

Finding Reaction Pathways of Type $A + B \rightarrow X$: Toward Systematic Prediction of Reaction Mechanisms

Satoshi Maeda^{*,†,‡} and Keiji Morokuma^{*,‡,§}[†]The Hakubi Center, Kyoto University, Kyoto 606-8302, Japan[‡]Fukui Institute for Fundamental Chemistry, Kyoto University, Kyoto 606-8103, Japan[§]Department of Chemistry and Cherry L. Emerson Center for Scientific Computation, Emory University, Atlanta, Georgia 30322, United States Supporting Information

ABSTRACT: In these five decades, many useful tools have been developed for exploring quantum chemical potential energy surfaces. The success in theoretical studies of chemical reaction mechanisms has been greatly supported by these tools. However, systematic prediction of reaction mechanisms starting only from given reactants and catalysts is still very difficult. Toward this goal, we describe the artificial force induced reaction (AFIR) method for automatically finding reaction paths of type $A + B \rightarrow X (+ Y)$. By imposing an artificial force to given reactants and catalysts, the method can find the reactive sites very efficiently. Further pressing by the artificial force provides approximate transition states and product structures, which can be easily reoptimized to the corresponding true ones. This procedure can be executed very efficiently just by minimizing a single function called the AFIR function. All important reaction paths can be found by repeating this cycle starting from many initial orientations. We also discuss perspectives of automated reaction path search methods toward the above goal.

I. INTRODUCTION

Finding reaction paths as well as underlying transition states (TSs) has been one of the primary tasks in theoretical studies of the chemical reaction mechanism. Transition state is a key concept in the transition-state theory (TST),¹ in which the reaction path is introduced rather abstractly as paths connecting a reactant and a product through a TS. By early development of semiempirical and ab initio theories, TS structures began to be located explicitly as first-order saddle points on potential energy surfaces (PESs).^{2–6} The intrinsic reaction coordinate (IRC) was introduced in 1970 as the center line of idealized reaction paths, defined as a mass-weighted steepest descent path starting from a first-order saddle point.⁷ Calculation of IRC on the basis of an ab initio theory was accomplished in 1977,⁸ by the development of the analytical gradient method,^{9,10} and paths for the $\text{HNC} \rightarrow \text{HCN}$ and $\text{H}^- + \text{CH}_4 \rightarrow \text{CH}_4 + \text{H}^-$ reactions were the first examples of ab initio IRCs. Many useful tools have since been developed for finding TSs and IRCs,^{11,12} and by using such tools, numerous reaction mechanisms have been elucidated.^{13–16}

It should be noted that our focus in this paper is on local chemical bond rearrangements involving a few to several tens of atoms. Global optimization, conformation sampling, crystal structure sampling, etc.,^{17–23} are outside the scope of this paper.

Once a TS is obtained, the associated IRC can nowadays be computed automatically by using one of the advanced steepest descent path integration methods.^{8,24–27} The most complicated step is locating the TS. A variety of geometry optimization techniques, such as the gradient minimization,⁴ the Broyden optimization,^{28,29} the eigenvector following (EF),³⁰ the geometry direct inversion in the iterative subspace,³¹ and the rational function optimization³² (RFO), have been a great help to locate

the exact saddle point starting from a guessed TS structure. When estimating a TS geometry and/or a reaction mechanism, qualitative chemical theories, such as the frontier molecular orbital (FMO) theory^{33,34} and the Woodward–Hoffmann rules,^{35,36} are very helpful and have been frequently considered. TS optimization often requires a very good initial guess. Hence, results may depend on the provided guess. It may happen that important paths are missed in the search, if one fails to provide a proper guess of the corresponding TS geometry. An absence of an important path may cause serious disagreement between a theoretical conclusion and experimental data. Moreover, a theoretical prediction may be meaningless by the lack of a single important path. Therefore, an automated search that does not rely on any initial guess is desired for reliable determination as well as prediction of the reaction mechanism.

