	1.62 (39)	0.18 (8)	0.46 (22)	0.11 (2)	0.22 (2)
300	partial-protein $R_{\text{temp}} = 15 \text{ \AA}$	yes	0.42 (5)	1.03 (5)	0.15 (6)	0.37 (3)	0.11 (0)	0.19 (1)
300	partial-protein $R_{\text{temp}} = 15 \text{ \AA}$	no	0.57 (5)	1.31 (4)	0.19 (6)	0.48 (3)	0.14 (1)	0.28 (1)
Ser-Proteinase [143–151]								
9320	full-protein	yes	0.57 (13)	1.33 (22)	0.15 (7)	0.29 (11)	0.12 (2)	0.17 (2)
300	partial-protein $R_{\text{temp}} = 13 \text{ \AA}$	yes	0.44 (1)	1.12 (3)	0.10 (2)	0.22 (1)	0.09 (1)	0.15 (1)
300	partial-protein $R_{\text{temp}} = 13 \text{ \AA}$	no	0.69 (1)	1.51 (3)	0.14 (1)	0.26 (2)	0.13 (1)	0.20 (1)
Proteinase Loop 1 [128–137]								
12 393	full-protein	yes	1.04 (20)	2.47 (46)	0.24 (9)	0.54 (9)	0.13 (4)	0.22 (6)
300	partial-protein $R_{\text{temp}} = 13 \text{ \AA}$	yes	0.59 (2)	2.00 (12)	0.10 (2)	0.34 (10)	0.08 (1)	0.13 (3)
300	partial-protein $R_{\text{temp}} = 13 \text{ \AA}$	no	0.74 (3)	2.19 (12)	0.12 (1)	0.37 (11)	0.09 (1)	0.17 (4)
Proteinase Loop 2 [188–196]								
12 393	full-protein	yes	0.72 (27)	1.64 (36)	0.17 (6)	0.50 (15)	0.10 (1)	0.24 (6)
300	partial-protein $R_{\text{temp}} = 13 \text{ \AA}$	yes	0.48 (33)	1.29 (39)	0.17 (10)	0.43 (14)	0.09 (1)	0.18 (2)
300	partial-protein $R_{\text{temp}} = 13 \text{ \AA}$	no	0.63 (44)	1.50 (50)	0.12 (1)	0.37 (11)	0.09 (1)	0.17 (4)

^a N_w and N_w^{pbc} denote the number of water molecules used in the partial- and full-protein models, respectively. Results for the partial-protein model were obtained using the spherical restraining method with a radius parameter of $R_{\text{cap}} = 20 \text{ \AA}$ in all cases. The superpose column indicates whether RMSD values were minimized by superposing structures (see text). Other parameters are defined in the caption of Table 3.

run to a length of 5 ns. Five independent 5 ns production runs are carried out for each system investigated.

Other important simulation details are as follows. The velocity form of the Verlet algorithm⁶¹ is used to integrate the equations of motion with a time step of 1 fs. The RATTLE⁶² algorithm is used to fix all bonds involving hydrogen atoms in the loop and to maintain the rigid geometry of the TIP3P water molecules. The temperature is maintained using a Berendsen thermostat⁶³ (weak coupling method) with a time constant of 0.1 ps. No distance-based cutoffs are applied to the nonbonded [Lennard-Jones (LJ) and Coulombic] interactions.

As mentioned in the Introduction, the explicit water partial-protein results are compared with results obtained from MD calculations carried out with the GBSA implicit solvation model of Still and co-workers,³⁵ as implemented within TINKER (using the same simulation parameters described above).

II.5. The Full-Protein Model and Simulations. Starting with the PDB coordinates (with added hydrogens and disulfide bonds), the entire protein was solvated in a rectangular box, giving a 10 \AA (11 \AA for ser-proteinase) buffer distance to each wall of the box, as implemented in LEaP. All of the crystallographic waters for ser-proteinase, and some of the waters for proteinase (the interior waters), were kept from the PDB files. Counterions (Na^+ or Cl^-) were added to make the overall system charge neutral. The resulting numbers of water molecules are given for each protein in Table 1 (denoted N_w^{pbc}).

To eliminate any bad contacts/strains, the entire system is energy minimized with harmonic positional restraints ($k = 100 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$) applied to all protein atoms. This is followed by a second minimization under weaker positional restraints ($k = 10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$). The coordinates resulting from these minimizations are used as a starting point for the MD simulations. They are also taken as the “X-ray reference

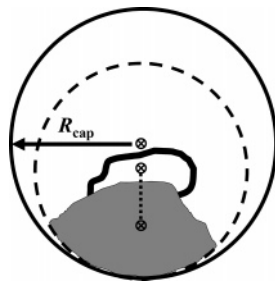


Figure 2. Two-dimensional diagram of the spherical water restraining region (the “SPH restraint”). The loop is represented as the heavy black curve, and the protein template is the region shown in gray. The dashed circle (radius = R_{temp}), defining the edge of the template, is the same as that in Figure 1 and is shown here for convenience. Three positions are marked with the symbol \otimes in the figure. These are, starting from the bottom, \mathbf{x}_{cm} , \mathbf{x}_{cmb} , and \mathbf{x}_{sph} . \mathbf{x}_{cm} is the center of mass of all of the protein atoms considered explicitly in the model (the loop and template atoms), while \mathbf{x}_{cmb} (also shown in Figure 1) is the center of mass of the loop backbone. \mathbf{x}_{cm} and \mathbf{x}_{cmb} are connected by a dotted line, which defines the vector direction (pointing from \mathbf{x}_{cm} to \mathbf{x}_{cmb}) that is used to determine the position of \mathbf{x}_{sph} . (That is, \mathbf{x}_{sph} is shifted away from the template, see eq 1.) Water molecules are contained within a spherical region defined by the distance R_{cap} measured from \mathbf{x}_{sph} . This containment region is represented by the large outer circle. Note that, generally, $R_{\text{cap}} > R_{\text{temp}}$, and therefore, the edge of this circle (sphere in 3D) is shifted to meet the (bottom) edge of the template so as to keep the majority of the water molecules on the “loop side” of the model system.

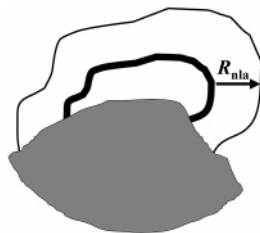


Figure 3. A two-dimensional diagram of the nearest-loop-atom-based restraining region (the “NLA restraint”). The loop is represented as the heavy black curve, and the protein template is the region shown in gray. Water molecules experience a restoring force only when the distance to the *nearest* loop atom becomes greater than a value, R_{nla} . For this reason, the boundary of the surrounding containment region mimics the shape of the loop itself, as shown in the figure. Note that the loop side-chain atoms are also considered (as nearest atoms) in the implementation.

coordinates”, used in the RMSD calculations. Several stages of MD equilibration are performed in addition to the production runs. All MD simulations are carried out under periodic boundary conditions, with a bath temperature of 300 K. Most of these simulations are also run under constant pressure (p) conditions, where p in all of these cases is set to 1 atm.

In the first stage of equilibration, the system is simulated for 10 ps at constant volume with the protein atoms under positional restraints ($k = 10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$). The next stage consists of 40 ps of constant pressure simulation, again with

the protein atoms under positional restraints ($k = 10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$). This is followed by another 40 ps of constant pressure simulation under weaker positional restraints ($k = 2 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$). In the final equilibration stage, the positional restraints are removed and the system is again simulated at constant pressure for 40 ps. The production MD simulations (constant T and p) are run to a length of 2 ns. Five independent 2 ns production runs are carried out for each protein.

Other important simulation details are as follows. The leapfrog form of the Verlet algorithm is used to integrate the equations of motion with a time step of 2 fs. The SHAKE algorithm⁶⁴ is used to fix all bonds involving hydrogen atoms in the protein and to maintain the rigid geometry of the TIP3P water molecules. Berendsen coupling methods⁶³ are applied to maintain constant temperature and pressure, both with time constants of 1 ps. Coulombic interactions are modeled using particle mesh Ewald electrostatics⁶⁵ with a real space cutoff of 8 Å. (LJ interactions are also cutoff at 8 Å, with a long-range correction added to the energy and pressure.)

II.6. Calculation of RMSD Values. An important gauge of behavior in this investigation is the RMSD of the loop atoms, measured with respect to the X-ray reference coordinates (\mathbf{X}_{ref}). We report two RMSD measures: the RMSD of the loop backbone atoms [which is denoted as RMSD-(BB)] and the RMSD of the loop side-chain atoms [denoted as RMSD(SC)]. (Corresponding RMSD fluctuations, $\sigma(\text{BB})$ and $\sigma(\text{SC})$, will also be reported.) In all cases, only the heavy atoms are considered.

The methods used to calculate RMSD in the partial-protein and full-protein models are somewhat different. Because of the fixed template, RMSD values for the partial-protein model can be straightforwardly calculated in a fixed coordinate system. That is, given a coordinate set \mathbf{X}_i for any structure i (sampled in the production runs), the (squared) distances of the loop atoms in \mathbf{X}_i are simply measured from their positions in \mathbf{X}_{ref} . (There is no superposing of structures.) In the case of the full-protein model, the value taken is the minimized RMSD resulting from superposing \mathbf{X}_i and \mathbf{X}_{ref} . Here, these superpositions are based on minimizing the RMSD of just the loop atoms and not the entire protein coordinate set. More specifically, for RMSD(BB), only the backbone atoms are superposed, and for RMSD(SC), only the side-chain atoms are superposed.

The quantities defined in this section can be applied for either the backbone or the side-chain atoms (or all of the loop atoms, etc.). Therefore, we will temporarily drop the “(BB)” and “(SC)” for compactness in the equations. In the discussion of the results, however, we will typically refer to the specific quantities (defined in eqs 2–6) by including this more detailed (BB) or (SC) notation.

We calculate RMSD_i values as averages for the entire run (trajectory) i . Thus, for a single configuration \mathbf{X}_t , we have the “instantaneous” value, RMSD'_t , which is superscripted with t for clarity, and the average value is therefore

$$\text{RMSD}_i = \frac{1}{n} \sum_{t=1}^n \text{RMSD}'_t \quad (2)$$

where n is the total number of configurations (snapshots) collected (evenly in time) over the course of the MD trajectory i . For each system, there are five independent runs, and in the tables, we provide values for the five-run average that, for simplicity, are denoted just RMSD. The standard deviation of the RMSD_i values (eq 2) for the five runs is also reported in the tables (in parentheses). These standard deviations can be helpful because, at times, there can be considerable variability in the results for individual runs (i). This, for example, can be due to changes in conformational microstates, which occur on time scales that are too long to be exhibited in all runs.

Even if the average RMSD_i is small for a given model, desirable behavior should also be manifested in the correct fluctuation properties. Therefore, the fluctuations in the instantaneous RMSD' values (about the average RMSD_i) are also a useful property and are calculated (for run i) as follows:

$$\sigma_i = \left[\frac{1}{n} \sum_{i=1}^n (\text{RMSD}' - \text{RMSD}_i)^2 \right]^{1/2} \quad (3)$$

σ_i is (among other things) a reflection of the local motion of the system. If the system remains in a single conformational microstate, σ_i will converge to a well-defined value (as the loop atoms simply execute local motion confined within that microstate). If, on the other hand, the system moves to another microstate (e.g., a major torsional change in the loop backbone), there will be a significant jump in the RMSD' values as they will now tend to oscillate about a new average value. The fluctuations within the new microstate may not be that different from the previous one. However, σ_i calculated according to eq 3 will be shifted significantly, because of the large overall spread of RMSD' values when both microstates are included.

Given the above points, it is helpful to also calculate fluctuations by window averaging. This is done by first defining

$$\sigma_{m(j)i} = \left[\frac{1}{m} \sum_{i=j}^{j+m-1} (\text{RMSD}' - \text{RMSD}_{m(j)i})^2 \right]^{1/2} \quad (4)$$

where

$$\text{RMSD}_{m(j)i} = \frac{1}{m} \sum_{i=j}^{j+m-1} \text{RMSD}' \quad (5)$$

$\sigma_{m(j)i}$ is a value for the short-time-averaged RMSD fluctuations, where $m < n$. The average is taken over the j th window, consisting of m consecutive snapshots (configurations) recorded during the MD run. (There are $n - m + 1$ such windows.) All possible (contiguous) m -step windows are then averaged to give

$$\sigma_i^w = \frac{1}{(n - m + 1)} \sum_{j=1}^{n-m+1} \sigma_{m(j)i} \quad (6)$$

thus defining the “window-averaged RMSD fluctuations”, σ_i^w . In this work, σ_i^w is calculated using time windows of 200 ps (i.e., the m steps cover a period of 200 ps).

We will report both fluctuation definitions. σ_i^w has the property of “smoothing over” the fluctuation effects of moving into (or perhaps flipping between) different microstates and, thus, more faithfully characterizes local motions. The gross changes resulting from different microstates (if they occur) will show up more strongly in σ_i^w and will also be reflected in the average RMSD. As for the reported RMSD values, the fluctuations will be averaged over five runs, and the standard deviations over the five runs appear in the tables in parentheses.

III. Results and Discussion

III.1. Solvation Properties of the Partial-Protein Model.

Before discussing RMSD results for the individual loops, it is important to discuss some of the general aspects of solvation that we have observed within the partial-protein model. The number of water molecules can be very small, and it is thus helpful to note some of the differences in behavior compared to when larger numbers of water molecules are used. In the partial-protein model, water molecules experience two obviously different environmental influences compared to those in a bulk water environment. Most importantly is the contact/interaction with protein surface atoms. Furthermore, there is also the inevitable exposure to a vacuum due to the modest number of water molecules employed, coupled with the chosen boundary conditions (i.e., nonperiodic boundaries).

III.1.1. Analysis of Surface Coverage. One of the important general characteristics of the present partial-protein solvation model is that there is typically plenty of “extra room” for the water molecules within the allotted restraining region. We mentioned in section II.3 that parameters for both the SPH and NLA restraint methods have been chosen such that the restraining volume is somewhat large for the given N_w . This is further evidenced in our simulations by the fact that, at any instant, very few water molecules are experiencing a boundary restoring force (and by small values for the average restraining potential, in general). This is especially true for small N_w values, where the water molecules will typically migrate to charged and polar groups on the loop and nearby template, often leaving nonpolar regions bare (as described in ref 59). It is reasonable to assume that the screening/bridging of interactions with charged and polar groups is one of the most important solvating effects provided by the water molecules. It is thus expected that it is better to allow the water molecules to spread out (within reason) such that they can access the more strongly interacting protein atoms, rather than attempting to confine them to a much smaller volume in an effort, for example, to keep them at a density that is closer to the bulk density for water.

A good way to gain a sense for the behavior of the water molecules within these models is to identify those molecules that are considered to belong to the surface region of the protein, separately from those that reside farther away from the protein. Specifically, we choose to define a “protein surface water” as one whose center of mass is a distance of 3.3 Å or less from any protein atom. The total number of these surface waters found (at any given instant) is denoted as N_{surf} . One particularly insightful way to analyze the nature

of these models is, thus, to monitor the number of surface water molecules (N_{surf}), or the fraction of protein surface water (N_{surf}/N_w), as the total number of water molecules (N_w) is varied. In Table 1, we provide some values for $\langle N_{\text{surf}} \rangle$ and $\langle N_{\text{surf}}/N_w \rangle$ accumulated from simulations of the loop [64–71] of RNase A, modeled under the SPH solvent restraint. (Details of the behavior of the loop itself are deferred until section III.3.) It is seen that, when N_w is very small, nearly all of the molecules are directly on the surface of the protein. For example, $\langle N_{\text{surf}}/N_w \rangle$ is nearly 90% when $N_w = 50$. It is not until $N_w = 200$ that this ratio reaches 50%, thus corresponding (on average) to a situation where the protein surface waters are surrounded by a second outer layer of water. At $N_w = 300$, roughly two-thirds of the water molecules are outside the inner hydrating layer. It is also important to note the trends in $\langle N_{\text{surf}} \rangle$ itself, where it is seen that the protein surface appears to saturate with about 100 water molecules at $N_w = 200$ (i.e., $\langle N_{\text{surf}} \rangle$ remains at about 100 for $N_w = 300$). This also implies, conversely, that even for $N_w = 120$ (with $\langle N_{\text{surf}} \rangle = 79$), significant bare regions remain on the protein surface.

III.1.2. Diffusion Properties and Residence Times. It is interesting to address some of the dynamical aspects of the solvation and to examine, in particular, how these properties are affected as N_w is increased within the partial-protein model. To do this, we have calculated diffusion constants for the water molecules using the Einstein relation $\langle r^2 \rangle = 6Dt$, where D is the diffusion constant and $\langle r^2 \rangle$ is the average squared distance that a particle will move in time t . Because the model is a finite system, the ratio $\langle r^2 \rangle/t$ will go to zero at long times. Therefore, we estimate D from $\langle r^2 \rangle$ values after a period of 10 ps (i.e., we take $D = \langle r^2 \rangle/6t$ at $t = 10$ ps). This is a compromise between the effect of ballistic (nondiffusive) motion at very short times (less than 1 ps) and the onset of nonlinearity in $\langle r^2 \rangle$ versus t , which is observed as $\langle r^2 \rangle$ begins to approach the size of the system (at $t > \sim 30$ ps).

In Table 1, we show diffusion constants for the partial-protein model as the number of water molecules is increased. D_{all} is the value of D that is calculated using all N_w molecules. D_{surf} , on the other hand, is the diffusion constant calculated for just the protein surface waters. (Specifically, a molecule is included in the calculation of D_{surf} if it is a distance of 3.3 Å or less from any protein atom at the beginning of the 10 ps interval.) It is seen that the value of D_{all} systematically decreases as N_w decreases. Part of the reason for this is the high fraction of surface waters exhibited in the models with small N_w values (e.g., $N_w = 50$ or 70). An important observation from other simulation studies of protein hydration^{66–72} is that water molecules on the surface of the protein diffuse significantly more slowly than water molecules in the bulk. These studies have shown that D for protein surface water is lower (than the bulk value) by about a factor of 2 or more (see, for example, refs 66 and 67). In our calculations, the value of D_{all} at $N_w = 300$ (4.96×10^{-5} cm²/s) is approaching values that are typical of bulk TIP3P water. (Commonly calculated values at $T = 300$ K and $p = 1$ atm are about 5×10^{-5} cm²/s but can vary depending on modeling details.⁷³) In contrast, the value for D_{surf} ($2.61 \times$

10^{-5} cm²/s) is much lower (by about a factor of 2), and thus, it is in good agreement with the findings of the previous studies. The value of D_{surf} for $N_w = 200$ (2.53×10^{-5} cm²/s) is nearly the same as the value at 300, suggesting that the protein surface waters behave quite similarly in both models. This is despite the differences in D_{all} , which are thus mostly attributable to the difference in the relative amount of surface molecules ($\langle N_{\text{surf}}/N_w \rangle$).

Though there is good agreement for the cases of $N_w = 200$ and 300, it is important to note, however, that D_{surf} becomes significantly lower as N_w is decreased further. Though there is still similarity in D_{surf} for the case of $N_w = 120$, D_{surf} at 50 and 70 molecules, however, is roughly half the value of that at 200 or 300. The important distinction in these models ($N_w = 50$ and 70) is that the water molecules generally lack neighboring water from a second layer (the $\langle N_{\text{surf}}/N_w \rangle$ values are 0.888 and 0.825). In view of the general observation of a lowered D for water in the first solvation shell of a protein, it is thus noted that the lack of a second solvation shell serves to lower D further. It is interesting to note, on the other hand, that despite the significantly lower D values for the case of $N_w = 50$ and 70, the stability of the loops (discussed in the next sections) can often be surprisingly good at these very low hydration levels.

Inherent in the diffusion properties is information on the time scales of solvation. Specifically, these values can provide insight on residence times (τ) for water molecules near the surface of the protein. In earlier experimental (NMR) work,⁷⁴ an upper bound for residence times of protein surface water was placed at around 500 ps. In much better agreement with the simulation literature, more recent experimental work⁷⁵ has placed typical residence times roughly around 25 ps. Residence times have been investigated in simulation studies on a variety of solvated proteins such as BPTI,⁶⁸ myoglobin,⁶⁹ lysozyme,^{70,71} and azurin.⁷² Here, we will only briefly make some comparisons. In Brunne et al.,⁶⁸ detailed studies were carried out to determine the residence time of hydrating water molecules in specific regions on the protein surface (i.e., near specific types of atoms/groups). They found that the residence time of a surface water molecule is (on average) about 30 ps. (Specific results would vary depending on the nearby protein atoms—backbone atoms, side-chain atoms, charged, polar, nonpolar, etc.)

As a very rough comparison, we can estimate residence times (the time for a water molecule to leave the neighborhood of a solute protein atom) simply from the diffusion results. We take the residence time as the time for a water molecule to diffuse about one and a half molecular diameters, specifically, 4.5 \AA (thus, $\tau = (4.5 \text{ \AA})^2/6D$). These residence times are given in Table 1, where τ_{surf} is the residence time calculated for a protein surface water and τ_{all} is calculated for all N_w water molecules. For the case of $N_w = 200$ or 300, the residence time for surface molecules is about 13 ps, which is in reasonable agreement with the 30 ps given by Brunne et al.,⁶⁸ especially when one accounts for the different modeling conditions. The modeling temperature in Brunne et al. was lower (277 K) to mimic NMR experimental conditions. Furthermore, these authors employed the SPC/E water model,⁷⁶ which is known to give a lower (more

accurate) value for the bulk diffusion coefficient compared to TIP3P. Indeed, calculations in ref 71 using both the SPC/E and TIP3P models showed that the TIP3P residence times were a factor of 2 shorter (roughly 14 ps [TIP3P] as opposed to 27 ps [SPC/E]). (It should also be noted that the diffusion constants and residence times will also vary depending on the distances chosen to define a “protein surface water”.)

In correspondence with their lowered D_{surf} values, the residence times for small N_w values (50 and 70) are longer. Though these values for τ_{surf} are closer to the values in some of the other studies, they should be interpreted as being “long” for the TIP3P water model (and therefore, they show a specific behavioral property of the partial-protein model at low hydration levels). It is thus expected that they would become much longer if a different water model was used, such as SPC/E or TIP4P,⁴⁷ both of which give more accurate diffusion properties.

One of the points discussed by Brunne et al.⁶⁸ was their, perhaps unintuitive, observation that the residence times near charged atoms were lower than those for polar, and even nonpolar, atoms. Though we did not carry out the detailed analysis as in that investigation, we did measure the diffusion time away from one specific charged group, the NH_3 group on the Lys side chain of the RNase loop, and for $N_w = 200$ and 300, we also find a decreased residence time (about 10 ps). Interestingly, this effect reverses itself for the case of $N_w = 50$. The value in this case is about 37 ps, which is longer than the average residence time for surface waters at this hydration level (N_w). The authors remarked that the shorter residence times for water molecules near charged groups must be related to the effects caused within the surrounding water. Obviously, the lack of outer layers in the case of small N_w values might suggest the possibility for different behavior. Here, at low hydration levels, arguments can more plainly be interpreted in terms of energetic benefits because more subtle entropic considerations (associated with surrounding water molecules) are less prevalent.

III.2. Some Properties of the Loops Studied. We now focus on the behavior of the protein loops. The four primary surface loops studied (ranging in size from 8 to 10 amino acid residues), and the related proteins, are presented in Table 2. The 3D structures of these proteins, taken from the PDB, have been determined with 1.5–1.8 Å resolution. The B factors of the loops of RNase A and the two loops 1 and 2 of proteinase are relatively small, where the maximal values obtained for the side chains are 35, 19, and 25, respectively; for ser-proteinase, the B factors of the backbone atoms of five residues range within 20–28, that is, still relatively low, while for some of the side-chain atoms, no significant electron density has been observed. It should be pointed out that side chains with a well-defined structure in the crystal environment (i.e., small B factors) might still be flexible in solution, the environment that is expected to better be described by our models.

While our tests require loops with well-defined structures, it is also imperative to verify that these loops are not stretched, as a stretched loop is insensitive to the model applied. Therefore, we present in Table 2 the ratio $R = \text{length of the stretched loop}/\text{distance between its ends}$, which is

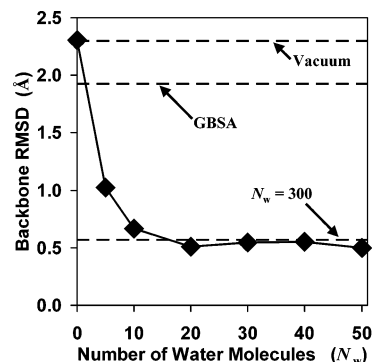


Figure 4. Plot of the average backbone RMSD [RMSD(BB)] as a function of the number of water molecules, N_w , for the loop [64–71] of RNase A. The dashed lines indicate the RMSD(BB) values obtained for 300 water molecules, the GBSA implicit solvation model, and simulation in vacuum.

calculated between the C^α atoms of the first and last residues of the loop. The length (in Å) of the extended structure is obtained using the expressions $6.046(n_{\text{res}}/2 - 1) + 3.46$ and $6.046(n_{\text{res}} - 1)/2$ for an even and odd number of residues, n_{res} , respectively; the factors 6.046 and 3.46 Å are taken from Flory’s book⁷⁷ (Chapter VII, p 251). To a large extent, R reflects the conformational freedom of the loop’s backbone and, to a lesser extent, also that of the side chains; the larger R is, the greater the flexibility; indeed, the R values of the four loops are relatively large, ranging from 3.2 to 4.9. Notice, however, that the conformational freedom depends also on the structure of the surrounding protein template and the template–loop interactions. Typically, surface loops are hydrophilic and often charged; therefore, our chosen loops are predominantly polar, where those of RNase A and ser-proteinase each contain one charged residue (bold-faced in Table 2) and loops 1 and 2 of proteinase have three and two charged residues, respectively.

III.3. The Loop of RNase A. We discuss, first, the partial-protein model results for the loop [64–71] of RNase A. Figure 4 is a “convergence plot” of (the backbone average) RMSD(BB) as the number of water molecules, N_w , is increased from 0 (vacuum) to 50. (All points are for the case of the SPH solvent restraint method and $R_{\text{temp}} = 15$ Å.) Also marked in the figure is RMSD(BB) for $N_w = 300$ (the largest N_w studied for the partial-protein model), as well as the result for GBSA. Though RMSD(BB) = 2.30 Å for the vacuum simulations is large (which is not unexpected), the figure suggests that only a handful of water molecules is necessary to stabilize it. RMSD(BB) is quite low for as little as $N_w = 20$ (0.51 Å), and it is, furthermore, in excellent agreement (converged) with all larger values of N_w .

More extensive results, covering a wider range of modeling conditions, are presented in Table 3. Here, RMSD(BB) ranges between 0.51 and 0.67 Å for all N_w values between 20 and 300, further suggesting that the backbone behavior is reasonably reproducible and, thus, insensitive to increased levels of hydration. These (backbone) results appear, as well, to be relatively insensitive to the number of environmental protein atoms incorporated into the model (i.e., the template size R_{temp}) and the water containment method (SPH or NLA)

and its associated restraining distances (R_{cap} or R_{nla}) (within the ranges tested). We note briefly that (for $N_w \geq 20$) the ranges of the average backbone fluctuations (over five runs), $\sigma(\text{BB})$, are small, 0.14–0.24 Å (0.19 Å for $N_w = 300$); the range of the corresponding $\sigma^w(\text{BB})$ is small as well, 0.13–0.15 Å (0.14 Å for $N_w = 300$). The above discussion suggests that, as far as the backbone is concerned, already, $N_w = 50$ (or less) is adequate.

It should be pointed out, however, that the standard deviations, for some of the runs in Table 3, are relatively high. This is due to individual runs that sample (“escape to”) different conformational microstates. These transitions are manifested by a significant change in one or more of the (backbone) torsion angles (typically 90° or more). The large free-energy barriers associated with these transitions make the time scale (~ 1 ns or more) too long to straightforwardly sample/average over various possible conformations (microstates) within typical MD simulation runs. This is a common (and unavoidable) difficulty in testing and assessing potentially flexible regions in protein models. (For example, these “escapes” were also exhibited in the fully solvated, full-protein model.)

We take, as an example of this behavior, the set of trajectories for $N_w = 300$. Here, four runs fell within the range $0.53 \leq \text{RMSD}_i(\text{BB}) \leq 0.55$ Å. However, for one run, the (cumulative average) $\text{RMSD}_i(\text{BB})$ (eq 2) grows systematically as a function of time for $t > 3$ ns, becoming 0.66 Å for 5 ns, and would have increased further if the simulation had been continued (tending toward about 1 Å), meaning that the loop had transferred to a different microstate. Similar deviant runs were observed for $N_w = 120$, in sets 5, 7, and 8, and in sets 12 ($N_w = 70$) and 16 ($N_w = 30$) (numbering rows [sets] from the top of Table 3). It should be noted, however, that, for $N_w \geq 50$, 73 runs (out of 80 total, i.e., 91%) lead to very low $\text{RMSD}_i(\text{BB})$ values within 0.48–0.60 Å, with the seven most deviant runs still only averaging to about 0.85 Å. The number of “escaped” runs does not seem to depend on N_w , as one escaped run is found for $N_w = 50$, one for $N_w = 70$, four for $N_w = 120$, and one for $N_w = 300$.

Moving to the side-chain properties, we note, in general, that the side-chain $\text{RMSD}(\text{SC})$ values are relatively small. The values range from 1 to 1.4 Å (with 1.31 Å for $N_w = 300$), and thus, the difference between most runs is also relatively small (about 0.2 Å). These $\text{RMSD}(\text{SC})$ values are lower, for example, than the values obtained for other loops but larger, of course, compared to the backbone values. The side chains seem to show a dependence on the template size and slightly on N_w . For $R_{\text{temp}} = 15$ Å, $\text{RMSD}(\text{SC})$ decreases slightly from 1.31 Å (with a very small standard deviation) for $N_w = 300$ to 1.13 Å for $N_w = 200$, to 1.02 and 1.09 Å for $N_w = 120$, and to 1.02 and 1.00 Å for $N_w = 70$. On the other hand, for $R_{\text{temp}} = 14$ and 16 Å, the $\text{RMSD}(\text{SC})$ values are larger. In general, the corresponding average side-chain fluctuations, $\sigma(\text{SC})$ and $\sigma^w(\text{SC})$, tend to decrease as N_w decreases. (See also the discussion for proteinase, loop 1.) Also, on average, $\sigma(\text{SC})$ and $\sigma^w(\text{SC})$ for $N_w = 50$ and 70 appear to give somewhat better agreement with larger N_w values when the NLA solvent restraint is used.

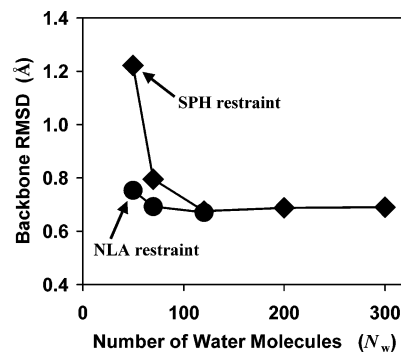


Figure 5. Plot of the average backbone RMSD [$\text{RMSD}(\text{BB})$] as a function of the number of water molecules, N_w , for the loop [143–151] of ser-proteinase. The diamonds mark values obtained using the spherical water restraining method (marked “SPH restraint” in the figure). The circles are for values obtained using the nearest-loop-atom-based restraint (marked “NLA restraint”).

The GBSA results, $\text{RMSD}(\text{BB}) = 1.93$ Å and $\text{RMSD}(\text{SC}) = 2.71$ Å, are significantly larger than those based on explicit water and are not much better than the vacuum results (see also Figure 4). It should be pointed out that, in two of the GBSA runs, $\text{RMSD}_i(\text{BB})$ and $\text{RMSD}_i(\text{SC})$ are still increasing significantly after 5 ns.

III.4. The Loop of Ser-proteinase. The results for the loop [143–151] of ser-proteinase are provided in Table 4. The $\text{RMSD}(\text{BB})$ values for $N_w = 300, 200,$ and 120 are very similar, ranging from 0.64 to 0.69 Å with very small standard deviations (≤ 0.03 Å) for each set of five runs. The corresponding $\text{RMSD}(\text{SC})$ values are only slightly more dispersed and can still be considered as very close, ranging from 1.39 to 1.54 Å with a maximal standard deviation (over the five runs) of 0.11 Å. The average backbone fluctuations, $\sigma(\text{BB})$, are again very close, ranging from 0.12 to 0.14 Å, and the same applies to the average side-chain fluctuations, $\sigma(\text{SC})$, that vary between 0.26 and 0.29 Å; the corresponding ranges for $\sigma^w(\text{BB})$ and $\sigma^w(\text{SC})$ are again narrow, 0.11–0.13 and 0.19–0.20 Å. These results, which were calculated for different templates ($R_{\text{temp}} = 12$ –15 Å), and with both solvent restraint methods (SPH and NLA), suggest that, already, $N_w = 120$ is sufficient to produce the results of full solvation.

Achieving adequate solvation for this loop becomes more problematic for $N_w < 120$, and it can, furthermore, depend on the modeling conditions. This is clearly shown in Figure 5, a convergence plot of $\text{RMSD}(\text{BB})$ as a function of N_w (all for the case of $R_{\text{temp}} = 13$ Å). While the points show convergence for $N_w = 120$ –300 (as discussed above), the results for $N_w = 50$ and 70 using the SPH solvent cap clearly begin to diverge. Interestingly, the NLA restraint appears to maintain adequate solvation to lower N_w values. The contrast of the two solvent restraint methods at $N_w = 50$ is fairly significant. For the SPH restraint ($R_{\text{cap}} = 17$ Å), the results $\text{RMSD}(\text{BB}) = 1.28$ Å and $\text{RMSD}(\text{SC}) = 1.88$ Å (Table 4) are significantly larger than the 0.69 and 1.51 Å obtained, respectively, for $N_w = 300$. While, on the other hand, the NLA restraint at $N_w = 50$ ($R_{\text{nla}} = 7$ Å) is much closer, with $\text{RMSD}(\text{BB}) = 0.75$ and $\text{RMSD}(\text{SC}) = 1.55$ Å; only $\sigma(\text{BB}) = 0.21$ and $\sigma(\text{SC}) = 0.32$ Å (for NLA) are larger than the

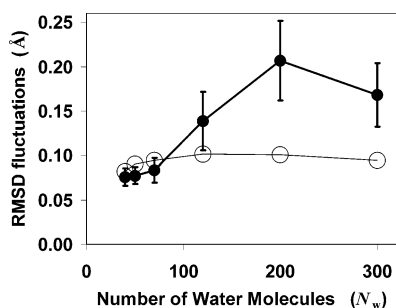


Figure 6. Plot of the window-averaged RMSD fluctuations for backbone and side-chain atoms [$\sigma^w(\text{BB})$ and $\sigma^w(\text{SC})$, respectively] as a function of the number of water molecules, N_w , for loop 1 of proteinase [128–137]. The values obtained for the side-chain RMSD fluctuations appear as solid circles and include error bars (the standard deviation of five trials). The backbone RMSD fluctuations appear as large open circles with a lighter trend line.

0.14 and 0.26 Å respectively obtained for $N_w = 300$. (The window-averaged fluctuations are fairly close, however, with $\sigma^w(\text{BB}) = 0.15$ and $\sigma^w(\text{SC}) = 0.19$ for (NLA) $N_w = 50$, and 0.13 and 0.20 Å, respectively, for $N_w = 300$.)

It should be pointed out that the GBSA results for RMSD(BB) and RMSD(SC) are actually equal to the corresponding $N_w = 300$ values, while the results $\sigma(\text{BB}) = 0.18$, $\sigma(\text{SC}) = 0.31$, $\sigma^w(\text{BB}) = 0.16$, and $\sigma^w(\text{SC}) = 0.24$ Å are slightly larger than their counterparts for $N_w = 300$.

III.5. Loop 1 of Proteinase. The results for loop 1 of proteinase [128–137] are summarized in Table 5. The table reveals that the backbone of this loop is very stable, where $\text{RMSD}(\text{BB}) \sim 0.70$ Å (with the standard deviation smaller than 0.05) already for $N_w \geq 20$. For $N_w = 10, 5$, and 0, $\text{RMSD}(\text{BB})$ increases to 0.89, 0.84, and 1.08 Å with relatively large standard deviations (of the five runs), 0.11, 0.17, and 0.13 Å, with maximal values of 1.01, 1.13, and 1.16 Å, respectively, where the first two maximal values have not been converged after 5 ns and are growing. The results for $\sigma(\text{BB})$ are small and similar for most N_w values: 0.12 Å for $N_w \geq 120$, 0.10 Å (on average) for $N_w = 70$ and 50, and 0.09–0.11 Å for $10 \leq N_w \leq 40$. Similar behavior is observed for $\sigma^w(\text{BB})$.

The $\text{RMSD}(\text{SC})$ values for this loop are significantly larger than those for the loop of ser-proteinase; that is, the side chains have moved significantly from their X-ray structure. For $N_w \geq 120$, $\text{RMSD}(\text{SC})$ ranges from 2.16 to 2.31 Å; for $N_w = 70$, the range is similar except in one case where $\text{RMSD}(\text{SC}) = 2.41$ Å is slightly larger. As N_w decreases further, $\text{RMSD}(\text{SC})$ increases moderately, becoming 2.65 Å for $N_w = 0$.

Though $\text{RMSD}(\text{SC})$ appears to be relatively converged at small N_w values, the side-chain fluctuations show a significant increase as N_w is increased. These trends are shown in Figure 6, which is a plot of the window-averaged side-chain fluctuations, $\sigma^w(\text{SC})$, as a function of N_w (all for the case of the SPH solvent restraint method and $R_{\text{temp}} = 13$ Å). $\sigma^w(\text{SC})$ is consistently small for all $N_w \leq 70$ compared to the higher solvation levels at $N_w = 200$ or 300. [Note, in contrast, that $\sigma^w(\text{BB})$, which is also given in the figure, appears to be converged for all N_w values shown.] To more clearly see

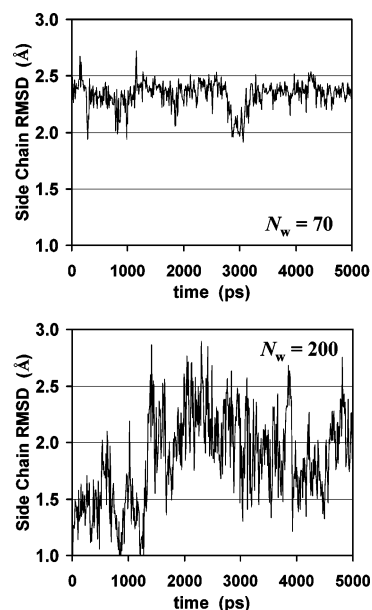


Figure 7. Instantaneous RMSD values of the side-chain atoms [$\text{RMSD}'(\text{SC})$] as a function of time for loop 1 of proteinase [128–137]. The upper plot is for a typical 5 ns trajectory for the case of $N_w = 70$. The lower plot is a typical trajectory with $N_w = 200$.

how these results are manifested in the trajectories, we have plotted, as an example, the instantaneous (snapshot) values of $\text{RMSD}'(\text{SC})$ over the course of a typical 5 ns run for $N_w = 70$ and compare that with a typical run for $N_w = 200$. These plots are shown in Figure 7. {Note that the y-axis scales [for $\text{RMSD}'(\text{SC})$] are the same in both plots.} Though the $\text{RMSD}'(\text{SC})$ values for these two runs are similar, on average, the oscillations in these values (even over short times) show very different amplitudes. That is, the $N_w = 200$ run appears to visit a more diverse array of states, and even within those states, the atomic fluctuations are more broad, meaning higher entropy than in the $N_w = 70$ case.

Some additional trends in the side-chain fluctuations are as follows. The table shows that the $\sigma(\text{SC})$ results for $N_w \geq 70$ decrease as the template radius R_{temp} is increased and N_w is decreased. Also, the NLA restraint leads to higher (i.e., better agreement with large N_w) $\sigma(\text{SC})$ and $\sigma^w(\text{SC})$ values than the spherical cap (SPH). Thus summarizing, for $R_{\text{temp}} = 13$ Å, we obtained almost the same $\sigma(\text{SC})$ values, 0.37, 0.40, and 0.30 Å, for $N_w = 300, 200$, and 120, respectively, and a slightly lower value, 0.20 Å, for $N_w = 70$ with the NLA restraint. The corresponding $\sigma^w(\text{SC})$ values, 0.17, 0.21, 0.14, and 0.11 Å, are also close. The results for $\sigma(\text{SC})$ and $\sigma^w(\text{SC})$ for $N_w \leq 50$ are significantly smaller than the corresponding values for $N_w = 300$. Therefore, $N_w = 120$ (perhaps less with the NLA restraint) is necessary to solvate this loop.

It is noted that the GBSA values, $\text{RMSD}_i(\text{BB}) = 0.90$ Å and $\text{RMSD}_j(\text{SC}) = 3.07$ Å (from two different runs), are not converged after 5 ns, but they are in an increasing trend. Thus, the GBSA result, $\text{RMSD}(\text{BB}) = 0.79$ Å, is not converged, and the corresponding GBSA result, $\text{RMSD}(\text{SC}) = 2.88$ Å, that is already significantly larger (by 0.7 Å) than the 2.19 Å obtained for $N_w = 300$ is not converged either.

III.6. Loop 2 of Proteinase. The results for loop 2 of proteinase [188–196] are summarized in Table 6. The RMSD(BB) results in the table are similar for all of the N_w values. However, it should be pointed out that, for $N_w = 300$, one MD run escaped from the X-ray microstate, leading to $\text{RMSD}_i(\text{BB}) = 1.42 \text{ \AA}$, where from 3 to 5 ns $\text{RMSD}_i(\text{BB})$ is still increasing. Similar behavior is observed for single runs of the sets of $N_w = 70$ ($R_{\text{temp}} = 11 \text{ \AA}$ and $R_{\text{cap}} = 17 \text{ \AA}$), where $\text{RMSD}_i(\text{BB}) = 0.62 \text{ \AA}$; $N_w = 70$ ($R_{\text{temp}} = 13 \text{ \AA}$ and $R_{\text{cap}} = 17 \text{ \AA}$), where $\text{RMSD}_i(\text{BB}) = 1.40 \text{ \AA}$; and $N_w = 50$ ($R_{\text{temp}} = 13 \text{ \AA}$ and $R_{\text{nla}} = 7 \text{ \AA}$), where $\text{RMSD}_i(\text{BB}) = 0.73 \text{ \AA}$. This suggests that the X-ray microstate of this loop may not be overwhelmingly stable (i.e., competing conformational microstates), as this instability is independent of the number of waters, N_w , occurring for $N_w = 50$ and 70 as well as for $N_w = 300$. Moreover, this behavior was exhibited in the full-protein model as well (see below). When the contribution of the “escaped” runs is omitted, all of the RMSD(BB) results are very close to 0.5 \AA , and the fluctuations $\sigma(\text{BB})$ and $\sigma^w(\text{BB})$ are close to 0.12 and 0.15 \AA , respectively.

The instability of the X-ray microstate is demonstrated even more strongly by the behavior of the side chains. While, for $N_w \geq 120$, the RMSD(SC) values are relatively close, ranging from 1.42 to 1.69 \AA , the corresponding standard deviations are large, suggesting that the individual values, $\text{RMSD}_i(\text{SC})$, are very different. Indeed, for the two sets of $N_w = 120$ ($R_{\text{temp}} = 15$), the minimum and maximum $\text{RMSD}_i(\text{SC})$ values are 1.39 and 2.22 and 1.27 and 2.23 \AA . Moreover, in some cases, the $\text{RMSD}_i(\text{SC})$ values have not been converged after 5 ns. For example, for $N_w = 300$, one MC run has led to a still unconverged value of $\text{RMSD}_i(\text{SC}) = 2.37 \text{ \AA}$, where for both $N_w = 200$ and 120 ($R_{\text{temp}} = 15 \text{ \AA}$ and $R_{\text{cap}} = 18 \text{ \AA}$) two unconverged $\text{RMSD}_i(\text{SC})$ values occurred. A similar picture is observed for $N_w = 70$ and 50.

Even though this loop is not stable, it is evident that similar results are obtained for $N_w \geq 120$ and, for R_{nla} , also for $N_w = 70$. This is also demonstrated by the results for $\sigma(\text{BB})$ and $\sigma^w(\text{BB})$ that are close for these runs, that is, within the ranges 0.18–0.12 and 0.12–0.09 \AA , respectively. The ranges of the side-chain fluctuations, $\sigma(\text{SC})$ and $\sigma^w(\text{SC})$, are also small, 0.43–0.25 and 0.23–0.17 \AA , respectively.

It is of interest to point out that the GBSA results are significantly different from those obtained with explicit water. Thus, not only is $\text{RMSD}(\text{BB}) = 1.16 \text{ \AA}$ considerably larger than the RMSD(BB) values obtained for explicit water but the standard deviation of the GBSA set is large because of elevated $\text{RMSD}_i(\text{BB})$ values within the range 0.67–1.71 \AA ; the same occurs also for the side chains, where $\text{RMSD}(\text{SC}) = 2.86 \text{ \AA}$ is significantly larger than the corresponding values obtained for the explicit water, where the $\text{RMSD}_i(\text{SC})$ values for GBSA range within 2.31–3.63 \AA .

III.7. Partial Study of Four More Loops. While the above study suggests that a relatively small number of waters is sufficient to solvate a loop, one would like to strengthen this conclusion by evidence from a larger number of loops. However, because of the extensive calculations required, we decided to carry out only partial studies of four extra loops, which indeed provide supportive evidence. We first treated

the seven-residue loop [244–250] (ITTIYQA) of peptidase (5cpa) with a flexibility ratio, $R = 2.7$. Defining $R_{\text{temp}} = 13 \text{ \AA}$ and using the spherical water restraint (SPH) with $N_w = 70$ waters, we obtained the relatively small $\text{RMSD}(\text{BB}) = 0.69 \text{ \AA}$, as the average of five MD runs. The second loop is of seven residues [57–63] (EAKEHC) of RNase H (2rn2), with a flexibility ratio $R = 1.6$, where again $R_{\text{temp}} = 13 \text{ \AA}$ and $N_w = 70$. Here, the SPH restraint led to $\text{RMSD}(\text{BB}) = 1.02 \text{ \AA}$, as two deviating MD runs contributed $\text{RMSD}_i(\text{BB})$ values of 1.57 and 1.34 \AA . However, the NLA restraint, which has been found to perform better for small N_w values, led to $\text{RMSD}(\text{BB}) = 0.72 \text{ \AA}$. Therefore, this loop is expected to stabilize with the SPH restraint at $N_w = 120$, similar to the case observed for ser-proteinase.

We also studied a seven-residue loop in porcine amylase (1pif) [304–310] (GHGAGGS) with a flexibility ratio $R = 3.2$ and the same loop in human amylase (1smd), where S is replaced by A and the flexibility ratio is $R = 2.3$. In the pig amylase, we used a template of $R_{\text{temp}} = 15 \text{ \AA}$ with an SPH restraint. For $N_w = 70$, only two runs were generated, which led to $\text{RMSD}(\text{BB}) = 0.47 \text{ \AA}$, whereas for $N_w = 200$, the five MD runs led to $\text{RMSD}(\text{BB}) = 0.45 \text{ \AA}$. For the human amylase, we obtained $\text{RMSD}(\text{BB}) = 0.73 \text{ \AA}$ using $R_{\text{temp}} = 15 \text{ \AA}$ and the NLA restraint with $N_w = 70$ waters.

III.8. Results for the Full-Protein Model. The RMSD results for the full-protein model appear in Table 7 together with the corresponding $N_w = 300$ results obtained for the partial-protein model. However, because the RMSD was calculated differently for the two models, and to make the comparison between them on the same footing, we have recalculated the RMSD of the partial-protein model in the same way as that for the full-protein model (marked as “yes” in the “superpose” column of the table). The table reveals that, for all loops, the RMSD values (and fluctuations) of the full-protein model are always larger than the corresponding results of the partial-protein model. This effect is to be expected, on one hand, because of the nonfixed coordinates (of the nonloop atoms), thus promoting greater flexibility. On the other hand, however, there should be a mild but consistent effect to reduce the RMSDs because of the use of (minimized) superposition. This latter effect appears to reduce the backbone RMSD values by roughly 0.15 \AA upon comparison of the superposed and nonsuperposed values for the partial-protein model in Table 7.

Not only are all the averages of the full-protein model larger than those of the partial model, but also the corresponding standard deviations (appearing in parentheses), which should be considered in the comparisons between the averages, are as well. Thus, for ser-proteinase, the values $\text{RMSD}(\text{BB}) = 0.57(13)$ and $0.44(1) \text{ \AA}$ are equal within the standard deviations and all runs, on average, span the same microstate, where the most deviant single run for the full-protein model, $\text{RMSD}_i(\text{BB}) = 0.80 \text{ \AA}$, leads to the corresponding large standard deviation; this run also contributes to the large fluctuation [$\sigma_i(\text{BB}) = 0.28 \text{ \AA}$] and its large standard deviation (0.07 \AA). A similar picture is seen for the side chains where $\text{RMSD}(\text{SC}) = 1.33(22)$ and $1.12(3) \text{ \AA}$ are equal within the standard deviation, where one run contributes most significantly, $\text{RMSD}_i(\text{SC}) = 1.70$, $\sigma_i(\text{SC})$

= 0.49 Å, and $\sigma_i^w(\text{SC}) = 0.20$ Å. Notice that the differences between the $\sigma^w(\text{BB})$ and $\sigma^w(\text{SC})$ values of the two models are small. In summary, for this loop, ignoring the effect of the run with largest results, both models lead to close results, $\text{RMSD}(\text{BB}) = 0.52$ and 0.44 Å, $\text{RMSD}(\text{SC}) = 1.23$ and 1.12 , $\sigma(\text{BB}) = 0.12$ and 0.10 , and $\sigma(\text{SC}) = 0.25$ and 0.22 Å.

Quite similar behavior is observed for RNase A, where only the results of $\text{RMSD}(\text{SC})$ of the two models differ significantly and are not covered by their standard deviations. Here again, the results for one trajectory i deviate significantly from the results of the other runs of the full-protein model, leading to $\text{RMSD}_i(\text{BB}) = 0.83$, $\text{RMSD}_i(\text{SC}) = 2.27$, $\sigma_i(\text{BB}) = 0.32$, and $\sigma_i(\text{SC}) = 0.82$ Å. Ignoring this run, the results for the two models are quite comparable, $\text{RMSD}(\text{BB}) = 0.55$ and 0.42 Å, $\text{RMSD}(\text{SC}) = 1.47$ and 1.03 , $\sigma(\text{BB}) = 0.15$ and 0.15 , $\sigma(\text{SC}) = 0.38$ and 0.37 , $\sigma^w(\text{BB}) = 0.11$ and 0.11 , and $\sigma^w(\text{SC}) = 0.22$ and 0.19 Å.

The results for loop 2 of proteinase for both models have relatively large standard deviations, reflecting differences among the results of the five runs. Thus, while the averages of the full-protein model are in most cases larger than those of the partial model, the differences are not large [e.g., 0.7 vs 0.5 Å for $\text{RMSD}(\text{BB})$ and 1.6 vs 1.3 Å for $\text{RMSD}(\text{SC})$], where the average values are always covered by the error bars.

For loop 1 of proteinase, the results of the two models show the most disagreement among the four loops, where the error bars in most cases do not cover the average values. However, even in this case, the results are not very different, 1.0 vs 0.6 Å for $\text{RMSD}(\text{BB})$ and 2.5 vs 2 Å for $\text{RMSD}(\text{SC})$.

IV. Conclusions

We have shown that, for the present loops described in the framework of the partial-protein model, the results, in general, become less dependent on the parameters of the model as the number of waters is increased. Relatively small numbers of water molecules (120 and sometimes less) lead to results for RMSD and its fluctuations that are very similar to those obtained for 300 waters. It is expected that (similarly) ~ 12 waters per residue will be found adequate for other loops; however, this number should be checked for each individual loop. (We have already noted in the Introduction that Steinbach and Brooks have studied the effect of increasing the number of water molecules on the protein structure; examples of similar convergence studies performed on ions, water, and small molecules appear in refs 78 and 79). We have also found that, for a small number, N_w , of waters, the NLA restraint leads to slightly better results than the SPH restraint. The good performance obtained here with a relatively small number of waters is in accord with the free-energy calculations of Beglov and Roux,⁶⁰ who (originally) applied the NLA restraint to the alanine dipeptide and tripeptide molecules and have found good agreement with calculations based on bulk solvation. As expected, the RMSD (and fluctuation) values for the full-protein model are somewhat larger than their counterparts for the partial model. Indeed, the differences are not large, and it is not

clear whether they stem from using more complete solvation (with particle mesh Ewald) or from modeling the entire protein with unfixed coordinates.

Still, the present partial-protein model can be made more realistic (1) by allowing residues neighboring the loop ends also to move, (2) by relaxing the fixed template atoms, by only restraining them harmonically to their X-ray positions, and (3) by increasing the template size; such changes would make the protein atom treatment in the present model more similar to the stochastic boundary MD approximation.⁵² However, while, in principle, the partial-protein model with implicit solvation (such as GBSA) is inferior to that with explicit solvation, the long-range electrostatic interactions of the latter model are still not treated correctly. A more rigorous treatment is provided by sophisticated hybrid models where the region of interest is described by explicit solvent and the effect of the remote region by the reaction field of continuum solvation.^{78–85} However, because of the complex and varying geometry of the *actual* outer surface of the protein–water system (e.g., this surface/boundary is not simply the boundary of the SPH or NLA restraining region), most of these techniques would be difficult to apply to the present partial-protein model, especially at small N_w values (see discussions in refs 84 and 85).

We intend to use the partial-protein model to study mobile loops that take part in binding processes. As mentioned in the Introduction, in the free protein, such a loop typically resides in an open (o) flexible microstate or it undergoes *intermediate flexibility*, that is, populates several microstates in thermodynamic equilibrium. Upon ligand binding, the loop moves to a structurally different (and less flexible) bound (b) microstate, sometimes creating a “lid” above the active site, thus protecting it from water. Several questions are of interest, for example: (i) Is the process of a selected-fit type? That is, is the microstate of the bound loop already included within those visited by the free protein (or otherwise the process is of an induced-fit type)? (ii) What is the loss in loop entropy in going from the open to the bound microstates, and what is the corresponding free-energy difference? (The backbone entropy can, in some cases, be compared with results obtained from NMR.) To study these problems, one would have to carry out MD simulations that cover both the bound and open microstates; such simulations are expected to become extremely long and, hence, prohibitive with the full-protein model.

However, with the partial-protein model (but not as easily with the full model), one can use replica-exchange or multicanonical techniques to carry out a conformational search more efficiently than with long MD simulations at constant temperature, and differences in free energies can be obtained from the relative duration of the trajectory in the microstates of interest. The feasibility of this approach (for the partial-protein model) is mainly due to the increased exchange acceptance that is concurrent with smaller system sizes. Still, the transition of a loop between microstates by simulations is typically difficult because of high energy barriers; therefore, procedures for calculating the *absolute* free energy are expected to be very effective, because they would lead to $\Delta F = F_o - F_b$ and $\Delta S = S_o - S_b$ by

subtracting the values obtained from two separate simulations for the open and bound microstates without the need to “cover” the latter by a long trajectory. One such method, called HSMC or HSMD, was developed by us and has been applied thus far to argon, TIP3P water,⁸⁶ self-avoiding walks on a lattice,⁸⁷ and peptides,^{88,89} and we intend to extend it to the present partial-protein model as well.

Acknowledgment. This work was supported by NIH grants R01 GM66090 and R01 GM61916.

References

- (1) Kwasigroch, J. M.; Chomilier, J.; Mornon, J. P. *J. Mol. Biol.* **1996**, *259*, 855.
- (2) Martin, A. C.; Toda, K.; Stirk, H. J.; Thornton, J. M. *Protein Eng.* **1995**, *8*, 1093.
- (3) Oliva, B.; Bates, P. A.; Querol, E.; Aviles F. X.; Sternberg M. J. *J. Mol. Biol.* **1997**, *266*, 814.
- (4) Fetrow, J. S. *FASEB J.* **1995**, *9*, 708.
- (5) Getzoff, E. D.; Geysen, H. M.; Rodda, S. J.; Alexander, H.; Tainer, J. A.; Lerner, R. A. *Science* **1987**, *235*, 1191.
- (6) Rini, J. M.; Schulze-Gahmen, U.; Wilson, I. A. *Science* **1992**, *255*, 959.
- (7) Constantine, K. L.; Friedrichs, M. S.; Wittekind, M.; Jamil, H.; Chu, C. H.; Parker, R. A.; Goldfarb, V.; Mueller, L.; Farmer, B. T. *Biochemistry* **1998**, *37*, 7965.
- (8) Bates, P. A.; Sternberg, M. J. *Proteins* **1999**, Suppl 3, 47.
- (9) Mosimann, S.; Meleshko, R.; James, M. N. *Proteins* **1995**, *23*, 301.
- (10) Sali, A. *Curr. Opin. Biotechnol.* **1995**, *6*, 437.
- (11) Petrey, D.; Xiang, Z.; Tang, C. L.; Xie, L.; Gimpepev, M.; Mitros, T.; Soto, C. S.; Goldsmith-Fischman, S.; Kernysky, A.; Schlessinger, A.; Koh, I. Y. Y.; Alexov, E.; Honig, B. *Proteins* **2003**, *53*, 430.
- (12) Crasto, C. J.; Feng, J. *Proteins* **2001**, *42*, 399.
- (13) Leszczynski, J. F.; Rose, G. D. *Science* **1986**, *234*, 849.
- (14) Donate, L. E.; Rufino, S. D.; Canard, L. H.; Blundell, T. L. *Protein Sci.* **1996**, *5*, 2600.
- (15) Fichteler, T.; Dengler, U.; Schomburg, D. *J. Mol. Biol.* **1995**, *253*, 114.
- (16) Chothia, C.; Lesk, A. M. *J. Mol. Biol.* **1987**, *196*, 901.
- (17) Chothia, C.; Lesk, A. M.; Tramontano, A.; Levitt, M.; Smith-Gill, S. J.; Air, G.; Sheriff, S.; Padlan, E. A.; Davies, D.; Tulip, W. R. *Nature* **1989**, *342*, 877.
- (18) Gō, N.; Scheraga, H. A. *Macromolecules* **1970**, *3*, 178.
- (19) Dudek, M. J.; Scheraga, H. A. *J. Comput. Chem.* **1990**, *11*, 121.
- (20) Bruccoleri, R. E.; Karplus, M. *Biopolymers* **1987**, *26*, 137.
- (21) Summers, N. L.; Karplus, M. *J. Mol. Biol.* **1990**, *216*, 991.
- (22) Tappura, K. *Proteins* **2001**, *44*, 167.
- (23) Moul, J.; James, M. N. *Proteins* **1986**, *1*, 146.
- (24) Fine, R. M.; Wang, H.; Shenkin, P. S.; Yarmush, D. L.; Levinthal, C. *Proteins* **1986**, *1*, 342.
- (25) Shenkin, P. S.; Yarmush, D. L.; Fine, R. M.; Wang, H. J.; Levinthal, C. *Biopolymers* **1987**, *26*, 2053.
- (26) Higo, J.; Collura, V.; Garnier, J. *Biopolymers* **1992**, *32*, 33.
- (27) Rosenfeld, R.; Zheng, Q.; Vajda, S.; DeLisi, C. *J. Mol. Biol.* **1993**, *234*, 515.
- (28) Zheng, Q.; Rosenfeld, R.; Vajda, S.; DeLisi, C. *J. Comput. Chem.* **1993**, *14*, 556.
- (29) Caralacci, L.; Englander, S. W. *J. Comput. Chem.* **1996**, *17*, 1002.
- (30) Das, B.; Meirovitch, H. *Proteins* **2001**, *43*, 303.
- (31) Das, B.; Meirovitch, H. *Proteins* **2003**, *43*, 470.
- (32) Mas, M. T.; Smith, K. C.; Yarmush, D. L.; Aisaka, K.; Fine, R. M. *Proteins* **1992**, *14*, 483.
- (33) Smith, K. C.; Honig, B. *Proteins* **1994**, *18*, 119.
- (34) Wesson, L.; Eisenberg, D. *Protein Sci.* **1992**, *1*, 227.
- (35) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem A* **1997**, *101*, 3005.
- (36) Rapp, C. S.; Friesner, R. A. *Proteins* **1999**, *35*, 173.
- (37) de Bakker, P. I. W.; DePristo, M. A.; Burke, D. F.; Blundell, T. L. *Proteins* **2003**, *51*, 21.
- (38) DePristo, M. A.; de Bakker, P. I. W.; Lovell, S. C.; Blundell, T. L. *Proteins* **2003**, *51*, 41.
- (39) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins* **2004**, *55*, 351.
- (40) Ghosh, A.; Rapp, C. S.; Friesner, R. *J. Phys. Chem.* **1998**, *102*, 10983.
- (41) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517.
- (42) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.
- (43) Zhang, C.; Liu, S.; Zhou, Y. *Protein Sci.* **2004**, *13*, 391.
- (44) Xiang, X.; Soto, C. S.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7432.
- (45) Tanner, J. J.; Nell, L. J.; McCammon, J. A. *Biopolymers* **1992**, *32*, 23.
- (46) Szarecka, A.; Meirovitch, H. *J. Phys. Chem. B.* **2006**, *110*, 2869.
- (47) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (48) Alder, B. J.; Wainwright, T. E. *J. Chem. Phys.* **1959**, *31*, 459.
- (49) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585.
- (50) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
- (51) Ponder, J. W. *TINKER, Software Tools for Molecular Design*, version 4.2; Washington University: St. Louis, MO, 2004.
- (52) Brooks, C. L., III; Brünger, A. T.; Karplus, M. *Biopolymers* **1985**, *24*, 843.
- (53) Meirovitch, H.; Hendrickson, T. F. *Proteins* **1997**, *29*, 127.
- (54) Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A. *Science* **1986**, *235*, 574.
- (55) Merz, K. M., Jr. *J. Am. Chem. Soc.* **1991**, *113*, 406.
- (56) Miyamoto, S.; Kollman, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 8402.

- (57) Essex, J. W.; Severance, D. L.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem.* **1997**, *101*, 9663.
- (58) Smith, R. H., Jr.; Jorgensen, W. L.; Tirado-Rives, J.; Lamb, M. L.; Janssen, P. A. J.; Michejda, C. J.; Kroeger Smith, M. B. *J. Med. Chem.* **1998**, *41*, 5272.
- (59) Steinbach, P. J.; Brooks, B. R. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 9135.
- (60) Beglov, D.; Roux, B. *Biopolymers* **1995**, *35*, 171.
- (61) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637.
- (62) Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24.
- (63) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- (64) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
- (65) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.
- (66) Levitt, M.; Sharon, R. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 7557.
- (67) Makarov, V. A.; Feig, M.; Andrews, B. K.; Pettitt, B. M. *Biophys. J.* **1998**, *75*, 150.
- (68) Brunne, R. M.; Liepinsh, E.; Otting, G.; Wuthrich, K.; van Gunsteren, W. F. *J. Mol. Biol.* **1993**, *231*, 1040.
- (69) Makarov, V. A.; Andrews, B. K.; Smith, P. E.; Pettitt, B. M. *Biophys. J.* **2000**, *79*, 2966.
- (70) Sterpone, F.; Ceccarelli, M.; Marchi, M. *J. Mol. Biol.* **2001**, *311*, 409.
- (71) Marchi, M.; Sterpone, F.; Ceccarelli, M. *J. Am. Chem. Soc.* **2001**, *124*, 6787.
- (72) Luise, A.; Falconi, M.; Desideri, A. *Proteins* **2000**, *39*, 56.
- (73) Feller, S. C.; Pastor, R. W.; Rojnukarin, A.; Bogusz, S.; Brooks, B. R. *J. Phys. Chem.* **1996**, *100*, 17011.
- (74) Otting, G.; Liepinsh, E.; Wuthrich, K. *Science* **1991**, *254*, 974.
- (75) Modig, K.; Liepinsh, E.; Otting, G.; Halle, B. *J. Am. Chem. Soc.* **2004**, *126*, 102.
- (76) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.
- (77) Flory, P. J. *Statistical Mechanics of Chain Molecules*; Hasner: New York, 1988.
- (78) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050.
- (79) Sham, Y. Y.; Warshel, A. *J. Chem. Phys.* **1998**, *109*, 7940.
- (80) King, G.; Warshel, A. *J. Chem. Phys.* **1989**, *91*, 3647.
- (81) Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1992**, *97*, 3100.
- (82) Lounnas, V.; Ludemann, S. K.; Wade, R. C. *Biophys. Chem.* **1999**, *78*, 157.
- (83) Alper, H.; Levy, R. M. *J. Chem. Phys.* **1993**, *99*, 9847.
- (84) Lee, M. S.; Salsbury, F. R., Jr.; Olson, M. A. *J. Comput. Chem.* **2004**, *25*, 1967.
- (85) Lee, M. S.; Olson, M. A. *J. Phys. Chem. B* **2005**, *109*, 5223.
- (86) White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2004**, *121*, 10889.
- (87) White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2005**, *123*, 214908.
- (88) Cheluvaraja, S.; Meirovitch, H. *J. Chem. Phys.* **2005**, *122*, 054903.
- (89) Cheluvaraja, S.; Meirovitch, H. *J. Phys. Chem. B* **2005**, *109*, 21963.

Atomic Charge Parameters for the Finite Difference Poisson–Boltzmann Method Using Electronegativity Neutralization

Qingyi Yang and Kim A. Sharp*

*Johnson Research Foundation and Department of Biochemistry and Biophysics,
University of Pennsylvania, Philadelphia, Pennsylvania 19104*

Received January 6, 2006

Abstract: An optimization of Rappe and Goddard's charge equilibration (QEq) method of assigning atomic partial charges is described. This optimization is designed for fast and accurate calculation of solvation free energies using the finite difference Poisson–Boltzmann (FDPB) method. The optimization is performed against experimental small molecule solvation free energies using the FDPB method and adjusting Rappe and Goddard's atomic electronegativity values. Using a test set of compounds for which experimental solvation energies are available and a rather small number of parameters, very good agreement was obtained with experiment, with a mean unsigned error of about 0.5 kcal/mol. The QEq atomic partial charge assignment method can reflect the effects of the conformational changes and solvent induction on charge distribution in molecules. In the second section of the paper we examined this feature with a study of the alanine dipeptide conformations in water solvent. The different contributions to the energy surface of the dipeptide were examined and compared with the results from fixed CHARMM charge potential, which is widely used for molecular dynamics studies.

Introduction

Accurate representation of the charge distribution within molecules is essential for most atomic resolution modeling techniques. It is crucial for techniques such as finite difference Poisson–Boltzmann (PB) and Generalized Born (GB) methods, since their primary output is electrostatic potentials, energies, or forces. A significant barrier to many modeling studies is that of encountering a 'new' cofactor, ligand, or functional group, for which there are no parameters. Modeling cannot continue until parameters have been obtained. Bond stretch, angle, etc. parameters can usually be transferred from chemically similar groups, but atomic charges are a less local, more context dependent property. There is thus a continual need for rapid determination of molecular charges for use with specific modeling techniques. There are several alternative ways of representing molecular charge distributions. The most accurate uses the nuclear coordinate positions plus the full electron density (electronic

orbital) distribution obtained from quantum mechanical (QM) calculations. This representation requires significant memory and computation, and it is not practical for most macromolecular modeling. Thus a reduced representation is required. The most common is point atomic charges, usually centered at the nuclei. For the point atomic charge representation, several approaches have been used to obtain parameter sets suitable for modeling. Again, QM methods are the most fundamental and, in principle, the most accurate. There are, however, two uncertainties encountered with this approach. First, there are different ways to obtain atomic charges from the orbitals, including Mulliken population analysis, apportioning the electron density by locating saddle points in the distribution, and fitting to the Coulomb potential at a set of points surrounding the molecule. The restrained electrostatic potential (RESP) fitting¹ is widely used since it fits to the principal property that the charges will actually be used for: The potential distribution around the molecule. Whatever method is used, however, the point charge representation is an approximation, even if the QM calculation were exact.

* Corresponding author phone: (215)573-3506; fax: (215)898-4217; e-mail: sharpk@mail.med.upenn.edu.

The second problem is producing QM based charges that mimic the condensed phase. Inclusion of, for example, a solvent reaction field in the QM calculations helps, but charges usually need to be further scaled or fine-tuned for maximum accuracy in modeling macromolecules in solution. Thus, even atomic charges that start with QM calculations are empirical. Another approach is to just parametrize charges by fitting calculated electrostatic properties to experiment. This can be done completely from scratch, as was the PARSE amino acid set for the finite difference PB (FDPB) method.² This worked well for the 20 amino acid side chains and the peptide backbone, with a restricted set of functional group types. It is difficult to extend this approach to more complicated groups/molecules without some starting guesses, however. Even parametrized charge sets usually require some preliminary calculations as a starting point.

Two important criteria for empirical atomic charge parametrization are what theoretical or experimental quantities are to be fit and what method(s) the charges are going to be used with. Atomic charges for molecular dynamics (MD) force fields are usually obtained through a combination of QM calculations and fitting to experiment and other theoretical calculations. In contrast, the OPLS set was fit to solvation data using free energy perturbation methods,³ and this set would be preferred for solvation energy calculations using the same methods. The original PARSE set was fit to experimental solvation free energies obtained from partition data, using the FDPB method.² Several methods have been parametrized to reproduce RESP fit charges from a particular level of QM calculations.^{4,5} Different semiempirical QM methods have been assessed by comparison to solvation free energies calculated by free energy perturbation.^{6,7} Many of these different charge sets are simultaneously in use for different types of simulations, evidence that different charge sets work better with particular modeling methods, for particular molecules or for particular quantities. This is not surprising given that atomic charges are an abstraction, and all sets are empirical to some degree. Each of the above parametrization strategies has its advantages for particular applications. Given this, we decided that with the wide use of the PB model for macromolecular electrostatics, it would be useful to develop a rapid procedure for generating atomic charges for use specifically with the FDPB method. Our criteria were first that the method be simple and could be applied to diverse cofactors and functional groups with the same small set of parameters. Second, that it be rapid and require little user input. Given the approximate nature of atomic charges, methods that consume a lot of time or that require the user to explore multiple options are not cost-effective. Third, given the empirical nature of atomic charge sets, we wanted to parametrize specifically for a given method, FDPB. As with the original PARSE charge set, we chose solvation free energy data to parametrize against: Solvation free energies can be measured reliably, the electrostatic component can be accurately extracted, and it bears on the important solvation component of macromolecules for which FDPB is often used. However, the original PARSE strategy of ground up parametrization for each chemical subgroup type is difficult to extend to more

complicated groups, and it becomes prone to subjectivity. On the other hand, QM based methods are computationally expensive, require several steps and choices of strategy (e.g. level of theory, basis sets, fitting methods), and in the end results in sets that are still partly empirical. We chose a middle course, selecting a charge equalization (CE) through an electronegativity method. Several implementations of this method have been described,^{8–13} it is easy to implement, and it is quite flexible. Gilson et al. showed that the equalization method can be made to mimic more rigorous QM calculations well by the careful choice of atomic types and that the method can account for different resonance forms. In this work, however, we focus on reproducing experimental solvation free energies rather than QM calculations. We chose to use the charge equalization method of Rappe and Goddard.¹¹ They developed a very simple but general scheme to generate charge distributions for use in molecular dynamics simulation based only on molecular geometry and atomic properties. The input parameter data are just atomic ionization energy (IP), electron affinity (EA), available from standard tabulations of element properties, and covalent radius (R), obtained from crystallographic data. The original method was parametrized for gas (vacuum) phase.

In this study we describe an optimization of Rappe and Goddard's charge equilibration method (which they call QEq) for FDPB solvation energies by a combination of adjusting the element data and making a small addition to the algorithm to increase the number of the functional groups it can handle. The method proposed here is thus designed to be a successor to and extension of the PARSE charge set. In the first section of this paper we show the results of the parametrization on small molecule solvation free energies. Following a minimalist strategy we used a rather small set of parameters, a small training set and a large test set, since we believe this is likely to result in a robust set of charges. The QEq method of Rappe and Goddard potentially provides a significant enhancement of the electrostatic treatment in molecular dynamics because it allowed the atomic charge distribution to respond to the molecule's geometry changes, i.e., it is effectively a polarizable potential function, and it is very fast. While polarizable potentials for MD applications are not our primary focus, they are currently under active development and testing.^{14–17} In the second section of this paper we applied our parametrized QEq method to a study of the alanine dipeptide conformations. We studied the Ramachandran energy map and compared the results of the QEq method with polarization to the fixed charge CHARMM potential¹⁸ which is widely used for MD. Both sets of maps were calculated with and without a FDPB solvation contribution.

Methods

Calculation of Partial Atomic Charges. Given an atom's ionization potential (IP) and electron affinity (EA) Rappe and Goddard defined an atomic scale chemical potential by taking the derivative of the total electrostatic energy with respect to that atom's charge, Q_A , leading to¹¹

$$\chi_A = \chi_A^0 + J_{AA}^0 Q_A + \sum_{B \neq A} J_{AB} Q_B \quad (1)$$

where $\chi_A^0 = (\text{IP} + \text{EA})/2$ is the electronegativity, and $J_{AA}^0 = (\text{IP} - \text{EA})$ is the Coulomb repulsion (self-Coulomb integral) between two electrons in A's outer orbital. J_{AB} is the interatomic electrostatic energy between atoms A and B. We followed Rappe and Goddard in representing the electron density of the atoms with Slater type orbitals

$$\phi(r, n) = B_n r^{n-1} e^{-\zeta r} \quad (2)$$

where r is distance from the nucleus, n is the row number of the element in the periodic table, B_n is a normalization constant, and ζ defines the size of the atom (orbital) using the covalent radius of an atom, R_A , through the relation

$$\zeta_A = (2n + 1)/(4R_A) \quad (3)$$

For the special case of hydrogen, which has only one electron, the size is charge dependent: $\zeta_H = \zeta_H^0 + Q_H$. J_{AB} is computed from the overlap integral of two Slater type orbitals with the form of eq 2:

$$J_{AB} = \int \int \phi_A(r, n_A) \frac{1}{|\mathbf{r} - \mathbf{s}|} \phi_B(s, n_B) \delta \mathbf{r} \delta \mathbf{s} \quad (4)$$

J_{AB} has the form of the Coulomb potential $1/R$ for large separations. For small separations as the electron density distributions of A and B overlap the interaction is shielded and J_{AB} plateaus to a constant value at $r = 0$. Given eq 1, the atomic charges of a molecule or group with N atoms are obtained by equating all the chemical potentials

$$\chi_i = \chi_N, \quad i = 1, N - 1 \quad (5)$$

subject to a user specified condition on the total charge of the molecule or group given by

$$\sum Q_i = Q_{\text{net}} \quad (6)$$

which leads to a simple matrix equation for the charges

$$C_{ij} Q_j = -D_i \quad (7)$$

where

$$C_{1j} = 1, D_1 = Q_{\text{net}} \quad (8a)$$

$$C_{ij} = J_{ij} - J_{1j}, D_i = \chi_i^0 - \chi_1^0, \text{ for } i \neq 1 \quad (8b)$$

To evaluate the J_{AB} terms required for the elements C_{ij} , eq 4 is integrated numerically by transforming into elliptical coordinates with atoms A and B at the foci and using a Gaussian quadrature routine.¹⁹ Equation 7 is then solved using the linear algebra routines from Numerical Recipes.¹⁹ For each hydrogen atom, initially Q_H is set to zero, and $\zeta_H = \zeta_H^0$. After solving eq 6, ζ_H is updated using $\zeta_H = \zeta_H^0 + Q_H$ with the current estimate of Q_H . Then eq 7 is resolved. This procedure is repeated to convergence (usually taking about 6–8 cycles). The algorithm was implemented as a Fortran 77 program called QEQUIL. The entire charge determination algorithm typically takes less than 0.1 s on a 2.5 GHz processor for compounds with up to several dozen

Table 1. Input Atomic Parameters

element	atomic		J (eV)	radius ^a (Å)	χ (eV)	
	no.	row			original	optimized
H	1	1	13.8904	0.371	4.528	4.498
C	6	2	10.126	0.759	5.343	5.723
N	7	2	11.76	0.715	6.899	8.599
O	8	2	13.364	0.669	8.741	8.961
F	9	2	14.498	0.706	10.874	6.374
S	16	3	8.972	0.947	6.928	6.268
Cl	17	3	9.892	0.994	8.564	5.464
Br	35	4	8.85	1.141	7.79	5.790
I	53	5	7.524	1.333	6.822	5.622

^a Covalent radius used for charge equilibration only.

atoms. Output is in the modified DelPhi PDB format with charge and radii in the occupancy and b-factor columns, so it can be read directly by DelPhi and GRASP.^{20,21} The program QEQUIL is available from the authors upon request.

Solvation Free Energies. Solvation free energies were calculated using the FDPB/SA method as described previously,² using the FDPB package DELPHI. Radii were taken from the PARSE set, for H = 1.0 Å, C = 1.9 Å, O = 1.6 Å, N = 1.65 Å, S = 1.9 Å. Radii for the halogens were taken from the AMBER parm99 potential function.²² Since we aimed to produce a second generation equivalent of the PARSE charges, radii were used as is, and we did not attempt to optimize them also. The covalent radii tabulated by Rappe and Goddard and in Table 1 are used only for the charge determination phase, as described by eqs 2 and 3, not for solvation free energy calculations. Solvent accessible surface (SAS) areas were calculated using the program SURFCV,²³ and the nonpolar contribution to solvation, $\Delta G^{\text{np}}(\text{calc})$, is obtained by multiplying the SAS by the 'hydrophobic' coefficient $\gamma = 5 \text{ cal/mol/Å}^2$. The electrostatic component of the solvation free energy, $\Delta G^{\text{elec}}(\text{calc})$, is computed as the difference in free energy of the molecule between vacuum (exterior dielectric constant $\epsilon=1$) and water (exterior dielectric constant $\epsilon=80$). Solute molecules were assigned a dielectric $\epsilon = 2$, a commonly accepted value that accounts for small solute polarizability. Am1-bcc charges were obtained using the commands "bcctype" and "bcc" within the antechamber module of AMBER V7.0.²⁴ AMSOL charges were obtained using the AMSOL v 7.1 program,²⁵ using the class IV CM2 point charge model. AMSOL/GB solvation calculations were performed using the SM5.42R solvation model in the AMSOL v 7.1 program.

Training Set of Solutes. The 14 polar amino acid side-chain analogues and the peptide backbone analogue N, methyl-acetamide, were used as the core of the training set. The omitted amino acid side chains G, A, P, V, L, and I are purely nonpolar and so do not affect parametrization of charges. Experimental vapor-to-water transfer free energies, $\Delta G^{\text{solv}}(\text{expt})$, were taken from Wolfenden.²⁶ For the solvation calculations the neutral form of ionizable residues was used, to conform to the experimental measurements.²⁶ Eight other compounds with a range of functional groups were also used for training (Table 3). Solvation energies for these were taken from Cabani et al.²⁷ The nonpolar part of the solvation free energy was assumed to be accurately represented by the

Table 2. Atom Pair Shielding Factors

atom pair	S_{AB}	atom pair	S_{AB}
C–N (amide)	0.66	C–O (aldehyde)	1.61
C=N (nitrile)	1.63	C–O (general)	1.03
C–N (general)	1.10	C≡C (-yne)	2.33
N=O (nitro)	2.33		

current SAS method and ‘hydrophobic’ coefficient γ , so γ was not reparametrized here. Thus to show more clearly the effects of charge parametrization we tabulate the calculated electrostatic contribution to solvation free energy, $\Delta G^{\text{elec}}_{\text{(calc)}}$, and the ‘experimental’ electrostatic contribution to solvation free energy defined as $\Delta G^{\text{elec}}_{\text{(expt)}} = \Delta G^{\text{sol}}_{\text{(expt)}} - \Delta G^{\text{np}}_{\text{(calc)}}$. Our parametrization was done against data for the vacuum-to-water transfer process for several reasons. First, the most common application of the FDPB method is for calculation of the net hydration free energy, which is the quantity obtained from vapor-water transfer data, so we wanted to optimize for this process. Second such data are of high experimental reliability, there being only one solvent to contend with. We did not aim to produce a general solvation parametrization for use with a range of organic solvents, such as the AMSOL set. This, while of more general applicability, would be less accurate for just water solvent applications.

Input Structures. Small molecules for the training and test sets were built with ISIS/DRAW 2.5 (MDL interactive systems, San Leandro, CA), and their conformations were optimized in InsightII (Accelrys, San Diego, CA). The lowest energy conformer was used for solvation calculations, except for the study of the alanine dipeptide. For the alanine

dipeptide, conformations with different Φ and Ψ values were built in CHARMM,¹⁸ using a torsion angle grid of 10° from -180° to 180° . Experimental solvation free energies were taken from Radzicka and Wolfenden²⁶ for the side-chain and backbone analogues and from Cabani²⁷ for the other compounds.

Parametrization Strategy. To minimize the number of adjustable parameters, we initially chose to optimize by adjusting only the atomic electronegativities, χ_i . Electrostatic solvation free energies for the training set were first calculated with the original parameters of Rappe and Goddard¹¹ (Table 1). To guide the initial direction of parametrization, the resulting data were examined for systematic deviations that could be attributed to specific element electronegativities. For example, compounds that contained sulfur had systematically overestimated electrostatic solvation contributions. The mean squared error in electrostatic solvation free energy, averaged over the training set of 23 compounds in Table 3, was then minimized by adjusting each electronegativity in turn, keeping the others constant. The resulting roughly optimized electronegativities were then used as input into a systematic optimization of the mean squared error in electrostatic solvation free energy by using a grid search over χ_i for the nine elements involved, H, C, O, N, S, F, Cl, Br, and I. The χ_i 's were varied over a range of 3.500–11.500 using a step size of 0.015.

Satisfactory optimization could be achieved by adjusting the χ_i 's alone. However, since the charge equilibration methods are intrinsically limited in dealing with pi- and delocalized bonds,²⁸ it is unlikely that the best results can be obtained by optimizing just the individual atomic quantities, χ_i 's. Similar considerations imply that adjusting the J_{AA}

Table 3. Training Set Solvation Free Energies (kcal/mol) Using Original and Optimized Parameter Sets

ID	molecule ^a	SAS (Å ²)	$\Delta G^{\text{np}}_{\text{(calc)}}$	$\Delta G^{\text{sol}}_{\text{(expt)}}$	$\Delta G^{\text{elec}}_{\text{(expt)}}$	$\Delta G^{\text{elec}}_{\text{(calc)}}$	
						original	optimized
1	1-heptyne	278	2.25	0.60	-1.65	-1.91	-2.87
2	fluoromethane	133	1.53	-0.22	-1.53	-10.95	-1.77
3	1-chloropropane	208	1.90	-0.27	-2.18	-12.04	-2.16
4	1-bromopropane	224	1.98	-0.56	-2.54	-8.53	-2.66
5	1-iodopropane	233	2.03	-0.59	-2.61	-5.74	-2.62
6	propanenitrile	198	1.85	-3.85	-5.7	-2.89	-5.60
7	1-nitropropane	224	1.98	-3.34	-5.33	-12.02	-5.43
8	pentanal	303	2.38	-3.03	-5.41	-8.69	-5.38
9	N-propylguanidine (arg)	320	2.46	-10.92	-13.38	-5.87	-13.60
10	acetamine (asn)	226	1.99	-9.72	-11.71	-7.05	-11.63
11	acetic acid (asp)	222	1.97	-6.70	-8.67	-8.83	-7.79
12	methylthiol (cys)	202	1.87	-1.24	-3.11	-4.70	-2.91
13	propionamide (gln)	260	2.16	-9.42	-11.58	-7.62	-12.06
14	propionic acid (glu)	256	2.14	-6.47	-8.61	-9.46	-8.45
15	methylimidazole (his)	272	2.22	-10.25	-12.47	-5.81	-12.23
16	N-butylamine (lys)	286	2.29	-4.38	-6.67	-4.46	-8.32
17	methyl ethyl sulfide (met)	272	2.22	-1.49	-3.71	-5.81	-3.99
18	toluene (phe)	306	2.39	-0.76	-3.15	-1.54	-3.09
19	methanol (ser)	180	1.76	-5.08	-6.84	-8.23	-7.28
20	ethanol (thr)	218	1.95	-4.90	-6.85	-8.17	-7.31
21	methylindole (trp)	348	2.60	-5.91	-8.51	-2.84	-6.23
22	p-cresole (tyr)	318	2.45	-6.13	-8.58	-7.58	-7.30
23	N-methylacetamide (backbone)	266	2.19	-10.08	-12.27	-7.63	-13.15

^a Amino acid side-chain/backbone analogue in brackets.

values would simply introduce more adjustable parameters without producing qualitatively improved fits, since both contribute in a linear fashion to the effective electronegativity in eq 1. To overcome this limitation, we introduced one additional parameter type, a shielding factor S_{AB} that scales the electrostatic interaction between two atoms, as described by a modified version of eq 1:

$$\chi_A = \chi_A^0 + f_{AA}^0 Q_A + \sum_{B \neq A} S_{AB} J_{AB} Q_B \quad (9)$$

This allows us to incorporate the effect of bond delocalization or atomic-neighbor, bond-dipole, and atomic context type effects in a simple way. Shielding factors were initially set to 1 and then optimized using a grid search. Significantly improved fits could be obtained by introduction of several shielding factors, most involving atom pairs with higher order/resonance bonds (Table 2). The shielding factors for all other atom pairs were 1 and thus could be omitted.

Molecular Mechanics of Alanine Dipeptide. Alanine dipeptide was built using CHARMM, with the CHARMM27 potential.²⁹ Conformations for the Ramachandran plots were generated using a 10° increment grid of phi and psi angles. For each conformation the dipeptide was minimized subject to constraints on the phi and psi angles using the CHARMM27 steepest descent minimizer by 600 steps, using a dielectric of 1, and a nonbond cutoff of 80.0 Å. The internal CHARMM energy terms were taken from the final minimized structures. Solvation energies were calculated using the FDPB method as described above from the minimized structures using either the CHARMM27 or QEq charges. The geometric center of each peptide bond was used as the origin for the calculation of that peptide's dipole moment as follows

$$p = \sum q_i \cdot r_i \quad (10)$$

where the sum runs over the C, O, N, and H atoms.

Results and Discussion

Charge Parametrization. Figure 1 shows the results of QEQUIL on the training set of 23 compounds using the original parameters of Rappe and Goddard (Table 1). The correlation is very poor at $r^2 = 0.01$, with a mean unsigned error of 3.5 kcal/mol. Examination of the direction of error for individual compounds reveals systematic errors: For example N and S containing groups have systematically over- and underestimated free energies, respectively (Figure 2). Systematic parametrization of the electronegativities improved the overall correlation to > 0.85 . Inclusion of an atomic pair shielding factor improved results still further, with a final $r^2 = 0.96$, a slope very close to unity (0.99), and a mean unsigned error of 0.5 kcal/mol, giving a good fit between experiment and calculated values (Figures 1 and 2). Optimized parameters in Table 1 show that most of the adjustment of nonhalogen elements occurred in the N and S elemental negativity values, with smaller adjustments in those for H, C, and O. To obtain the best fit seven pair shielding factors were required to be significantly different from the null value of 1 (Table 2).

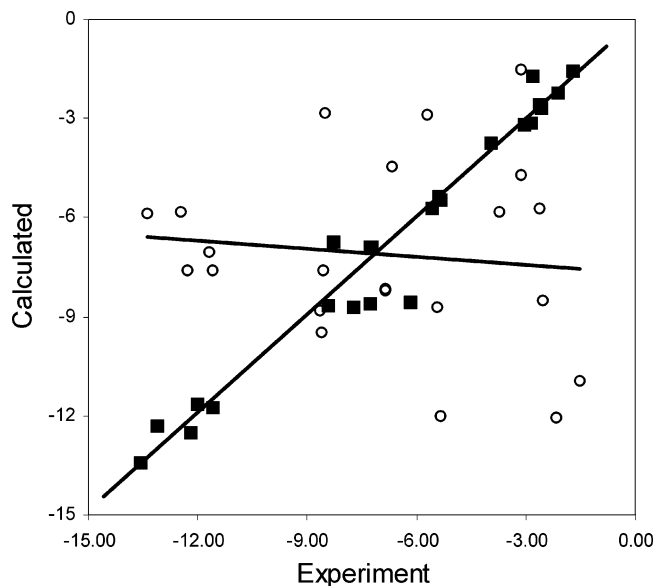


Figure 1. Comparison of experimental and calculated solvation free energies for the training set in Table 3. Energies are in kcal/mol. (o): Original parameter set. (■): Optimized parameter set. Lines are the least-squares fits.

The major error in the amino acid analogues was from the Lys, Trp, and Tyr side-chain analogues (Figure 2). The delocalized electron systems of Trp and Tyr are understandably less well represented in any classical method, such as the one used here.²⁸ An option would be to add further bond shielding factors, although this was not done here since we wanted to keep the number of adjustable parameters to a minimum. Aside from these deviations, the modified parameters show excellent agreement for side-chain analogues with carboxyl, thiol, amide, and OH moieties. Table 3 gives solvation free energy values for the original and optimized parameters, and the additional breakdown into different terms. In total, nine original element parameters were optimized, along with seven shielding parameters, for a total of 16 parameters on a 23 compound training set. This is a very modest number of parameters for charge sets; they often have dozens of fitted parameters.