In many cases, one knows both a reactant and a product. Thus, double-end approaches are very useful to look for the TS(s) and a minimum energy path connecting the known end points. There have been considerable efforts for developing such approaches, e.g., the synchronous transit method,³⁷ the saddle optimization method,³⁸ the self-penalty walk method,³⁹ the nudged elastic band (NEB) method,⁴⁰ the string method,⁴¹ the growing string method,⁴² and others.^{11,12} These double-ended methods considerably reduced the difficulty concerned with the initial guess of TS. However, these methods usually assume at first that the reaction path connects these two points along the straight line and tend to provide to a path closest to this straight line. Hence, it is not guaranteed that the most important (in general the lowest)

Received: April 27, 2011

Published: June 24, 2011

path is obtained when there is more than one path. This problem can be avoided by providing appropriate intermediate geometries if available. In other words, double-end methods usually require a guess of reaction mechanism, that is, a series of important geometries, including a reactant, a product, and some key intermediates as well.

There are some methods which can find a reaction path automatically starting with only the reactant if reaction variables are given. The single coordinate driving (SCD), which has frequently been employed since very early studies of reaction paths,^{5,6} was employed also in automated path sampling.⁴³ The metadynamics⁴⁴ method is very powerful in calculations of free energy barriers for given reaction variables and has been successfully applied to many problems, including chemical reactions.⁴⁵ The fast marching method can very quickly give a minimum energy path for a given set of reaction variables.⁴⁶ Although no input of product or TS is necessary in these methods, one has to select a small number of well-chosen, chemically relevant collective variables. In this sense, these methods also require a guess of reaction mechanism.

There are several approaches completely free from the issue of initial guess. Three such methods, the gradient extremal following (GEF) method,^{47–56} the reduced gradient following (RGF, also called Newton trajectory) method,^{57–63} and the anharmonic downward distortion following (ADDF) method,^{64–66} were applied to global mapping of the whole topography of the PES of H_2CO on the basis of quantum mechanical calculations. These methods automatically located all local minima and TSs relevant to all chemical bond rearranging reactions, i.e., $\text{H}_2\text{CO} \leftrightarrow \text{H}_2 + \text{CO}$, $\text{H}_2\text{CO} \leftrightarrow \text{HCOH}$, $\text{HCOH} \leftrightarrow \text{H}_2 + \text{CO}$, $\text{HCOH} \leftrightarrow \text{COH}_2$, and $\text{COH}_2 \leftrightarrow \text{H}_2 + \text{CO}$, starting from the local minimum of H_2CO .^{56,57,65} Thus, these methods have the ability to find all paths for unimolecular reactions automatically. However, the former two are computationally too expensive to be applied to systems larger than H_2CO . Only the ADDF method has been practically used successfully in global reaction route mapping for systems consisting of ~ 10 atoms.⁶⁷ Several options are available in the ADDF method regarding applications to larger systems and nonadiabatic reactions.^{68–70} There are some valley filling type approaches, such as the isopotential contour following⁷¹ method and the chemical flooding⁷² method. These can find only a single reaction path starting from a valley of a given reactant molecule, although this single path often corresponds to the lowest barrier path.

The goal of this study is development of a tool by which all important reaction paths can be obtained automatically just by providing multiple reactant species. Reactions among them, denoted $\text{A} + \text{B} \rightarrow \text{X}$ (+ Y) type in this paper, are of essential importance in organic chemistry, where B can be a catalyst. Nevertheless, among the above fully automated search methods without any initial guess, the only practical ones, the ADDF method as well as the valley filling methods are designed to find paths of type $\text{A} \rightarrow \text{X}$ (+ Y) starting from local minima. Global optimization, conformation sampling, cluster structure sampling, etc. usually treat paths of type $\text{A} \rightarrow \text{X}$, e.g., protein folding occurring between two conformers of a peptide. Hence, methods successfully applied to such a problem are not readily applicable to paths of type $\text{A} + \text{B} \rightarrow \text{X}$. Although there are shallow potential minima of weak complexes between A and B, such potential wells are often located in very floppy regions of PES, far away from TSs for bond reorganization. Since topology of PES involved in initial association states between A and B is very different from that in