The FDPB method for solvation is often used in conjunction with the MD force fields such as CHARMM. The MD is typically used either to generate an ensemble of protein structures for postprocessing with FDPB³⁰ or more recently with FDPB treatment of solvent forces integrated into the MD simulation.^{31–33} We wanted to see how accurate unmodified CHARMM charges were with the same FDPB protocol used for QEq charges. If CHARMM charges could be used as is for solvation energy calculations, it would make integrated MD/FDPB calculations easier and require fewer parameters. Since CHARMM charges were available only for amino acids, results for these charges refer only to the 15 side-chain and backbone analogue compounds of the training set in Table 3. The results were significantly better than unoptimized QEq charges, with a correlation coefficient of 0.85 (Table 6) and a reasonable slope of 0.87. There is, however, a significant mean error of 2.1 kcal/mol so that use of the same set of charges for MD and solvation via FDPB would be less accurate than using two sets.

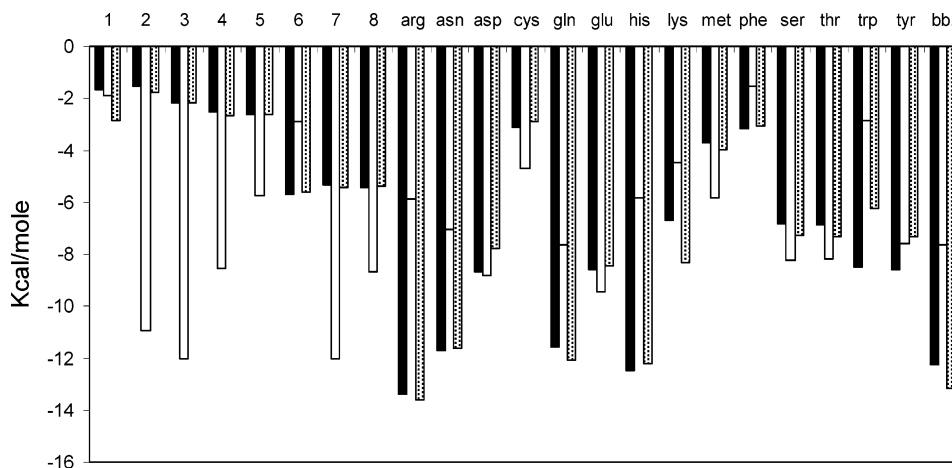


Figure 2. Comparison of experimental and calculated solvation free energies for the training set, including amino acid side-chain analogues (three letter amino acid code) and the backbone analogue NMA (bb). Other compound keys are given in Table 3. Filled bars: experiment. Unfilled bars: original parameters. Shaded bars: optimized parameters.

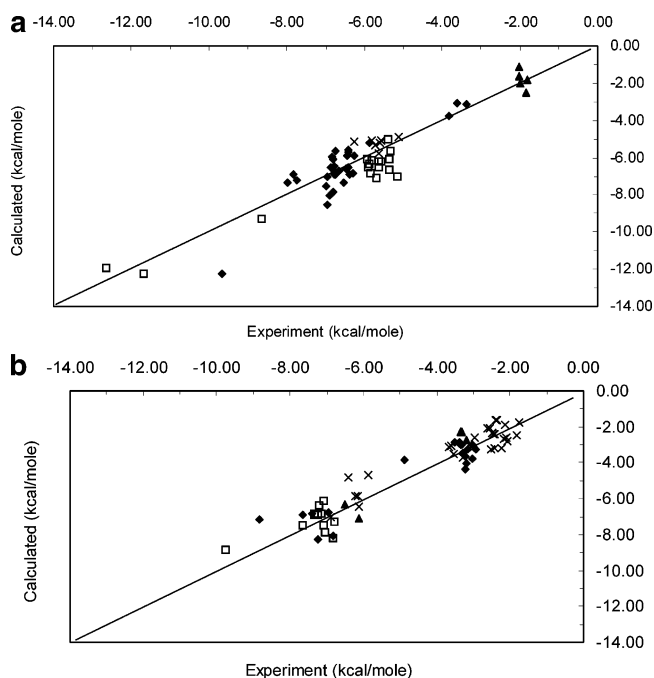


Figure 3. Comparison of experimental and calculated solvation free energies for the test set. $X = Y$ line is shown on each figure. (a) Compounds containing $-OH$, NH_n , and SH (\blacklozenge), $C=O$, $COOH$, and CHO (\square), CN and NO_2 (\times), alkynes (\blacktriangle). (b) Compounds containing aromatic groups (\blacklozenge), pyridines (\square), halogenated alkanes and alcohols (\times), halogenated aromatics (\blacktriangle).

For a test set, we used a different set of 127 compounds with a variety of functional groups that includes those used in the original PARSE parametrization, and for which there are reliable solvation data.²⁷ Figure 3 shows a comparison of experimental and calculated electrostatic solvation free energies, with compounds grouped into broad categories based on the main type of functional group. Table 4 gives an additional breakdown of the energy terms for each compound. The overall agreement was good, with $R^2 = 0.90$, a calculated vs experimental slope of 0.98, and a mean unsigned error of 0.61 kcal/mol. Overall the method does

well considering the small number of parameters and the fact that the test set is five times the size of the training set. Compound classes that contributed more to the mean error include the ketones and pyridines. The error from the latter is understandable as this is a large class of compounds, and the functional group did not occur in the training set. Compounds with alcohol, amide, thiol, sulfide, aldehyde, and benzyl groups were all accurately represented, as were the majority of halogenated compounds.

Comparison with Other Charge Sets. A number of parametrized atomic charged sets have been described in the literature using different approaches and designed for different methods/applications. We compared to two recent and widely used charge sets to see if these could be used with the FDPB method as is: the SM5.42R generated by AMSOL^{4,6} and the AM1-bcc set.^{5,7} These are both QM derived sets. AM1-bcc was parametrized on 2755 compounds so as to reproduce high level QM charges with a more computationally tractable semiempirical QM method, but they have also been reported to give good solvation energies with the explicit water/free energy perturbation method.⁷ The AMSOL set was specifically designed to produce good solvation energies with the generalized Born (GB) solvent model for a wide range of solvents, including water. Our calculations were done using exactly the same FDPB protocol as for the QEq charges, since we wanted to compare charges using the same method. AMBER radii²² were used with the AM1-bcc set in order to compare these charges on a more equal footing. Both previous charge sets gave significantly larger errors and, in the case of AMSOL charges, a poorer slope (Table 6). Figure 4 compares these charge sets graphically for compounds with a representative range of chemical types, to show where systematic differences, in terms of functional group type, occurred. We emphasize that the AM1-bcc and AMSOL charge sets were not parametrized specifically for the FDPB solvation application, and obviously they perform better with the methods they were designed for. However, the results reiterate the point that

Table 4. Solvation Free Energies (kcal/mol) for Test Set

code	molecule	main functional group(s) ^a	SAS (Å ²)	$\Delta G^{\text{TP}}(\text{calc})$	$\Delta G^{\text{Solv}}(\text{expt})$	ΔG^{elec}	
						(expt)	(calc)
1	propanol	OH	250	2.11	-4.83	-6.94	-7.04
2	butanol	OH	282	2.27	-4.72	-6.99	-7.51
3	isopropyl alcohol	OH	246	2.09	-4.76	-6.85	-6.54
4	2-butanol	OH	278	2.25	-4.58	-6.83	-5.99
5	3-methyl-1-butanol	OH	306	2.39	-4.42	-6.81	-7.84
6	ammonia	NH ₃	140	1.56	-4.31	-5.87	-5.23
7	methylamine	NH ₂	190	1.81	-4.57	-6.38	-6.91
8	ethylamine	NH ₂	222	1.97	-4.50	-6.47	-6.62
9	propylamine-N-butylamine	NH ₂	254	2.13	-4.39	-6.52	-7.32
10	dimethylamine	NH ₂	230	2.01	-4.29	-6.30	-6.84
11	diethylamine	NH ₂	298	2.35	-4.07	-6.42	-6.52
12	ethylthiol	SH	240	2.06	-1.30	-3.36	-3.12
13	dimethyl sulfide	CH-S-CH	242	2.07	-1.54	-3.61	-3.06
14	diethyl sulfide	CH-S-CH	308	2.40	-1.43	-3.83	-3.79
15	acetone	>C=O	244	2.08	-3.85	-5.93	-6.07
16	2-butanone	>C=O	276	2.24	-3.64	-5.88	-6.36
17	2-pentanone	>C=O	304	2.38	-3.53	-5.91	-6.50
18	3-pentanone	>C=O	306	2.39	-3.41	-5.80	-6.16
19	3-methyl-2-butanone	>C=O	306	2.39	-3.24	-5.63	-6.20
20	2,4-dimethyl-3-pentanone	>C=O	346	2.59	-2.74	-5.33	-5.66
21	acetaldehyde	HCO	206	1.89	-3.50	-5.39	-4.40
22	propionaldehyde	HCO	240	2.06	-3.44	-5.50	-4.72
23	butyric acid	COOH	284	2.28	-6.36	-8.64	-9.31
24	benzene	arom	258	2.15	-0.87	-3.02	-3.01
25	ethylbenzene	arom	326	2.49	-0.80	-3.29	-3.48
26	pyridine	CH=N-CH=	250	2.11	-4.70	-6.81	-8.10
27	4-methylpyridine	CH=N-CH=	290	2.31	-4.93	-7.24	-8.30
28	2-methylpyridine	CH=N-CH=	290	2.31	-4.63	-6.94	-6.75
29	phenol	OH (arom)	272	2.22	-6.62	-8.84	-7.15
30	pentanol	OH	314	2.43	-4.47	-6.90	-8.06
31	4-methyl-2-pentanol	OH	332	2.52	-3.73	-6.25	-5.89
32	2-pentanol	OH	310	2.41	-4.39	-6.80	-6.65
33	2-methyl-2-butanol	OH	302	2.37	-4.43	-6.80	-6.11
34	2,3-dimethylisobutyl alcohol	OH	324	2.48	-3.92	-6.40	-5.68
35	3-pentanol	OH	310	2.41	-4.35	-6.76	-6.49
36	hexanol	OH	346	2.59	-4.36	-6.95	-8.56
37	3-hexanol	OH	342	2.57	-4.08	-6.65	-6.71
38	2-methyl-3-pentanol	OH	334	2.53	-3.89	-6.42	-5.61
39	4-heptanol	OH	374	2.73	-4.01	-6.74	-6.92
40	2-methylpropanol	OH	276	2.24	-4.52	-6.76	-6.89
41	2-methyl-2-propanol	OH	272	2.22	-4.51	-6.73	-5.64
42	2-methyl-2-pentanol	OH	332	2.52	-3.93	-6.45	-5.93
43	4-methyl-2-pentanone	>C=O	330	2.51	-3.06	-5.57	-6.23
44	4-heptanone	>C=O	366	2.69	-2.93	-5.62	-6.54
45	5-nonanone	>C=O	430	3.01	-2.67	-5.68	-7.07
46	2-hexanone	>C=O	340	2.56	-3.29	-5.85	-6.84
47	butanal	HCO	272	2.22	-3.18	-5.40	-5.00
48	hexanal	HCO	336	2.54	-2.81	-5.35	-6.08
49	heptanal	HCO	368	2.70	-2.67	-5.37	-6.65
50	octanal	HCO	400	2.86	-2.29	-5.15	-7.03
51	N-methylformamide	amide	230	2.01	-10.00	-12.01	-10.79
52	cyclopentanol	OH (cyclic)	278	2.25	-5.49	-7.74	-7.23
53	cyclohexanol	OH (cyclic)	302	2.37	-5.47	-7.84	-6.89
54	cycloheptanol	OH (cyclic)	324	2.48	-5.49	-7.97	-7.37
55	pyrrolidine	5ring NH	260	2.16	-5.48	-7.64	-6.88
56	piperidine	5ring NH	284	2.28	-5.11	-7.39	-6.85
57	propylbenzene	alkyl arom	358	2.65	-0.53	-3.18	-4.03
58	<i>o</i> -xylene	alkyl arom	324	2.48	-0.90	-3.38	-2.85
59	<i>m</i> -xylene	alkyl arom	332	2.52	-0.83	-3.35	-2.97

Table 4. (Continued)

code	molecule	main functional group(s) ^a	SAS (Å ²)	$\Delta G^{\text{np}}(\text{calc})$	$\Delta G^{\text{soliv}}(\text{expt})$	ΔG^{elec}	
						(expt)	(calc)
60	<i>p</i> -xylene	alkyl arom	332	2.52	-0.81	-3.33	-2.97
61	naphthalene	alkyl arom	328	2.50	-2.39	-4.89	-3.83
62	1,2,4-trimethylbenzene	alkyl arom	360	2.66	-0.86	-3.52	-2.83
63	methylethylbenzene	alkyl arom	352	2.62	-0.30	-2.92	-3.25
64	butylbenzene	alkyl arom	390	2.81	-0.40	-3.21	-4.37
65	methylpropylbenzene	alkyl arom	380	2.76	-0.45	-3.21	-3.68
66	dimethylethylbenzene	alkyl arom	370	2.71	-0.44	-3.15	-3.27
67	dimethylpropylbenzene	alkyl arom	396	2.84	-0.18	-3.02	-3.80
68	3-methylpyridine	-CH=N-CH=	288	2.30	-4.77	-7.07	-7.49
69	2-ethylpyridine	-CH=N-CH=	320	2.46	-4.33	-6.79	-7.27
70	3-ethylpyridine	-CH=N-CH=	320	2.46	-4.60	-7.06	-7.85
71	4-ethylpyridine	-CH=N-CH=	318	2.45	-4.37	-6.82	-8.23
72	2,4-dimethylpyridine	-CH=N-CH=	326	2.49	-4.86	-7.35	-6.93
73	3,4-dimethylpyridine	-CH=N-CH=	316	2.44	-5.22	-7.66	-7.47
74	3,5-dimethylpyridine	-CH=N-CH=	324	2.48	-4.84	-7.32	-6.82
75	2,3-dimethylpyridine	-CH=N-CH=	308	2.40	-4.83	-7.23	-6.92
76	2,5-dimethylpyridine	-CH=N-CH=	326	2.49	-4.72	-7.21	-6.38
77	2,6-dimethylpyridine	-CH=N-CH=	326	2.49	-4.60	-7.09	-6.12
78	dimethylethylpyridine	-CH=N-CH=	364	2.68	-4.46	-7.14	-6.86
79	1,2-ethanediol	OH-OH	226	1.99	-7.66	-9.65	-12.27
80	1-propyne	-yne	173	1.72	-0.31	-2.03	-1.15
81	1-butyne	-yne	200	1.86	-0.16	-2.02	-1.61
82	1-pentyne	-yne	229	2.01	0.01	-1.99	-2.01
83	1-hexyne	-yne	253	2.12	0.29	-1.84	-2.49
84	1-buten-3-yne	-yne	196	1.84	0.04	-1.80	-1.85
85	fluoromethane	F, -alkane	133	1.53	-0.22	-1.75	-1.77
86	chloroethane	Cl, alkane	183	1.78	-0.63	-2.41	-1.60
87	2-chloropropane	Cl, alkane	206	1.89	-0.25	-2.14	-1.87
88	1-chlorobutane	Cl, alkane	234	2.03	-0.14	-2.17	-2.66
89	1-chloropentane	Cl, alkane	260	2.16	-0.07	-2.23	-3.20
90	2-chloropentane	Cl, alkane	256	2.14	0.07	-2.07	-2.82
91	3-chloropentane	Cl, alkane	254	2.13	0.04	-2.09	-2.63
92	chloroethene	Cl, alkene	180	1.76	-0.59	-2.35	-1.60
93	3-chloror-1-propene	Cl, alkene	208	1.90	-0.57	-2.47	-2.35
94	chlorobenzene	Cl, arom	243	2.07	-1.12	-3.20	-2.74
95	bromoethane	Br, alkane	197	1.85	-0.70	-2.54	-2.06
96	2-bromopropane	Br, alkane	219	1.96	-0.48	-2.43	-2.38
97	1-bromobutane	Br, alkane	249	2.11	-0.41	-2.51	-3.26
98	1-bromo-2methylpropane	Br, alkane	243	2.08	-0.03	-2.10	-2.63
99	bromobenzene	Br, arom	257	2.15	-1.46	-3.61	-3.04
100	1-bromo-4-methylbenzene	Br, arom	282	2.27	-1.39	-3.66	-3.12
101	1-bromo-2-ethylbenzene	Br, arom	296	2.34	-1.19	-3.53	-3.53
102	1-bromo-2-(1-methylethyl)benzene	Br, arom	318	2.45	-0.85	-3.30	-3.70
103	iodoethane	I, alkane	207	1.89	-0.72	-2.62	-2.06
104	2-iodopropane	I, alkane	227	1.99	-0.46	-2.46	-2.39
105	1-iodobutane	I, alkane	259	2.15	-0.26	-2.41	-3.22
106	acetonitrile	C≡N	173	1.72	-3.89	-5.61	-5.20
107	butanenitrile	C≡N	225	1.99	-3.65	-5.63	-5.79
108	nitroethane	NO ₂	199	1.86	-3.71	-5.57	-5.11
109	2-nitropropane	NO ₂	223	1.98	-3.14	-5.12	-4.90
110	nitrobenzene	NO ₂ , arom	255	2.14	-4.12	-6.26	-5.17
111	1-methyl-2-nitrobenzene	NO ₂ , arom	273	2.23	-3.59	-5.82	-5.10
112	1-methyl-3-nitrobenzene	NO ₂ , arom	281	2.27	-3.45	-5.72	-5.31
113	1,1-difluoroethane	halo-alkane	171	1.71	-0.11	-1.82	-2.46
114	1,4-dimethylpiperazine	cyclic amine	262	2.17	-7.57	-9.74	-8.84
115	1,1-dichlorobutane	halo-alkane	280	2.26	-0.70	-2.96	-2.59
116	1,3-dichlorobenzene	halo-arom	294	2.33	-0.98	-3.31	-2.30
117	1,4-dichlorobenzene	halo-arom	294	2.33	-1.01	-3.34	-2.28
118	3-hydroxybenzaldehyde	arom, HCO	261	2.17	-9.51	-11.68	-12.23

Table 4. (Continued)

code	molecule	main functional group(s) ^a	SAS (Å ²)	$\Delta G^{\text{np}}(\text{calc})$	$\Delta G^{\text{solv}}(\text{expt})$	ΔG^{elec}	
						(expt)	(calc)
119	4-hydroxybenzaldehyde	arom, HCO	262	2.17	-10.47	-12.64	-11.92
120	2-chloropyridine	halo, CH=N-CH	250	2.11	-4.39	-6.51	-6.34
121	3-chloropyridine	halo, CH=N-CH	251	2.11	-4.01	-6.13	-7.13
122	1,1'-thiobis(2-chloroethane)	halo-alkane	325	2.49	-3.92	-6.41	-4.83
123	2,2,2-trifluoroethanol	halo, OH	198	1.85	-4.30	-6.15	-5.88
124	1,1,1-trifluoropropan-2-ol	halo, OH	220	1.96	-4.16	-6.12	-6.41
125	2,2,3,3-tetrafluoropropan-1-ol	halo, OH	230	2.01	-4.88	-6.89	-7.04
126	2,2,3,3,3-pentafluoropropan-1-ol	halo, OH	241	2.06	-4.15	-6.22	-5.84
127	1,11,3,3,3-hexafluoropropan-2-ol	halo, OH	247	2.10	-3.77	-5.86	-4.71

^a Arom: aromatic, halo: halogenated.

for the current state of empirical charge sets, it is essential to have a set designed specifically for each application/method.

Although we compared with AMSOL and AM1-bcc charge sets to answer the question of whether it was necessary to parametrize a new charge set for the FDPB method, we also wanted to compare the different charge sets using the solvation methods for which they were designed. When the AMSOL charges are used with the AMSOL radii and method (GB), the results are much better, with a mean unsigned error of only 0.65 kcal/mol and a $R^2 = 0.88$, comparable to the optimized QEq charges on this test set. In making this comparison, it should be noted that the AMSOL charges were parametrized on more than 200 compounds, and so its training set undoubtedly includes some of the common organic compounds in our test set. The true free R^2 for AMSOL for our test set would be somewhat lower, and the unbiased mean unsigned error would be higher, than the values in Table 6.

For the AM1-bcc set, simulations with explicit solvent and the free energy perturbation (FEP) method were used to calculate relative solvation energies of 40 compounds.⁷ The mean unsigned error was 0.69 kcal/mol, giving somewhat worse accuracy. In this comparison it should be noted that the set of compounds used was different from those used here, so relative accuracies may vary depending on the mix of compounds used. However, due to the order of magnitude more computation required to do explicit water FEP simulations it is not practical to do the 127 test compounds used here. Also the method gives relative free energies of solvation, i.e., differences between two closely related compounds such as methane and methanol. AMSOL/GB and QEq/FDPB methods both give absolute solvation free energies and are implicit solvent models with modest computational requirements. These two are easier to compare directly on the same test set.

The test set was chosen to be large relative to the training set (a ratio of about 5:1) in order to obtain a good idea of the robustness of the QEq charge calculation method. However, the variety of compounds one has in a large test set depends on the availability of reliable experimental data. Therefore for experimental reasons certain functional groups are over-represented, such as alcohols, and some under-represented. As the test set is made larger, this imbalance tends to increase. Since any charge set/solvation energy

method is likely to perform better on some types of compounds than others, this affects the assessment of accuracy and the comparison between different sets. To counter this we pruned the test set so that each major functional group or group combination occurring in the original set of 127 was represented just twice if possible, but at least once, resulting in 52 test compounds. Table 5 lists this nonredundant test set and the calculated and experimental electrostatic solvation free energies. The non-redundant test set was used to compare the different charge sets, QEq, AMSOL, and AM1-bcc, in a more compound-unbiased way. The results are summarized in Table 6. QEq charges used with FDPB and AMSOL charges used with the GB radii and GB method have about the same accuracy, with mean unsigned errors of 0.50 and 0.57 kcal/mol, respectively, and slopes very close to unity. Again, use of the AMSOL and AM1-bcc charge sets with a method different from the one used for parametrization gives significantly poorer results.

In all these comparisons of calculated and experimental solvation energies, we put more emphasis on the mean unsigned error and the best fit slope and less on the correlation coefficient R^2 . First, the mean unsigned error gives what the user is most interested in, an estimate of the method's accuracy in application. Second, in our judgment a large deviation from unity of the best fit slope indicates a poor charge set even if the correlation coefficient is good. With a poor slope there are clearly systematic errors due to over- or under-representation of some factor. This is more than likely to give large errors of *unknown sign* since this factor is altered in an unknown way when the method is used outside the training/test set compounds. From this perspective, the QEq charge set looks robust, with a slope of almost exactly 1. The fact that the QEq method can achieve as good results as AMSOL with far fewer training compounds and parameters is encouraging in terms of the simplicity and robustness of the QEq method.

Alanine Dipeptide Conformational Energy Map Analysis. The Rappe-Goddard algorithm for determining atomic charges is both rapid and takes account of conformation. The conformation dependence comes from the pairwise atomic Slater-Coulomb term J_{AB} in eq 1, which depends inversely on the distance between the A and B atoms. Examination of eq 1 shows that if Q_B is of opposite sign to Q_A , then at constant χ_A the J_{AB} term will tend to increase the magnitude

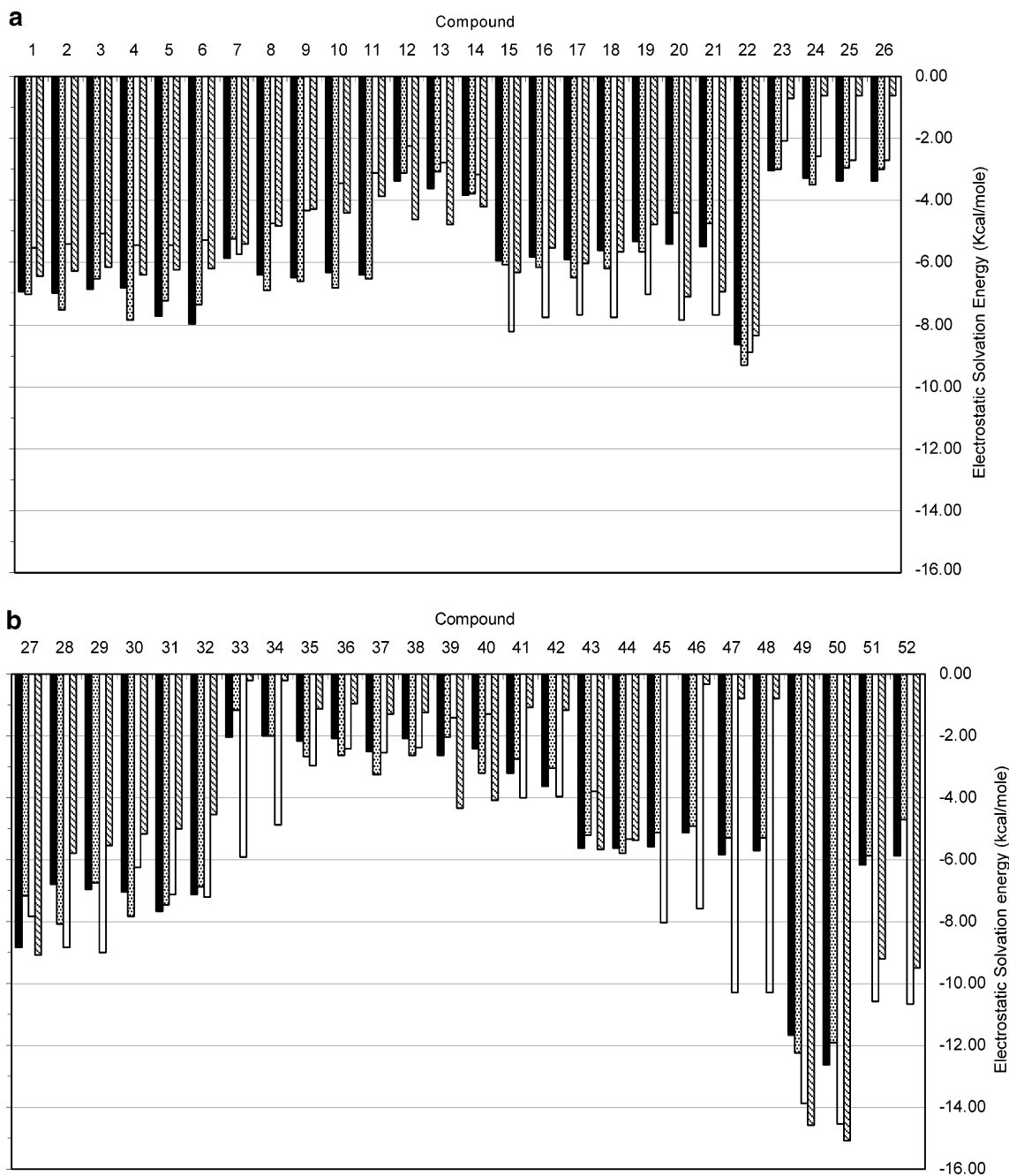


Figure 4. Comparison of experimental and calculated solvation free energies for nonredundant test set. Filled bars: experiment. Shaded bars: QEQUL charges. Unfilled bars: AMSOL charges. Striped bars: AM1-bcc charges. The FDPB method was used for all three charge sets: (a) compounds 1–26 and (b) compounds 27–52. See Table 5 for compound key.

of Q_A , while if Q_B and Q_A are of the same sign, then the J_{AB} interaction will decrease the magnitude of Q_A . Q_A will have the same effect on Q_B . In other words, favorable intramolecular Coulombic interactions will tend to produce more polarization or larger magnitude partial charges. As pointed out in the original QEq paper, this conformation dependence could be used to implement a polarizable force field in MD simulations. The resulting conformation dependent polarization will also affect the solvation in a conformation dependent way, particularly when using say a FDPB implicit solvent treatment with the MD simulations. To examine the relationship between polarizable charges, conformation, and solvation energy with the QEq charge set, a conformational study of the alanine dipeptide was performed. Alanine dipeptide

is the small compound most often used to evaluate most new molecular mechanics methods and for new solvation models designed for use in molecular mechanics of peptides and proteins.³⁴ The molecule has two planar peptide groups one at the N-terminus, the other at the C-terminus. It also has a side chain and two principle degrees of freedom, the phi and psi torsion angles which determine the relative orientations of the two peptide groups. Analysis of the effect of force field, charge, and solvent is typically displayed in terms of energy surfaces in the dipeptide's phi-psi space, i.e., as Ramachandran type plots. In this work the energies are plotted relative to the lowest point on each energy surface (which is thus set to 0 kcal/mol). This makes it easier to compare relative contributions of each factor at each phi/psi

Table 5. Nonredundant Functional Group Test Set

ID	name	functional group(s)	Gexp, ele	GEQ,ele
1	propanol	-OH linear alkane	-6.94	-7.04
2	butanol	-OH linear alkane	-6.99	-7.51
3	isopropyl alcohol	-OH branched alkane	-6.85	-6.54
4	3-methyl-1-butanol	-OH branched alkane	-6.81	-7.84
5	cyclopentanol	-OH cyclic alkane	-7.74	-7.23
6	cycloheptanol	-OH cyclic alkane	-7.97	-7.37
7	ammonia	NH ₃	-5.87	-5.23
8	methylamine	-H ₂ linear alkane	-6.38	-6.91
9	ethylamine	-H ₂ linear alkane	-6.47	-6.62
10	dimethylamine	-NH ₂ branched alkane	-6.30	-6.84
11	diethylamine	-NH ₂ branched alkane	-6.42	-6.52
12	ethylthiol	-SH	-3.36	-3.12
13	dimethyl sulfide	-S-	-3.61	-3.06
14	diethyl sulfide	-S-	-3.83	-3.79
15	acetone	>C=O in middle	-5.93	-6.07
16	3-pentanone	>C=O in middle	-5.80	-6.16
17	2-pentanone	>C=O not in middle	-5.91	-6.50
18	3-methyl-2-butanone	>C=O branched alkane	-5.63	-6.20
19	2,4-dimethyl-3-pentanone	>C=O branched alkane	-5.33	-5.66
20	acetaldehyde	-HCO	-5.39	-4.40
21	propionaldehyde	-HCO	-5.50	-4.72
22	butyric acid	-COOH	-8.64	-9.31
23	benzene	aromatic ring	-3.02	-3.01
24	ethylbenzene	alkylated aromatic	-3.29	-3.48
25	<i>m</i> -xylene	bi-alkylated aromatic	-3.35	-2.97
26	<i>p</i> -xylene	bi-alkylated aromatic	-3.35	-2.97
27	phenol	aromatic, OH	-8.84	-7.15
28	pyridine	-CH=N-CH=	-6.81	-8.10
29	2-methylpyridine	CH=N-CH= linear alkylation	-6.94	-6.75
30	3-ethylpyridine	CH=N-CH= linear alkylation	-7.06	-7.85
31	3,4-dimethylpyridine	-CH=N-CH= multi-alkylation	-7.66	-7.47
32	dimethylethylpyridine	-CH=N-CH= multi-alkylation	-7.14	-6.86
33	1-propyne	-CCH with linear chain	-2.03	-1.15
34	1-pentyne	-CCH with linear chain	-1.99	-2.01
35	1-chlorobutane	halogenated alkane, Cl	-2.17	-2.66
36	3-chloropentane	halogenated alkane, Cl	-2.09	-2.63
37	1-bromobutane	halogenated alkane, Br	-2.51	-3.26
38	1-bromo-2-methylpropane	halogenated alkane, Br	-2.10	-2.63
39	iodoethane	halogenated alkane, I	-2.62	-2.06
40	1-iodobutane	halogenated alkane, I	-2.41	-3.22
41	chlorobenzene	halogenated, aromatic	-3.20	-2.74
42	bromobenzene	halogenated, aromatic	-3.61	-3.04
43	acetonitrile	-C≡N	-5.61	-5.20
44	butanenitrile	-C≡N	-5.63	-5.79
45	nitroethane	-NO ₂	-5.57	-5.11
46	2-nitropropane	-NO ₂	-5.12	-4.90
47	1-methyl-2-nitrobenzene	-NO ₂ , aromatic	-5.82	-5.31
48	1-methyl-3-nitrobenzene	-NO ₂ , aromatic	-5.72	-5.10
49	3-hydroxybenzaldehyde	aromatic -CHO	-11.68	-12.23
50	4-hydroxybenzaldehyde	aromatic -CHO	-12.64	-11.92
51	2,2,2-trifluoroethanol		-6.15	-5.88
52	1,11,3,3,3-hexafluoropropan-2-ol	fluorinated alcohol	-5.86	-4.71

conformation. Figure 5 shows such a plot for the nonelectrostatic contribution to the dipeptide energy using the CHARMM bonded and van der Waals parameters. Approximate centers of commonly defined conformation regions are indicated in the figure. The resulting energy surface recapitulates the original Ramachandran analysis based on

steric considerations, with low-energy regions for negative values of phi that span the beta and alpha conformations. Figure 6a shows the electrostatic contribution from fixed CHARMM22 charges. Adding this contribution to the nonelectrostatic contribution yields the in vacuo energy surface, Figure 6b. Figure 6c shows the solvation free energy

Table 6. Summary of Comparisons of Experimental vs Calculated Free Energies

molecule set (no. of compounds) ^a	charge set	method ^b	R^2	slope	mean unsigned error (kcal/mol)
training (23)	QEq (unoptimized)	FDPB	0.01	0.08	3.52
training(15)	CHARMM	FDPB	0.85	0.87	2.1
training (23)	QEq (optimized)	FDPB	0.96	0.99	0.50
test (127)	QEq (optimized)	FDPB	0.90	0.98	0.61
test (127)	AMSOL	FDPB	0.48	0.83	1.60
test (127)	AM1-bcc	FDPB	0.60	1.02	1.77
test (127)	AMSOL	AMSOL/GB	0.88	1.08	0.65
test (40)	AM1-bcc ^c	FEP	n/a	n/a	0.69
nonredundant(52)	Qeq (optimized)	FDPB	0.93	0.98	0.50
nonredundant(52)	AMSOL	FDPB	0.54	0.98	1.71
nonredundant(52)	AM1-bcc	FDPB	0.65	1.17	1.75
nonredundant(52)	AMSOL	AMSOL/GB	0.91	1.10	0.67

^a Training, test, and nonredundant test molecules are listed in Tables 3–5, respectively. ^b FDPB: finite difference Poisson–Boltzmann, AMSOL/GB: generalized Born with AMSOL parametrization, FEP: free energy perturbation using explicit solvent. ^c Taken from the original AM1-bcc parametrization paper.⁷ Mean unsigned error for relative (differences in) solvation free energies for 40 compound pairs listed therein.

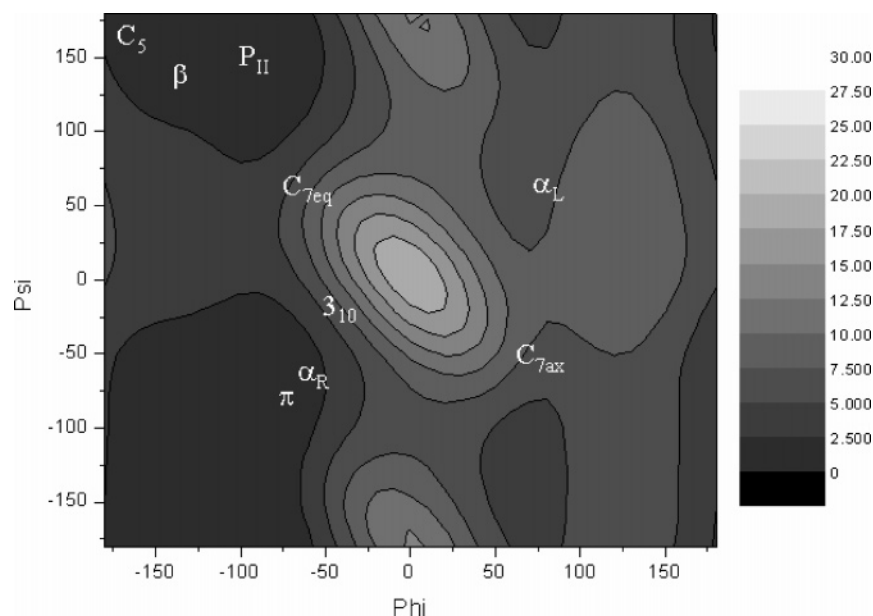


Figure 5. Ramachandran (ϕ , ψ) plot for the alanine dipeptide. Gray scale indicates the nonelectrostatic energy (sum of bond stretch, angle, torsion, and van der Waals energy terms) using the CHARMM force field. Energies are relative to the lowest point on the map (by definition at 0 kcal/mol). Approximate locations of the center of commonly referred to conformations are labeled on the map.

of the dipeptide with these fixed charges calculated using the FDPB solvation model, which, added to the in vacuo energy, yields the total energy in water. The total energy surface is shown in Figure 6d. The resulting low-energy regions are broadly the same as with just the nonelectrostatic terms, but the allowed regions are somewhat more tightly constrained around the canonical alpha and beta regions.

The corresponding energy surfaces for electrostatic and solvation energy terms using the polarizable QEq charges are shown in Figure 7a–d. The nonelectrostatic contribution is the same as with the fixed CHARMM charges (Figure 5). Summing the nonelectrostatic and electrostatic terms gives the in vacuo energy for the QEq charge model (Figure 7b), and adding in the solvation terms gives the total energy surface for the QEq charges (Figure 7d).

Considering first the internal electrostatic energy term, for the fixed CHARMM charges the energy surface shows a

broad trough lying along the $\phi + \psi = 0^\circ$ line. Analyzing this in terms of the N-terminal, or ϕ -angle dependent peptide group, and the C-terminal, or ψ -angle dependent peptide group (each composed of their O \rightarrow C and N \rightarrow H dipoles), the trough roughly corresponds to configurations with antiparallel peptide group alignments or, in electrostatic terms, antiparallel alignments of the dipoles associated with these two groups. This results in a favorable interaction between the peptide groups. This trough includes the internally H-bonded C7ax (62, -65) and C7eq (-65, 69) configurations. The solvation map is roughly the inverse of the internal electrostatic map, with a peak along the $\phi + \psi = 0^\circ$ line, but overall the solvating map is flatter than the internal electrostatic map. Interpreting the map again in terms of two separate dipoles arising from the two peptide groups, the trough along $\phi + \psi = -180^\circ$ corresponds to configurations with parallel alignments of the dipoles. The

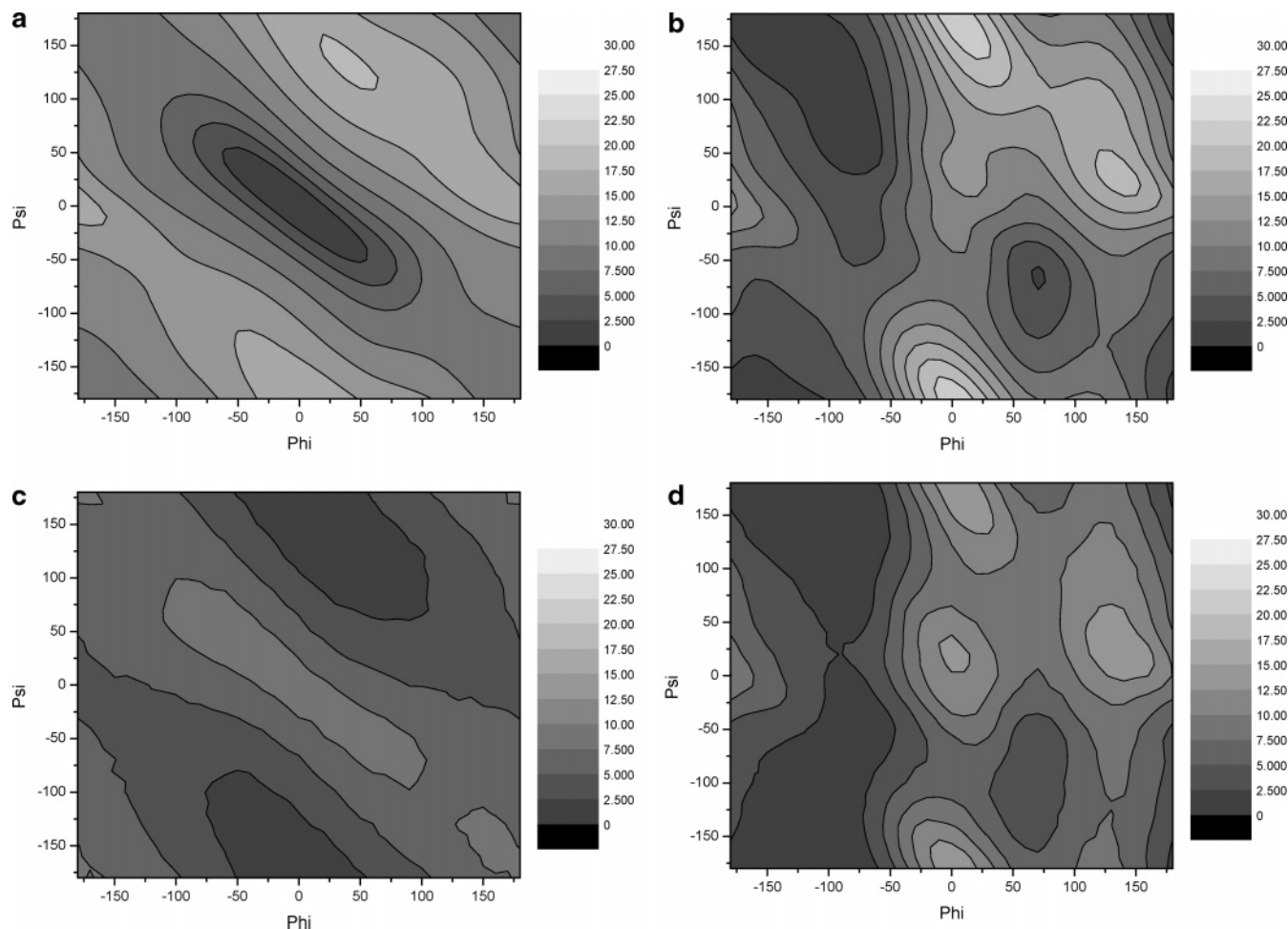


Figure 6. Ramachandran (ϕ , ψ) plot for the alanine dipeptide. Gray scale indicates (a) electrostatic part of internal energy using CHARMM charges, (b) total in vacuo energy (sum of electrostatic and nonelectrostatic energies), (c) solvation energy using the CHARMM charges and the FDPB method, and (d) total energy (sum of nonelectrostatic, electrostatic, and solvation terms). Energies are relative to the lowest point on the map.

two peptide dipoles tend to add with respect to the overall molecular dipole, producing the largest solvent reaction field and hence the most negative solvation free energy. Thus for fixed charges the solvation energy term cancels a lot of the internal electrostatics, giving a total energy surface flatter than, but similar to, the nonelectrostatic internal energy map.

The internal electrostatic energy term in the polarizable QEq model is similar to that of the fixed charged model, but the trough of favorable energy is shorter, less deep, and centered at approximately $(-60, 20)$ rather than at $(0, 0)$. This shrinking and shifting of the minimum region is understandable in terms of the conformation dependent charge polarization, principally the changing magnitude of the phi and psi peptide dipoles (Figure 8 (parts a and b, respectively)). The region where the sum of phi and psi dipole magnitudes is largest lies in the same region as the minimum in the internal electrostatic energy. The shorter trough in the electrostatic energy surface produces an in vacuo energy surface that, while very similar to the CHARMM in vacuo energy surface for $\phi < -50^\circ$, is higher (more unfavorable) in the other regions ($\phi > -50^\circ$).

The polarizable charges produce a qualitatively different solvation map from the fixed charges (Figure 7c). The map is much flatter. Note that the energy scale in this figure is a

factor of 5 smaller. Again this can be interpreted in terms of the magnitude and alignment of the two principle dipolar groups. Conformations with a parallel dipolar alignment have less favorable Coulombic interaction between the two peptide groups and thus significantly less polarized charges. This results in smaller magnitude dipoles, principally not only for the psi dependent group (Figure 8b) but also for the phi dependent group (Figure 8a). This dipole magnitude effect runs in the opposite direction to the dipole orientation effect, the magnitudes being smaller when the dipoles are parallel, and larger when they are 'antiparallel', so the effects on the solvation reaction field tend to cancel, leading to a much flatter map.

Neither the fixed charge solvation energy map nor the polarizable charge solvation energy map can be adequately explained in terms of the total dipole moment of the dipeptide (results not shown), but only in terms of the individual peptide dipole moments. This indicates that representing the molecular charges as a single dipole is a bad approximation for describing solvation when the separation of charged groups is significant compared to the size of cavity formed in solution, as it is in the alanine dipeptide. One must account for the higher order poles, e.g. quadrupoles, etc., or at least describe the distribution as two separate dipole centers.

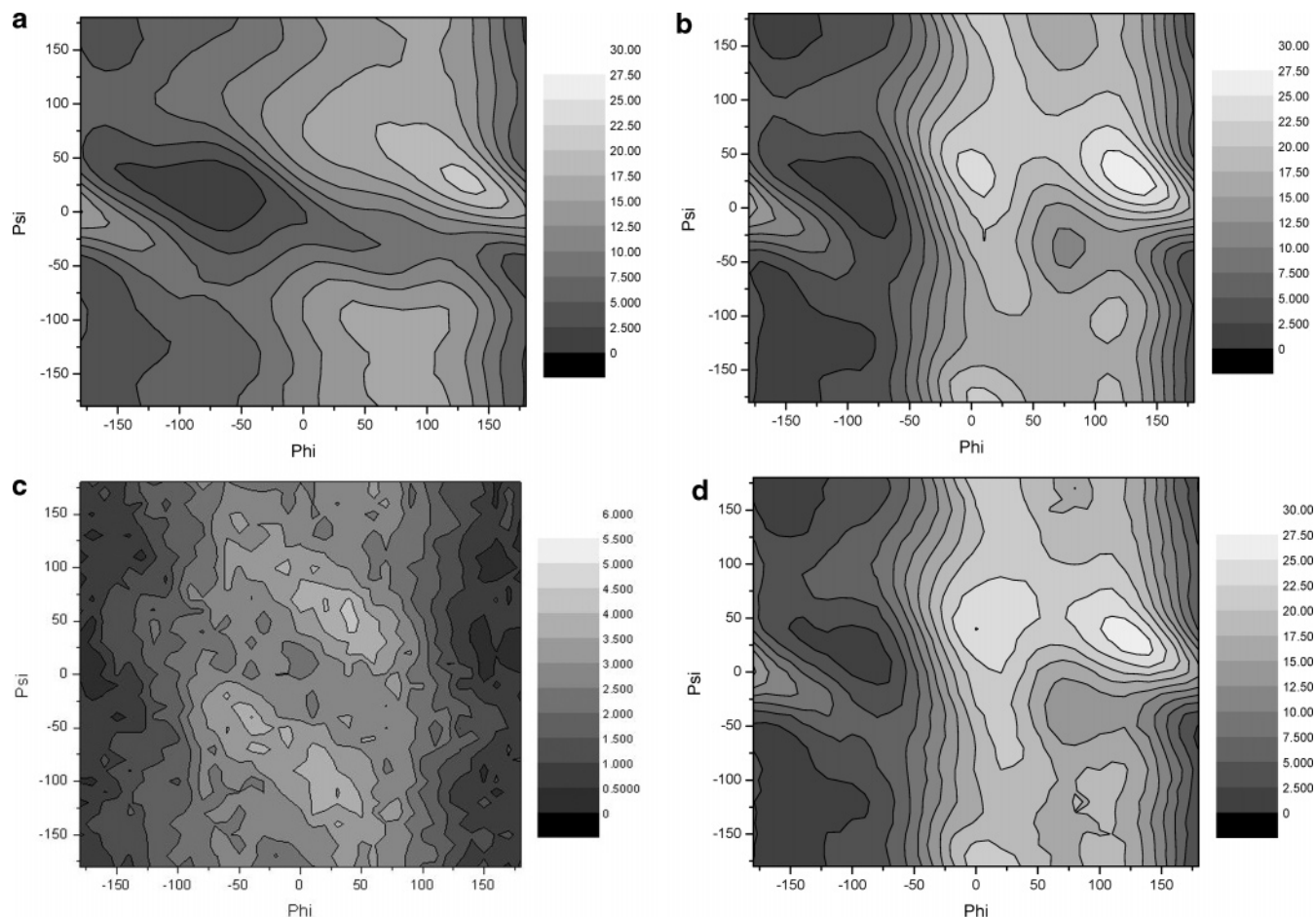


Figure 7. Ramachandran (ϕ , ψ) plot for the alanine dipeptide. Gray scale indicates (a) electrostatic part of internal energy using conformation-dependent QEq charges determined by QEQUIL, (b) total in vacuo energy (sum of electrostatic and nonelectrostatic energies), (c) solvation energy using the QEq charges and the FDPB method, and (d) total energy (sum of nonelectrostatic, electrostatic, and solvation terms). Energies are relative to the lowest point on the map

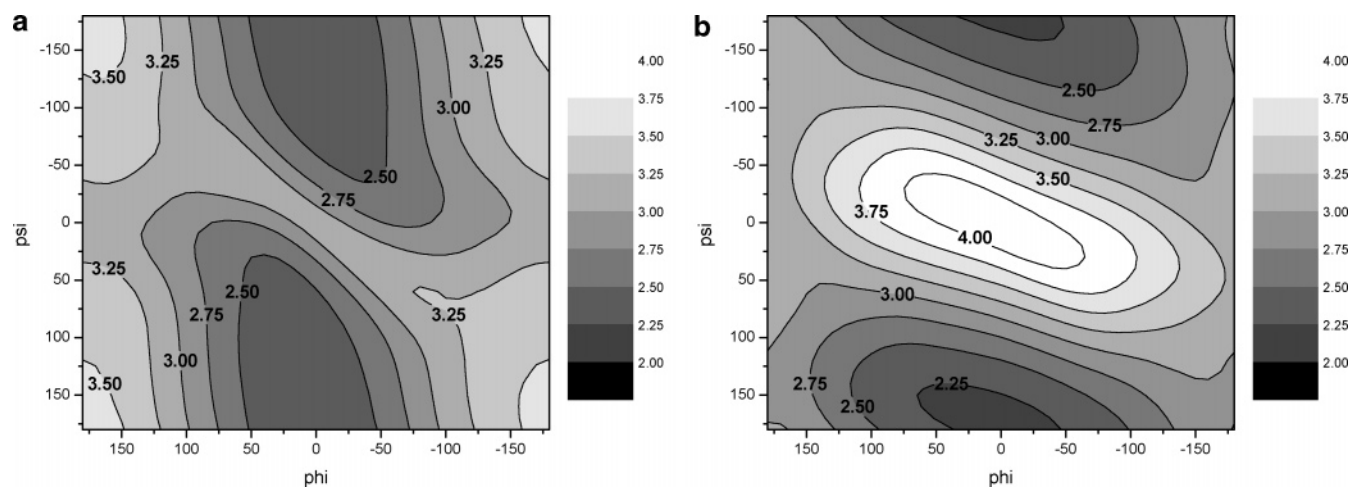


Figure 8. Ramachandran (ϕ , ψ) plot for the alanine dipeptide. Gray scale indicates the magnitude of the peptide moiety dipole moment in $e\text{\AA}$, using conformation-dependent QEq charges: (a) N-terminal (ϕ) peptide and (b) C-terminal (ψ) peptide.

Comparing the total energy surface for the two charge models (Figures 6d and 7d) the QEq model surface has a larger range of energies and a smaller favorable region in the vicinity of the two main secondary structure types, beta sheet (β) and alpha helical (α_R), particularly around the latter. In fact the α_R region lies more on a low saddle in the energy surface than a minimum. The less favorable energy in the

alpha region obtained with the polarizable model is larger attributable to the shorter trough in the corresponding internal electrostatic energy surface with QEq charges (Figure 7a vs Figure 6a). The larger difference between high and low regions with the QEq model principally results from higher energies in the region with $\phi > -50^\circ$. Again, this is attributable to the shorter trough in the corresponding internal

electrostatic energy surface compared to the fixed charge model. Since the $\phi > -50^\circ$ tends to be less populated in folded proteins and in MD simulations (with the exception of Pro and Gly residues) the large difference in energy surfaces in this region between the two models is unlikely to result in as large a difference in conformational preferences. Conformational preference differences between the two models are more sensitive to the more populated beta and alpha regions, and here the differences between the two charge models are much smaller. Whether the QEq charge mode provides an improved description of the peptide backbone energy surface in proteins cannot be determined from examination of the energy surface in ϕ - ψ space alone but must be examined using actual molecular dynamics simulations with say an integrated FDPB solvent treatment. The development of the rapid and solvation-optimized method for charge determination described here should make it easier to perform such simulations in the future. We note, however, that to do accurate molecular dynamics such a charge set would have to be implemented in the MD code in a fully consistent manner including polarization forces for both solvation and internal interactions, a significant undertaking given the complexity of current MD packages.

Conclusions

Partial charges sets are required for most atomic level simulations and calculations. Unfortunately, with the current state of technology no one charge set is adequate for all applications and simulation methods. In this study we describe an optimization and extension of Rappe and Goddard's charge equilibration through electronegativity neutralization method for calculating atomic charges (QEq). The optimized charges are designed to be used specifically with the FDPB method for calculating solvation, and they are in effect a successor to the PARSE charge set. The latter, being effectively parametrized by hand, cannot easily be extended to new functional groups. The method is designed to use a small set of element data in order to handle a wide range of chemical functional groups. The method was optimized using 23 compounds of which 15 were amino acid side-chain and backbone analogues in order to maximize accuracy when applied to peptides and proteins. The method uses 16 adjustable parameters. The mean unsigned error compared to experiment was 0.50 kcal/mol. Testing the method on a larger set of compounds (127) gave a somewhat larger unsigned error of 0.61 kcal/mol and, just as importantly, a best fit slope vs experimental data of close to unity. The fact that we use a large ratio of test to training compounds and that we get a good slope indicates that the method is quite accurate and robust. As with any parametrized charge set, some functional groups are treated better than others. Some functional groups systematically contribute more to the mean error than others, for example ketones, whose magnitudes are systematically overestimated. Pyridines as a class also contribute significantly to the mean error, but are both over- and underestimated, and so contribute little to the overall (functional group independent) systematic error. Systematic error due to particular functional groups can be exacerbated or hidden depending on the

composition of the test set. Using a nonredundant functional group set of compounds, the mean unsigned error for the QEq charges was reduced from 0.61 kcal/mol to 0.50 kcal/mol. This significantly smaller value for the mean unsigned error indicates that larger test sets, with over-representation of the more common functional groups, may not be better tests of charge sets in general. Since the slopes for test sets using the QEq charges are very close to one, there is little systematic error averaged over the range of functional groups examined here, i.e., there is no systematic under- or overestimation of solvation energy due to some factor common to different functional groups.

Since the QEq program calculates charges in a rapid and conformation dependent manner, we performed a preliminary investigation of the method's potential for implementing a polarizable charge set in MD simulations. We examined the ϕ - ψ energy surface for the alanine dipeptide using the QEq program, calculated the energy surface for different electrostatic components, and compared them to a standard fixed charge set used for MD (CHARMM). The polarizable charges produce a much less conformation dependent solvation reaction field than the fixed charges due to compensating effects from charge polarization. Overall the ϕ - ψ energy surface with polarizable charges has a larger energy difference between high and low regions and somewhat smaller allowed regions around the beta and alpha regions, especially the latter. Testing whether the polarizable charge model can provide an improved description of the peptide energy surface in molecular dynamics simulations of proteins is an obvious future direction for this work. Regardless of its potential for a polarizable force field, the QEq charge determination as implemented in the QEq program does provide an accurate way of generating charges for a range of functional groups, specifically for use with the widely used finite difference Poisson-Boltzmann method of calculating of solvation energies.

Acknowledgment. This work was supported by Grant MCB02-35440 from the NSF. We thank Ninad Prabhu, Ryan Coleman, and John Skinner for helpful discussions.

References

- (1) Cornell, W.; Cieplak, P.; Bayly, C.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620–9631.
- (2) Sitkoff, D.; Sharp, K.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.
- (3) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.
- (4) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 1820–1831.
- (5) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.
- (6) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **1998**, *102*, 3257–3271.
- (7) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.
- (8) Gasteiger, J.; Marsili, M. *Tetrahedron Lett.* **1978**, *34*, 3181.
- (9) Gasteiger, J.; Marsili, M. *Tetrahedron Lett.* **1980**, *36*, 3219–3221.

- (10) Tai, K.; Grant, J. A.; Scheraga, H. A. *J. Phys. Chem.* **1990**, *94*, 4732–4739.
- (11) Rappe, A.; Goddard, W. A. *J. Phys. Chem.* **1991**, *95*, 3358–3363.
- (12) Bultinck, P.; Lahorte, P.; De Proft, F.; Geerlings, P.; Waroquier, M.; Tollenaere, J. P. *J. Phys. Chem.* **2002**, *106*, 7887–7894.
- (13) Gilson, M. K.; Gilson, H. S. R.; Potter, M. J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1982–1997.
- (14) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141–6151.
- (15) Patel, S.; Brooks, C. L., III *J. Comput. Chem.* **2004**, *25*, 1–15.
- (16) Patel, S.; Alexander, D.; Mackerell, J.; Brooks, C. L., III *J. Comput. Chem.* **2004**, *25*, 1504–1514.
- (17) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (18) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
- (19) Press, W.; Flannery, B.; Teukolsky, S.; Vetterling, W. *Numerical Recipes*; Cambridge University Press: Cambridge, 1986.
- (20) Gilson, M.; Sharp, K. A.; Honig, B. *J. Comput. Chem.* **1988**, *9*, 327–335.
- (21) Nicholls, A.; Sharp, K. A.; Honig, H. *Proteins* **1991**, *11*, 281–296.
- (22) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (23) Sridharan, S.; Nicholls, A.; Sharp, K. A. *J. Comput. Chem.* **1995**, *16*, 1038–1044.
- (24) Case, D.; Pearlman, D.; Caldwell, J. W.; Cheatham, T.; Wang, J.; Ross, C.; Simmerling, T.; Darden, T.; Merz, K.; Stanton, A.; Chenn, J.; Vincent, M.; Crowley, V.; Crowley, V.; Tsui, V.; Gohlke, H.; Radmer, R.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G.; Singh, U.; Weiner, P.; Kollman, P. A. *AMBER 7*; UCSF, 2002.
- (25) Hawkins, G. D.; Giesen, D.; Lynch, G.; Chambers, C.; Rossi, I.; Storer, J.; Li, J.; Zhu, T.; Thompson, J.; Winget, P.; Lynch, B.; Rinaldi, D.; Liotard, D.; Cramer, C. J.; Truhlar, D. G. *Amsol v. 7.1*; Regents of the University of Minnesota, 2004.
- (26) Radzicka, A.; Wolfenden, R. *Biochemistry* **1988**, *27*, 1664–1670.
- (27) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Soln. Chem.* **1981**, *10*, 563–595.
- (28) Moran, A.; Mukamel, S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 506–510.
- (29) MacKerell, A. D.; Brooks, B.; Brooks, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; John Wiley & Sons: Chichester, 1998; Vol. 1, pp 271–277.
- (30) Schiffer, C.; Caldwell, J.; Kollman, P.; Stroud, R. *Mol. Simul.* **1993**, *10*, 121.
- (31) Lu, Q.; Luo, R. *J. Chem. Phys.* **2003**, *119*, 11035–11047.
- (32) Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23*, 1244–1253.
- (33) Prabhu, N. V.; Zhu, P.-J.; Sharp, K. A. *J. Comput. Chem.* **2004**, *25*, 2049–2064.
- (34) Prabhu, N.; Sharp, K. *Chem. Rev.* **2005**, *106*, 1616–1623.

CT060009C

JCTC

Journal of Chemical Theory and Computation

Molecular Modeling the Reaction Mechanism of Serine-Carboxyl Peptidases

Ksenia Bravaya,[†] Anastasia Bochenkova,[†] Bella Grigorenko,[†] Igor Topol,[‡] Stanley Burt,[‡] and Alexander Nemukhin^{*,†,§}

Department of Chemistry, M. V. Lomonosov Moscow State University, Moscow 119992, Russian Federation, Advanced Biomedical Computing Center, National Cancer Institute at Frederick, Frederick, Maryland 21702, and Institute of Biochemical Physics, Russian Academy of Sciences, Moscow 119997, Russian Federation

Received February 17, 2006

Abstract: We performed molecular modeling on the mechanism of serine-carboxyl peptidases, a novel class of enzymes active at acidic pH and distinguished by the conserved triad of amino acid residues Ser-Glu-Asp. Catalytic cleavage of a hexapeptide fragment of the oxidized B-chain of insulin by the *Pseudomonas* sedolisin, a member of the serine-carboxyl peptidases family, was simulated. Following motifs of the crystal structure of the sedolisin-inhibitor complex (PDB accession code 1NLU) we designed the model enzyme–substrate (ES) complex and performed quantum mechanical–molecular mechanical calculations of the energy profile along a reaction route up to the acylenzyme (EA) complex through the tetrahedral intermediate (TI). The energies and forces were computed by using the PBE0 exchange–correlation functional and the basis set 6-31+G** in the quantum part and the AMBER force field parameters in the molecular mechanical part. Analysis of the ES, TI, and AE structures as well as of the corresponding transition states allows us to scrutinize the chemical transformations catalyzed by sedolisin. According to the results of simulations, the reaction mechanism of serine-carboxyl peptidases should be viewed as a special case of carboxyl (aspartic) proteases, with the nucleophilic water molecule being replaced by the Ser residue. The catalytic triad Ser-Glu-Asp in sedolisin functions differently compared to the well-known triad Ser-His-Asp of serine proteases, despite the structural similarity of sedolisin and the serine proteases member, subtilisin.

Introduction

Serine-carboxyl peptidases or sedolisins^{1–9} are presently assigned to the family S53 of clan SB of serine proteinases (the MEROPS database, URL <http://merops.sanger.ac.uk>). Following the results of intense recent studies of sedolisins primarily by the methods of X-ray spectroscopy, it has been established that there is overall similarity of two protease families, sedolisins and serine protease type subtilisins:

practically all secondary structure elements found in the smaller subtilisins also are present in sedolisins. Although both subtilisins and sedolisins utilize the serine residue as the principal nucleophile, other members of the catalytic triad are different. The second member of the triad Ser-His-Asp in subtilisin, histidine, is substituted in sedolisin (with the triad Ser-Glu-Asp) by a topologically equivalent glutamic acid, while the third residue of the triad, aspartic acid in both enzymes, is contributed by topologically different parts of the structure.

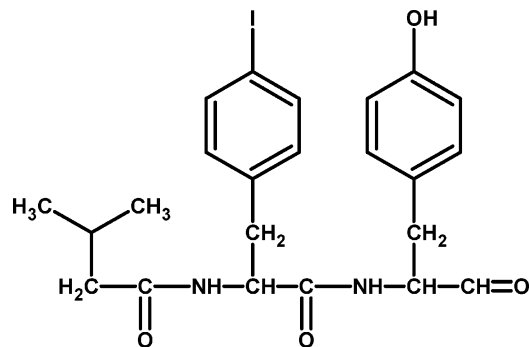
A fundamental question for serine-carboxyl peptidases is whether these enzymes with the triad Ser-Glu-Asp use the catalytic mechanism similar to that of serine proteases¹⁰ with

* Corresponding author e-mail: anem@lcc.chem.msu.ru.

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

[‡] National Cancer Institute at Frederick.

[§] Russian Academy of Sciences.

Chart 1. Chemical Formula of Pseudoiodotyrostatin Used To Generate a Complex 1NLU with Sedolisin⁹

the triad Ser-His-Asp. A practical reason for interest in these enzymes is explained by their potential use in medicine due to activity of the family's member kumamolisin-As as a collagen-degradating agent^{6,9} and in industry because of maximum activity at the comparatively low pH of 3–5 and stability of some of these enzymes at high temperatures up to 60 °C.

At present the natural substrates for serine-carboxyl peptidases are unknown, and the knowledge about their structure and function are learned from the experimental studies of the enzymes interacting with different inhibitors. All model inhibitors used in the X-ray studies were peptides with the terminal aldehyde group. In many of these complexes an inhibitor was covalently bound to the enzyme by the hemiacetal bond with the OH group of serine. A 1.3 Å resolution of the structure of the *Pseudomonas* sp. 101 sedolisin (PDB accession code 1NLU),^{7,8} in which sedolisin was complexed with two molecules of the inhibitor, pseudoiodotyrostatin (Chart 1), was an important contribution to the field. It was concluded that this structure, in which only one molecule of the inhibitor was covalently bound to Ser, could be viewed as representing the product of enzymatic cleavage of a hexapeptide substrate: the hemiacetal involving Ser287 would represent the complex (or a tetrahedral adduct) formed following nucleophilic attack of the Ser oxygen, while another molecule of pseudoiodotyrostatin would represent the amino product from cleavage.⁹

In this paper we describe the first theoretical simulations of the reaction energy profile and the analysis of the reaction intermediates for the serine-carboxyl peptidases taking the catalytic cleavage of peptide bonds by sedolisin as an important example. Prompted by the crystal structure 1NLU,⁸ we designed the model enzyme–substrate (ES) complex and performed calculations of the energy along a reaction route up to the acylenzyme (EA) complex through the tetrahedral intermediate (TI). The model structure corresponding to TI may be directly compared to the experimental moiety 1NLU assigned to “the product complex following enzymatic cleavage of a hexapeptide substrate” or a mimic of the tetrahedral adduct.⁷ Therefore, in simulations we constructed the entire reaction profile inspired by the single hint from the experimental studies.

Late in the course of our work, the paper of Guo et al.¹¹ was published in which the results of quantum mechanical/molecular mechanical and molecular dynamics simulations

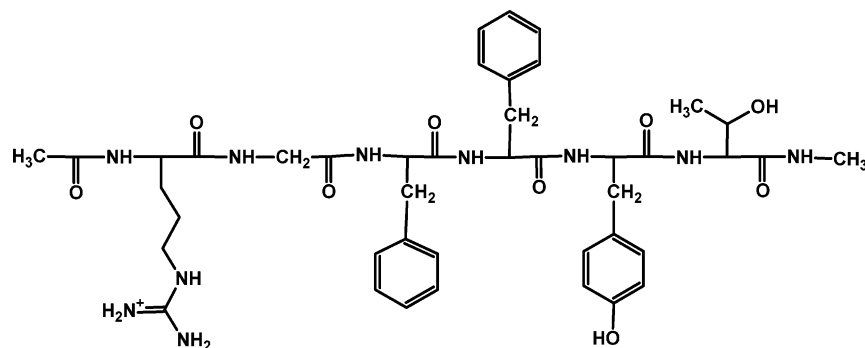
for the properties of possible tetrahedral adduct in serine-carboxyl peptidases were described. The authors considered the X-ray structure of kumamolisin-As with the inhibitor *N*-acetyl-isoleucyl-prolyl-phenylalanine covalently bound to the Ser residue (PDB accession code 1SIO) as a mimic of such adduct. Following the estimates of positions of protons at the nearby residues, Guo et al. concluded that the mode of stabilization of a hypothetical tetrahedral intermediate in serine-carboxyl peptidases may be different compared to that of serine proteases.¹¹ However, a very limited segment of a possible reaction coordinate was explored in those calculations to confirm the hypothesis on the reaction mechanism.

At preliminary steps of the modeling procedure we applied conventional molecular docking and molecular dynamics simulations, first of all, to determine a reasonable starting position of a model substrate trapped by the enzyme. After that, the reaction energy profile was constructed by using the combined quantum mechanical-molecular mechanical (QM/MM) method,^{12–19} which is becoming an important modern tool in studies of enzymatic mechanisms.^{20–33} Comprehensive analysis of chemical transformations for the serine protease prototype reactions by the results of previous QM/MM calculations^{29–33} provides a suitable basis for comparison.

Theoretical Approaches

Molecular docking calculations were carried out with the Autodock 3.0 program.³⁴ Molecular dynamics trajectories were computed with the parallel version of the NAMD 2.5 program suite.³⁵ The most demanding calculations at the combined quantum mechanical–molecular mechanical (QM/MM) level were performed by using the Intel-specific version of the GAMESS(US) program system,³⁶ PC GAMESS (Granovsky, A. A. URL <http://lcc.chem.msu.ru/gran/games>), specially adjusted for QM/MM calculations. In this program, the mechanical embedding QM/MM technique by Bakowies and Thiel¹⁶ as implemented by Kress and Granovsky was used. The conventional link hydrogen atom approach was applied to interface the QM and MM regions. The energy diagram for the reaction path from the enzyme–substrate complex (ES) to the acylenzyme complex (EA) through the tetrahedral intermediate (TI) was calculated in series of unconstrained and constrained energy minimizations in the QM/MM approximation. Electron polarization of the QM electron density by the protein environment was taken into account by the electronic embedding based QM/MM¹⁶ energy calculations at all stationary points.

As mentioned above, the structures that directly mimic enzyme–substrate complexes are not available from the X-ray studies of sedolisin with appropriate inhibitors. Therefore, selection of a model substrate and construction of the starting configuration of reagents was one of the most difficult tasks of our simulations. By assuming that a model substrate should resemble the covalently bound inhibitor molecule in the 1NLU complex, we choose the hexapeptide fragment corresponding to the residues 22–27 from the insulin oxidized B-chain (PDB accession code 1AIO) known to be hydrolyzed by sedolisin. According to the MEROPS database (URL <http://merops.sanger.ac.uk>) the cleavage of

Chart 2. Chemical Formula of the Model Hexapeptide Substrate for Sedolisin: Arg-Gly-Phe-Phe-Tyr-Thr

Arg-Gly-Phe-Phe+Tyr-Thr takes place between the Phe and Tyr residues. Therefore, the Phe-Phe fragment of a model substrate (Chart 2) may mimic the Tyr-iodo-Phe moiety (Chart 1) of the covalently bound inhibitor in S1–S2 substrate binding pockets in 1NLU. The concluding residues of the hexapeptide, Arg and Thr, were terminated by the CH₃ groups.

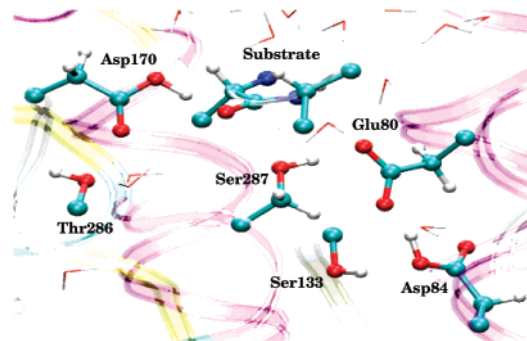
A starting configuration of the ES complex for calculations of the energy profile was created by using molecular docking, molecular dynamics, and the QM/MM approaches. To this goal we deleted the inhibitor molecules from the 1NLU structure and replaced it by the model hexapeptide Arg-Gly-Phe-Phe-Tyr-Thr.

The Lamarckian genetic algorithm was applied for the flexible docking procedure with the AutoDock program package.^{34,37,38} The grid maps of the dimension 80 × 80 × 80 were centered on the sedolisin active site. The grid spacing was 0.375 Å, allowing the ligand to explore configuration space within approximately 30 Å from the substrate-binding site. The parameters recommended for the blind docking of flexible ligands by Heteny and Spoel³⁹ were used. The initial population size used for genetic algorithm was 250, the number of energy evaluation was 10⁷, and the maximum number of generations was 1 × 10⁶. The default values of other parameters³⁴ were used in 100 docking searches. To facilitate the search of suitable positions of substrate, the side chain of Trp136 in the 1NLU structure was slightly shifted from its position in the crystal.

The molecular dynamics simulations (MD) for the enzyme–substrate complex were carried out with the CHARMM force field parameters within the NAMD 2.5 computer package.³⁵ The protein–substrate complex was buried inside a large cluster of water molecules. The MD trajectories of 80 ps length with a time step of 1 fs initiated from different starting geometry configurations were recorded at 300 K. The protein atoms lying farther than 5 Å from the active site residues were frozen during equilibration. Fifteen of the lowest energy structures obtained in the molecular docking procedure were used as starting geometry configurations. The structures mainly differed in the position of the Tyr26 and Thr27 side chains which occupied two different binding pockets. MD simulation starting from these configurations resulted in two principally different ES geometries, and the one with the lowest energy was used as a starting point for the following computations.

The resulting structure was analyzed in order to locate possible cavities inside the protein near the reaction center to be filled by the water molecules. The surface area of the enzyme which included the substrate binding site was also extensively solvated. The molecular system was thermally equilibrated at 300 K and gradually relaxed to 0 K. The atomic coordinates at the end of trajectories were considered as initial guesses for the subsequent QM/MM geometry optimization for the enzyme–substrate complex.

For QM/MM computations we selected ~2500 atoms comprising complete coverage of the active site approximately within 10 Å from the atoms of the catalytic residues. Forty-eight atoms of the active site were assigned to the QM-part as illustrated in Figure 1. The energies and forces were

**Figure 1.** The ball-and-stick representation of the groups included to the QM part of the reacting system.

computed by using the PBE0 exchange-correlation functional and the basis set 6-31+G** in the quantum part and the AMBER force field parameters in the molecular mechanical part.

Partial Hessian analysis was performed at all stationary points located on the potential energy surface. According to this procedure the evaluation of the whole QM/MM force constant matrix was carried out followed by diagonalization of its smaller part corresponding to the QM subsystem.

Results

The structure of enzyme–substrate complex obtained as a minimum energy configuration in the unconstrained QM/MM optimization of geometry parameters is shown in Figure 2.

We note the positions of protons along hydrogen bonds for the most important pairs: (i) Glu80 and Asp84 share a proton which resides on Asp at a very short distance (1.4

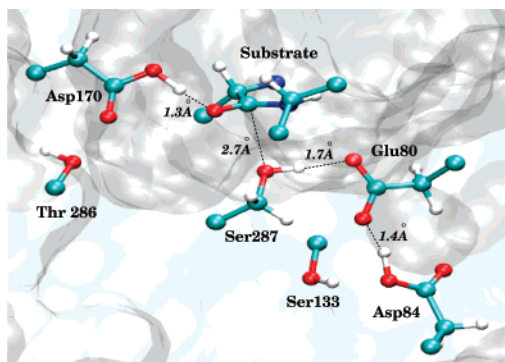


Figure 2. Geometry configuration of the enzyme–substrate complex.

Å) to Glu; (ii) the proton of Ser287 is in perfect position for the move to Glu80; and (iii) the proton of Asp170 is very close (1.3 Å) to the carbonyl oxygen of substrate. We also note a reasonable distance (2.7 Å) from oxygen of Ser287 to carbon of substrate which causes us to consider further developments along the lines typical for serine proteases. Thus, a gradual decrease of this coordinate, the O(Ser)–C(Sub) distance, should lead to the tetrahedral intermediate (TI).³³

Therefore, we selected the distance (R_{OC}) between the oxygen of Ser287 and the carbon of the substrate as a reaction coordinate for the first stage of the reaction. In a series of constrained minimizations (by keeping fixed values of R_{OC} and optimizing other internal coordinates) we succeeded in locating the equilibrium geometry configuration of the saddle point, or the first transition state TS1, with $R_{OC} = 1.7$ Å. The search of the stationary point corresponding to the tetrahedral intermediate was accomplished as an unconstrained minimization starting from some shorter than 1.7 Å values of R_{OC} . Remarkably, the computed minimum energy path specifies the first proton transfer from Asp170 to the carbonyl oxygen of the substrate occurring at the early values of reaction coordinate. The normal mode of the single imaginary frequency of the first saddle point corresponds to the concerted Asp170 proton transfer and C–O bond formation. Unlike the case of serine proteases, the principal nucleophile, Ser287, donates the proton to its partner in the catalytic triad Glu80 only near the TI configuration. This means that Asp170 plays a crucial role in the reaction mechanism of sedolisin catalysis as an acid activator for the carbonyl group of the substrate. Protonation of the carbonyl oxygen makes this carbonyl group more electrophilic for a subsequent nucleophilic attack by Ser287. In serine proteases, the proton transfer within the catalytic triad from Ser to His aims to activate a nucleophile (Ser) and initialize the chain of transformations, while the role of the oxyanion hole residue(s), analogous to Asp170 in sedolisin, is basically to compensate the developing negative charge on carbonyl oxygen.

The computed structure of the tetrahedral intermediate with the R_{OC} value of 1.5 Å is shown in Figure 3. We note that the third member of the catalytic triad, Asp84, remains protonated at this stage of the reaction.

The energy profile for the route from TI to acylenzyme (EA) was computed by another choice of a reaction

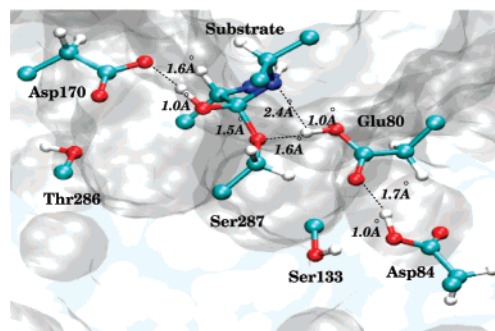


Figure 3. Geometry configuration of the tetrahedral intermediate.

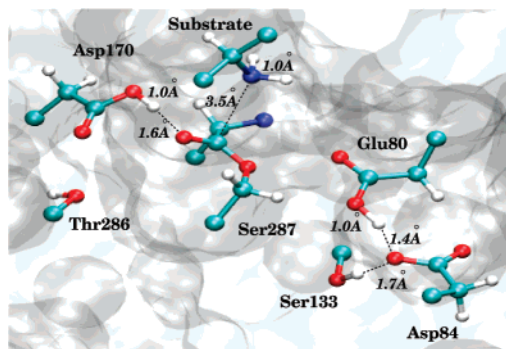


Figure 4. Geometry configuration of the acylenzyme complex.

coordinate. At this stage, the proton initially donated by Ser287 to Glu80 should be transferred to nitrogen of the scissile peptide bond of the substrate. We performed a series of constrained minimizations by gradually decreasing the corresponding H–N distance and locating the second transition state TS2. The normal mode of the single imaginary frequency of the second saddle point corresponds solely to this proton-transfer event. The downhill move toward the acylenzyme complex was carried out as an unconstrained minimization of all internal coordinates. The computed structure of AE is shown in Figure 4.

We note that the scissile peptide bond is cleaved at this point: the distance between initially covalently bound C and N atoms in the hexapeptide is now 3.5 Å. The Asp170 residue restores its protonation status (compare Figures 2 and 3). The Glu80 residue abstracts the proton from the third member of the catalytic triad, Asp84, after losing the primarily transferred proton for the half-product of the hydrolysis reaction.

The computed energy profile for the route from the enzyme–substrate complex (ES) to the acylenzyme complex (EA) through the tetrahedral intermediate (TI) for the peptide bond cleavage of the model hexapeptide Arg-Gly-Phe-Phe+Tyr-Thr by sedolisin is shown in Figure 5. According to these calculations, the activation energy barriers for the reaction are fairly low. The electronic embedding QM/MM results do not alter the mechanistic conclusions qualitatively but reveal additional stabilization of TI by the protein environment. The relative energy of TI is reduced to 5.5 kcal/mol, while the energy barriers are only slightly lowered by 2.8 and 0.8 kcal/mol for TS1 and TS2 saddle points, respectively.

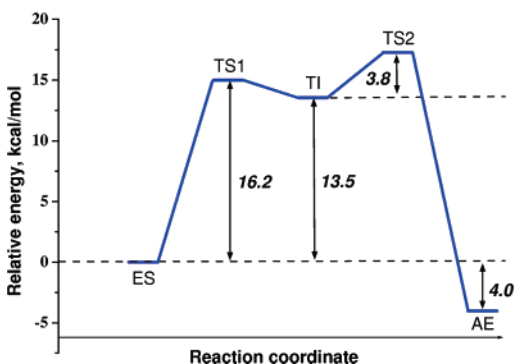


Figure 5. The computed energy diagram for the reaction path from the enzyme–substrate complex (ES) to the acylenzyme complex (EA) through the tetrahedral intermediate (TI) for the peptide bond cleavage of the hexapeptide Arg-Gly-Phe-Phe+Tyr-Thr by sedolisin.

Discussion and Conclusions

As mentioned in the Introduction, the experimentally resolved structure 1NLU refers to the complex of sedolisin with two inhibitor pseudoiodotyrosatin molecules, one of which is covalently bound to Ser287 mimicking a tetrahedral adduct. We can directly compare atomic coordinates of this experimental structure with those obtained in our simulations for the tetrahedral intermediate (TI). The left panel of Figure 6 shows a superposition of experimental and theoretical structures, where blue sticks designate the chains from the crystal moiety 1NLU, and red sticks refer to the calculation results. In the latter case we show the Phe (P1)–Phe (P2)–Gly (P3) fraction of the model substrate. Apparently, two structures are in good agreement, what can be quantitatively characterized by the RMSD values of 0.5 Å calculated for all heavy atoms of the residues in the active site Glu80, Asp84, Ser133, Asp170, Thr286, and Ser287. Even a superposition of substrate P1 and P2 chains on the inhibitor chains shows a remarkable similarity: RMSD is 0.5 Å for P1, and 1.6 Å for P2, and the groups of the model substrate occupy the same binding pockets as the groups of pseudoiodotyrosatin. Therefore, we can conclude that the results of simulations are consistent with the available experimental information.

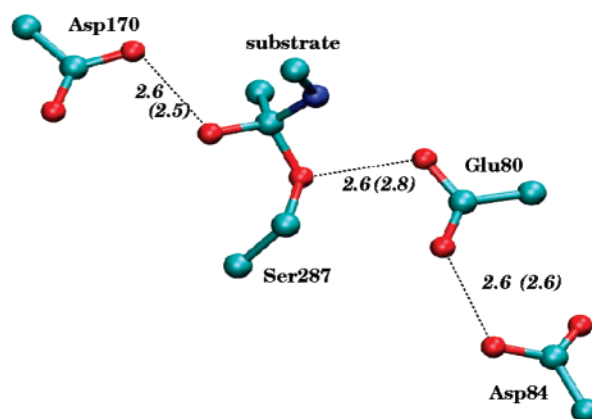
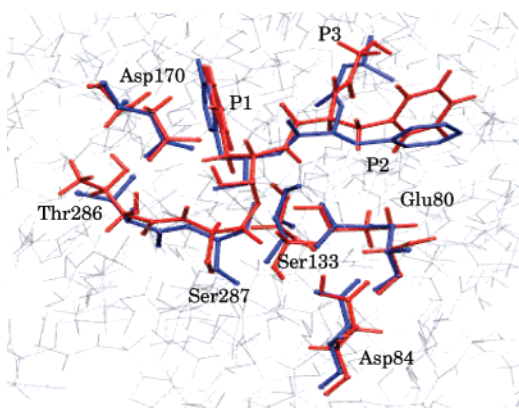


Figure 6. Comparison of the computed structure of TI and the experimental structure 1NLU. Left: superposition of the chains in the active site showing experimental data in blue and computational results in red. Right: equilibrium geometry configuration showing distances in Å; the values in parentheses refer to the calculation results.

The reaction mechanism of sedolisin which comes into view in the present work is summarized in Scheme 1. The substrate–enzyme complex (ES), acylenzyme (AE), tetrahedral intermediate (TI), and both transition states (TS1 and TS2) are shown. Step I is the nucleophilic attack of serine oxygen on a carbon atom of the substrate with an enhanced electrophilic character. Structure of the first transition state (Figure 7) clearly demonstrates an acid-based activation

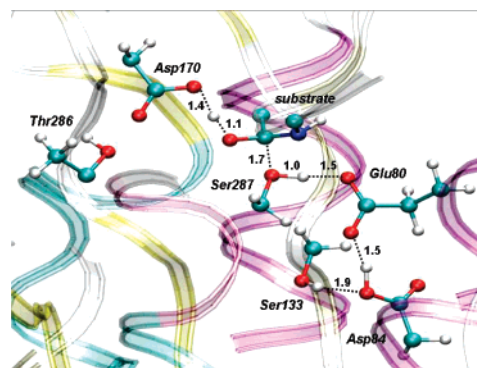
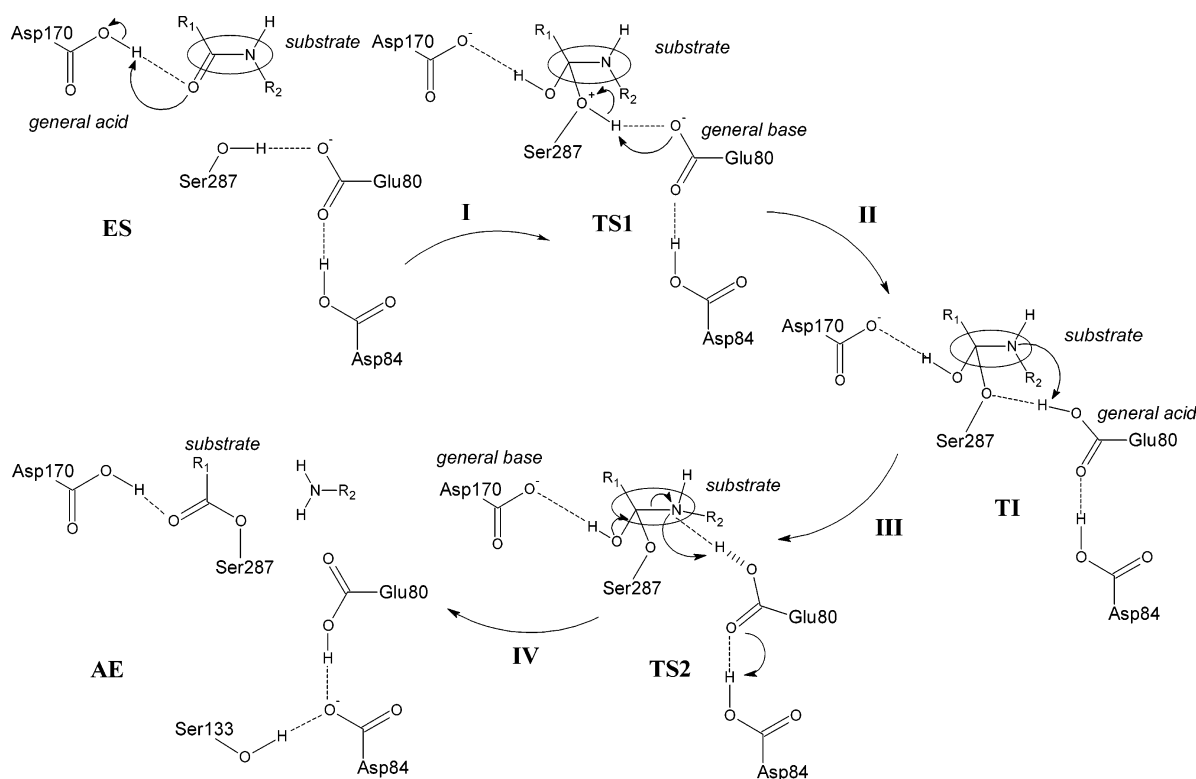


Figure 7. The computed structure of TS1. The distances are shown in Å.

mechanism of the enzyme. The normal-mode analysis at the TS1 configuration also supports the primary role of Asp170 residue, revealing the reaction coordinate being the concerted Asp170 proton transfer and C–O bond formation. Apparently, Glu80 does not play the role of His in the catalytic triad, if the mechanisms of serine-carboxyl peptidases and serine proteases are compared.

In step II, the negatively charged Glu80 acts as a general base to remove the proton from Ser287. A tetrahedral intermediate formed at this stage represents a hemiacetal bonded complex of substrate and the enzyme. Step III includes the activation of the leaving group by the protonated Glu80 which acts as the general acid donating the proton to nitrogen. This step also involves the reorientation of the hydrogen bond of Glu80 from Ser287 to nitrogen of the scissile peptide bond. A structure of the second transition state represents the initial stage of the proton transfer along

Scheme 1. Mechanism of the Catalytic Action of Sedolisin Proposed by the Results of Simulations

the line connecting O(Glu80) and N(substrate) atoms (Figure 8). The final step IV refers to the cleavage of the peptide bond accompanied by deprotonation of the hemiacetal complex, thus restoring the carbonyl group of acylenzyme. At this stage the negatively charged Asp170 serves as a general base. The proton transfer to the leaving group is facilitated by the simultaneous proton uptake from Asp84 to Glu80, which in turn is assisted by Ser133. It should be noted that the proton shuttling between Glu80 and Asp84 on the segments III and IV, which appears as a result of the present calculations, may be refined at a higher level computational scheme; however, it should not affect the general conclusions of our modeling.