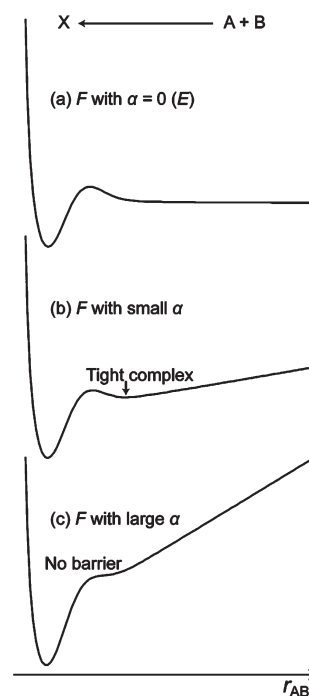


Figure 1. $F(Q)$ in eq 1: (a) with $\alpha = 0$, (b) with small α , and (c) with large α .

the product minimum of X, a method specially suited for such systems needs to be developed.

Recently, we have developed a method capable of solving the present problem.⁷³ The idea is pretty simple; just pressing given reactant molecules to each other by a constant force. Hence, we call the method the artificial force induced reaction (AFIR) method. We have applied this method to an organic multi-component reaction and to nonadiabatic spin flipping reactions^{74,75} and have demonstrated that the method has the ability to discover unexpected reaction paths automatically and systematically. In this paper, the AFIR method is described in detail. The reason why the method works very well in the automated reaction path search is illustrated with a very simple application to the reaction $\text{CO}_2 + \text{H} \rightarrow \text{HCO}_2/\text{HOCO}$. An application to the reaction between vinyl alcohol ($\text{H}_2\text{C} = \text{CH}-\text{OH}$) and formaldehyde (H_2CO), which has an implication to the aldol reaction, provided automatically many reaction pathways including the aldol reaction among others. As case studies of three and four component reactions, one or two water molecules were added to these reactants as potential molecular catalysts, and many paths involving the water molecule(s) as a proton-transfer agent were located. Finally, perspectives of automated reaction path search methods are discussed.

II. THEORY

AFIR Method. In the AFIR method,⁷³ two or more reactants are pressed to each other by a constant force. At first, it is illustrated for the simplest case in which both A and B are single atoms. A constant force can be applied between A and B just by adding a linear function of distance r_{AB} to the potential energy function $E(r_{\text{AB}})$:

$$F(r_{\text{AB}}) = E(r_{\text{AB}}) + \alpha r_{\text{AB}} \quad (1)$$

where α is a parameter of the strength of the force. Figure 1a–c shows the function of eq 1 with $\alpha = 0$ and with small and large α , respectively, for a typical diatomic potential energy curve $E(r_{AB})$. Along the curve in Figure 1a, there is a reaction barrier separating the reactant region A + B and the well of the product X. By imposing a weak force in Figure 1b, a local minimum for a tight complex is generated before the barrier. With a strong force, the barrier disappears, and the curve in Figure 1c is fully attractive. On $F(r_{AB})$ in Figure 1c, one can find the product X just by minimizing the function starting from A + B.

In designing a general representation of the force term for a system with multiple degrees of freedom, one should take into account some requirements: the function should be uniquely defined (automatically determined) at all geometries, the function should be differentiable up to the second order, and the function should not change anisotropies of each reactant as much as possible. The third requirement suggests that a best function would be as a sum of isotropic functions centered at each atom. Hence, we proposed the following AFIR function:

$$F(\mathbf{Q}) = E(\mathbf{Q}) + \alpha \frac{\sum_{i \in A} \sum_{j \in B} [(R_i + R_j)/r_{ij}]^p r_{ij}}{\sum_{i \in A} \sum_{j \in B} [(R_i + R_j)/r_{ij}]^p} \quad (2)$$

where $E(\mathbf{Q})$ is the PES that depends on the atomic coordinates $\mathbf{Q} = \{Q_k\}$, r_{ij} is a distance between the i th and j th atoms, and summations are taken over all pairs of atoms in the reactants A and B. Chemical bond reorganization usually occurs in a local reaction center. Thus, the inverse distance weighting⁷⁶ is employed so that the force is imposed only to closely interacting pairs, where p is a parameter (set to the standard value: $p = 6$) of the weight. In this weight, each inverse distance is scaled by the sum of covalent radii ($R_i + R_j$) of the i th and j th atoms, respectively, to take the difference in atomic size into account.

The parameter α for the strength of force can be rewritten as

$$\alpha = \frac{\gamma}{\left[2^{-1/6} - \left(1 + \sqrt{1 + \frac{\gamma}{\varepsilon}} \right)^{-1/6} \right]} R_0 \quad (3)$$

This equation represents an average force acting on two atoms in the energy range $0 < E < \gamma$, when E is formally represented by the Lennard-Jones potential (although the method is insensitive to what form of potential is used). The parameter γ is a model collision energy between two reactant particles; in the following and also in previous studies, γ is used instead of α for convenience. In practice we adopted R_0 and ε to be the values for argon clusters ($R_0 = 3.8164 \text{ \AA}$ and $\varepsilon = 1.0061 \text{ kJ/mol}$).

A minimization of $F(\mathbf{Q})$ gives an approximate structure for a product X starting from a reactant pair A + B. The minimization path of $F(\mathbf{Q})$ is called the AFIR path, along which an approximate TS geometry can be obtained as a highest point of $E(\mathbf{Q})$. Such an approximate TS geometry is then used as an initial guess for full optimization of the true TS geometry without the artificial force. Although eq 2 is the functional form we presently adopt, further improvements of the functional form might increase the efficiency and the accuracy in the future.

Algorithm. A flowchart of the automated search is shown in Figure 2. Four kinds of information are required in the input: a set of (separately optimized) reactant structures, a set of quantum chemical parameters (computation level, total charge, spin

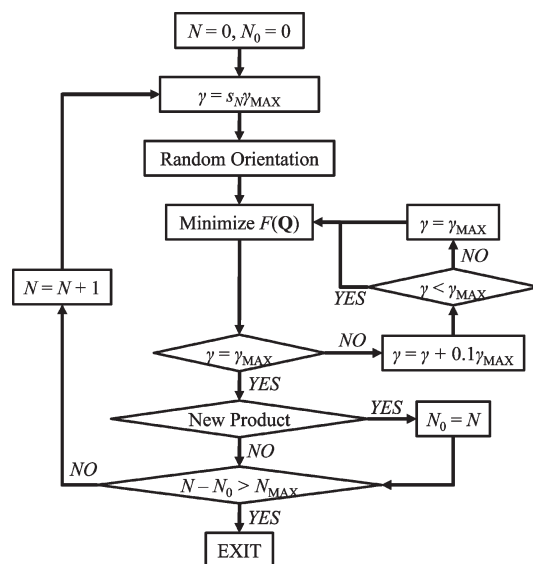


Figure 2. A flowchart of the automated reaction path search by the AFIR method.

multiplicity, etc.), a maximum model collision energy γ_{MAX} value (see below), and a stopping criterion N_{MAX} (also see below). γ_{MAX} essentially controls the highest energy that is searched in the present method. A choice of small value of γ_{MAX} restricts the search to the low-energy region that can be reached with the model collision energy γ_{MAX} , and the search is less expensive. On the other hand, a large value of γ_{MAX} forces the search to high-energy regions, and the search is more exhaustive but is more expensive. In the present study, the AFIR minimization is performed starting from many random orientations, although more systematic ways or use of an intuitive orientation for a known mechanism can also be considered.