The above considerations leave room for an assignment of the serine-carboxyl peptidases to a certain class of enzymes, if the mode of substrate activation is taken into account. Generally, the first stage may include nucleophile activation by a general base as in the case of serine proteases at neutral pH¹⁰ or an acid-catalyzed mechanism of the

substrate carbonyl group activation by its protonation at lower pH as for carboxyl (aspartic) peptidases.⁴⁰ In both cases, the active sites are designed to provide an additional activation of the reaction partner. In serine proteases, an oxyanion hole facilitates the peptide bond hydrolysis by stabilizing TI and, to a certain extent, by activating the substrate carbonyl group by hydrogen bonds.¹⁰ Despite an existing classification of serine-carboxyl peptidases as enzymes of serine-protease type, which is mainly inspired by the structural homology of sedolisins and subtilisins, the results obtained in this work favor the catalytic mechanism similar to that of acid-based hydrolysis of peptides.

Although tentative proposals on catalytic action of serine-carboxyl peptidases have been formulated by the authors of refs 8 and 11, their assumptions are consistent with an activation stage typical for the serine proteases. When considering the energy changes along a very short part of the reaction path (between the hemiacetal and aldehyde complexes), the authors of ref 11 hypothesize that the role of aspartic acid (Asp170 for sedolisin) is only to stabilize the tetrahedral adduct by donating its proton to the oxyanion actually formed as a result of the general base-activated nucleophilic attack. According to our results (Scheme 1) the protonated aspartate (Asp170 in sedolisin), when standing in a similar position as the oxyanion hole moieties in serine proteases, actively participates in the reaction by donating its proton to the peptide bond of the substrate.

The aspartic acid residue coupled with the glutamate partner (Asp170 and Glu80 in sedolisin) can be considered as a catalytic dyad analogue to the protonated and unprotonated Asp residues in aspartic proteases.⁴⁰ The important role of this dyad has been recognized in experimental site-directed mutagenesis studies of different members of the

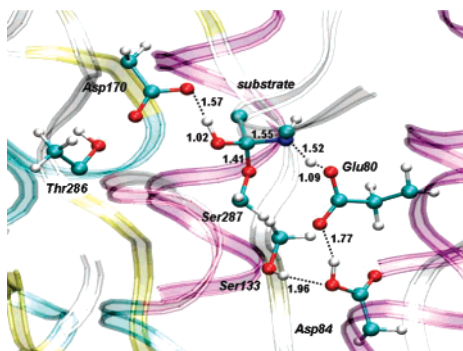


Figure 8. The computed structure of TS2. The distances are shown in Å.

sedolisin family S53. The pairs of residues Asp360/Glu272 in CLN2,^{41,42} Asp164/Glu78 in kumamolisin,⁴³ and Asp211/Glu86 in aorsin⁴⁴ are considered as structural analogues of the Asp170/Glu80 acidic pair in sedolisin. When analyzing kinetic parameters of mutant proteins Asp360/Ala vs Glu272/Ala of human tripeptidyl-peptidase (CLN2), the authors of ref 42 arrived at the conclusion that the catalytic efficiencies were greatly reduced compared to the wild-type enzyme. Similar observations were noticed for kumamolisin,⁴³ the Asp164/Ala and Glu78/Ala mutants of which did not exhibit proteolytic activity. In the case of aorsin, Asp211/Asn and Glu86/Gln mutations resulted in the loss of catalytic activity by 4 orders of magnitude.⁴⁴ Finally, Oyama et al. showed that Asp170/Ala and Asp169/Ala mutants in sedolisin and sedolisin-B, respectively, did not show any autocatalytic processing and proteinase activity.⁴⁵ The authors stressed that the Asp170 residue in serine-carboxyl peptidases should be considered as a catalytic residue.⁴⁵ Wlodawer et al. reported about an attempt to create a mutant of kumamolisin-As, in which the glutamate residue of the active site was replaced by a histidine in order to mimic the classical catalytic triad of serine proteases.⁷ The authors concluded that a normal catalytic triad could not be reconstructed in this case, thus underlining the uniqueness of the glutamate residue in serine-carboxyl peptidases. On the contrary, mutations of the oxyanion hole residues in serine proteases do not show a great impact on the enzyme catalytic activity.¹⁰

In summary, the computed energy profile for the reaction route from ES to AE for the enzymatic cleavage of a model substrate by sedolisin allows us to formulate conclusions on the reaction mechanism of serine-carboxyl peptidases. As specified in the literature, e.g., in refs 46 and 47, the entropic contributions may slightly change the activation barriers shown in Figure 5 by about 2 kcal/mol without altering the qualitative consequences. According to the results of simulations described in this work, the reaction mechanism of serine-carboxyl peptidases should be viewed as a special case of carboxyl (aspartic) proteases active at acidic pH with the nucleophilic water molecule being replaced by the Ser residue. Despite the structural similarity of sedolisins and subtilisins, the only feature of serine carboxyl peptidases common to serine proteases is the formation of a covalent substrate–enzyme bond at the stage of nucleophilic attack.

Acknowledgment. We thank Prof. A. Wlodawer and Prof. K. Oda for valuable discussions of the project. The substantial contributions of J. Kress and A. Granovsky to the QM/MM extended PC GAMESS version are greatly acknowledged. This work is supported in part by the grants from the Russian Federal Science and Innovation Agency (State contract 02.442.11.7435), from Russian Foundation for Basic Researches (project 04-03-32007), and from the Russian Academy of Sciences (Program 10 of the Division of Chemistry and Material Sciences). We thank the staff and administration of the Advanced Biomedical Computing Center for their support of this project. This project has been funded in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. 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 organizations imply endorsement by the U.S. Government.

References

- (1) Oda, K.; Sugitani, M.; Fukuhara, K.; Murao, S. *Biochim. Biophys. Acta* **1987**, *923*, 463–469.
- (2) Wlodawer, A.; Li, M.; Dauter, Z.; Gustchina, A.; Uchida, K.; Oyama, H.; Dunn, B. M.; Oda, K. *Nature Struct. Biol.* **2001**, *8*, 442–446.
- (3) Wlodawer, A.; Li, M.; Gustchina, A.; Oyama H., Dunn, B. M.; Oda, K. *Acta Biochim. Pol.* **2003**, *50*, 81–102.
- (4) Wlodawer, A. *Structure* **2004**, *12*, 1117–1119.
- (5) Comellas-Bigler, M.; Maskos, K.; Huber, R.; Oyama, H.; Oda, K.; Bode, W. *Structure* **2004**, *12*, 1313–1323.
- (6) Wlodawer, A.; Li, M.; Gustchina, A.; Tsuruoka, N.; Ashida, M.; Minakata, H.; Oyama, H.; Oda, K.; Nishino, T.; Nakayama, T. *J. Biol. Chem.* **2004**, *279*, 21500–21510.
- (7) Wlodawer, A.; Li, M.; Gustchina, A.; Dauter, Z.; Uchida, K.; Oyama, H.; Goldfarb, N.; Dunn, B. M.; Oda, K. *Biochemistry* **2001**, *40*, 15602–15611.
- (8) Wlodawer, A.; Li, M.; Gustchina, A.; Oyama, H.; Oda, K.; Beyer, B. B.; Celmente, J.; Dunn, B. M. *Biochem. Biophys. Res. Comm.* **2004**, *314*, 638–645.
- (9) Oda, K.; Nakatani, H.; Dunn, B. M. *Biochem. Biophys. Acta* **1992**, *1120*, 208–214.
- (10) Hedstrom, L. *Chem. Rev.* **2002**, *102*, 4501–4523.
- (11) Guo, H.; Wlodawer, A.; Guo, H. *J. Am. Chem. Soc.* **2005**, *127*, 15662–15663.
- (12) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.
- (13) Field, M. J.; Bash, P. A.; Karplus, M. A. *J. Comput. Chem.* **1990**, *11*, 700–733.
- (14) Gao, J. L.; Xia, X. F. *Science* **1992**, *258*, 631.
- (15) Stanton, R. V.; Hartsough, D. S.; Merz, K. M. *J. Comput. Chem.* **1995**, *16*, 113.
- (16) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580–10594.
- (17) Zhang, Y.; Lee, T.-S.; Yang W. *J. Chem. Phys.* **1999**, *110*, 46.
- (18) Murphy, R. B.; Philipp, D. M.; Friesner, R. A. *J. Comput. Chem.* **2000**, *21*, 1442–1457.
- (19) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569–585.
- (20) Alhambra, C.; Corchado, J. C.; Sanchez, M. L.; Gao, J.; Truhlar, D. G. *J. Am. Chem. Soc.* **2000**, *122*, 8197–8203.
- (21) Cui, Q.; Karplus, M. *J. Am. Chem. Soc.* **2002**, *124*, 3093.
- (22) Cisneros, G. A.; Liu, H.; Zhang, Y.; Yang, W. *J. Am. Chem. Soc.* **2003**, *125*, 10384–10393.
- (23) Guallar, V.; Baik, M.-H.; Lippard, S. J.; Friesner, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 6998–7002.
- (24) Shurki, A.; Warshel, A. *Adv. Protein. Chem.* **2003**, *66*, 249–379.
- (25) Li, G.; Cui, Q. *J. Phys. Chem. B* **2004**, *108*, 3342–3357.
- (26) Klähn, M.; Braun-Sand S.; Rosta, E.; Warshel, A. *J. Phys. Chem. B* **2005**, *109*, 15645–15650.

- (27) Wong, K. F.; Selzer, T.; Benkovic, S. J.; Hammes-Shiffer, S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6807–6812.
- (28) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. *Chem. Rev.* **2005**, *105*, 2279–2328.
- (29) Bentzien, J.; Muller, R. P.; Florián, J.; Warshel, A. *J. Phys. Chem. B* **1998**, *102*, 2293.
- (30) Torf, M.; Varnai, P.; Richards, W. G. *J. Am. Chem. Soc.* **2002**, *124*, 14780–14788.
- (31) Zhang, Y.; Kua, J.; McCammon, J. A. *J. Am. Chem. Soc.* **2002**, *124*, 10572–10577.
- (32) Molina, P. A.; Jensen, J. H. *J. Phys. Chem. B* **2003**, *107*, 6226–6233.
- (33) Nemukhin, A. V.; Grigorenko, B. L.; Rogov, A. V.; Topol, I. A.; Burt, S. K. *Theor. Chem. Acc.* **2004**, *111*, 36–48.
- (34) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
- (35) Kal, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. *J. Comput. Phys.* **1999**, *151*, 283–312.
- (36) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.
- (37) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. *J. Comput.-Aided. Mol. Des.* **1996**, *10*, 293–304.
- (38) Goodsell, D. S.; Olson, A. J. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 195–202.
- (39) Heteny, C.; Spoel, V. D. *Prot. Sci.* **2002**, *11*, 1729–1737.
- (40) Dunn, B. M. *Chem. Rev.* **2002**, *102*, 4431–4458.
- (41) Lin, L.; Sohar, I.; Lackland, H.; Lobel, P. *J. Biol. Chem.* **2001**, *276*, 2249–2255.
- (42) Walus, M.; Kida, E.; Wisniewski, K. E.; Golabek, A. A. *FEBS Lett.* **2005**, *579*, 1383–1388.
- (43) Comellas-Bigler, M.; Fuentes-Prior, P.; Maskos, K.; Huber, R.; Oyama, H.; Uchida, K.; Dunn, B. M.; Oda, K.; Bode, W. *Structure* **2002**, *10*, 865–876.
- (44) Lee, B. R.; Furukawa, M.; Yamashita, K.; Kanasugi, Y.; Kawabata, C.; Hirano, K.; Ando, K.; Ichishima, E. *Biochem. J.* **2003**, *371*, 541–548.
- (45) Oyama, H.; Abe, S.-I.; Ushiyama, S.; Takahashi, S.; Oda, K. *J. Biol. Chem.* **1999**, *274*, 27815–27822.
- (46) Florián, J.; Warshel, A. *J. Phys. Chem. B* **1998**, *102*, 719–734.
- (47) Kötting, C.; Gerwert, K. *Chem. Phys.* **2004**, *307*, 227–232.

CT6000686

Theoretical Study on the Structure and the Frequency of Isomers of the Naphthalene Dimer

Morihisa Saeki* and Hiroshi Akagi

*Quantum Beam Science Directorate, Japan Atomic Energy Agency, Tokai-mura,
Naka-gun, Ibaraki 319-1195, Japan*

Masaaki Fujii

*Chemical Resources Laboratory, Tokyo Institute of Technology, 4259 Nagatsuta-cho,
Midori-ku, Yokohama 226-8503, Japan*

Received November 11, 2005

Abstract: The structures of the naphthalene monomer and dimer were investigated with performing vibrational analysis. The MP2 optimization showed that the naphthalene monomer has a nonplanar geometry in the 6-31G, 6-31G*, 6-31+G*, and 6-311G basis sets, while it has a planar geometry in the 6-31G*(0.25) and Dunning's correlation consistent basis sets. The MP2/cc-pVDZ calculation showed the presence of the four stable isomers, which were part of the isomers in the previous MP2/6-31G* calculation (Walsh, T. R. *Chem. Phys. Lett.* **2002**, 363, 45). The presence of extra structures in the MP2/6-31G* calculation is attributed to a poor description of the potential energy surface, which is evident from the nonplanar structure of the monomer in the MP2/6-31G* calculation. The relative stability among the isomers in the MP2/cc-pVDZ calculation without counterpoise correction was maintained in both the single-point calculation at the MP2/aug-cc-pVDZ//MP2/cc-pVDZ level and the counterpoise-corrected optimization at the MP2/cc-pVDZ level. The relative stability among the isomers suggested an enhancement of the π - π interaction in the structure with lower symmetry, which could be explained using a molecular-orbital model. The vibrational analysis in MP2/cc-pVDZ without the counterpoise correction suggested that the isomers of the naphthalene dimer were distinguishable by the observation of the infrared spectrum in the low-frequency region (150–600 cm^{-1}).

Introduction

Intermolecular interactions between polycyclic aromatic hydrocarbon (PAH) molecules are unique, because the π -type molecular orbitals contribute them. The π orbital of one PAH molecule interacts with the π orbital or the C–H bonding of the other. The interaction with the π orbital is called a π - π interaction (also called a stacking interaction), while that with the C–H bonding is named a C–H $\cdots\pi$ interaction. The relative strength between these interactions determines the structure of PAH clusters.

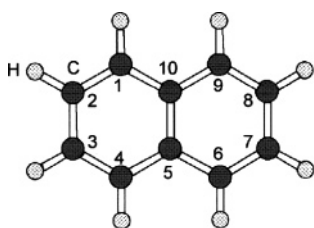
Much concerted effort between experiment and theory was made to study several dimers of PAH: benzene,^{1–14} naphthalene,^{1–3,15–21} and anthracene^{22,23} dimers. Especially, the structure of the benzene dimer has allured controversies for a long time. Klemperer et al. have suggested a structure dominated by the C–H $\cdots\pi$ interaction (T-shaped structure) using molecular beam electronic resonance spectroscopy.⁵ Bernstein et al., however, have claimed that the structure dominated by the π - π interaction (sandwich structure) may be more stable than the T-shaped structure based on their theoretical consideration.⁷ A final conclusion has been given by Felker et al.^{8–11} They observed the splitting of the vibrational band in the Raman spectrum, which was attributed

* Corresponding author phone: +81-29-282-6100; e-mail: saeki.morihisa@jaea.go.jp.

Table 1. Basis-Set Dependence of Symmetry, Energy (hartree), and C–C Bond Length (Å) of Naphthalene Monomer

	symmetry	energy	bond length ^b			
			C1–C2	C2–C3	C5–C10	C1–C10
6-31G	C_{2h}	–384.08453	1.393	1.432	1.446	1.436
			(+0.022)	(+0.020)	(+0.026)	(+0.014)
6-31G*	C_{2h}	–384.61466	1.379	1.415	1.432	1.419
			(+0.008)	(+0.003)	(+0.012)	(–0.003)
6-31+G*	C_{2h}	–384.63770	1.382	1.416	1.433	1.421
			(+0.011)	(+0.004)	(+0.013)	(–0.001)
6-311G	C_{2h}	–384.20668	1.388	1.428	1.441	1.432
			(+0.017)	(+0.016)	(+0.021)	(+0.010)
6-31G*(0.25)	D_{2h}	–384.36700	1.402	1.438	1.456	1.440
			(+0.031)	(+0.026)	(+0.036)	(+0.018)
cc-pVDZ	D_{2h}	–384.68161	1.389	1.423	1.441	1.427
			(+0.018)	(+0.011)	(+0.021)	(+0.005)
aug-cc-pVDZ	D_{2h}	–384.73937	1.392	1.425	1.444	1.428
			(+0.021)	(+0.013)	(+0.024)	(+0.006)
cc-pVTZ	D_{2h}	–385.04465	1.377	1.411	1.430	1.415
			(+0.006)	(–0.001)	(+0.010)	(–0.007)
exptl ^a			1.371	1.412	1.420	1.422

^a Experimental bond length is obtained from ref 30. ^b The value in the parentheses indicates the difference between the calculated and experimental bond length.

**Figure 1.** Definition of the label of the naphthalene monomer.

to a structure dominated by the C–H $\cdots\pi$ interaction. A measurement of the depolarization ratio of the Raman band by Ebata et al. has supported the conclusion given by Felker et al.¹² The discrepancy between the experiment and the calculation indicates the difficulty to accurately calculate the π – π and C–H $\cdots\pi$ interactions.

The information concerning the structure of the naphthalene dimer is not rich compared with that of the benzene dimer. Saigusa et al. have investigated the electronic structure of the naphthalene dimer using a resonant two-photon ionization (R2PI).^{15,16} The electronic spectrum was shown to be broad and structureless. Considering that sharp bands were observed in the spectrum of the benzene dimer,⁴ the structure of the naphthalene dimer is assumed to be different from that of the benzene dimer. There have been many theoretical studies about the structure of the naphthalene dimer.^{1–3,17–21} These studies suggest that the π – π interaction was adequately estimated by a calculation including the electron correlation, diffuse orbital, and the basis set superposition error (BSSE) correction. However, considerations of both the electron correlation and the diffuse orbital burden the calculation of the naphthalene dimer, because it is a large system containing 20 carbon atoms and 16 hydrogen atoms. Gonzalez and Lim expressed the dilemma as follows:¹ “Unfortunately, because of the large size of these species (*benzene, naphthalene and anthracene dimer*), the basis sets that can be employed to obtain conformational geometries and energies from ab initio calculations are rather limited in size. Moreover, because dispersion effects are important in determining the geometries of vdW dimer, SCF calculations are completely inadequate and methodologies that include electron correlation must be employed”. Con-

cerning the dilemma, most researchers have not calculated the frequencies of the naphthalene dimer, although vibrational analysis is essential to ensure that the optimized structures are located in the local minimum.

Recently Walsh has optimized the geometry of the naphthalene dimer while calculating the frequencies at MP2/6-31G* level.²⁰ He found a new structure of naphthalene dimer belonging to the C_2 point group, which was the most stable among the isomers. However, as described in this paper, the naphthalene monomer has a nonplanar geometry in the MP2/6-31G* calculation. For assurance of the stability of the new structure, the isomers of the naphthalene dimer should be investigated using the computational level whose monomer is calculated to be planar. First, we optimized the structure of the monomer using various basis sets and determined the computational level adequate for the calculation of naphthalene. At the determined level, the isomers of naphthalene dimer were investigated with performing the vibrational analysis. The relative stability among the obtained isomers was explained using a molecular-orbital model. Based on the result of the vibrational analysis of the monomer and the dimer, we discussed the vibrational modes available to experimentally distinguish the isomers.

Computational Method

The used program was the GAUSSIAN 98²⁴ and the GAUSSIAN 03 package.²⁵ Optimizations of the naphthalene monomer and dimer were carried out by the MP2 method with the frozen core approximation, because the electron correlation must be considered for precisely estimating the π – π interaction.^{1–3,18–21,26} The initial structure of the dimer employed in this study is the same as that in Walsh’s work.²⁰ The stability of the optimized structures was checked by a harmonic frequency analysis. The frequency was computed analytically. If the optimized structure had one or more imaginary frequencies with fixed symmetry, it was reoptimized with the lower symmetry. The procedure was iterated until the true local minimum structure was obtained. In estimating the binding energy of the dimer, BSSE was corrected using the counterpoise (CP) method.²⁷ In addition to these procedures, we optimized the geometry of the

naphthalene dimer with considering the CP correction. The geometrical optimization with CP was performed using “Counterpoise” keyword in GAUSSIAN 03. This keyword runs the optimization process developed by Simon et al.²⁸ All of the computations were carried out on an NEC SX-7 computer at Research Center for Computational Science, Okazaki, Japan.

Results

Basis-Set Dependence of the Structure of the Naphthalene Monomer.

The structure of the naphthalene monomer was optimized with 6-31G, 6-31G*, 6-31+G*, 6-311G, 6-31G*(0.25), cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ basis sets. The 6-31G*(0.25) basis set is 6-31G* with the exponent on the d function reduced to 0.25 (0.80 in the 6-31G*) and was developed to describe the $\pi-\pi$ interaction.^{26,29} The calculated symmetry, energy, and bond length of naphthalene are listed in Table 1, together with the experimental bond length in ref 30. The label of the carbon atom used for the geometrical parameter is defined in Figure 1. The optimized geometry belongs to the C_{2h} point group in the MP2/6-31G calculation, although it should belong to the D_{2h} point group. This means that the naphthalene molecule is distorted out of the plane. The distortion is kept in calculations with the 6-31G*, 6-31+G*, 6-311G basis sets, while the naphthalene monomer has the planar geometry in 6-31G*(0.25), cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ. The results suggested that a correct calculation of naphthalene is accomplished by the 6-31G*(0.25) and Dunning’s correlation consistent basis sets at the MP2 level. The comparison between the calculation and the experiment suggests that the calculated bond lengths are drastically improved from MP2/6-31G*(0.25) to MP2/cc-pVDZ. The bond lengths in MP2/aug-cc-pVDZ show a worse agreement with the experimental values than those in MP2/cc-pVDZ. It means that the structure of the monomer cannot be refined by the addition of the diffuse functions to the cc-pVDZ. On the other hand, there is very good agreement of the bond lengths between the MP2/cc-pVTZ calculation and the experiment. The agreement with the experiment becomes better with the ordering of 6-31G*(0.25) < aug-cc-pVDZ < cc-pVDZ < cc-pVTZ. The MP2/cc-pVTZ level is most adequate for the calculation of naphthalene but consumes computational resources very much. Thus, we selected the MP2/cc-pVDZ level for the calculation of the naphthalene dimer.

Frequency of the Naphthalene Monomer in MP2/cc-pVDZ. The naphthalene molecule has 48 normal modes, whose symmetries are $9a_{1g}+8b_{3g}+3b_{1g}+4b_{2g}+4a_{1u}+4b_{3u}+8b_{1u}+8b_{2u}$. The frequencies in the MP2/cc-pVDZ calculation are listed in Table 2. The types of the vibrational modes are noted as $r(\text{CH})$ for CH stretching, $R(\text{CC})$ for CC stretching, $\beta(\text{CH})$ for CH in-plane bending, $\alpha(\text{CCC})$ for CCC in-plane bending, $\epsilon(\text{CH})$ for CH out-of-plane bending, and $\tau(\text{CCC})$ for CCC out-of-plane bending vibration. The assignment of the vibrational types follows Ellinger’s work.³¹ The experimental frequencies were obtained from ref 32.

The scaling factor is populated between 0.9 and 1.1 in all modes, except for the ν_{23} mode (~ 1.5). The extraordinary value in the ν_{23} mode may be attributed to an incorrect

Table 2. Calculated and Experimental Frequency (cm^{-1}) of the Vibrational Modes of the Naphthalene Monomer

mode	symmetry	frequency		scaling factor ^d
		calcd	exptl ^b	
$r(\text{CH})$ Type ^a				
ν_1	a_{1g}	3239	3060	0.9447
ν_2	a_{1g}	3209	3031	0.9445
ν_{10}	b_{3g}	3224	3092	0.9591
ν_{11}	b_{3g}	3203	3060	0.9554
ν_{33}	b_{1u}	3225	3065	0.9504
ν_{34}	b_{1u}	3204	3058	0.9544
ν_{41}	b_{2u}	3238	3090	0.9543
ν_{42}	b_{2u}	3207	3027	0.9439
averaged scaling factor				0.9508 (0.0059)
$\beta(\text{CH})$ Type ^a				
ν_6	a_{1g}	1170	1145	0.9786
ν_{15}	b_{3g}	1161	1158	0.9974
ν_{36}	b_{1u}	1406	1389	0.9879
averaged scaling factor				0.9880 (0.0094)
$\alpha(\text{CCC})$ Type ^a				
ν_9	a_{1g}	514	512	0.9961
ν_{16}	b_{3g}	931	936	1.0054
ν_{17}	b_{3g}	506	506	1.0000
ν_{39}	b_{1u}	803	747	0.9303
ν_{40}	b_{1u}	356	359	1.0084
ν_{48}	b_{2u}	618	618	1.0000
averaged scaling factor				0.9900 (0.0296)
$R(\text{CC}) + \beta(\text{CH})$ Type ^a				
ν_4	a_{1g}	1490	1460	0.9799
ν_7	a_{1g}	1051	1025	0.9753
ν_{44}	b_{2u}	1495	1361	0.9104
ν_{45}	b_{2u}	1255	1209	0.9633
ν_{46}	b_{2u}	1170	1138	0.9726
averaged scaling factor				0.9603 (0.0285)
$R(\text{CC}) + \beta(\text{CH}) + \alpha(\text{CCC})$ Type ^a				
ν_3	a_{1g}	1625	1577	0.9705
ν_5	a_{1g}	1458	1376	0.9438
ν_8	a_{1g}	771	758	0.9831
ν_{12}	b_{3g}	1688	1624	0.9621
ν_{13}	b_{3g}	1484	1438	0.9690
ν_{14}	b_{3g}	1255	1239	0.9873
ν_{35}	b_{1u}	1637	1595	0.9743
ν_{37}	b_{1u}	1277	1265	0.9906
ν_{38}	b_{1u}	1139	1125	0.9877
ν_{43}	b_{2u}	1562	1509	0.9661
ν_{47}	b_{2u}	1042	1008	0.9674
averaged scaling factor				0.9729 (0.0138)
$\epsilon(\text{CH})$ Type ^a				
ν_{18}	b_{1g}	924	943	1.0206
ν_{19}	b_{1g}	721	717	0.9945
ν_{21}	b_{2g}	951	980	1.0305
ν_{22}	b_{2g}	849	876	1.0318
ν_{25}	a_{1u}	933	970	1.0397
ν_{26}	a_{1u}	837	841	1.0048
ν_{29}	b_{3u}	932	958	1.0279
ν_{30}	b_{3u}	785	782	0.9962
averaged scaling factor				1.0182 (0.0174)
$\tau(\text{CCC})$ Type ^a				
ν_{20}	b_{1g}	377	386	1.0239
ν_{23}	b_{2g}	535	846	1.5813
ν_{24}	b_{2g}	451	461	1.0222
ν_{27}	a_{1u}	558	581	1.0412
ν_{28}	a_{1u}	182	195	1.0714
ν_{31}	b_{3u}	463	476	1.0281
ν_{32}	b_{3u}	168	176	1.0476
averaged scaling factor				1.0391 (0.0187) ^c

^a The type of the vibrational modes are noted as $r(\text{CH})$ for CH stretching, $R(\text{CC})$ for CC stretching, $\beta(\text{CH})$ for CH in-plane bending, $\alpha(\text{CCC})$ for CCC in-plane bending, $\epsilon(\text{CH})$ for CH out-of-plane bending, and $\tau(\text{CCC})$ for CCC out-of-plane bending vibration. ^b Experimental value is obtained from ref 32. ^c The ν_{23} mode is neglected in averaging of the scaling factor. ^d Averaged scaling factors are calculated in the respective type. The value in the parentheses indicates standard deviation.

assignment, because the calculated frequencies of other modes have good agreement with the experimental ones. The

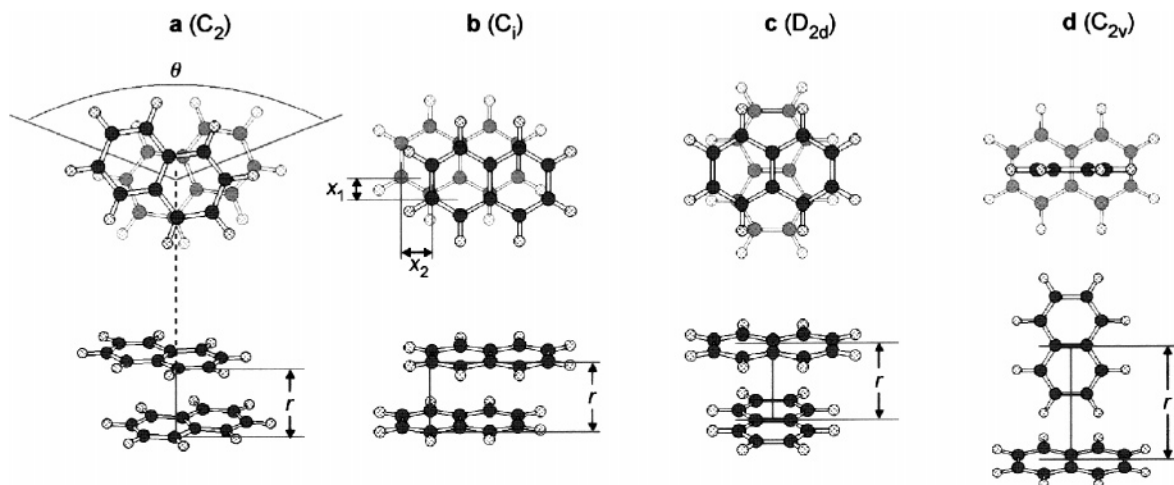


Figure 2. Optimized isomers of the naphthalene dimer in the MP2/cc-pVDZ calculation. The symbol in the parenthesis indicates the symmetry of each structure. The geometries and energies of these structures are summarized in Table 3.

Table 3. Dependence of Geometry and Binding Energy D_0 (kcal/mol) on Computational Method in Structures **a–d**^a

method	a			b				c		d	
	r	θ	D_0	r	x_1	x_2	D_0	r	D_0	r	D_0
	Geometry without CP										
MP2/cc-pVDZ	3.24	136	-5.30	3.20	0.93	1.25	-5.00	3.25	-4.53	5.86	-3.55
MP2/aug-cc-pVDZ//MP2/cc-pVDZ			-10.95				-10.34		-9.48		-5.68
	Geometry with CP										
MP2/cc-pVDZ	3.46	135	-6.24	3.40	1.03	1.31	-5.89	3.48	-5.38	6.03	-3.80

^a The unit of the geometrical parameters is degree in θ and angstrom in r , x_1 , and x_2 . The geometry of the naphthalene dimer was optimized without and with the CP correction in MP2/cc-pVDZ. Single-point calculation at the MP2/aug-cc-pVDZ level was performed using the optimized geometry in MP2/cc-pVDZ without CP. All binding energies are CP corrected.

modes of the $r(\text{CH})$, $\beta(\text{CH})$, $\alpha(\text{CCC})$, $R(\text{CC})+\beta(\text{CH})$, and $R(\text{CC})+\beta(\text{CH})+\alpha(\text{CCC})$ types are in-plane vibration, while those of the $\epsilon(\text{CH})$ and $\tau(\text{CCC})$ types are out-of-plane vibration. The averaged scaling factor suggests that the calculated frequency is overestimated in the in-plane vibrations and is underestimated in the out-of-plane ones. The overestimation of the frequency of the out-of-plane vibrations has been found in the calculation of benzene.^{33,34} In the harmonic frequency analysis the calculated frequency should be overestimated, because the anharmonicity is neglected. The underestimation in the out-of-plane vibration indicates a poor description of the potential surface along the out-of-plane direction.

Stable Isomers of the Naphthalene Dimer. Figure 2 shows stable structures of the naphthalene dimer in the MP2/cc-pVDZ calculation. The geometries and energies of structures **a–d** are summarized in Table 3. Walsh has suggested eight isomers were stable in the MP2/6-31G* calculation.²⁰ However, the MP2/cc-pVDZ calculation elucidated that structures **a–d** were located in the local minimum, while the others were in transition states. We compared the calculated geometry of the naphthalene molecules between the monomer and the dimer and found that the geometrical parameters in the monomer are almost maintained in the dimer. The naphthalene molecules are bound by the π - π interaction in structures **a–c** and by the C-H $\cdots\pi$ interaction in structure **d**. The counterpoise-corrected binding energy is 4.5–5.3 kcal/mol in structures **a–c**, while it is 3.6 kcal/mol in structure **d**. The binding energy ordering is **a** > **b** > **c** > **d**. Structure **a**, which

corresponds to the isomer newly found by Walsh,²⁰ is most stable among structures **a–d**. There is little difference of the intermolecular distance r among structures **a–c**.

To investigate the basis-set dependence of the relative stability among the isomers, we performed the single-point calculation at the MP2/aug-cc-pVDZ level using the optimized geometry in MP2/cc-pVDZ. As shown in Table 3, the ordering of the CP-corrected binding energy in the MP2/cc-pVDZ calculation is kept in MP2/aug-cc-pVDZ//MP2/cc-pVDZ. The binding energy in structures **a–c** increases by 5.0–5.7 kcal/mol from MP2/cc-pVDZ to MP2/aug-cc-pVDZ, while that in structure **d** increases by only 2.1 kcal/mol. It indicates that the addition of diffuse functions has a larger effect on the π - π interaction than on the C-H $\cdots\pi$ interaction.

In clusters the optimization with the CP correction may refine the geometry optimized without CP, because the estimation of the intermolecular interaction is improved.²⁸ Using the obtained isomers in MP2/cc-pVDZ without CP as the initial geometry, we investigated the geometrical change with the CP correction. As shown in Table 3, the intermolecular distance r was elongated by ~ 0.2 Å in every structure when the CP correction was considered. In principle, since the BSSE causes the intermolecular interactions to be artificially too attractive, the CP correction should make the cluster less stable.²⁸ Consequently, the intermolecular distance will be longer when the cluster is optimized with the CP correction. The ordering of the binding energy is the same between geometry with CP and without CP. The binding energy of the geometry with CP is larger than that

Table 4. Calculated Frequency (cm^{-1}) of the Intramolecular Vibrational Modes of the Naphthalene Dimer in the MP2/cc-pVDZ Calculation without CP^a

mode	a	b	c	d	mode	a	b	c	d
<i>r</i> (CH) Type (Scaling Factor 0.9508)									
	3077(b)	3079(a _u)	3074(a ₁)	3095(a ₁)		3077(a)	3079(a _g)	3074(b ₂)	3084(b ₂)
	3073(b)	3074(a _u)	3073(e)	3079(a ₁)		3073(a)	3074(a _g)	3073(e)	3078(b ₂)
	3063(a)	3063(a _u)	3062(e)	3077(a ₁)		3062(b)	3063(a _g)	3062(e)	3066(b ₁)
	3060(a)	3061(a _g)	3062(b ₁)	3066(a ₂)		3060(b)	3061(a _u)	3061(a ₂)	3064(b ₂)
	3052(a)	3048(a _g)	3050(b ₂)	3052(a ₁)		3051(b)	3047(a _u)	3049(a ₁)	3050(b ₂)
	3044(b)	3046(a _u)	3048(e)	3049(a ₁)		3044(a)	3046(a _g)	3048(e)	3048(b ₁)
	3043(a)	3042(a _u)	3043(e)	3047(a ₂)		3042(b)	3041(a _g)	3043(e)	3046(a ₁)
	3039(a)	3040(a _g)	3043(a ₂)	3045(b ₂)		3039(b)	3040(a _u)	3042(b ₁)	3043(b ₂)
<i>β</i> (CH) Type (Scaling Factor 0.9880)									
ν_6^+	1154(b)	1155(a _u)	1153(a ₁)	1160(a ₁)	ν_6^-	1152(a)	1154(a _g)	1152(b ₂)	1155(a ₁)
ν_{15}^+	1143(b)	1144(a _g)	1145(b ₁)	1149(b ₂)	ν_{15}^-	1143(a)	1143(a _u)	1143(a ₂)	1146(a ₂)
ν_{36}^+	1387(a)	1387(a _u)	1387(e)	1389(b ₂)	ν_{36}^-	1387(b)	1387(a _g)	1387(e)	1389(b ₁)
<i>α</i> (CCC) Type (Scaling Factor 0.9900)									
ν_9^+	508(a)	509(a _g)	509(b ₂)	511(a ₁)	ν_9^-	508(b)	508(a _u)	508(a ₁)	509(a ₁)
ν_{16}^+	921(a)	920(a _g)	920(a ₂)	920(b ₂)	ν_{16}^-	919(b)	920(a _u)	920(b ₁)	919(a ₂)
ν_{17}^+	499(a)	503(a _u)	499(b ₁)	502(b ₂)	ν_{17}^-	499(b)	500(a _g)	499(a ₂)	501(a ₂)
ν_{39}^+	794(b)	794(a _u)	794(e)	795(b ₁)	ν_{39}^-	793(a)	793(a _g)	794(e)	795(b ₂)
ν_{40}^+	351(a)	351(a _u)	351(e)	353(b ₂)	ν_{40}^-	351(b)	351(a _g)	351(e)	352(b ₁)
ν_{48}^+	610(b)	610(a _u)	610(e)	612(a ₁)	ν_{48}^-	609(a)	610(a _g)	610(e)	611(b ₂)
<i>R</i> (CC)+ <i>β</i> (CH) Type (Scaling Factor 0.9603)									
ν_4^+	1430(b)	1431(a _u)	1431(a ₁)	1431(a ₁)	ν_4^-	1429(a)	1430(a _g)	1430(b ₂)	1430(a ₁)
ν_7^+	1009(a)	1008(a _g)	1009(b ₂)	1008(a ₁)	ν_7^-	1008(b)	1008(a _u)	1008(a ₁)	1008(a ₁)
ν_{44}^+	1453(b)	1451(a _g)	1449(e)	1439(b ₂)	ν_{44}^-	1451(a)	1450(a _u)	1449(e)	1438(a ₁)
ν_{45}^+	1205(a)	1206(a _u)	1206(e)	1206(a ₁)	ν_{45}^-	1205(b)	1205(a _g)	1206(e)	1206(b ₂)
ν_{46}^+	1122(a)	1121(a _g)	1122(e)	1124(b ₂)	ν_{46}^-	1122(b)	1120(a _u)	1122(e)	1123(a ₁)
<i>R</i> (CC)+ <i>β</i> (CH)+ <i>α</i> (CCC) Type (Scaling Factor 0.9729)									
ν_3^+	1576(b)	1577(a _g)	1575(b ₂)	1579(a ₁)	ν_3^-	1574(a)	1575(a _u)	1575(a ₁)	1578(a ₁)
ν_5^+	1420(b)	1420(a _u)	1420(a ₁)	1419(a ₁)	ν_5^-	1420(a)	1419(a _g)	1419(b ₂)	1418(a ₁)
ν_8^+	750(a)	749(a _g)	750(a ₁)	750(a ₁)	ν_8^-	749(b)	749(a _u)	749(b ₂)	749(a ₁)
ν_{12}^+	1638(b)	1638(a _g)	1638(a ₂)	1641(b ₂)	ν_{12}^-	1638(a)	1638(a _u)	1638(b ₁)	1640(a ₂)
ν_{13}^+	1440(a)	1441(a _g)	1442(b ₁)	1442(a ₂)	ν_{13}^-	1440(b)	1441(a _u)	1441(a ₂)	1442(b ₂)
ν_{14}^+	1217(a)	1218(a _g)	1218(b ₁)	1221(a ₂)	ν_{14}^-	1217(b)	1218(a _u)	1218(a ₂)	1219(b ₂)
ν_{35}^+	1587(a)	1588(a _u)	1589(e)	1591(b ₂)	ν_{35}^-	1586(b)	1587(a _g)	1589(e)	1590(b ₁)
ν_{37}^+	1240(a)	1240(a _u)	1241(e)	1243(b ₁)	ν_{37}^-	1240(b)	1240(a _g)	1241(e)	1241(b ₂)
ν_{38}^+	1105(b)	1105(a _u)	1105(e)	1109(b ₂)	ν_{38}^-	1104(a)	1105(a _g)	1105(e)	1107(b ₁)
ν_{43}^+	1518(b)	1518(a _g)	1517(e)	1518(a ₁)	ν_{43}^-	1518(a)	1518(a _u)	1517(e)	1517(b ₂)
ν_{47}^+	1014(b)	1014(a _u)	1013(e)	1013(b ₂)	ν_{47}^-	1013(a)	1013(a _g)	1013(e)	1013(a ₁)
<i>ε</i> (CH) Type (Scaling Factor 1.0182)									
ν_{18}^+	926(a)	928(a _u)	927(e)	940(b ₂)	ν_{18}^-	925(b)	927(a _g)	927(e)	940(b ₁)
ν_{19}^+	728(b)	730(a _u)	727(e)	736(b ₁)	ν_{19}^-	726(a)	725(a _g)	727(e)	735(b ₂)
ν_{21}^+	957(b)	957(a _u)	955(e)	967(b ₁)	ν_{21}^-	955(a)	955(a _g)	955(e)	965(a ₂)
ν_{22}^+	856(a)	858(a _u)	854(e)	866(b ₁)	ν_{22}^-	852(b)	853(a _g)	854(e)	865(a ₂)
ν_{25}^+	938(a)	938(a _g)	937(a ₂)	948(a ₂)	ν_{25}^-	936(b)	936(a _u)	934(b ₁)	945(a ₂)
ν_{26}^+	842(a)	845(a _g)	845(b ₁)	853(a ₂)	ν_{26}^-	840(b)	843(a _u)	839(a ₂)	852(a ₂)
ν_{29}^+	933(b)	936(a _g)	937(b ₂)	947(a ₁)	ν_{29}^-	933(a)	933(a _u)	934(a ₁)	947(b ₁)
ν_{30}^+	789(a)	792(a _g)	795(a ₁)	800(b ₁)	ν_{30}^-	787(b)	788(a _u)	788(b ₂)	799(a ₁)
<i>τ</i> (CCC) Type (Scaling Factor 1.0391)									
ν_{20}^+	387(b)	387(a _u)	387(e)	393(b ₁)	ν_{20}^-	386(a)	385(a _g)	387(e)	388(b ₂)
ν_{23}^+	512(b)	524(a _g)	519(e)	569(b ₁)	ν_{23}^-	511(a)	523(a _u)	519(e)	533(a ₂)
ν_{24}^+	456(a)	459(a _u)	455(e)	469(a ₂)	ν_{24}^-	453(b)	455(a _g)	455(e)	466(b ₁)
ν_{27}^+	565(a)	569(a _g)	563(a ₂)	580(a ₂)	ν_{27}^-	565(b)	567(a _u)	563(b ₁)	576(a ₂)
ν_{28}^+	205(a)	207(a _g)	213(b ₁)	197(a ₂)	ν_{28}^-	203(b)	201(a _u)	182(a ₂)	189(a ₂)
ν_{31}^+	473(a)	474(a _g)	472(b ₂)	480(b ₁)	ν_{31}^-	469(b)	469(a _u)	472(a ₁)	477(a ₁)
ν_{32}^+	188(a)	187(a _g)	181(a ₁)	182(a ₁)	ν_{32}^-	185(b)	180(a _u)	176(b ₂)	180(b ₁)

^a The frequencies are scaled by the factor calculated in the monomer.

Table 5. Calculated Frequency (cm^{-1}) of the Intermolecular Vibrational Modes of the Naphthalene Dimer in the MP2/cc-pVDZ Calculation without CP^a

a	b	c	d
22(a)	8(a _u)	2(b ₁)	5(b ₁)
45(b)	39(a _g)	15(e)	22(a ₂)
64(a)	54(a _g)	15(e)	34(b ₂)
95(b)	81(a _u)	92(e)	60(b ₂)
103(a)	100(a _u)	92(e)	62(a ₁)
106(a)	110(a _g)	94(a ₁)	69(b ₁)

^a The frequencies are not scaled.

without CP by 0.85–0.94 kcal/mol in structures **a–c** and by 0.25 kcal/mol in structure **d**. The effect of the CP correction on the binding energy is more enhanced in the structures dominated by the π – π interaction.

Frequencies of the Vibrational Modes of the Naphthalene Dimer. The vibrational analysis of the naphthalene dimer was performed using only the geometry optimized without the CP correction, because GAUSSIAN 03 does not provide the vibrational analysis with CP. The calculated frequencies of the vibrational modes of the naphthalene dimer are listed in Table 4 for the intramolecular vibrations and in Table 5 for the intermolecular ones. Under an assumption that the vibrational motions within the naphthalene moieties is hardly disturbed by the intermolecular interactions, the intramolecular modes of the dimer are described as a linear combination of the modes of the monomer. The assignment of the intramolecular modes is also given in Table 4. The linear combination of one mode of the monomer leads to the formation of two modes of the dimer. We denoted the modes with higher frequency as ν_j^+ and those with lower frequency as ν_j^- . A detailed assignment was avoided in the $r(\text{CH})$ type, because various modes of the monomer are mixed in the vibration of the dimer. The frequencies of the intramolecular modes are scaled by the averaged factor in the monomer (Table 2), while those of the intermolecular modes are not scaled.

Discussion

Relative Stability among Isomers of Naphthalene Dimer.

As shown in Figure 2, there are four stable isomers in the MP2/cc-pVDZ optimization with the vibrational analysis, although there have been eight stable isomers in MP2/6-31G*.²⁰ The presence of extra structures in the MP2/6-31G* calculation is attributed to a poor description of the potential energy surface, which is evident from the nonplanar structure of the naphthalene monomer in MP2/6-31G* (Table 1). Thus, for the calculation of the naphthalene dimer we should employ 6-31G*(0.25) or Dunning's correlation consistent basis sets in the MP2 method.

The ordering of the binding energy of the naphthalene dimer is **a** > **b** > **c** > **d** in the MP2/cc-pVDZ calculation without the CP correction. The ordering is kept in the single point calculation at the MP2/aug-cc-pVDZ//MP2/cc-pVDZ level. Moreover, the CP-corrected optimization does not change the ordering in MP2/cc-pVDZ without CP. The calculated results ensure us that structure **a** is most stable among the isomers in the MP2 method. Tsuzuki et al. have

estimated the binding energy of the isomers including structures **b** and **c** in the CCSD(T)/6-31G* calculation (structures **a** and **d** were neglected in their work).²¹ Comparing the binding energy between CCSD(T)/6-31G* and MP2/6-31G*, they showed the overestimation of the binding energy in the MP2 calculation. However, the results of the CCSD(T) calculation suggested structure **b** is more stable than structure **c**. This ordering of the binding energy agrees with our results. Thus, we assume that the ordering of the binding energy in MP2 is maintained in CCSD(T).

The symmetry of structures **a** (C_2) and **b** (C_i) is lower than that of structure **c** (D_{2d}), while the ordering of the binding energy is **a** > **b** > **c**. It means that the π – π interaction is enhanced in the structures with lower symmetry. The enhancement of the π – π interaction can be explained based on a molecular-orbital model (Figure 3). To simplify our explanation, we compare only structures **a** and **c**. The same discussion is true for structure **b**. In this model, the molecular orbitals of the naphthalene dimer are formed by the interaction between the highest occupied molecular orbitals (HOMOs) of the naphthalene monomers and by that between the lowest unoccupied molecular orbitals (LUMOs). The interaction between HOMO and LUMO is neglected because they have different symmetry. Considering the overlap between the molecular orbitals of the naphthalene molecules, the molecular orbitals of structure **a** are composed of a bonding orbital between HOMOs (**A1**), an antibonding orbital between HOMOs (**A2**), a bonding orbital between LUMOs (**A3**), and an antibonding orbital between LUMOs (**A4**). The electrons are occupied in orbitals **A1** and **A2** but are unoccupied in orbitals **A3** and **A4**. The occupied orbitals in structure **a** can interact with the unoccupied ones, because the symmetry is the same between orbitals **A1** and **A3** (a symmetry) and between orbitals **A2** and **A4** (b symmetry). The interaction of the occupied orbitals with the unoccupied ones contributes to the stabilization of structure **a**. On the other hand, the molecular orbitals of structure **c** are composed of a bonding orbital between HOMOs (**C1**), an antibonding orbital between HOMOs (**C2**), and nonbonding orbitals originating from LUMOs (**C3**). The electrons are occupied in orbitals **C1** and **C2** but are unoccupied in orbitals **C3**. In structure **c**, the occupied orbitals (b₁ and a₂ symmetry) do not interact with the unoccupied ones (e symmetry), because the symmetry of the orbitals is different between orbitals **C1** and **C3** and between orbitals **C2** and **C3**. Thus, structure **c** is not stabilized by the interaction between the occupied and unoccupied orbitals. Based on the above discussion, we conclude that the lowering of the symmetry enables the occupied orbitals to interact with the unoccupied ones.

Comparison of the Infrared Spectrum among the Isomers. As shown in Figure 2, the naphthalene dimer has four stable isomers. Experimentally, the isomers are distinguished based on the observed vibrational spectrum. For this purpose, we have investigated the difference of the calculated vibrational spectrum among the isomers. Table 4 suggests that the difference of the frequency is most evident in the $\tau(\text{CCC})$ -type modes. Figure 4 shows the calculated infrared spectrum of structures **a–d** in the region of 150–600 cm^{-1} , where the vibrational bands of the $\tau(\text{CCC})$ and $\alpha(\text{CCC})$ type

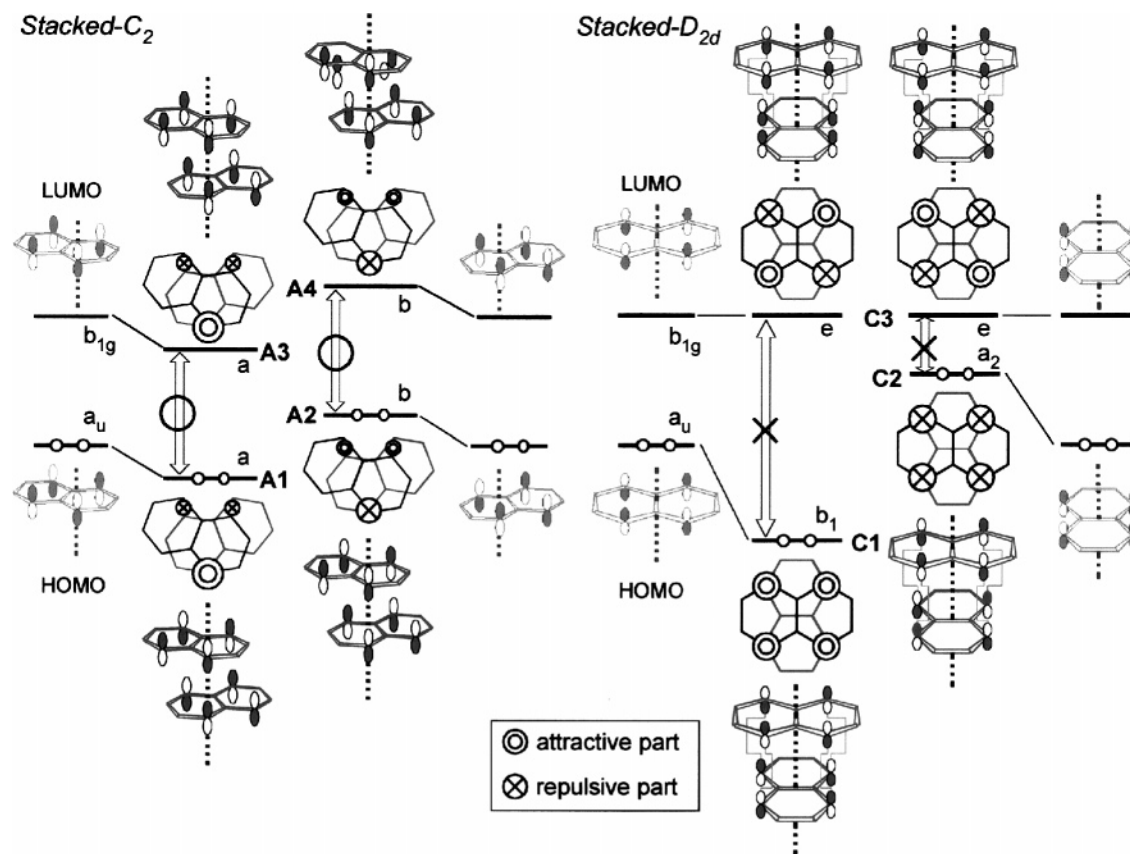


Figure 3. Schematic diagram of molecular-orbital interactions in structures **a** and **c**. The orbitals of the naphthalene dimer are formed by the interaction between HOMOs and by that between LUMOs. The orbitals with the same phase form the attractive part, while those with different phases form the repulsive one. Considering the condition of the overlap between molecular orbitals, we attributed **A1**, **A3**, and **C1** to bonding orbitals, **A2**, **A4**, and **C2** to antibonding orbitals, and **C3** to nonbonding orbitals.

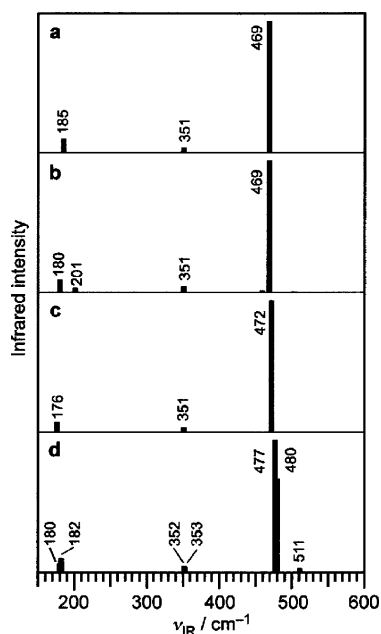


Figure 4. Calculated infrared spectrum of structures **a–d** in the low-frequency region. The bands around 180, 200, and 470 cm^{-1} are assigned to the $\tau(\text{CCC})$ type, while those around 350 and 510 cm^{-1} are assigned to the $\alpha(\text{CCC})$ one.

are observed. The most intense band is located around 470 cm^{-1} and is assigned to the ν_{31}^{\pm} mode. The band is degenerated in structures **a–c** and is split in structure **d**. In

addition, the frequency of the band in structure **c** is blue-shifted from that in structures **a** and **b**. Structures **a** and **b** are distinguishable from the bands around 201 cm^{-1} , which are assigned to the ν_{28}^{-} mode. The band is observed in structure **b** and is absent in structure **a**. We can distinguish the isomers of the naphthalene dimer by the infrared spectrum in the low-frequency region.

Table 5 suggested that the frequencies of the intermolecular modes are below 110 cm^{-1} and largely depend on the isomers. The calculated infrared intensities, however, are $<1\%$ of the intensity of the ν_{31}^{\pm} mode (not shown). The vibrational bands of the intermolecular modes are difficult to observe in the infrared spectrum. We assume that the intermolecular modes are observable in the vibronic spectrum of the ground state.

Conclusions

We investigated the structures of the naphthalene monomer and dimer while performing a vibrational analysis. The MP2 optimization showed the naphthalene monomer has the nonplanar geometry in the 6-31G, 6-31G*, 6-31+G*, and 6-311G basis sets, while it has the planar geometry in the 6-31G*(0.25) and Dunning's correlation consistent basis sets. Based on the result of the monomer, we employed the MP2/cc-pVDZ level for calculation of the naphthalene dimer. The MP2/cc-pVDZ optimization showed the presence of structures **a–d** in Figure 2, which were part of the stable structures in the MP2/6-31G* calculation. The presence of

extra structures in the MP2/6-31G* calculation is attributed to a poor description of the potential energy surface, which is evident from the nonplanar structure of the monomer in MP2/6-31G*. The calculation of the naphthalene dimer should be performed using the 6-31G*(0.25) or Dunning's correlation consistent basis sets in the MP2 method.

The ordering of the binding energy is $\mathbf{a} > \mathbf{b} > \mathbf{c} > \mathbf{d}$ in the MP2/cc-pVDZ optimization without the CP correction. The relative stability among the isomers was maintained in both the single-point calculation at the MP2/aug-cc-pVDZ//MP2/cc-pVDZ level and the CP-corrected optimization at the MP2/cc-pVDZ level. Thus, we concluded that structure \mathbf{a} is most stable among the isomers. Structure \mathbf{a} has lower symmetry than structure \mathbf{c} . It indicates that the π - π interaction is enhanced by lowering the symmetry. A discussion based on the molecular-orbital model elucidated that in the naphthalene dimer the symmetry lowering enhances the interaction between the occupied and unoccupied orbitals.

The intramolecular vibrations of the naphthalene dimer were assigned as a linear combination of the vibrational modes of the naphthalene monomer. Based on the results of the vibrational analysis, we concluded that the isomers are experimentally distinguishable from the infrared spectrum in the low-frequency region (150–600 cm^{-1}).

Acknowledgment. We would like to thank Dr. O. Dopfer for his comments about the ab initio MO calculation of the π - π interaction. We are also thankful to Dr. A. Yokoyama for valuable discussions.

References

- (1) Gonzalez, C.; Lim, E. C. *J. Phys. Chem. A* **2000**, *104*, 2953.
- (2) Tsuzuki, S.; Uchimaru, T.; Matsumura, K.; Mikami, M.; Tanabe, K. *Chem. Phys. Lett.* **2000**, *319*, 547.
- (3) Sato, T.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2005**, *123*, 104307.
- (4) Hopkins, J. B.; Powers, D. E.; Smally, R. E. *J. Phys. Chem.* **1981**, *85*, 3739.
- (5) Steed, J. M.; Dixon, T. A.; Klemperer, W. *J. Chem. Phys.* **1979**, *70*, 4940.
- (6) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Chem. Phys.* **1990**, *93*, 5893.
- (7) Schauer, M.; Bernstein, E. R. *J. Chem. Phys.* **1985**, *82*, 3722.
- (8) Henson, B. F.; Hartland, G. V.; Venturo, V. A.; Herts, R. A.; Felker, P. M. *Chem. Phys. Lett.* **1991**, *176*, 91.
- (9) Henson, B. F.; Hartland, G. V.; Venturo, V. A.; Felker, P. M. *J. Chem. Phys.* **1992**, *97*, 2189.

- (10) Venturo, V. A.; Felker, P. M. *J. Chem. Phys.* **1993**, *99*, 748.
- (11) Schaeffer, M. W.; Maxton, P. M.; Felker, P. M. *Chem. Phys. Lett.* **1994**, *224*, 544.
- (12) Ebata, T.; Ishikawa, S.; Ito, M.; Hyodo, S. *Laser Chem.* **1994**, *14*, 85.
- (13) Engkvist, O.; Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Chem. Phys.* **1999**, *110*, 5758.
- (14) Špirko, V.; Engkvist, O.; Soldán, P.; Selzle, H. L.; Schlag, E. W.; Hobza, P. *J. Chem. Phys.* **1999**, *111*, 572.
- (15) Saigusa, H.; Sun, S.; Lim, E. C. *J. Phys. Chem.* **1992**, *96*, 2083.
- (16) Saigusa, H.; Lim, E. C. *Acc. Chem. Res.* **1996**, *29*, 171.
- (17) East, A. L. L.; Lim, E. C. *J. Chem. Phys.* **2000**, *113*, 8981.
- (18) Song, J. K.; Han, S. Y.; Chu, I.; Kim, J. H.; Kim, S. K.; Lyapustina, S. A.; Xu, S.; Nilles, J. M.; Bowen, K. H., Jr. *J. Chem. Phys.* **2002**, *116*, 4477.
- (19) Lee, N. K.; Park, S.; Kim, S. K. *J. Chem. Phys.* **2002**, *116*, 7902. Lee, N. K.; Park, S.; Kim, S. K. *J. Chem. Phys.* **2002**, *116*, 7910.
- (20) Walsh, T. R. *Chem. Phys. Lett.* **2002**, *363*, 45.
- (21) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M. *J. Chem. Phys.* **2004**, *120*, 647.
- (22) Gonzalez, C.; Lim, E. C. *Chem. Phys. Lett.* **2000**, *322*, 382.
- (23) Piuze, F.; Dimicoli, I.; Mons, M.; Millié, P.; Brenner, V.; Zhao, Q.; Soep, B.; Tramer, A. *Chem. Phys.* **2002**, *275*, 123.
- (24) Frisch, M. J.; et al. *Gaussian 98, revision A.11*; Gaussian, Inc.: Pittsburgh, PA, 1998.
- (25) Frisch, M. J.; et al. *Gaussian 03, revision C.01*; Gaussian, Inc.: Wallingford CT, 2004.
- (26) Šponer, J.; Leszczynski, J.; Hobza, P. *J. Mol. Struct.: Theochem* **2001**, *573*, 43.
- (27) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
- (28) Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024.
- (29) Kroon-Batenburg, L. M. J.; Van Duijneveldt, F. B. *J. Mol. Struct.: Theochem* **1985**, *121*, 185.
- (30) Bastiansen, O.; Skancke, P. N. *Adv. Chem. Phys.* **1961**, *3*, 323.
- (31) Pauzat, F.; Talbi, D.; Miller, M. D.; DeFrees, D. J.; Ellinger, Y. *J. Phys. Chem.* **1992**, *96*, 7882.
- (32) Krainov, E. P. *Opt. Spektrosk.* **1964**, *16*, 415 and 763.
- (33) Martin, J. M. L.; Taylor, P. R.; Lee, T. J. *Chem. Phys. Lett.* **1997**, *275*, 414.
- (34) Yagi, K.; Hirao, K.; Taketsugu, T.; Schmidt, M. W.; Gordon, M. S. *J. Chem. Phys.* **2004**, *121*, 1383.

CT050278N

Energy Minimization of Crystal Structures Containing Flexible Molecules

Panagiotis G. Karamertzanis and Sarah L. Price*

*Department of Chemistry, University College London, 20 Gordon Street,
London, United Kingdom WC1H 0AJ*

Received March 24, 2006

Abstract: This paper proposes a new methodology for the accurate minimization of crystal structures of flexible molecules. The intramolecular contributions to the crystal energy are calculated from ab initio calculations and appear well-balanced with the intermolecular interactions being evaluated via a conformation-dependent distributed multipole model in conjunction with an empirical repulsion–dispersion potential model. The validity of the methodology was initially tested by minimizing the experimental crystal structures of a set of flexible molecules. In a more stringent test, the methodology was used to refine the low-energy structures found in rigid-body crystal structure prediction studies of the diastereomeric salt pair (*R*)-1-phenylethylammonium (*R/S*)-2-phenylpropanoate and the antiepileptic drug carbamazepine. The refinement improved the relative stability of the known forms and their ranking in the list of hypothetically generated structures by leading to energetically more favorable hydrogen-bond geometries and dispersion interactions.

1. Introduction

Most crystal structure prediction algorithms rely on the generation of a large number of hypothetical crystal structures and their subsequent energy minimization based on some force field.¹ Recent crystal structure prediction blind tests^{2–4} organized by the Cambridge Crystallographic Data Centre revealed that, for a limited number of crystallographically independent molecules and conformational degrees of freedom, the experimentally determined polymorphs usually appear somewhere in the list of putative low-energy structures. Thus, with the exceptions of flexible molecules, complicated asymmetric units,^{5,6} and the occurrence of a rare space group, the problem seems to be not the search methodology but the selection of a few stable structures from the multitude of minima⁷ with sufficiently close packing arrangements. A critical comparative study of all participants' submissions⁸ in the latest blind test for crystal structure prediction⁴ showed that the energy models used in the modeling of the organic solid state are not yet sufficiently reliable for this task.

The exact energy ranking of the putative structures, and consequently the success of the prediction, depends on the force-field parametrization. When the energy penalty involved in a structurally significant change in molecular conformation is comparable to the improvement this can produce in the binding energy of a molecular cluster, the force field needs to contain terms for both intra- and intermolecular contributions. The accuracy of the intermolecular potential can be greatly improved by modeling the electrostatic interactions with distributed multipoles derived directly from the wave function.^{9,10} For rigid molecules, the use of distributed multipoles^{9,11,12} offers a significant improvement in the reproduction of hydrogen-bonded crystals and the ranking of hypothetical crystal structures¹³ in comparison with atomic charge models. However, in the case of flexible molecules, the benefits of realistic anisotropic intermolecular energy models are diminished when the latter are combined with empirical intramolecular force fields, which often lead to nonphysical distortion of the molecular geometry that prevents the accurate reproduction of the known crystal structures¹⁴ and their favorable ranking with respect to hypothetical structures.¹⁵ This failure can be attributed to the inaccurate force-field parametrization, as

* Corresponding author tel.: +44 (0)20 7679 4622; fax +44 (0)-20 7679 7463; e-mail: s.l.price@ucl.ac.uk.

the inter- and intramolecular models are often derived separately and thus there is no guarantee that they will be sufficiently well-balanced to model the deformations of the molecular structure caused by the packing forces within the crystal.¹⁶

Poor accuracy often arises from various assumptions associated with the desire to achieve a compromise between accuracy on one side and computational cost and ease of implementation on the other. The charge distribution is only approximately transferable between different conformations due to local effects and the through-space polarization when the rotation around single bonds alters the relative positions of polar and polarizable parts of the molecule.^{17–19} Thus, it is often not sufficient to fix the multipoles in their local axes system for modeling molecular clusters²⁰ and crystal structures.¹⁴ Thus, the ab initio recalculation of the electrostatic model following any significant conformational change is necessary.^{21,22} Fortunately, electronic structure calculations can provide both the intermolecular electrostatic model and the deformation energy from the gas-phase optimal geometry, avoiding the inaccuracies of empirical intramolecular force fields. Although computationally expensive, this approach has successfully been applied in crystal structure prediction studies for glycol,²³ glycerol,²³ and a series of six monosaccharides.²⁴

This paper describes a hybrid computational methodology for the accurate lattice energy minimization of flexible molecules by combining a realistic electrostatic model for the intermolecular interactions based on distributed multipoles with ab initio intramolecular energies. The approach implemented in the program DMAflex extends the applicability of the hybrid approach originally applied by van Eijck et al.^{23,24} to alcohols by considering a wider range of functional groups. The accuracy of the methodology is first assessed by its ability to reproduce (section 3) the lattice geometric parameters and conformational degrees of freedom for a set of experimentally determined crystal structures (Tables 1 and 2). Some of these crystal structures were chosen because they were poorly reproduced in a previous investigation of the ability of an empirical force field to model the crystal structures of flexible organic molecules of pharmaceutical complexity.¹⁴

The usefulness of the methodology in crystal structure prediction is investigated by refining the hypothetical crystal structures generated earlier by rigid-body search methodologies, to assess the effect of packing-induced molecular distortions on their relative stability (section 3.2). The systems considered (Table 2) are the diastereomeric salt pair (*R*)-1-phenylethylammonium (*R/S*)-2-phenylpropanoate, which exhibited significant sensitivity of the lattice energy to the ions' conformations,²⁵ and the antiepileptic drug carbamazepine for which spectroscopic and theoretical investigations indicate that a rigid model²⁶ may not be sufficient for reliable crystal structure prediction. The reranking of the putative crystal structures due to the packing-induced molecular distortions is shown to make a significant difference to the realism of the predictions.

2. Computational Methodology

Crystal structure prediction is generally based on the assumption that the experimentally determined polymorphs correspond to local minima in the crystal energy surface E^{crys} , which is usually partitioned into an intramolecular, ΔE^{intra} , and an intermolecular, U^{inter} , energy contribution:

$$E^{\text{crys}} \equiv \Delta E^{\text{intra}}(\boldsymbol{\theta}) + U^{\text{inter}}(\mathbf{X}, \boldsymbol{\theta}) \quad (1)$$

with ΔE^{intra} being the energy increase due to the deformation of the in vacuo molecular geometries in the crystalline solid. \mathbf{X} denotes the degrees of freedom that define the intermolecular contacts (referred to as lattice variables hereafter) for given molecular conformations, that is, lattice lengths, lattice angles, and the position and Euler angles for each of the M crystallographically independent molecular entities. For space groups other than triclinic, the presence of symmetry relations reduces the dimensionality of vector \mathbf{X} . The vector $\boldsymbol{\theta}$ denotes the set of intramolecular degrees of freedom, that is, $\sum_{j=1}^M (3N_j - 6)$ bond lengths, bond angles, and torsion angles, where N_j is the number of atoms for the molecular entity j .

The minimization of the crystal energy for flexible molecules is technically challenging because of the difficulties associated with the calculation of its derivatives with respect to the intramolecular degrees of freedom, as the multipole moments are an unknown function of the latter. However, the derivatives of the crystal energy with respect to the lattice variables \mathbf{X} have been computed analytically and implemented in rigid-body lattice modeling packages such as DMAREL.²⁷ A feasible way to exploit the availability of these analytical gradients is to reformulate the crystal energy minimization as a bilevel optimization problem:

$$\min_{\boldsymbol{\theta}} [\Delta E^{\text{intra}}(\boldsymbol{\theta}) + \min_{\mathbf{X}} U^{\text{inter}}(\mathbf{X}; \boldsymbol{\theta})] \quad (2)$$

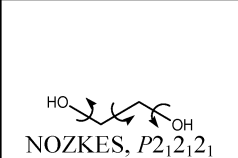
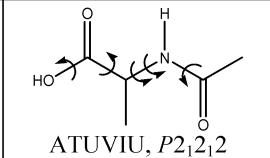
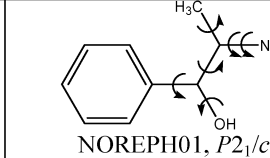
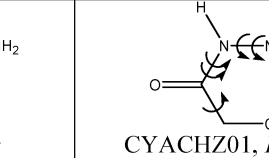
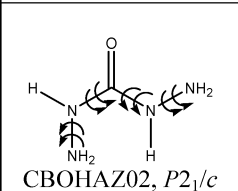
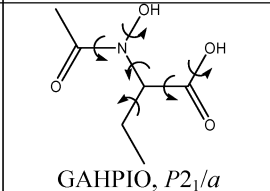
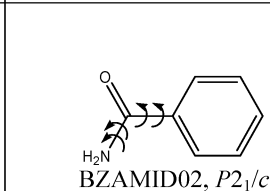
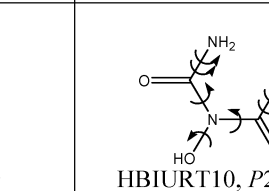
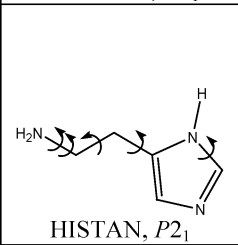
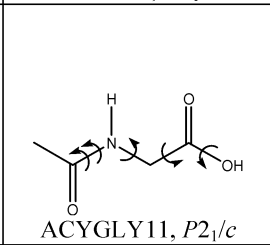
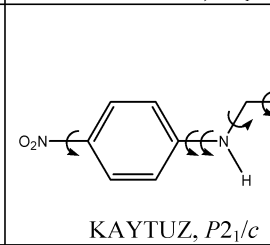
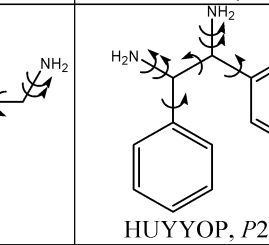
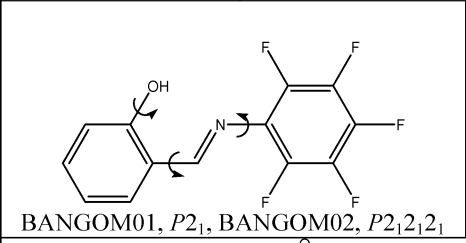
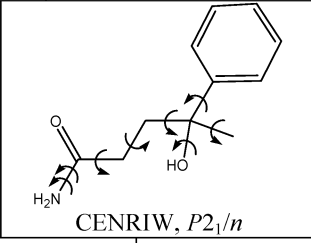
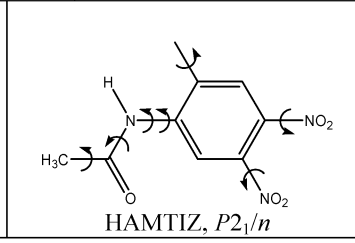
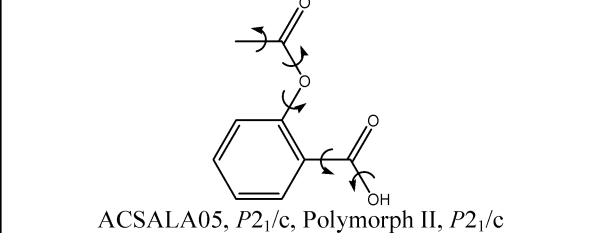
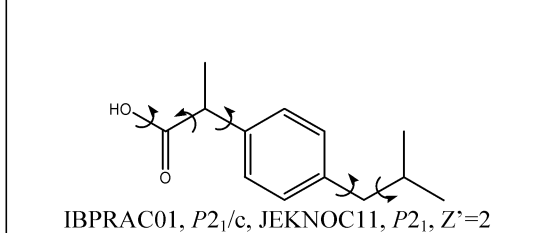
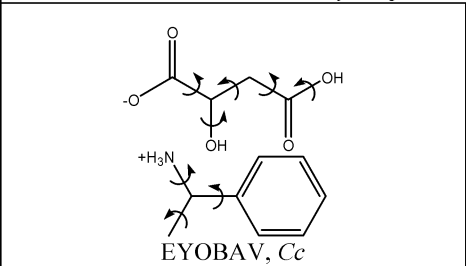
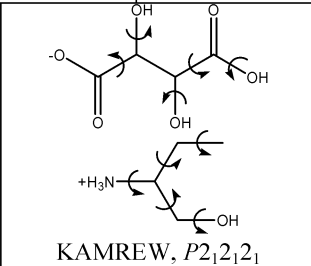
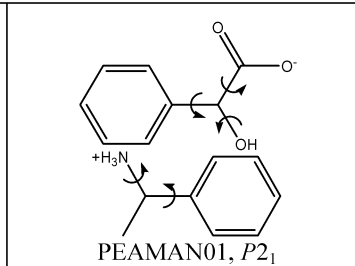
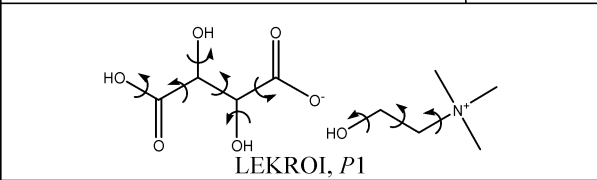
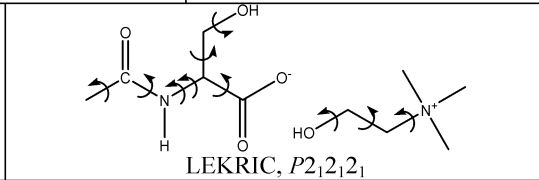
where the relative position and orientation of the molecular entities in the crystal are optimized at the inner minimization and the conformational degrees of freedom at the outer.

In molecular crystals, the intermolecular bonding energies are significantly weaker than the energy of typical covalent energies, and thus, the packing-induced molecular distortions are often limited.^{28,29} More importantly, the number F of intramolecular degrees of freedom that can deviate appreciably from their in vacuo values (referred to as flexible hereafter) is usually much smaller than the dimensionality of $\boldsymbol{\theta}$. Thus, an approximation to the minimization problem is

$$\min_{\boldsymbol{\theta}^f} \{ \Delta E^{\text{intra}}(\boldsymbol{\theta}^f) + \min_{\mathbf{X}} U^{\text{inter}}[\mathbf{X}; \boldsymbol{\theta}(\boldsymbol{\theta}^f)] \} \quad (3)$$

where $\boldsymbol{\theta}^f$ is the set of flexible degrees of freedom, such as torsion around single bonds. Assuming one crystallographically independent molecular entity for notational brevity, the rigid degrees of freedom, such as bond lengths, defined as the complement set $\boldsymbol{\theta}^r \equiv \boldsymbol{\theta} / \boldsymbol{\theta}^f$, are optimized at each step of the outer minimization problem via a constrained ab initio

Table 1. Molecular Structures with the Flexible Intramolecular Degrees of Freedom That Were Optimized within the Crystal Energy Minimization Indicated^a

 NOZKES, $P2_12_12_1$	 ATUVIU, $P2_12_12$	 NOREPH01, $P2_1/c$	 CYACHZ01, $P2_1/c$
 CBOHAZ02, $P2_1/c$	 GAHP10, $P2_1/a$	 BZAMID02, $P2_1/c$	 HBIURT10, $P2_12_12_1$
 HISTAN, $P2_1$	 ACYGLY11, $P2_1/c$	 KAYTUZ, $P2_1/c$	 HUYYP01, $P2_12_12_1$
 BANGOM01, $P2_1$, BANGOM02, $P2_12_12_1$	 CENRIW, $P2_1/n$	 HANTIZ, $P2_1/n$	
 ACSALA05, $P2_1/c$, Polymorph II, $P2_1/c$		 IBPRAC01, $P2_1/c$, JEKNOC11, $P2_1$, $Z'=2$	
 EYOBV, Cc	 KAMREW, $P2_12_12_1$	 PEAMAN01, $P2_1$	
 LEKROI, $P1$		 LEKRIC, $P2_12_12_1$	

^a CSD reference codes given for the lowest temperature/lowest R-factor structures with preference to neutron determinations (Z' shown when greater than 1). Double arrows indicate the independent rotation of two fragments around the same single bond [for example, for KAYTUZ, both torsions C(Ph)–C(Ph)–N–C and C(Ph)–C(Ph)–N–H were considered flexible].

optimization that also provides the molecular deformation energy:

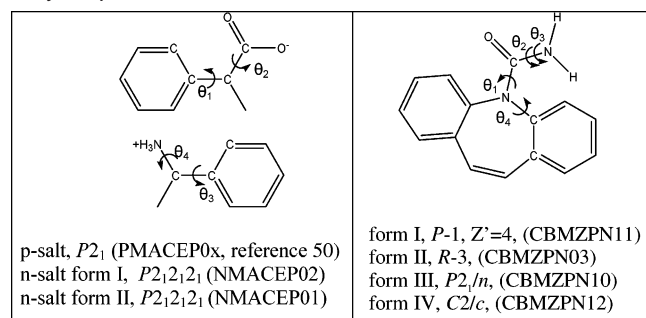
$$\theta^r(\theta^f) = \arg \min_{\theta^r} E^{\text{intra}}(\theta^f, \theta^r)$$

$$\Delta E^{\text{intra}}(\theta^f) = \min_{\theta^r} E^{\text{intra}}(\theta^f, \theta^r) - E^{\text{vac}} \quad (4)$$

where E^{vac} is the global minimum in vacuo molecular energy. This constant only needs to be computed once to compare the crystal energy to the experimentally determined heat of sublimation.^{30,31}

The inner optimization problem in eq 3 is solved with the rigid-body crystal structure modeling program DMAREL.²⁷

Table 2. Molecular Structures and CSD Reference Codes for the Known Polymorphs of (*R*)-1-Phenylethylammonium-(*R/S*)-2-phenylpropanoate^a (left) and Carbamazepine^a (right) for Which DMAflex Has Been Used for the Refinement of Putative Rigid-Body Crystal Structures in Addition to the Experimentally Determined Polymorphs



^a The set of flexible torsions, defined by the explicitly labeled atoms, comprise the rotation of the phenyl θ_1 and θ_3 , carboxylate θ_2 , and ammonium θ_4 groups for 1-phenylethylammonium-2-phenylpropanoate and the rotation of the carboxamide group θ_1 , the independent rotation of the two amide hydrogen atoms θ_2 and θ_3 , and the torsion angle θ_4 defining the tilting of the carboxamide group with respect to the seven-member ring for carbamazepine.

For all of the results reported in this paper, the electrostatic interaction model comprises atomic multipoles up to hexadecapole,³² derived through a distributed multipole analysis^{9,10,32} of the molecular charge density computed at the MP2/6-31G(d,p) level of theory for the isolated molecule. The multipoles are computed each time the inner minimization problem is solved for the corresponding molecular geometry θ to account for their conformational dependence. As one of the aims of this paper is to assess the reranking of the results of rigid-body searches due to conformational relaxation, we employed the same repulsion–dispersion potential. Thus, the repulsion–dispersion interactions were modeled with an empirical exp-6 potential with the parameters for the atomic types C, N, O, F, and H_C (hydrogen connected to carbon) taken from Williams et al.^{33–35} and for H_{N,O} (hydrogen connected to nitrogen or oxygen) from a reparametrization of this force field in conjunction with atomic multipoles.¹² The contributions from the slowly convergent charge–charge, charge–dipole, and dipole–dipole interactions were accurately summed with the Ewald summation³⁶ technique, while higher multipole and repulsion–dispersion contributions were evaluated in direct space. Because quadrupole–charge interactions are only conditionally convergent (their distance dependence is R^{-3}), a cutoff distance of 60.0 Å between the molecular centers of mass was used, which ensured that oscillations in lattice energy due to molecules coming in and out of the cutoff region³⁷ were sufficiently small.

The outer minimization problem is solved with the downhill simplex algorithm.³⁸ The initial simplex comprises $F + 1$ points; the elements of the first point are set equal to the flexible degrees of freedom in the starting crystal structure θ_o^f , and the other F points are set equal to $\theta_i^f = \theta_o^f + \lambda_i \mathbf{e}_i$, where \mathbf{e}_i are unit vectors. The characteristic length scales λ_i should be inversely proportional to the sensitivity of the crystal energy to the corresponding degree of freedom. For

bond angles or torsion angles which exhibit large intramolecular gradients or that may significantly affect the geometry of structure-defining interactions (and hence the intermolecular lattice energy), the parameters λ_i should be smaller than those with weaker lattice energy gradients, such as the rotation of methyl groups with no intra- and intermolecular steric repulsions. Generally, smaller values are preferable, although this may increase the number of iterations because of expansion of the first few simplices, to avoid failure of the inner lattice energy minimization because of severe conformational changes that may lead to close intermolecular contacts or saddle points. For all of the minimizations reported in this paper, we only consider torsion angles around single bonds (and in one case carboxylate bond angles, as explained in section 3.2.1) and set the length scale λ_i equal to 3° in all cases. A minimization is considered converged when the decrease in crystal energy in the terminating step is smaller than $1 \times 10^{-4} E^{\text{cryst}}$. For the systems discussed in this paper, this translates to a numerical accuracy in crystal energy of approximately 0.01–0.05 kJ mol⁻¹ depending on whether the crystal is molecular or ionic. For some crystals, the flexible degrees of freedom could only be converged to 1–2°, because of crystal energy oscillations on the order of a few tenths of a kilojoule per mole. This proved to result from minor discontinuities of the multipole moments with conformation generated by the distributed analysis method, as explained in the discussion.

The intramolecular energy and the rigid degrees of freedom are computed as a function of the flexible degrees of freedom at each iteration of the outer minimization by solving the optimization problem of eq 4 using the Gaussian³⁹ suite of programs. For all test cases considered, this optimization problem was first solved at the SCF/6-31G(d,p) level of theory. To investigate the effect of electron correlation, a subset of the crystal structures was also modeled by using a second-order Møller–Plesset perturbation expansion with the same basis set. At each outer minimization iteration, the new molecular conformation obtained by solving the minimization problem (eq 4) needs to be inserted in the lattice. This is achieved by a least-squares overlap of the non-hydrogen atoms in the new molecular conformation and the one found in the crystal structure that had the lowest crystal energy in all previous outer minimization steps.

The methodology is first validated by the minimization of 32 experimentally determined crystal structures (section 3) reported for the 24 compounds shown in Tables 1 and 2. These crystal structures were retrieved from the Cambridge Structural Database (CSD)⁴⁰ by searching for good-quality determinations of both neutral and ionic compounds having a diverse variety of functional groups and degrees of flexibility, ranging from 3 to 16 flexible degrees of freedom. The sample includes some crystal structures previously modeled with an empirical intramolecular force field¹⁴ (BANGCOM01, CENRIW, BANGCOM02, ACSALA05, IBPRAC01, JEKNOC11, PEAMAN01, LEKROI, and LEKRIC), where in several cases the molecular conformations significantly deformed from the experimental on energy minimization. The atom types were restricted to carbon,

oxygen, nitrogen, hydrogen, and fluorine, for which the parametrization of the repulsion–dispersion potential^{12,33–35} has been extensively tested.

The accuracy of reproduction of the crystal structures was assessed on the basis of the root-mean-square (RMS_{15}) deviation of a 15-molecule coordination sphere between the experimental and minimized crystal structures⁴¹ (hydrogen atoms omitted in the comparison) and the errors in the reproduction of the conformational degrees of freedom, density, and lattice lengths and angles. The hydrogen atoms were not included in the comparison because of the apparent foreshortening of X–H bond lengths in their X-ray determinations. The modeling accuracy for the hydrogen atom positions can be more reliably established by the reproduction of the torsion angles involving hydrogen atoms, which are reported separately from other conformational degrees of freedom, and should be deduced in the light of X-ray limitations. To ensure that the occurrence of poor reproduction is not due to inaccuracies in the intermolecular potential, the structures were also minimized with the molecular geometries held rigid at their experimental conformations (Expt), with the hydrogen positions of the X-ray structures adjusted to standard neutron bond lengths.⁴² A second rigid-body minimization was performed with the molecular conformations replaced with the ab initio optimized ones with the flexible degrees of freedom constrained to their experimental values (ConOpt). The RMS_{15} value for a ConOpt minimization also includes the effect of any minor deviations in the rigid degrees of freedom due to errors in their ab initio and experimental determinations or genuine deformations by the packing forces. A DMAflex refinement is considered successful if it leads to only a small increase in RMS_{15} compared with the ConOpt minimization.

3. Results

3.1. Ability to Reproduce Known Crystal Structures as Energy Minima. The errors in lattice lengths and angles and conformational degrees of freedom from the DMAflex minimization of the experimentally determined crystal forms are shown in Table 3. Although there is significant variation in the quality of reproduction, there are no cases for which the minimization leads to excessive distortions of the molecular conformations or unit cell geometries. This agreement with the experimental structures constitutes a substantial improvement over the previous study, which employed atomic multipole moments with empirical intramolecular force fields.¹⁴ When the intramolecular energies are modeled at the HF/6-31G(d,p) level of theory, the average RMS_{15} discrepancy for the 32 structures considered is only 0.222 Å, which is generally within the uncertainties in energy minimization due to the neglect of thermal effects.

The accuracy of the crystal structure reproductions with DMAflex are not significantly worse than the reproductions with the molecular geometry fixed at the experimental conformation. The 31 Expt rigid-body minimizations had an average RMS_{15} error of 0.171 Å (Table 3) and, hence, demonstrate that the errors in the intermolecular force field are not significant. This success can be partially attributed to the use of distributed multipoles, which ensures the

accurate modeling of hydrogen-bonded systems.^{11,43} The worst reproduction was observed in the case of KAYTUZ ($\text{RMS}_{15} = 0.476$ Å), as the nitrogen–nitrogen distance of the hydrogen-bonded amide groups was overestimated with this model potential by 0.4 Å. Another case where the intermolecular potential is less accurate and gives rise to the worst DMAflex reproduction is that of HUYYP, where the unusually elongated $\text{NH}\cdots\text{H}$ distance (3.25 Å predominantly along the *a* crystallographic axis) is severely underestimated.

The rigid-body minimization with the ab initio optimized conformations with the flexible torsion angles held at the experimental values (ConOpt) has an average RMS_{15} error of 0.188 Å (Table 3) at the HF/6-31G(d,p) level of theory, which is comparable to the rigid-body reproduction accuracy with the experimental conformations (Expt). This confirms that packing forces within the crystals may only appreciably affect the flexible degrees of freedom identified in Tables 1 and 2. The other, less deformable degrees of freedom, can generally be reasonably predicted by isolated molecule, ab initio calculations at a relatively modest level of theory, such as HF/6-31G(d,p). It is encouraging that the simultaneous relaxation of the flexible intramolecular degrees of freedom by DMAflex does not substantially deteriorate the quality of reproduction despite increasing the number of minimization variables.

As shown in Table 3, overall, DMAflex reproduces the molecular conformations in the known crystals very well with small errors in the flexible torsion angles ($\Delta\theta$ and $\Delta\theta_{\text{H}}$), including subtle details such as the pyramidalization of amino groups. For example, the significant pyramidalization of both amino groups in HUYYP, due to the balance of hydrogen bonding, $\text{N}-\text{H}\cdots\pi$, and steric interactions, is accurately reproduced, despite the 0.4 Å underestimation of the *a* cell length. One case where errors in the molecular geometry lead to poor overall reproduction is CYACHZ01, where an overestimation of the $(\text{N}\equiv)\text{C}-\text{C}-\text{C}-\text{N}$ angle leads to an elongation of the *a* cell length because of steric repulsion from the nitrile group protruding from the amide plane. For KAMREW, the accuracy in the modeling of the hydrogen-bonding motif depends strongly on the level of theory for the intramolecular energy. The MP2 model is satisfactory, whereas the SCF model gives a large error in the torsion angle for the 3-hydroxyl group, which forms a nonphysical, bifurcated intermolecular hydrogen bond to a tartrate carboxyl and an ammonium-butanol hydroxy oxygen acceptor.

With the exception of KAMREW, the inclusion of electron correlation on the intramolecular energy estimates has little effect on the reproduction of the crystal structures by DMAflex. For the 15 systems also studied at the MP2/6-31G(d,p) level, the average RMS_{15} error was 0.229 Å compared with 0.179 Å when electron correlation was neglected. This is primarily because the two methods produce somewhat different intramolecular energy surfaces. This is corroborated by the similarity of the ConOpt RMS_{15} values at the HF/6-31G(d,p) and MP2/6-31G(d,p) levels of theory indicating that the differences in the rigid degrees of freedom in the molecular structures are minimal. It is expected that the method used to evaluate the intramolecular energy contributions $\Delta E^{\text{intra}}(\theta^f)$ will be more important (but less

Table 3. Reproduction of Crystal Structures by Simultaneous Optimization of the Flexible Torsion Angles and Lattice Variables Contrasted with Rigid-Body Minimizations Using Experimental Information for the Molecular Conformation

structure	intra-molecular energy ^a	flexible-molecule							rigid-body				
		conventional unit cell					$\Delta\theta^c$ (deg)		$\Delta\theta_H^d$ (deg)		RMS ₁₅ ^e (Å)	RMS ₁₅ (Å)	Expt ^g
		a(Å)	b(Å)	c(Å)	angles ^b (deg)	density (g cm ⁻³)	max	mean	max	mean	RMS ₁₅ ^e (Å)	ConOpt ^f	
NOZKES		5.013	6.915	9.271		1.283							
	HF	-1.76%	+0.35%	+1.55%		-0.16%	3.28	3.28	8.39	6.22	0.122	0.143	
	MP2	-1.48%	+0.61%	+0.74%		+0.16%	2.86	2.86	12.29	3.22	0.133	0.165	0.255
ATUVIU		10.388	11.545	5.743		1.265							
	HF	-0.99%	+2.12%	-0.17%		-0.95%	4.82	2.92	4.68	4.18	0.161	0.129	
	MP2	-1.27%	+1.86%	+0.49%		-1.11%	6.14	3.67	7.01	4.36	0.160	0.137	0.191
NOREPH01		12.507	8.771	8.130	β 106.20	1.173							
	HF	-0.99%	-3.74%	+1.55%		+2.64%	0.67	0.65	3.10	2.25	0.249	0.235	
	MP2	+0.56%	-3.69%	+0.87%		+2.39%	0.29	0.23	3.96	2.81	0.253	0.220	0.166
CYACHZ01		7.247	8.678	7.855	β 116.80	1.493							
	HF	+7.81%	-0.62%	-0.82%		+3.58	11.43	6.94	6.12	3.15	0.329	0.205	
	MP2	+10.56%	-0.09%	-1.73%		+4.11	16.99	8.68	3.69	2.36	0.436	0.201	0.151
CBOHAZ02		3.618	8.789	12.487	β 106.43	1.571							
	HF	-1.27%	-4.65%	+1.35%		+4.32	7.13	4.35	4.23	2.43	0.242	0.210	
	MP2	-1.24%	-3.60%	+3.11%		+5.73	11.44	8.49	10.27	6.80	0.287	0.190	0.202
GAHPHO		14.003	5.425	10.495	β 93.70	1.345							
	HF	+2.24%	+1.14%	-3.00%		-1.21	8.72	2.94	7.34	5.19	0.235	0.179	
	MP2	+3.71%	+0.94%	-3.88%		-3.17	9.99	5.41	15.49	9.23	0.331	0.223	0.085
BZAMIDO02		5.529	5.033	21.343	β 88.73	1.355							
	HF	-2.06%	+0.54%	+3.68%		-1.25	1.30	1.30	5.01	2.69	0.194	0.229	
	MP2	-1.56%	+1.29%	+3.64%		-1.50	1.65	1.65	7.54	4.14	0.206	0.238	0.196
HBIURT10		10.868	11.698	3.603		1.727							
	HF	-0.02%	-2.40%	-2.14%		+4.69%	6.80	4.49	8.22	5.39	0.217	0.191	
	MP2	+0.85%	-2.80%	-2.00%		+4.11%	19.57	16.28	14.34	7.37	0.266	0.203	0.140
HISTAN		7.249	7.634	5.698	β 104.96	1.212							
	HF	-2.47%	+0.17%	+0.61%		+0.19	4.12	2.18	10.88	8.09	0.131	0.138	
	MP2	-1.92%	+0.35%	+0.68%		+0.33	4.64	2.70	9.12	7.93	0.121	0.121	0.146
ACYGLY11		4.859	11.546	14.633	β 138.29	1.424							
	HF	+0.40%	-1.76%	+3.94%		+0.81	3.00	1.00	0.00	0.00	0.207	0.209	
	MP2	-0.21%	-1.37%	+4.22%		+0.64	4.96	2.35	3.49	1.78	0.227	0.225	0.177
KAYTUZ		10.668	8.958	10.308	β 115.75	1.356							
	HF	-0.16%	-2.00%	+5.72%		+3.54	4.13	2.73	7.54	5.05	0.321	0.215	0.476
HUYUOP		5.145	12.326	18.536		1.200							
	HF	-8.55%	+3.12%	+3.95%		+2.00%	6.82	2.48	2.19	1.42	0.499	0.404	0.321
BANGOM01		12.738	7.263	6.039	β 98.15	1.724							
	HF	+0.60%	+2.64%	+1.49%		+0.75	3.72	2.92	1.29	1.29	0.280	0.375	0.269
BANGOM02		12.101	7.373	12.890	β 95.89	1.667							
	HF	+1.14%	-1.82%	+1.18%		+1.57	3.03	1.88	4.49	4.49	0.251	0.204	0.149
CENRIW		24.215	6.981	6.147	β 91.70	1.236							
	HF	-0.64%	+1.35%	-1.92%		-0.77	3.18	2.29	8.97	4.19	0.173	0.200	0.119
HAMTIZ		12.569	4.853	17.266	β 99.16	1.528							
	HF	+1.38%	+1.06%	+1.28%		+0.46	3.11	1.36	3.89	2.75	0.119	0.125	0.093
ACSALA05		11.186	6.540	11.217	β 96.07	1.466							
	HF	+2.52%	+0.83%	+0.97%		+0.38	1.43	0.91	9.74	5.97	0.135	0.145	
	MP2	+0.29%	+2.87%	+1.70%		+0.38	4.27	2.03	8.01	5.08	0.164	0.153	0.150
acetylsalicylic acid, polymorph II		12.095	6.491	11.323	β 111.51	1.447							
	HF	+1.37%	+0.79%	+0.51%		-0.22	2.98	1.81	0.38	0.23	0.113	0.125	
	MP2	-0.47%	+2.94%	+1.54%		+0.67	7.22	3.31	6.27	3.20	0.190	0.132	0.105
IBPRAC01		14.397	7.818	10.506	β 99.7	1.176							
	HF	+2.76%	0.72%	+0.11%		-0.71	4.93	3.65	0.37	0.37	0.192	0.186	0.150
JEKNOC11		12.456	8.036	13.533	β 112.86	1.098							
	HF	-1.40%	+0.97%	-1.24%		-1.53	6.04	2.67	5.44	3.52	0.164	0.127	0.189
EYOBVAV		7.537	15.035	11.662	β 106.81	1.340							
	HF	+0.49%	+0.76%	-0.39%		+1.13	8.42	4.61	6.22	0.35	0.159	0.100	
	MP2	+0.77%	+0.82%	+0.36%		+1.18	7.25	3.96	5.26	1.92	0.166	0.128	0.079
KAMREW		7.296	9.484	16.020		1.433							
	HF	-4.43%	-1.83%	+2.80%		+3.70%	6.60	2.25	75.13	15.45	0.255	0.151	
	MP2	-2.47%	-1.52%	+1.82%		+2.30%	6.39	4.40	11.03	4.41	0.199	0.164	0.105
PEAMAN01		8.322	6.801	12.885	β 91.74	1.245							
	HF	-1.26%	+2.19%	+0.82%		+2.66	11.86	7.92	13.39	7.06	0.256	0.176	
	MP2	-0.32%	+2.73%	+0.23%		+2.86	12.30	8.78	9.79	5.50	0.284	0.188	0.134
LEKROI					α 95.42								
					+0.44								
					β 99.48	1.435							
	HF	+1.75%	+2.90%	-1.88%		-0.30	4.37	2.17	12.04	6.46	0.185	0.152	0.120
					γ 108.99								
					+0.15								
LEKRIC		9.997	10.347	12.680		1.268							
	HF	+4.30%	-1.33%	-1.15%		-1.74%	8.08	3.95	14.06	8.16	0.209	0.169	0.209
Systems for Which Crystal Structure Prediction Was also Performed													
PMACEPOx (p-salt)		11.008	6.539	12.160	β 116.01	1.146							
	HF	+0.35%	+0.95%	-1.07%		-0.44%	9.77	7.67	5.01	5.01	0.234	0.164	0.190
NMACEPO2 (n-salt form I)		5.797	15.444	17.073		1.179							
	HF	+5.26%	+1.06%	-4.26%		-1.78%	16.40	11.23	6.28	6.28	0.382	0.176	0.112
NMACEPO2 (n-salt form II)		5.941	15.469	17.501		1.121							
	HF	+5.08%	+0.70%	-5.29%		-0.27%	9.92	7.69	4.42	4.42	0.349	0.283	0.200

Table 3. Continued

structure	intra-molecular energy ^a	flexible-molecule							rigid-body				
		conventional unit cell				density (g cm ⁻³)	$\Delta\theta^c$ (deg)		$\Delta\theta_H^d$ (deg)		RMS ₁₅ ^e (Å)	RMS ₁₅ (Å)	
		a(Å)	b(Å)	c(Å)	angles ^b (deg)		max	mean	max	mean		ConOpt ^f	Expt ^g
CBMZPN11 (form I)					α 84.12 +0.94								
HF	5.171	20.574	22.245	β 88.01 -1.01	1.339	-1.87%	6.73	2.60	15.34	8.35	0.248	0.139	0.137
				γ 85.19 +0.94									
CBMZPN03 ^h (form II)	HF	35.454	35.454	5.253		1.235					0.113		
		+0.55%	+0.55%	-2.08%		+0.97%	1.34	0.81					
CBMZPN10 (form III)	HF	7.537	11.156	13.912	β 92.86	1.343							
		+1.87%	-0.37%	-1.70%	-0.28	+0.22%	1.21	0.79	4.88	4.33	0.169	0.176	0.177
CBMZPN12 (form IV)	HF	26.609	6.927	13.957	β 109.72	1.233							
		+0.54%	+0.49%	+2.68%	+1.55	+2.35%	0.60	0.48	4.98	4.12	0.198	0.183	0.122

^a Intramolecular energies and molecular geometries were computed (eq 4) with the 6-31G(d,p) basis set at the HF and MP2 levels. ^b Only angles not determined by space group symmetry are given. ^c Maximum/average absolute change of non-hydrogen torsion angles during refinement. ^d Maximum/average absolute change of hydrogen torsion angles during refinement. ^e Root-mean-square overlap of the 15-molecule coordination sphere of the experimental and minimized structures for the flexible-molecule lattice energy minimization (DMAflex). ^f Root-mean-square overlap of the 15-molecule coordination sphere of the experimental and minimized structures for rigid-body lattice energy minimization with the ab initio optimized conformations with the flexible torsions frozen to their experimental values (ConOpt). ^g Root-mean-square overlap of the 15-molecule coordination sphere of the experimental and minimized structures for rigid-body lattice energy minimization with the experimental conformations and XH bond lengths adjusted to standard neutron values for X-ray determinations (Expt). ^h Hydrogen atom positions have not been experimentally determined.

practically feasible) for large molecules where the intramolecular dispersion between distant functional groups determines the intramolecular energy surface.⁴⁴

In addition to reproducing the crystal and molecular structures, the DMAflex methodology should also provide realistic values for the energies and relative stabilities of known polymorphic forms. For the three polymorphic systems discussed in this section, aspirin form I (AC-SALA05) is predicted to be +0.18 kJ mol⁻¹ less stable than the metastable form II⁴⁵ (-0.08 kJ mol⁻¹ at the MP2 level), *n*-salicylidene-pentafluoroaniline form I (BANGOM01) is 2.20 kJ mol⁻¹ less stable than form II (BANGOM02), and the racemic form of the anti-inflammatory agent ibuprofen (IBPRAC01) is 5.93 kJ mol⁻¹ more stable than its stereoisomer (JEKNOC11). Although we are not aware of quantitative experimental relative stabilities for comparison, the predicted stability differences are plausible,⁴⁶ more so than earlier predictions with empirical force fields.¹⁴ The crystal energies of the known forms of the polymorphic systems (*R*)-1-phenylethylammonium-*(R/S)* 2-phenylpropanoate and carbamazepine were also predicted to be within a few kilojoules per mole and will be discussed in the context of crystal structure prediction in the next section.

3.2. Reranking of Hypothetical Crystal Structures.

3.2.1. Diastereomeric Salt Pairs: The Case of (*R*)-1-Phenylethylammonium (*R/S*)-2-phenylpropanoate. The separation of enantiomeric pairs is a challenging and important aspect of the pharmaceutical and fine chemical industries,⁴⁷ as their physical properties are identical and, in most cases, crystallization produces racemic crystals.⁴⁸ A frequently used separation process relies on the addition of a carefully chosen optically pure resolving acid or base that will produce a diastereomeric salt pair with sufficiently different solubilities (free energies).⁴⁹ In a recent publication,²⁵ we reported a methodology for the crystal structure prediction of such systems by performing rigid-body searches for low-lattice-energy structures for the *p*-salt (*R,S*) and *n*-salt (*R,R*) of the diastereomeric salt pair system 1-phenylethyl-

ammonium-2-phenylpropanoate.²⁵ In these searches, the ion conformations had the torsion angles θ_1 , θ_2 , and θ_3 (Table 2) constrained to values suggested by a statistical analysis of the CSD. The known *p*-salt structure⁵⁰ was predicted at the global minimum (columns 1 and 2 of Table 4). However, the thermodynamically stable form I of the *n*-salt⁵¹ was ranked 12th and was 7.4 kJ mol⁻¹ less stable than the global conformational minimum, which corresponded to the metastable form II⁵¹ (columns 1 and 2 of Table 5). Furthermore, it was predicted that the packing of *p*-salt structures is energetically favored by 11.6 kJ mol⁻¹ compared to that of *n*-salt structures, which does not agree with solution calorimetry measurements that suggest that the enthalpy of the *p*-salt is 3.9 kJ mol⁻¹ higher²⁵ than the enthalpy of the *n*-salt at 25 °C. Thus, although the search found all known forms within the low-energy region, the relative stability of the putative minima was not sufficiently accurate to assess the efficiency of (*R*)-1-phenylethylamine to resolve racemic 2-phenylpropanoic acid mixtures. In this section, we report the refinement of the rigid-body predictions for the 20 lowest-energy structures for each diastereomeric salt, by considering the effect of the packing forces on the most flexible degrees of freedom θ_1 , θ_2 , θ_3 , and θ_4 (Table 2) at the HF/6-31G-(d,p) level. These angles can vary by more than 40° with up to a 5 kJ mol⁻¹ increase in intramolecular energy.²⁵ The DMAflex refinement reproduces the three known enantiomorphous forms satisfactorily (Table 3), although the errors in lattice lengths are greater in the case of *n*-salt polymorphs. As demonstrated in the Supporting Information, further improvements in the reproduction accuracy will also require developments in the intermolecular potential, while small improvements are observed when the carboxylate angle is also included in the flexible degrees of freedom.

The refinement of the rigid-body putative structures considerably changes their relative energies, as shown in Figure 1 and Tables 4 and 5. In the case of the *p*-salt, the experimentally determined structure still corresponds to the global minimum, but for example, the 20th most stable

Table 4. Effect of Conformational Relaxation on the Relative Stability of (*R*)-1-Phenylethylammonium (*S*)-2-phenylpropanoate (p-salt) Putative Crystal Structures^a

rigid-body search		flexible-ion refinement										
rank, space group	U^b (kJ mol ⁻¹)	rank	ΔE^c		$U^d + \Sigma \Delta E^c$ (kJ mol ⁻¹)	\hat{V}^e (Å ³)	RMS ^f (Å)	anion			cation	
			anion	cation				θ_1 (deg)	θ_2 (deg)	RMS ^f (Å)	θ_3 (deg)	θ_4 (deg)
1, P2₁	-645.25	1	1.22	0.05	-655.00	394.95	0.118	-52.10	87.90	0.030	75.00	-64.84
								(-61.87)	(92.69)		(66.56)	(-69.85)
2, P2 ₁ 2 ₁ 2	-642.61	3	2.45	1.12	-650.22	368.99	0.232	-63.18	84.97	0.201	61.84	-61.40
3, P2 ₁ 2 ₁ 2 ₁	-636.89	10	4.28	0.98	-644.66	417.90	0.325	-69.40	89.07	0.151	66.85	-71.01
4, C2	-636.83	6	1.66	0.01	-647.82	416.34	0.129	-50.01	91.44	0.025	78.62	-62.65
5, P2 ₁ 2 ₁ 2 ₁	-634.97	13	4.35	0.91	-641.24	430.90	0.308	-67.73	93.01	0.151	66.66	-70.21
6, P2 ₁ 2 ₁ 2 ₁	-632.42	7	3.51	3.49	-647.17	413.47	0.310	-69.02	80.77	0.290	57.57	-78.62
7, P2 ₁ 2 ₁ 2 ₁	-631.42	14	1.74	0.01	-641.07	430.03	0.132	-50.47	91.69	0.022	78.46	-62.71
8, P2 ₁ 2 ₁ 2 ₁	-630.52	11	2.93	5.10	-643.38	402.43	0.192	-56.75	96.11	0.382	50.53	-78.85
9, C2	-630.40	9	2.43	5.86	-644.69	399.25	0.168	-54.56	94.50	0.438	45.53	-73.81
10, C2	-629.63	15	0.02	0.96	-639.98	427.20	0.014	-47.50	74.32	0.146	67.25	-71.17
11, C2	-628.26	4	1.88	0.01	-647.95	418.15	0.139	-51.28	92.49	0.018	75.50	-62.40
12, P2 ₁ 2 ₁ 2 ₁	-628.03	8	4.30	5.21	-646.42	396.15	0.212	-44.51	103.63	0.404	48.32	-75.40
13, C2	-627.58	5	1.71	0.00	-647.86	417.19	0.131	-49.93	91.75	0.002	76.96	-62.66
14, P2 ₁ 2 ₁ 2 ₁	-627.39	17	2.04	2.32	-638.24	411.04	0.196	-60.14	86.92	0.233	61.37	-75.66
15, P2 ₁ 2 ₁ 2 ₁	-627.23	20	3.44	0.97	-633.99	391.84	0.280	-66.36	87.94	0.174	63.85	-60.26
16, C2	-627.18	19	7.05	0.45	-635.47	415.30	0.414	-73.57	104.82	0.128	67.44	-63.53
17, P2 ₁ 2 ₁ 2 ₁	-626.69	12	4.73	1.24	-641.27	430.50	0.325	-68.74	94.18	0.181	64.52	-70.96
18, C2	-626.49	16	4.60	0.52	-639.77	373.32	0.378	-73.57	76.30	0.136	66.79	-62.43
19, C2	-626.14	18	1.54	0.02	-635.59	457.47	0.124	-56.55	82.07	0.029	71.64	-62.70
20, P2 ₁ 2 ₁ 2 ₁	-625.46	2	0.71	0.68	-652.13	392.50	0.154	-35.28	71.14	0.132	67.88	-68.98

^a The row in bold corresponds to the experimentally determined form with the experimental values of the intramolecular degrees of freedom shown in parentheses. ^b Where intermolecular lattice energy was calculated with a 15 Å cutoff distance for the repulsion–dispersion and higher multipole moment interactions; the rigid-body ion conformations were fixed at the MP2/6-31G(d,p) conformational minimum with the phenyl and carboxylate rotation constrained to the CSD average values and the atomic multipoles derived from the MP2/6-31G(d,p) charge density (ref 25). ^c HF/6-31G(d,p) intramolecular energy for the ion conformations at the crystal energy minimum. ^d Intermolecular lattice energy at the crystal lattice energy minimum with atomic multipoles derived from the MP2/6-31G(d,p) charge density and 60 Å cutoff distance for the repulsion–dispersion and higher multipole moment interactions; because of the differences in cutoff distance and level of theory for the determination of rigid degrees of freedom, the starting energies for the DMAflex refinement may differ from the energies in the second column by up to 5 kJ mol⁻¹. ^e Cell volume per ion pair. ^f All-atom ion root-mean-square discrepancy from the HF/6-31G(d,p) global conformational minimum ($\theta_1 = -46.84^\circ$, $\theta_2 = 72.64^\circ$, $\theta_3 = 76.82^\circ$, and $\theta_4 = -62.67^\circ$) due to the effect of the packing forces.

Table 5. Effect of Conformational Relaxation on the Relative Stability of (*R*)-1-Phenylethylammonium (*R*)-2-Phenylpropanoate (n-salt) Putative Crystal Structures

rigid-body search		flexible-ion refinement										
rank, space group	U^b (kJ mol ⁻¹)	rank	ΔE^c		$U^d + \Sigma \Delta E^c$ (kJ mol ⁻¹)	\hat{V}^e (Å ³)	RMS ^f (Å)	anion			cation	
			anion	cation				θ_1 (deg)	θ_2 (deg)	RMS ^f (Å)	θ_3 (deg)	θ_4 (deg)
1, P2₁2₁2₁	-633.73	3	5.94	1.54	-646.25	403.04	0.389	72.86	-97.14	0.166	66.54	-74.92
								(76.26)	(-87.39)		(56.62)	(-70.50)
2, P2 ₁ 2 ₁ 2 ₁	-632.30	7	2.18	3.49	-642.20	394.06	0.237	64.04	-79.09	0.333	52.55	-67.48
3, P2 ₁ 2 ₁ 2 ₁	-631.13	4	5.76	1.71	-646.16	403.37	0.388	73.10	-95.07	0.165	67.02	-76.11
4, P2 ₁ 2 ₁ 2 ₁	-628.95	10	0.91	0.79	-639.80	428.16	0.096	49.27	-86.37	0.142	67.20	-69.51
5, P2 ₁	-628.31	12	1.94	2.14	-639.56	414.20	0.140	50.95	-92.93	0.195	64.90	-77.39
6, P2 ₁ 2 ₁ 2 ₁	-628.11	15	3.05	1.65	-637.10	403.42	0.263	65.35	-86.48	0.230	60.09	-66.51
7, P2 ₁ 2 ₁ 2 ₁	-627.84	19	4.92	0.93	-635.12	405.22	0.284	64.08	-101.28	0.173	64.53	-67.23
8, P2 ₁ 2 ₁ 2 ₁	-627.43	6	0.44	0.59	-644.17	417.16	0.108	54.82	-74.94	0.171	89.46	-64.42
9, P2 ₁ 2 ₁ 2	-627.38	8	3.77	1.93	-641.65	374.47	0.325	70.08	-80.34	0.167	64.92	-49.93
10, P2 ₁ 2 ₁ 2 ₁	-627.16	13	2.30	0.00	-637.60	398.84	0.155	51.95	-94.77	0.004	76.60	-62.28
11, C2	-627.11	14	4.22	1.64	-637.45	414.32	0.341	70.95	-84.07	0.182	65.19	-74.63
12, P2₁2₁2₁	-626.35	1	2.02	0.01	-655.88	389.15	0.222	59.80	-56.88	0.021	75.25	-62.36
								(71.13)	(-73.28)		(69.28)	(-56.08)
13, P2 ₁ 2 ₁ 2 ₁	-625.70	11	0.40	1.21	-639.63	372.43	0.101	54.26	-75.82	0.196	62.79	-67.53
14, P2 ₁	-625.50	2	0.24	1.43	-651.42	385.96	0.056	44.17	-65.93	0.191	63.89	-71.87
15, P2 ₁ 2 ₁ 2 ₁	-625.36	9	7.54	0.35	-639.80	397.69	0.523	81.24	-102.82	0.106	68.98	-60.31
16, P2 ₁ 2 ₁ 2 ₁	-625.35	16	5.63	2.12	-636.70	409.15	0.416	75.57	-88.33	0.169	64.94	-48.83
17, P2 ₁	-625.20	17	1.16	1.12	-635.76	384.77	0.164	58.71	-79.81	0.181	64.18	-69.20
18, P2 ₁	-625.13	18	1.24	1.13	-635.75	384.68	0.170	59.16	-79.79	0.183	64.02	-69.07
19, P2 ₁ 2 ₁ 2 ₁	-625.03	5	0.32	0.35	-644.44	415.13	0.095	40.45	-76.50	0.107	69.30	-65.91
20, P2 ₁	-624.90	20	3.82	2.83	-634.66	422.97	0.331	70.50	-78.86	0.275	58.00	-75.20

^a The rows in bold correspond to the experimentally determined forms with the experimental values of the intramolecular degrees of freedom shown in parentheses. ^{b–f} As in Table 4.

structure in the rigid-body search becomes the second, and its energy relative to the global minimum is reduced from 19.8 to 2.9 kJ mol⁻¹. In the case of the n-salt, the changes in the ranking order are even more pronounced. The simultaneous relaxation of the intramolecular degrees of freedom brings the 12th most stable structure to the global minimum, which is in accord with experimental measure-

ments, as it corresponds to the thermodynamically stable polymorph. Furthermore, the metastable polymorph n-salt II becomes the third most stable structure, 9.6 kJ mol⁻¹ higher in energy. The second most stable minimum on the crystal energy surface, which was ranked 14th in the rigid-body search, has a distinct packing of the same hydrogen-bonding ladder as in the global minimum. The relaxation of

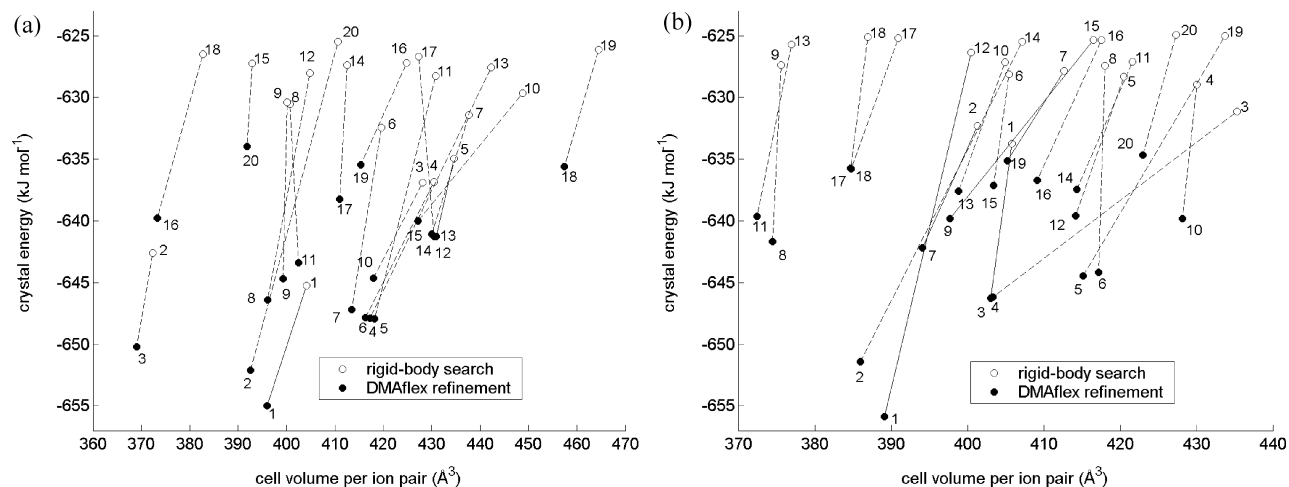


Figure 1. Refinement (full circles) of the 20 most stable crystal structures identified by the rigid-body search (open circles, ref 25) for the (a) p-salt and (b) n-salt of the diastereomeric salt pair (*R*)-1-phenylethylammonium (*R/S*)-2-phenylpropanoate. The solid lines correspond to experimentally determined polymorphs and the dashed lines to hypothetical low-energy crystal structures.

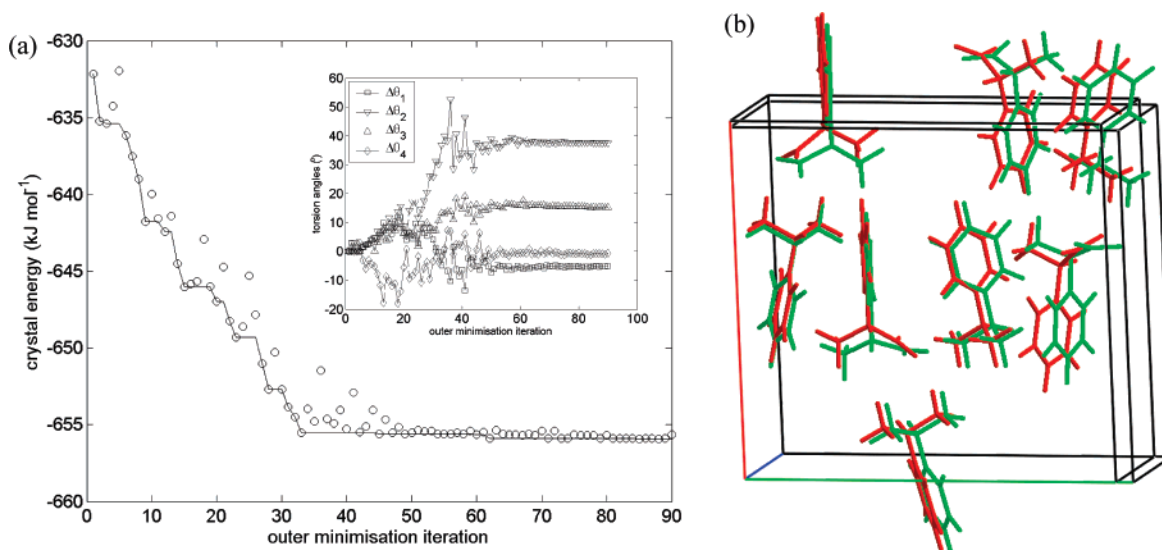


Figure 2. Simultaneous relaxation of the lattice variables and ion conformations starting at the 12th n-salt rigid-body minimum (corresponds to n-salt I polymorph). (a) The evolution of the crystal energy as a function of the outer minimization iteration (open circles; continuous line corresponds to the lowest-energy achieved up to the corresponding step). The inset illustrates the changes in the flexible torsion angles during refinement. (b) The overlay of the starting rigid-body structure (green) and flexible-ion lattice energy minimum (red).

the molecular geometries reduces the number of putative structures as some of the rigid-body minima lead to the same minimum. In the case of the n-salt, the global and third rigid-body minima differed by an RMS_{15} of 0.73 \AA , but the DMAflex refinement led to the same minimum, involving significant adjustments in the cell lengths and volume (Table 5). Similarly, the following clusters of rigid-body minima led to the same crystal energy minimum: n-salt 17th and 18th; p-salt 4th, 11th, and 13th; and p-salt 5th and 17th. In addition to the reduction of the number of putative structures, the simultaneous relaxation of the ion conformations increases the energy range for the 20 putative structures to 21.0 kJ mol^{-1} and 21.2 kJ mol^{-1} for the p-salt and n-salt, respectively, which leads to greater energetic separation between the known and hypothetical structures.

The simultaneous relaxation of the ion conformations also improves the calculated relative stability of the diastereomeric

salts. The thermodynamically most stable n-salt polymorph becomes 0.9 kJ mol^{-1} more stable than the p-salt structure, which is in reasonable agreement with the experimental value of 3.9 kJ mol^{-1} on the basis of solution calorimetry measurements.²⁵

Figure 2 illustrates the changes that take place during a typical refinement for the rigid-body minimum corresponding to n-salt I polymorph. In accord with earlier observations regarding the significant sensitivity of the intermolecular lattice energy to the fine details of the ion conformations,²⁵ the crystal energy is reduced by approximately 23 kJ mol^{-1} , although the only significant change in conformation is in the rotation of the carboxylate group (θ_2), while the overall changes in the crystal structure are modest. Although the torsion angle changes for some structures in Tables 4 and 5 can exceed 20° , the average absolute changes for the torsion angles θ_1 , θ_2 , and θ_3 during refinement are 9.83° , 10.07° ,

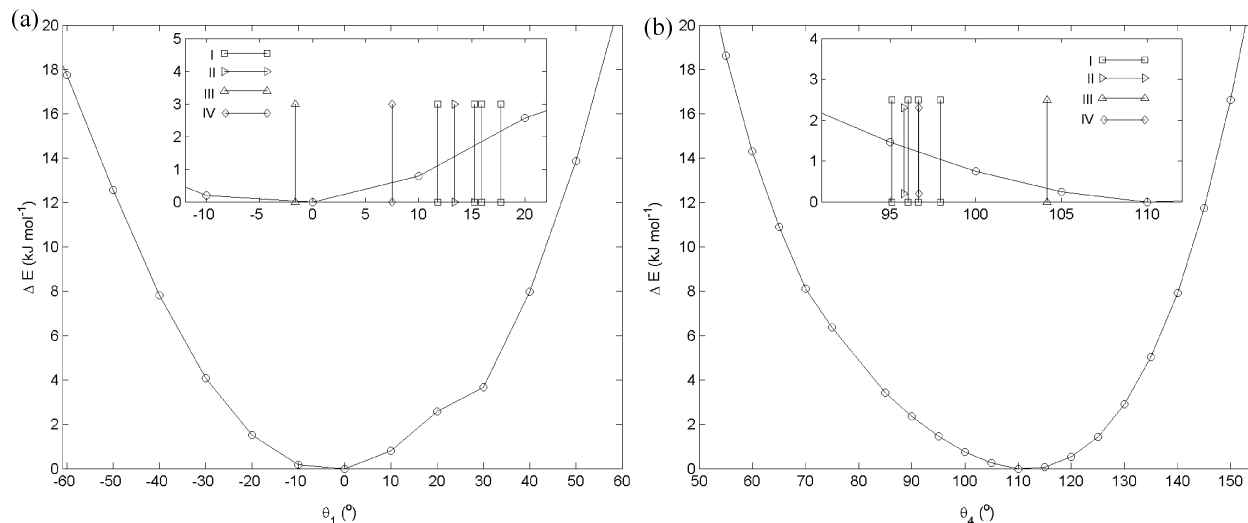


Figure 3. One-dimensional intramolecular energy variation as a function of the (a) rotation θ_1 and (b) tilting θ_4 of the carboxamide group with respect to the seven-member ring at the B3LYP/6-31G(d,p) level of theory for carbamazepine. The conformational energy profiles were computed by constraining the scanning torsion and optimizing the rest of the molecular geometry. The insets show a magnification over the low-energy regions with the values corresponding to experimental conformations indicated. The discontinuity at $\theta_1 = 30^\circ$ in part a is because the amine hydrogen atoms switch between two local minima (see Figure 4).

and 8.62° , respectively. The change in the ion torsion angles from the starting conformations appeared to be largely determined by the packing forces, as shown by the changes in the hydrogen-bonding geometries during the refinement (Tables S2 and S3 in Supporting Information) which approach their typical CSD values^{52–55} for the majority of hypothetical structures. For example, the significant reranking of the structure corresponding to the n-salt I polymorph is associated with a 0.15 \AA reduction in the average N–H \cdots O distance, while the H \cdots O=C angle approaches its ideal value for all three hydrogen bonds to a single cation. Although the changes in ion conformation are mainly driven by the improvement in the hydrogen-bond geometries, there is also a dispersion contribution to the stabilization because most structures become denser, on average by 2%.

3.2.2. Carbamazepine. Extensive research on the solid state of carbamazepine (Table 2), a drug for the treatment of epilepsy and trigeminal neuralgia, has led to the crystallographic characterization of four polymorphs and several solvates and cocrystals.^{26,56–63} A recent rigid-body crystal structure prediction study²⁶ [with the MP2/6-31G(d,p) optimized geometry] indicated that two energetically competitive C=O \cdots N–H hydrogen-bonding motifs [dimer $R_2^2(8)$ and chain $C(4)$ motifs] equally populate the set of low-energy minima. However, dimers are observed in all four polymorphs, solvates, and cocrystals. Whether the predicted stabilities of the hypothetical chain structures are artifacts of the rigid-body search²⁶ can be established by the DMAflex refinement of the most stable rigid-body putative crystal structures.

Carbamazepine comprises a rigid skeleton formed by two phenyl rings fused at the positions 3 and 4 of each side of a 5-azacycloheptene (azepine) ring, which appears conformationally locked in a boat configuration, as the angle between the two phenyl group planes varies little in the determined polymorphs (mean 54.05° , standard deviation 2.02°) and ab initio optimized conformations [53.77° and 48.75° for the

HF/6-31G(d,p) and B3LYP/6-31G(d,p) levels of theory, respectively]. However, the geometry of the pendant CONH₂ group will be determined by the delicate balance of inter- and intramolecular forces. A recent spectroscopic study⁶⁴ revealed that the polymorph-sensitive IR modes are localized to the latter group and show the greatest disparity from the theoretical spectra, suggesting that the crystalline forces perturb its in vacuo geometry and relative position with respect to the carbamazepine backbone. A frequency analysis for an isolated molecule at the B3LYP/6-311** $(2d,2p)$ level shows that the lowest and third vibrational modes at 56.08 cm^{-1} and 85.35 cm^{-1} predominately correspond to the tilting and rotation of the carboxamide group. A conformational analysis revealed that the rotation (θ_1) and tilting (θ_4) of the carboxamide group with respect to the seven-membered ring can both vary in a range of approximately $50\text{--}60^\circ$ with less than a 4 kJ mol^{-1} increase in intramolecular energy, as shown by the relaxed scans at the B3LYP/6-31G(d,p) level in Figure 3. This agrees with an analysis of 16 carbamazepine solvates and cocrystals (CSD version 5.27, November 2005), in which θ_1 was found to vary in a range of 22.1° with an average value of 9.8° , while θ_4 varied across 24.6° with an average value of 97.3° . This shows a wider variation but the same systematic packing effect on the molecular conformation shown for the four polymorphs in the insets of Figure 3.

The set of flexible degrees of freedom should also include the independent rotation of the two amide hydrogen atoms,⁶⁵ which will affect the hydrogen-bond geometries and, hence, the relative stability of the hypothetical crystal structures. The intramolecular energy surface for the rotation of the amide hydrogen atoms at the B3LYP/6-31G(d,p) level exhibits two conformational minima: in the global minimum, the hydrogen atoms are pointing toward the seven-membered ring, whilst the second minimum is marginally less stable (1 kJ mol^{-1}) with the hydrogen atoms pointing in the opposite direction (Figure 4). All experimental structures with hydrogen atoms exhibit almost planar amide group geometries,

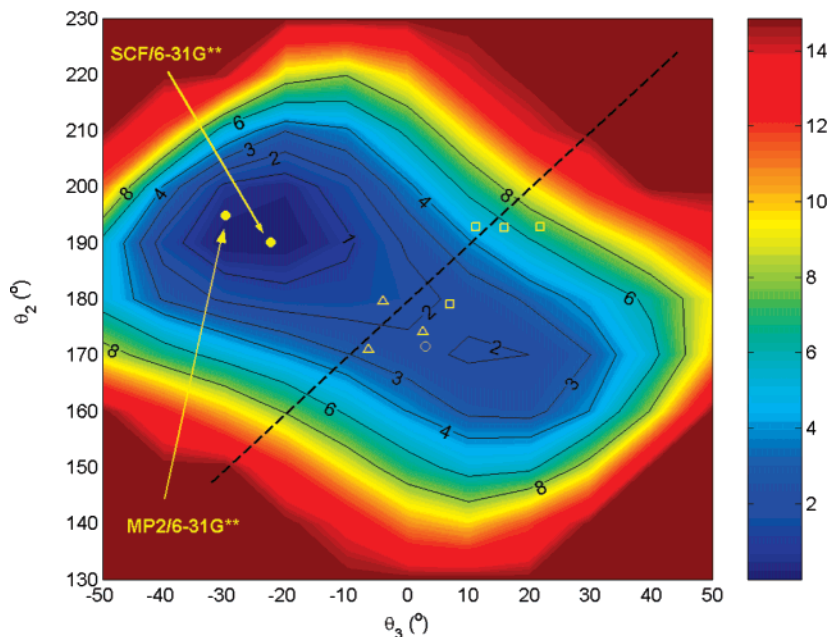


Figure 4. Two-dimensional intramolecular energy variation as a function of the independent rotation of the two amide hydrogen atoms (torsion angles θ_2 and θ_3) at the B3LYP/6-31G(d,p) level of theory for carbamazepine. The conformational energy surface was computed by constraining torsion angles θ_2 and θ_3 and optimizing the rest of the molecular conformation. Full circles correspond to ab initio global minima at different levels of theory, open triangles to the three determinations of the *P*-monoclinic polymorph (form III: CBMZPN10, CBMZPN01, and CBMZPN02), open squares to the four molecules present in the asymmetric unit of the triclinic polymorph (form I: CBMZPN11), and open circles to the *C*-monoclinic polymorph (form IV: CBMZPN12). Planar CNH₂ conformations lie on the black dashed line, which separates the two different pyramidalization configurations, with the top left having the hydrogen atoms pointing to the seven-membered ring.

although no definite conclusions can be made because of X-ray limitations.⁶⁶

The DMAflex reproduction quality for all four carbamazepine polymorphs is satisfactory (Table 3). The refined flexible torsion angles are within a few degrees of their values in the crystal and significantly different from those in the unconstrained ab initio minima (Figures 3 and 4). This suggests that these torsion angles are determined by a fine balance of intra- and intermolecular forces, which is well-described by our model. The energy reranking due to the refinement of the rigid-body search results improves the relative stability of the known structures (Figure 5). The global minimum still corresponds to a putative chain structure,²⁶ but the energy difference from the thermodynamically stable form III is reduced from 2.2 kJ mol⁻¹ to 0.9 kJ mol⁻¹. Moreover, form IV becomes the 11th most stable structure, 5.0 kJ mol⁻¹ higher in energy than the global minimum, whereas in the rigid-body search, it was ranked 25th (9 kJ mol⁻¹ above the global minimum). The rigid-body minimizations predicted that the triclinic form I and trigonal form II had lattice energies of +7.3 and +9.7 kJ mol⁻¹, respectively, relative to the global minimum; that is, their ranks would respectively be 18 and 30 if we assume that no other structures would be found under a more exhaustive search. When the molecular flexibility is accounted for, forms I and II become the 9th and 15th most stable structures, respectively, at +4.4 and +6.8 kJ mol⁻¹ higher in energy than the global minimum.

On the basis of differential scanning calorimetry and heat of solution measurements, it has been deduced that the most

stable polymorph is the monoclinic form III, followed by the triclinic form I (+1.34⁶⁰ to +3.00⁶⁷ kJ mol⁻¹), the *C*-centered monoclinic IV (+1.93⁶⁰ kJ mol⁻¹), and finally the loosely packed trigonal form II (+2.89⁶⁰ kJ mol⁻¹). The stability order follows the density order at room temperature, which is expected given that all of the forms exhibit the same hydrogen-bonded dimers through the carboxamide donor and acceptor and differ in the way the aromatic rings interact and the pattern of weak C–H···O interactions. After the refinement, the predicted stability order of the four polymorphs is in reasonable agreement with experimental evidence (I +3.3, IV +4.17, and II +5.9 kJ mol⁻¹ with respect to form III; Table 6), although the enthalpy differences are slightly overestimated.

It is encouraging that in a few cases the flexible molecule minimization inverts the direction of the amide hydrogen atoms. The DMAflex refinement of the 25th most stable structure (which corresponds to the known form IV) alters the torsion angles θ_2 and θ_3 by 27° and 38°, respectively, with the hydrogen atoms pointing away from the ring at the minimum in good agreement with the experimental values (see Table 6 and Figure S1 of the Supporting Information). The significant adjustment of the amide hydrogen atoms construes the considerable reranking of this structure, and hence, the DMAflex refinement seems to have overcome any bias from the initial rigid-body search. Overall, after refinement, the ranges in which the torsion angles θ_1 , θ_2 , θ_3 , and θ_4 of the search structures vary are 22.0, 32.8, 37.7, and 12.8° and are comparable to the ranges observed for the known polymorphs and solvate crystals. The changes in lattice

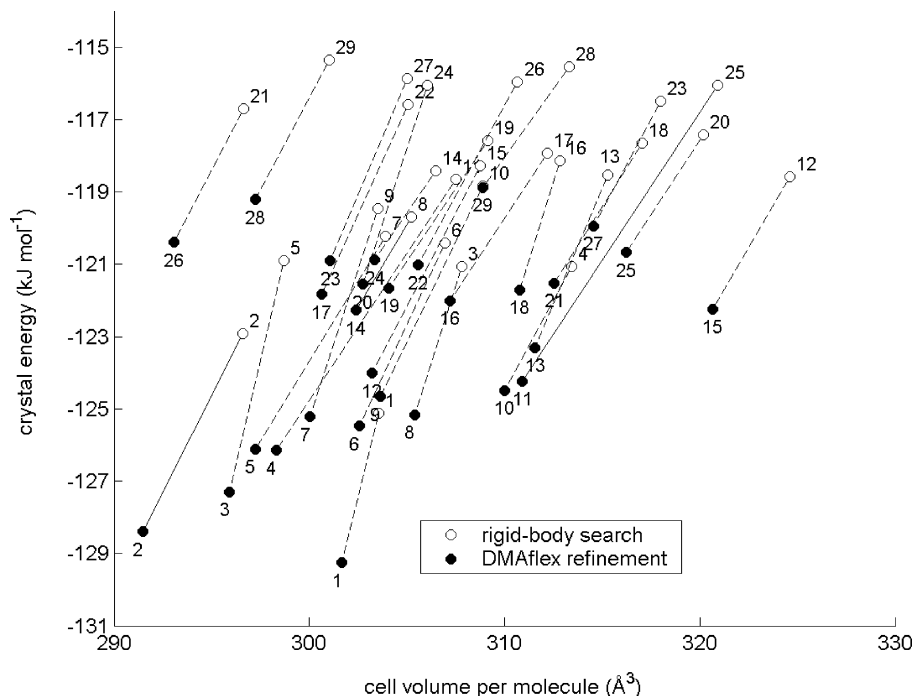


Figure 5. Refinement (full circles) of the 29 most stable crystal structures identified by the rigid-body search (open circles, ref 26) for carbamazepine. The solid lines correspond to experimentally determined polymorphs and the dashed lines to hypothetical low-energy crystal structures.

lengths are modest and do not exceed 0.5 Å, while the average reduction in cell volume is 1.4%. In contrast with the diastereomeric salt case, the refinement does not always improve the hydrogen-bond geometry (see Table S4 in the Supporting Information), which suggests that the changes in conformation are strongly driven by the stronger dispersion arising from a denser lattice.

4. Discussion

In molecular crystals, the intermolecular bonding energies are significantly weaker than the energy of typical covalent energies, and thus, the packing-induced molecular distortions are often limited.²⁸ Hence, using the *ab initio* optimized molecular structures is often a good approximation to the conformations in crystals, and crystal structure prediction studies using such molecular structures as rigid provide a good starting point.^{43,68,69} Nevertheless, this work has demonstrated that the consideration of molecular flexibility, such as NH_2 geometries, within the lattice energy minimization significantly alters the ranking of the putative structures. The DMAflex approach addresses this issue by using quantum mechanical calculations to estimate the molecular deformation energy. This eliminates the inaccuracies associated with the use of empirical intramolecular potentials due to their oversimplified functional form and the fact that they have often been parametrized in conjunction with different intermolecular models than those actually used in the modeling (or even use an electrostatic model derived for intermolecular interactions to evaluate intramolecular energy contributions^{14,70}). Moreover, DMAflex uses a realistic distributed multipole representation of the charge density for the intermolecular electrostatic contributions, to ensure that the orientation dependence of strong directional interactions,

such as hydrogen bonds and π - π stacking, is modeled accurately. The conformational dependence of the charge distribution, reflecting the through-space polarization effects due to changes in the relative positions of the functional groups, is automatically accounted for by the recalculation of the distributed multipoles at each outer minimization step. The minimization of a wide set of experimentally determined crystal structures demonstrates that the approach offers substantial accuracy improvements in comparison with existing methodologies.¹⁴

Although the DMAflex methodology constitutes a significant improvement over empirical inter- and intramolecular force fields, its applicability is restricted by the computational cost and the limited degree of molecular flexibility that can be practically considered without an analytical functional form for the derivatives of the crystal energy, E^{cryst} , with respect to conformation. The need to restrict the number of conformational variables that are optimized within the crystal energy minimization is satisfied by partitioning the intramolecular degrees of freedom into rigid and flexible, which is to some extent arbitrary and relies on chemical intuition. Moreover, the computational cost is significant even when the effect of the packing forces is restricted to a limited set of easily deformable torsion angles. For example, the optimization of the salt structure illustrated in Figure 2 involved two SCF optimizations and two MP2 charge density calculations for each of the 80 outer minimization steps required to reach convergence for the four conformational variables. The computing time for a crystal energy minimization depends on the molecular size and number of conformational degrees of freedom and, for the results reported in this paper, varied from a few hours to two weeks when executed serially on a modern workstation.

Table 6. Effect of Conformational Relaxation on the Crystal Structure Prediction of Carbamazepine^a

rigid-body search		flexible-molecule refinement							
rank, space group, graph set	U^b (kJ mol ⁻¹)	rank	ΔE^c	$U^b + \Delta E^c$ (kJ mol ⁻¹)	\bar{V}^e (Å ³)	θ_1 (deg)	θ_2 (deg)	θ_3 (deg)	θ_4 (deg)
1, <i>P2₁/c</i> , C4	-125.11	1	0.169	-129.25	301.68	6.16	-171.39	-18.15	101.50
2, <i>P2₁/c</i>, R₂²(8)	-122.92	2	1.538	-128.40	291.48	-1.27	175.77	0.87	102.96
form III						(-1.64)	(179.54)	(-4.01)	(104.17)
3, <i>P2₁/c</i> , C4	-121.06	8	0.171	-125.16	305.39	5.63	-171.39	-17.67	101.97
4, <i>P2₁2₁2₁</i> , C4	-121.06	10	0.128	-124.49	309.99	1.66	-168.59	-19.94	102.65
5, <i>P1</i> , R ₂ ² (8)	-120.90	3	2.604	-127.29	295.92	4.47	173.54	13.65	104.69
6, <i>Pbca</i> , C4	-120.42	6	0.574	-125.48	302.55	6.06	-172.74	-13.00	103.16
7, <i>P1</i> , R ₂ ² (8)	-120.24	5	2.416	-126.11	297.22	10.03	-178.06	5.84	99.60
8, <i>P1</i> , R ₂ ² (8)	-119.69	14	0.220	-122.28	302.39	3.83	-165.75	-21.96	104.26
9, <i>P2₁/c</i> , R ₂ ² (8)	-119.46	7	1.826	-125.21	300.05	5.43	-177.84	7.16	101.81
10, <i>P2₁/c</i> , C4	-118.85	9	0.697	-124.66	303.63	5.90	-174.06	-10.98	102.77
11, <i>Pna2₁</i> , C2	-118.64	4	3.374	-126.14	298.32	9.98	-158.92	-22.32	109.10
12, <i>P2₁/c</i> , C4	-118.58	15	0.120	-122.24	320.65	-0.13	-167.96	-20.81	105.67
13, <i>P1</i> , R ₂ ² (8)	-118.53	13	0.492	-123.30	311.56	7.88	-165.46	-20.21	98.57
14, <i>P1</i> , none	-118.42	24	0.197	-120.88	303.36	6.65	-168.04	-24.04	102.88
15, <i>P2₁/c</i> , R ₂ ² (8)	-118.27	12	2.391	-124.01	303.21	8.21	-179.89	7.82	101.26
16, <i>P2₁/c</i> , R ₂ ² (8)	-118.14	18	0.345	-121.71	310.77	8.57	-167.52	-20.44	97.94
17, <i>C2/c</i> , R ₂ ² (8)	-117.94	16	1.591	-122.01	307.19	-4.61	-162.75	-23.62	103.93
18, <i>P1</i> , R ₂ ² (8)	-117.66	21	0.328	-121.52	312.55	-0.23	-173.89	-17.49	106.52
19, <i>P2₁/c</i> , C4	-117.59	19	0.778	-121.66	304.08	3.45	-165.09	-23.85	98.27
20, <i>P2₁</i> , C4	-117.43	25	0.135	-120.67	316.22	0.33	-167.75	-22.79	104.34
21, <i>P2₁/c</i> , C4	-116.70	26	0.929	-120.39	293.07	-3.44	-170.52	-10.41	108.04
22, <i>P2₁/c</i> , C4	-116.59	17	0.996	-121.83	300.63	6.20	-173.72	-10.73	104.88
23, <i>P2₁/c</i> , R ₂ ² (8)	-116.49	27	0.241	-119.95	314.55	6.22	-166.98	-22.35	99.49
24, <i>P2₁/c</i> , R ₂ ² (8)	-116.05	20	1.920	-121.55	302.73	4.82	-179.56	3.97	103.94
25, <i>C2/c</i>, R₂²(8)	-116.05	11	1.820	-124.23	310.91	8.10	168.26	7.98	96.33
form IV						(7.50)	(171.51)	(3.00)	(96.62)
26, <i>P2₁/c</i> , R ₂ ² (8)	-115.97	22	2.550	-121.03	305.58	6.69	-178.19	5.34	104.32
27, <i>Pbca</i> , C4	-115.87	23	0.833	-120.91	301.07	-11.94	-172.71	-11.94	106.29
28, <i>Pbca</i> , C4	-115.53	29	0.216	-118.88	308.88	2.24	-166.37	-22.96	102.59
29, <i>Pna2₁</i> , C4	-115.35	28	0.953	-119.22	297.23	-3.02	-170.44	-10.08	106.46
Known Polymorphs out of the Scope of the Rigid-Body Search									
form II		3.03	-122.49	314.54		12.04	179.34	15.77	95.54
						(13.38)			(95.82)
form I		5.04	-125.09	298.51		13.56	-175.12	16.51	97.50
						(17.74)	(-167.15)	(21.74)	(97.23)
		3.05				11.46	-178.68	10.94	98.24
						(15.22)	(-167.20)	(11.28)	(96.05)
		1.35				11.40	-173.40	-7.11	97.88
						(11.79)	(178.98)	(7.03)	(96.66)
		3.11				9.18	177.29	20.49	97.17
						(15.91)	(-167.37)	(15.85)	(95.12)

^a The rows in bold correspond to the experimentally determined forms found in the search with the experimental values of the intramolecular degrees of freedom shown in parentheses. ^b Intermolecular lattice energy calculated with the MP2/6-31G(d,p) global conformational minimum ($\theta_1 = -0.46^\circ$, $\theta_2 = -165.19^\circ$, $\theta_3 = -29.80^\circ$, and $\theta_4 = 109.50^\circ$) and atomic multipoles derived from the MP2/6-31G(d,p) charge density with a 15 Å cutoff distance for the repulsion–dispersion and higher multipole moment interactions (ref 26). ^c HF/6-31G(d,p) intramolecular energy for the conformation at the crystal energy minimum. The flexible torsions at the HF/6-31G(d,p) global minimum are $\theta_1 = 2.98^\circ$, $\theta_2 = -169.85^\circ$, $\theta_3 = -22.36^\circ$, and $\theta_4 = 103.68^\circ$. ^d Intermolecular lattice energy at the crystal energy minimum with atomic multipoles derived from the MP2/6-31G(d,p) charge density and a 60 Å cutoff distance for the repulsion–dispersion and higher multipole moment interactions. ^e Cell volume per molecule.

The computational cost could be reduced if the rigid degrees of freedom were kept frozen to their in vacuo values $\theta^{\text{r,vac}}$ by replacing the ab initio optimization in eq 4 with a single-point intramolecular energy evaluation:

$$\Delta E^{\text{intra}}(\theta^f) = E^{\text{intra}}(\theta^f, \theta^{\text{r,vac}}) - E^{\text{vac}} \quad (5)$$

Although this approach appears computationally attractive, we have found that it overestimates the molecular energy if the flexible degrees of freedom deviate significantly from their in vacuo values. Moreover, the ab initio optimization of the rigid degrees of freedom does not prohibitively increase the computational cost because the conformational changes are sufficiently modest so that the previously converged conformation provides an excellent starting point. A really significant decrease in computational cost could be obtained by reducing the number of iterations, and hence the number of quantum mechanical calculations, needed to reach convergence. This could be achieved by using a

gradient-based optimization algorithm, such as a Broyden–Fletcher–Goldfarb–Shanno scheme, but requires the analytical evaluation of the lattice energy gradients with respect to conformation. This preliminary study demonstrates that such an algorithm will be a significant advancement toward the reliable prediction of the structure of crystals containing flexible molecules and would also allow the estimation of realistic vibrational modes on an atomistic level and, hence, free energies. At the same time, it should also be possible to optimize all degrees of freedom within the crystal energy minimization.

The desired goal of optimization of the cell and all atomic coordinates, avoiding the inter/intra and rigid/flexible partitioning could in principle be achieved by computing the crystal energy entirely at the quantum mechanical level with periodic density functional theory. However, in addition to the prohibitive computational cost,⁷¹ recent studies on the binding of molecular clusters^{72–74} and crystals^{71,75} questions

the ability of commonly used functionals to quantitatively predict the binding energy and geometries of dispersion-bound complexes, although empirical corrections to the van der Waals interaction energies have been recently proposed.^{76–78} An alternative approach to avoid the use of atom–atom potentials for the intermolecular energy by relying on numerical integration over the ab initio charge densities^{79,80} can, at present, only be applied for the evaluation of the lattice energy of rigid molecules. Hence, a hybrid methodology, such as DMAflex, appears to be the most currently viable approach to combine a realistic intermolecular force field with accurate models for the molecular geometry and intramolecular energy.

One problem that was overcome in the course of this study was that a few DMAflex optimizations ended in oscillations of less than 0.5 kJ mol⁻¹ in the crystal energy. This was due to discontinuities in the electrostatic interactions introduced by the distributed multipole analysis⁹ moving the charge density contributions from the product of primitives on different atoms to the nearest nucleus. This sometimes leads to contributions to moments being switched to different nuclei when two functional groups within the molecule are in close, but changing, proximity, as with the amino and hydroxyl groups in NOREPH01. The use of a recently proposed revised multipole analysis,¹⁰ which uses numerical integration for the diffuse functions, solves this problem and was used for the structure reproductions of the structures ATUVIU, NOREPH01, GAHP10, ACYGLY11, IBPRAC01, and JEKNOC11.

The importance of DMAflex refinements of crystal structures comes from their ability to reproduce the molecular deformations that lead to energetically more favorable hydrogen-bonding geometries and denser lattices than can be achieved using a rigid molecular geometry. However, it is essentially a local minimization method, and hence, it will only provide the minimum closest to the starting crystal structure. For significantly flexible systems, all low-energy minima need to be approximately located prior to refinement by using empirical force fields²³ or a set of rigid-body geometries,⁶⁸ whose number depends on the complexity of the intramolecular and crystal energy surfaces. In the case of carbamazepine, the refinement gave structures spanning the low-energy conformational space (Table 6) starting from just the ab initio optimized minimum. This would clearly not be sufficient to predict the conformational polymorphism of piracetam,⁶⁸ though the successful blind prediction of form IV could have been achieved with far fewer rigid-body searches had DMAflex refinement been used. The complexity of the E^{cryst} landscape in the case of the diastereomeric salt pair system suggests that more rigid-body searches and more DMAflex refinements over a wider energy range might well find additional structures that are energetically competitive with the known forms.

The realistic modeling of the molecular deformation under the crystalline forces constitutes an important development toward the reliable prediction of the thermodynamic stability of putative and known crystal structures. However, further improvements are necessary for the reliable prediction of thermodynamic stability: free energy differences of 3–4 kJ

mol⁻¹ correspond to a solubility ratio of 2:1, and so, such accuracy is needed for the theoretical screening of resolving agents. Although the stability order of the carbamazepine polymorphs is correctly reproduced, the global minimum still corresponds to a catemeric structure whose existence has not been experimentally confirmed despite the strenuous experimental screening.²⁶ We are currently investigating whether the development of ab initio repulsion–dispersion intermolecular potentials and the modeling of the polarization of the molecular charge density by the crystalline environment are likely to change the ranking of low-energy crystal structures even further. Further research is also needed in evaluating the accuracy of quantum mechanical estimates for the intramolecular energy, because of intramolecular basis set superposition errors and an inaccurate description of the dispersion interactions between distant functional groups.⁴⁴ For flexible systems, there is an even greater challenge in using sufficiently accurate energy models for realistic estimates of the entropic and zero-point energy contributions on one hand and thermal expansion on the other. However, the accurate modeling of the free energy alone is not likely to be sufficient for the successful prediction of the organic solid state, as concomitant polymorphism⁸¹ shows that crystallization is not always thermodynamically controlled. Nevertheless, the exact extent to which kinetic effects can determine the crystallization outcome cannot be assessed without accurate thermodynamic models that reliably rank theoretically derived structures and limit the subset of these that should be considered as potential polymorphs.^{13,24,78}

5. Conclusions

This paper presents a novel methodology for the lattice energy minimization of crystal structures which contain molecular entities whose conformation may be distorted under the packing forces. The proposed approach aims to address the limitations of rigid-body minimizations when the lattice energy is particularly sensitive to the molecular conformations, as is often the case in the modeling of strongly bound hydrogen-bonded crystals, such as salts.²⁵ The minimization of a wide set of experimentally determined crystal structures and the reminimization of putative structures earlier generated by rigid-body searches demonstrate that the approach offers substantial accuracy improvements in comparison to existing methodologies.

Acknowledgment. The authors acknowledge financial support from the Basic Technology Program of the Research Councils U. K. as part of the CPOSS project (<http://www.cposs.org.uk>) and are grateful for the use of UCL Research Computing services (C³).

Supporting Information Available: Details regarding the reproduction of the experimentally determined 1-phenylethylammonium 2-phenylpropanoate crystal structures (Table S1), the changes in cell and hydrogen-bond geometries during the DMAflex refinement of the most stable 1-phenylethylammonium 2-phenylpropanoate (Tables S2 and S3) and carbamazepine (Table S4) rigid-body minima, and the overlay of the experimental and rigid-body (Figure S1a) and

flexible molecule (Figure S1b) minima for form IV of carbamazepine. This information is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Verwer, P.; Leusen, F. J. J. Computer Simulation to Predict Possible Crystal Polymorphs. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1998; pp 327–365.
- (2) Lommerse, J. P. M.; Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Mooij, W. T. M.; Price, S. L.; Schweizer, B.; Schmidt, M. U.; Van Eijck, B. P.; Verwer, P.; Williams, D. E. *Acta Crystallogr., Sect. B* **2000**, *56*, 697–714.
- (3) Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Dzyabchenko, A.; Erk, P.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Lommerse, J. P. M.; Mooij, W. T. M.; Price, S. L.; Scheraga, H.; Schweizer, B.; Schmidt, M. U.; Van Eijck, B. P.; Verwer, P.; Williams, D. E. *Acta Crystallogr., Sect. B* **2002**, *58*, 647–661.
- (4) Day, G. M.; Motherwell, W. D. S.; Ammon, H. L.; Boerrigter, S. X. M.; Della Valle, R. G.; Venuti, E.; Dzyabchenko, A.; Dunitz, J. D.; Schweizer, B.; Van Eijck, B. P.; Erk, P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Leusen, F. J. J.; Liang, C.; Pantelides, C. C.; Karamertzanis, P. G.; Price, S. L.; Lewis, T. C.; Nowell, H.; Torrisi, A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; Verwer, P. *Acta Crystallogr., Sect. B* **2005**, *61*, 511–527.
- (5) Van Eijck, B. P.; Kroon, J. *Acta Crystallogr., Sect. B* **2000**, *56*, 535–542.
- (6) Van Eijck, B. P. *J. Comput. Chem.* **2002**, *23*, 456–462.
- (7) Dunitz, J. D.; Scheraga, H. A. *PNAS* **2004**, *101*, 14309–14311.
- (8) Van Eijck, B. P. *Acta Crystallogr., Sect. B* **2005**, *61*, 528–535.
- (9) Stone, A. J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047–1064.
- (10) Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.
- (11) Price, S. L. *J. Chem. Soc., Faraday Trans.* **1996**, *92*, 2997–3008.
- (12) Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. *J. Phys. Chem.* **1996**, *100*, 7352–7360.
- (13) Day, G. M.; Motherwell, W. D. S.; Jones, W. *Cryst. Growth Des.* **2005**, *5*, 1023–1033.
- (14) Brodersen, S.; Wilke, S.; Leusen, F. J. J.; Engel, G. *Phys. Chem. Chem. Phys.* **2003**, *5*, 4923–4931.
- (15) Mooij, W. T. M. Ab Initio Prediction of Crystal Structures. Ph.D. Thesis, Utrecht University, Utrecht, The Netherlands, 2000.
- (16) Gavezzotti, A. *Modell. Simul. Mater. Sci. Eng.* **2002**, *10*, R1–R29.
- (17) Koch, U.; Popelier, P. L. A.; Stone, A. J. *Chem. Phys. Lett.* **1995**, *238*, 253–260.
- (18) Price, S. L. *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 1755–1763.
- (19) Dudek, M. J.; Ponder, J. W. *J. Comput. Chem.* **1995**, *16*, 791–816.
- (20) Koch, U.; Stone, A. J. *J. Chem. Soc., Faraday Trans.* **1996**, *92*, 1701–1708.
- (21) Mooij, W. T. M.; Van Eijck, B. P.; Kroon, J. *J. Phys. Chem. A* **1999**, *103*, 9883–9890.
- (22) Mooij, W. T. M.; Van Eijck, B. P.; Kroon, J. *J. Am. Chem. Soc.* **2000**, *122*, 3500–3505.
- (23) Van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. *J. Comput. Chem.* **2001**, *22*, 805–815.
- (24) Van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. *J. Phys. Chem. B* **2001**, *105*, 10573–10578.
- (25) Karamertzanis, P. G.; Price, S. L. *J. Phys. Chem. B* **2005**, *109*, 17134–17150.
- (26) Florence, A. J.; Johnston, A.; Price, S. L.; Nowell, H.; Kennedy, A. R.; Shankland, N. *J. Pharm. Sci.* **2006**, in press.
- (27) Willock, D. J.; Price, S. L.; Leslie, M.; Catlow, C. R. A. *J. Comput. Chem.* **1995**, *16*, 628–647.
- (28) Allen, F. H.; Harris, S. E.; Taylor, R. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 247–254.
- (29) Osborn, J. C.; York, P. *J. Mol. Struct.* **1999**, *474*, 43–47.
- (30) Pertsin, A. J.; Kitaigorodsky, A. I. *The Atom-Atom Potential Method. Applications to Organic Molecular Solids*; Springer-Verlag: Berlin, 1987.
- (31) Gavezzotti, A. Crystal Symmetry and Molecular Recognition. In *Theoretical Aspects and Computer Modeling of the Molecular Solid State*; Gavezzotti, A., Ed.; John Wiley & Sons: Chichester, U. K., 1997.
- (32) Stone, A. J. *Chem. Phys. Lett.* **1981**, *83*, 233–239.
- (33) Cox, S. R.; Hsu, L. Y.; Williams, D. E. *Acta Crystallogr., Sect. A* **1981**, *37*, 293–301.
- (34) Williams, D. E.; Cox, S. R. *Acta Crystallogr., Sect. B* **1984**, *40*, 404–417.
- (35) Williams, D. E.; Houpt, D. J. *Acta Crystallogr., Sect. B* **1986**, *42*, 286–295.
- (36) Ewald, P. *Ann. Phys.* **1921**, *64*, 253.
- (37) Leslie, M. *Mol. Phys.* **2006**, in press.
- (38) Nelder, J. A.; Mead, R. *Comput. J.* **1965**, *7*, 308–313. Implemented in Numerical Recipes (<http://www.numerical-recipes.com>; accessed May 2006).
- (39) 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.9; Gaussian, Inc.: Pittsburgh, PA, 1998.
- (40) Allen, F. H. *Acta Crystallogr., Sect. B* **2002**, *58*, 380–388.
- (41) Chisholm, J. A.; Motherwell, S. *J. Appl. Crystallogr.* **2005**, *38*, 228–231.
- (42) Allen, F. H.; Kennard, O.; Watson, D. G. *J. Chem. Soc., Perkin Trans. 2* **1987**, S1–S19.

- (43) Price, S. L. *Adv. Drug Delivery Rev.* **2004**, *56*, 301–319.
- (44) van Mourik, T.; Karamertzanis, P. G.; Price, S. L. *J. Phys. Chem. A* **2006**, *110*, 8–12.
- (45) Vishweshwar, P.; McMahon, J. A.; Oliveira, M.; Peterson, M. L.; Zaworotko, M. J. *J. Am. Chem. Soc.* **2005**, *127*, 16802–16803.
- (46) Bernstein, J. *Polymorphism in Molecular Crystals*; Oxford Science Publications: Oxford, U. K., 2002.
- (47) Rekoske, J. E. *AIChE J.* **2001**, *47*, 2.
- (48) Jacques, J.; Collet, A.; Wilen, S. H. *Enantiomers, Racemates and Resolutions*; Wiley-Interscience: New York, 1981.
- (49) Pasteur, L. C. R. *Hebdomadae Seances Acad. Sci.* **1853**, *37*, 162.
- (50) Karamertzanis, P. G.; Hulme, A. T.; Anandamanoharan, P. R.; Cains, P. W.; Vickers, M.; Tocher, D. A. **2006**, in preparation.
- (51) Dufour, F.; Perez, G.; Coquerel, G. *Bull. Chem. Soc. Jpn.* **2004**, *77*, 79–86.
- (52) Taylor, R.; Kennard, O.; Versichel, W. *J. Am. Chem. Soc.* **1983**, *105*, 5761–5766.
- (53) Taylor, R.; Kennard, O.; Versichel, W. *Acta Crystallogr., Sect. B* **1984**, *40*, 280–288.
- (54) Taylor, R.; Kennard, O. *Acc. Chem. Res.* **1984**, *17*, 320–326.
- (55) Taylor, R.; Kennard, O. *Acta Crystallogr., Sect. B* **1983**, *39*, 133–138.
- (56) Lowes, M. M. J.; Caira, M. R.; Lotter, A. P.; van der Watt, J. G. *J. Pharm. Sci.* **1987**, *76*, 744–752.
- (57) Himes, V. L.; Mighell, A. D.; DeCamp, W. H. *Acta Crystallogr., Sect. B* **1981**, *37*, 2242–2245.
- (58) Rustichelli, C.; Gamberini, G.; Ferioli, V.; Gamberini, M. C.; Ficarra, R.; Tommasini, S. *J. Pharm. Biomed. Anal.* **2000**, *23*, 41–54.
- (59) Lang, M. D.; Kampf, J. W.; Matzger, A. J. *J. Pharm. Sci.* **2002**, *91*, 1186–1190.
- (60) Grzesiak, A. L.; Lang, M. D.; Kim, K.; Matzger, A. J. *J. Pharm. Sci.* **2003**, *92*, 2260–2271.
- (61) Fleischman, S. G.; Kuduva, S. S.; McMahon, J. A.; Moulton, B.; Walsh, R. D. B.; Rodriguez-Hornedo, N.; Zaworotko, M. J. *Cryst. Growth Des.* **2003**, *3*, 909–919.
- (62) Lang, M. D.; Grzesiak, A. L.; Matzger, A. J. *J. Am. Chem. Soc.* **2002**, *124*, 14834–14835.
- (63) Hilfiker, R.; Berghausen, J.; Blatter, F.; Burkhard, A.; De Paul, S. M.; Freiermuth, B.; Geoffroy, A.; Hofmeier, U.; Marcolli, C.; Siebenhaar, B.; Szlagiewicz, M.; Vit, A.; von Raumer, M. *J. Therm. Anal. Calorim.* **2003**, *73*, 429–440.
- (64) Strachan, C. J.; Howell, S. L.; Rades, T.; Gordon, K. C. *J. Raman Spectrosc.* **2004**, *35*, 401–408.
- (65) Cruz-Cabera, A. J.; Day, G. M.; Motherwell, W. D. S.; Jones, W. *Cryst. Growth Des.* **2006**, in press.
- (66) Speakman, J. C. In *Molecular Structure by Diffraction Methods*; Sim, G. A., Sutton, L. E., Eds.; The Chemical Society: London, 1973; p 203.
- (67) Chong-Hui, G.; Grant, D. J. W. *J. Pharm. Sci.* **2001**, *90*, 1277–1287.
- (68) Nowell, H.; Price, S. L. *Acta Crystallogr., Sect. B* **2005**, *61*, 558–568.
- (69) Ouvrard, C.; Price, S. L. *Cryst. Growth Des.* **2004**, *4*, 1119–1127.
- (70) Mooij, W. T. M.; Leusen, F. J. J. *Phys. Chem. Chem. Phys.* **2001**, *3*, 5063–5066.
- (71) Chisholm, J. A.; Motherwell, S.; Tulip, P. R.; Parsons, S.; Clark, S. J. *Cryst. Growth Des.* **2005**, *5*, 1437–1442.
- (72) van Mourik, T.; Gdanitz, R. J. *J. Chem. Phys.* **2002**, *116*, 9620–9623.
- (73) Tsuzuki, S.; Luthi, H. P. *J. Chem. Phys.* **2001**, *114*, 3949–3957.
- (74) Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. *Chem. Phys. Lett.* **2004**, *394*, 334–338.
- (75) Montanari, B.; Ballone, P.; Jones, R. O. *J. Chem. Phys.* **1998**, *108*, 6947–6951.
- (76) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149–5155.
- (77) Wu, Q.; Yang, W. T. *J. Chem. Phys.* **2002**, *116*, 515–524.
- (78) Neumann, M. A.; Perrin, M. A. *J. Phys. Chem. B* **2005**, *109*, 15531–15541.
- (79) Gavezzotti, A. *J. Phys. Chem. B* **2002**, *106*, 4145–4154.
- (80) Gavezzotti, A. *J. Phys. Chem. B* **2003**, *107*, 2344–2353.
- (81) Bernstein, J.; Davey, R. J.; Henck, J. O. *Angew. Chem., Int. Ed.* **1999**, *38*, 3441–3461.

CT600111S

JCTC

Journal of Chemical Theory and Computation

A Second Look at Canonical Sampling of Biomolecules Using Replica Exchange Simulation

Daniel M. Zuckerman* and Edward Lyman†

Department of Computational Biology, School of Medicine, and Department of Environmental & Occupational Health, Graduate School of Public Health, Suite 3064 BST3, 3501 Fifth Avenue, University of Pittsburgh, Pittsburgh, Pennsylvania 15213

Received February 10, 2006

Abstract: The replica exchange approach, also called parallel tempering, is gaining popularity for biomolecular simulation. We ask whether the approach is likely to be efficient compared to standard simulation methods for fixed-temperature equilibrium sampling. To examine the issue, we make a number of straightforward observations on how “fast” high-temperature molecular simulations can be expected to run, as well as on how to characterize efficiency in replica exchange. Although our conclusions remain to be fully established, on the basis of a range of results in the literature and some of our own work with a 50-atom peptide, we are not optimistic for the efficiency of replica exchange for the canonical sampling of biomolecules.

Because of growing interest in temperature-based sampling methods for biomolecules, such as replica exchange^{1–9} (see also ref 10), this letter aims to make some observations and raise some potentially important questions which we have not seen addressed sufficiently in the literature. Mainly, we wish to call attention to limits on the maximum speed-up to be expected from temperature-based methods and also note the need for careful quantification of sampling efficiency. Here, *we are strictly concerned with canonical sampling at a fixed temperature*, and *not* with conformational searching. Because potentially lengthy studies may be necessary to address the issues, we felt it would be useful to bring them to the attention of the broader community. Some of the observations we make below were noted previously in a

careful study of a one-dimensional system by Brown and Head-Gordon.¹¹

We will base our discussion around a generic replica exchange protocol, consisting of $M + 1$ levels spanning from the temperature T_0 , at which canonical sampling is desired, up to T_M . Replica exchange is motivated by the increased rate of barrier crossing possible at higher temperatures. We assume each level is simulated for a time t_{sim} , which implies a total CPU cost $(M + 1) \times t_{\text{sim}}$. In typical explicitly solvated peptide systems, $M \sim 20$, $T_0 \approx 300$ K, and $T_M \sim 450$ K.⁵ For typical M values, the relatively low *maximum* temperature (T_M) values reflect the well-known requirement for configuration-space overlap between neighboring levels of the temperature ladder:^{4,5} if temperature increments (and also T_M) become too large, the necessary exchanges between levels become rare because of low overlap. We note that a new exchange variant introduced by Berne and co-workers permits the use of “cold” solvent and larger temperature gaps,¹² but the issues we raise still apply to the new protocol, especially as larger solutes are considered.

While replica exchange is often thought of as an “enhanced sampling method”, what does that mean? Indeed, what is an appropriate criterion for judging efficiency? As our first observation, we believe that (**Obs. I**) efficiency can only mean a decrease in the total CPU usage—that is, *summed over all processors*—for a given degree of sampling quality. (We will defer the necessary discussion of assessing sampling quality and only assume such assessment is possible.) When the goal is canonical sampling at T_0 , after all, one has the option of running an ordinary parallel simulation at T_0 (e.g., ref 13) or perhaps $M + 1$ independent simulations.¹⁴ A truly efficient method must be a superior alternative to such “brute force” simulation.

Reports in the literature offer an ambiguous picture as to whether replica exchange attains efficiency for the canonical sampling of biomolecules (**Obs. II**). Sanbonmatsu and Garcia compared replica exchange to an equivalent amount of brute-force sampling, but their claim of efficiency is largely based on the alternative goal of enhancing sampling over the full range of temperatures, rather than for canonical sampling at T_0 .⁵ When the data solely for T_0 are examined, there is no clear gain, especially noting that assessment was based on principal components derived only from the replica exchange data. Another claim of efficiency, by Duan and co-workers,⁹ fails to include the full CPU cost of all $M + 1$ levels. When suitably corrected, there does appear to be a speed-up of perhaps a factor of 2 for $T_0 = 308$ K, but the system studied is considerably smaller (permitting larger temperature jumps) than would be possible in protein systems of interest. Another efficiency claim by Roe et al. also does not account for the

* Corresponding author phone: 412-648-3335; fax: 412-648-3163; e-mail: dmz@ccb.pitt.edu

† E-mail: elyman@ccb.pitt.edu (E.L.).

Table 1. High-Temperature Speed-up Factors Calculated Using Arrhenius Factors^a

	$\Delta E = 2k_B T_0$	$4k_B T_0$	$6k_B T_0$	$8k_B T_0$
$T_M = 400$ K	1.65	2.72	4.48	7.39
500 K	2.23	4.95	11.0	24.5
600 K	2.72	7.39	20.1	54.6

^a Speed-up factors are computed as the ratio $k_a(T_M)/k_a(T_0 = 300$ K) for the indicated energy barriers ΔE via eq 1. Energy barriers are given in units of $k_B T_0$. A rough estimate of the efficiency factor (the factor by which the total CPU usage is reduced) obtainable in an M -level parallel replica exchange simulation with maximum temperature T_M is the table entry divided by $M + 1$.

full CPU cost of all ladder levels.⁸ In a structural-glass system, replica exchange was found not to be helpful,¹⁵ although efficiency has been noted in spin systems.^{1,16} We emphasize that *biomolecular* replica exchange should indeed be efficient in certain cases (with high enough energy barriers, see below). At least one such instance has been noted by Garcia, using a suitable brute-force comparison system.¹⁷

The lack of clear-cut results in a much-heralded approach merits closer examination. What might be preventing efficiency gain? Or put another way, what is the maximum efficiency possible in a standard replica exchange simulation? The very construction of the method implies that, (**Obs. III**) in any parallel exchange protocol, the sampling “speed” at the bottom level—lowest T —will be controlled by the speed at which the top level—highest T —samples configuration space. A parallel exchange simulation can go no faster than its fastest level and, on average, will be slower. Therefore, given our interest in efficient canonical sampling at T_0 , the speed of the top level should exceed that of the bottom by *at least* a factor of $M + 1$. If not, the simulation does not “break even” in total CPU cost, as compared to brute-force canonical sampling at T_0 for the full length $(M + 1) \times t_{\text{sim}}$.

The basic temperature dependence of barrier-crossing rates is well-known (e.g., ref 18) and has important consequences for replica exchange. The Arrhenius factor indicates that the temperature-dependent rate k for crossing a particular barrier obeys

$$k_a(T) = k_0 \exp(\Delta S/k_B) \exp(-\Delta E/k_B T) \quad (1)$$

for a fixed-volume system, where k_0 is an unknown prefactor insensitive to temperature and is assumed constant; ΔE is the energy barrier, and ΔS is the entropy barrier—that is, “narrowing” of the configuration space—which must be expected in a multidimensional molecular system. Two observations are immediate: (**Obs. IV**) the entropic component of the rate is completely unaffected by an increase in temperature, and the possible speed-up due to the energetic part can easily be calculated.

Table 1 gives possible speed-ups for several energy barriers and temperatures, in units of $k_B T_0$ for $T_0 = 300$ K. Speed-ups are computed simply as the ratio $k_a(T_M)/k_a(T_0)$ for possible values of T_M . It is clear that, for modest barriers, the speed-up attainable even with a top temperature $T_M = 500$ K is only on the order of a typical number of replicas in replica exchange, $M \sim 20$. Thus, (**Obs. V**) if modest barriers ($< 8k_B T_0$) dominate a system’s dynamics, efficiency will be difficult to obtain via a typical replica exchange

simulation, because the speed-up factor noted in the table needs to be divided by $M + 1$.

We reiterate that our discussion is narrowly focused on equilibrium sampling. When, for instance, thermodynamic information over a range of temperatures is desired (e.g., ref 6), replica exchange may indeed be useful. Certainly, if temperatures lower than our T_0 (300 K) are studied, the speed-up factors given in the table will increase.

How high are barriers encountered in molecular systems? We can only begin to answer this question, but one must first be careful about which barriers matter. We believe that (**Obs. VI**) “local” barriers will matter the most; that is, the energy barriers actually encountered along a trajectory will dominate the sampling speed. Apparent barriers determined by projections onto arbitrary low-dimensional reaction coordinates would seem of uncertain value.¹⁹ (We note that Zwanzig has attempted to account for local roughness with an effective diffusion constant on a slowly varying landscape.²⁰)

Evidence from simulations and experiments is far from complete but indicates that (**Obs. VII**) energy barriers in molecular systems appear to be modest. Here, unless noted otherwise, $T_0 \approx 300$ K. In their extensive study of a tetrapeptide, Czerminski and Elber found barriers < 3 kcal/mol $\approx 5k_B T_0$ for the lowest-energy transition path.²¹ Equally interesting, they found approximately 1000 additional paths with similar energy profiles (differing by < 1 kcal/mol $< 2k_B T_0$)—suggesting what we might term a “pebbly” rather than “mountainous” energy landscape (see also ref 22). In our own work (unpublished) with the implicitly solvated 50-atom dileucine peptide,²³ increasing the temperature from 298 K to a series of values (400, 500, and 600 K), independently, led to hopping-rate increases by factors of (2.0, 3.8, and 4.6), respectively. The comparison of these speed-up factors—for hopping between the states defined in ref 24—with the table suggests that there is no barrier larger than $\sim 3k_B T_0$. Similarly, Sanbonmatsu and Garcia found that barriers for explicitly solvated met-enkephalin were small, on the order of $k_B T_0$.⁵ An experimental study has also suggested that barriers are modest ($< 6k_B T_0$).²⁵ Although this list is fairly compelling, we believe the question of barrier heights is far from settled. Further study should carefully consider local versus global barriers, as well as entropy versus energy components of barriers. (We purposely do not discuss barriers to protein folding, because our scope here is solely equilibrium fluctuations.)

Importantly, the goal of understanding efficiency implies the need for reliable means for assessing sampling. An ideal approach to assessment would survey all pertinent substates to ensure appropriate Boltzmann frequencies. Present approaches to assessment typically calculate one- or two-dimensional free energy surfaces (equivalently, projected probability distributions), which are evaluated visually. Principal components (e.g., refs 5 and 9) as well as “composite” coordinates such as the radius of gyration⁸ are popular coordinate choices. Yet we believe that (**Obs. VIII**) the use of low-dimensional sampling assessment is intrinsically limited, because it could readily mask structural diversity—that is, be consistent with substantially distinct conformational ensembles. Future work could usefully pursue higher-dimensional measures, which can always be numerically compared between independent simulations for sam-

pling assessment. In our own work, for instance, we have begun to use a histogram measure which directly reports on the structural distribution of an ensemble.²⁶

Finally, the possibility for improving replica exchange should be noted. A number of groups have made efforts to optimize the approach by examining temperature ladders and exchange acceptance ratios.^{27–30} Such efforts are important and may benefit from the discussion presented here. The work of Brown and Head-Gordon¹¹ and our perspective as reflected in the table suggest that a reduction of the number of replicas M could prove useful (for a fixed T_M). However, the extent to which M can be reduced will depend on a concomitant increase in the exchange attempt frequency, whose cost is hardware-dependent and less amenable to analysis. This point, interestingly, was raised in the 1990 J-walking paper.¹⁰

In conclusion, we have attempted to tie together a number of straightforward observations which reflect concerns about the effectiveness of the replica exchange simulation method, when the goal is single-temperature canonical sampling for biomolecular systems. On the basis of typical current implementations of the replica exchange approach, it is far from clear that the approach is one that should be widely adopted for canonical sampling. The concerns we have outlined suggest that other simulation strategies, such as Hamiltonian exchange³¹ and resolution exchange,^{24,32} may merit consideration—as well as scrutiny. Alternative temperature-based schemes (e.g., refs 10, 11, 12, and 33) also should be considered. We emphasize that our goal has been to raise questions more than to answer them.

ACKNOWLEDGMENT.

The authors wish to thank Rob Coalson, Juan de Pablo, Ron Elber, Angel García, Hagai Meirovitch, Rohit Pappu, and Robert Swendsen for very useful conversations. We gratefully acknowledge support from the NIH, through Grants ES007318 and GM070987. We also greatly appreciate support from the Department of Computational Biology and the Department of Environmental & Occupational Health.

REFERENCES

- (1) Swendsen, R. H.; Wang, J.-S. *Phys. Rev. Lett.* **1986**, *57*, 2607–2609.
- (2) Geyer, C. J. Markov Chain Monte Carlo Maximum Likelihood. In *Proceedings of the 23rd Symposium on the Interface*; Keramidas, E.

- M., Ed.; Computing Science and Statistics Interface Foundation of North America: Fairfax Station, VA, 1991.
- (3) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.
- (4) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140–150.
- (5) Sanbonmatsu, K. Y.; Garcia, A. E. *Proteins* **2002**, *46*, 225–234.
- (6) Paschek, D.; Garcia, A. E. *Phys. Rev. Lett.* **2004**, *93*, 238105.
- (7) Rathore, N.; Chopra, M.; de Pablo, J. J. *J. Chem. Phys.* **2005**, *122*, 024111.
- (8) Roe, D. R.; Hornak, V.; Simmerling, C. J. *Mol. Biol.* **2005**, *352*, 370–381.
- (9) Zhang, W.; Wu, C.; Duan, Y. *J. Chem. Phys.* **2005**, *123*, 154105.
- (10) Frantz, D. D.; Freeman, D. L.; Doll, J. D. *J. Chem. Phys.* **1990**, *93*, 2769–2784.
- (11) Brown, S.; Head-Gordon, T. *J. Comput. Chem.* **2002**, *24*, 68–76.
- (12) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13749–13754.
- (13) Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. *J. Comput. Phys.* **1999**, *151*, 283–312.
- (14) Caves, L. S. D.; Evanseck, J. D.; Karplus, M. *Protein Sci.* **1998**, *7*, 649–666.
- (15) De Michele, C.; Sciortino, F. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2002**, *65*, 051202.
- (16) Wang, J.-S.; Swendsen, R. *Prog. Theor. Phys. Suppl.* **2005**, *157*, 317–323.
- (17) Garcia, A. E. 2005. Personal communication.
- (18) Atkins, P.; de Paula, J. *Physical Chemistry*, 7th ed.; Freeman: New York, 2002.
- (19) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964–1977.
- (20) Zwanziq, R. W. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2029–2030.
- (21) Czerminski, R.; Elber, R. *J. Chem. Phys.* **1990**, *92*, 5580–5601.
- (22) Hyeon, C.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 10249–10253.
- (23) All-atom dileucine peptide (ACE-Leu₂-NME) simulations were performed using TINKER version 4.2, running Langevin Dynamics (91 ps⁻¹ friction coefficient, 1.0 fs time step), with the OPLS-AA force field and GBSA implicit solvent.
- (24) Lyman, E.; Ytreberg, F. M.; Zuckerman, D. M. *Phys. Rev. Lett.* **2006**, *96*, 028105.
- (25) Nevo, R.; Brumfeld, V.; Kapon, R.; Hinterdorfer, P.; Reich, Z. *EMBO Rep.* **2005**, *6*, 482–486.
- (26) Lyman, E.; Zuckerman, D. M. *Biophys. J.* **2006**, In press. Archived version: <http://www.arXiv.org/abs/physics/0601104> (accessed June 2, 2006).
- (27) Earl, D. J.; Deem, M. W. *J. Phys. Chem. B* **2005**, *108*, 6844–6849.
- (28) Earl, D. J.; Deem, M. W. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910–3916.
- (29) Kone, A.; Kofke, D. *J. Chem. Phys.* **2005**, *122*, 206101.
- (30) Trebst, S.; Troyer, M.; Hansmann, U. H. E. *J. Chem. Phys.* **2006**, *124*, 174903.
- (31) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042–6051.
- (32) Lyman, E.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2006**, In press.
- (33) Neal, R. M. *Stat. Comput.* **2001**, *11*, 125–139.

CT0600464