At the first step of the N th cycle, the model collision energy γ (see eq 3) is set to $s_N \gamma_{MAX}$, where s_N is a random number in the range: $\{0,1\}$. Then, a random orientation of the reactant species is generated. $F(\mathbf{Q})$ is minimized starting from the selected random orientation. Once all convergence criteria (maximum absolute gradient g_{MAX} , root-mean-square gradient g_{rms} , maximum absolute displacement d_{MAX} , root-mean-square displacement d_{rms}) are met, γ is compared to γ_{MAX} . If $\gamma < \gamma_{MAX}$, then γ is increased by $0.1\gamma_{MAX}$, and minimization of $F(\mathbf{Q})$ is repeated with the updated γ starting from the latest geometry. Once a local minimum with $\gamma = \gamma_{MAX}$ is obtained, the structure is examined: (1) If no new chemical bond is seen between the reactant species, then the structure is discarded, (2) if the structure is already found in a previous cycle, then the structure is discarded, or (3) the structure is registered to a list of products, and N_0 is updated to N . This cycle is repeated while $N - N_0 \leq N_{MAX}$.

A parallel implementation of this algorithm is very easy; just distributing each minimization task (from the $\gamma = s_N \gamma_{MAX}$ step to the $N = N + 1$ step in Figure 2) to each CPU core. In this study, we actually implemented this procedure in parallel.

The AFIR method is implemented in a local developmental version of the GRRM program.^{64–66} The minimum search was performed by using the RFO method.³² When a step size of the RFO method exceeded a given trust radius d_{TR} (about the control of d_{TR} is discussed below), the minimization step was replaced by a step of the trust radius method,⁷⁷ which gives a

quasi-Newton step of the size d_{TR} . In this study, the exact Hessian matrix was computed in every 50 minimization steps and was updated by the BFGS^{78–81} method, which is the default setting of the GRRM program and can be changed by users. The convergence criteria, i.e., g_{MAX} , g_{rms} , d_{MAX} , and d_{rms} were set to 6.0×10^{-5} hartree \AA^{-1} , 4.0×10^{-5} hartree \AA^{-1} , 3.0×10^{-4} \AA , and 2.0×10^{-4} \AA , respectively, in the final optimization with $\gamma = \gamma_{\text{MAX}}$. Looser criteria with 10 times these values were used when $\gamma < \gamma_{\text{MAX}}$. All optimized structures without the artificial force, i.e., true TSs and true products, also met the tighter optimization criteria, where the RFO optimizer implemented in the GRRM program was employed in the TS optimization as well as in the product geometry optimization.

The initial random geometries were generated as follows:

- (A) Determine the random orientation of all the fragments:
 - (1) Make three random vectors with their origins at the center-of-mass of one fragment.
 - (2) These vectors are orthonormalized.
 - (3) The three vectors are considered as new *xyz* axes and the original geometry is represented by the new axes.
 - (4) perform steps 1 to 3 for all the fragments.
- (B) Now determine the random positions of the centers-of-mass of all the fragments:
 - (1) Among *M* fragments, one fragment is selected randomly and its center-of-mass position is placed at the origin.
 - (2) Then, another fragment is chosen randomly from the remaining *M* – 1 fragments; this second fragment is placed at a random position around the first fragment. Since this procedure may put two fragments too close to each other, the center-of-mass position of the second fragment is moved randomly (with a step size of 0.05 \AA) until the minimum distance from the first fragment exceeds a threshold (the sum of the covalent radii plus 0.8 \AA).
 - (3) Repeat this procedure for the third, fourth, .. and the last fragments.

There are some technical subtleties for a better implementation of this algorithm. The treatment starting with small γ and gradually increasing it by $0.1\gamma_{\text{MAX}}$ is one of these. With this trick, one can set the γ_{MAX} parameter to a relatively large value so that many reaction paths with a variety of barrier heights can be obtained in one search. For low-barrier paths, a very large γ allows the AFIR minimization to pass through a high-energy region far from the corresponding true TS. Hence, the smaller γ tends to give a better approximate TS. When the same product is reached, the maximum energy point along the previous path is compared with the one along the present path, and then, the path with a lower barrier is retained and the other is discarded. With this trick, a good approximate TS is obtained along low-barrier AFIR paths even if γ_{MAX} is very large.

Another point to consider in order to obtain a good approximate TS is a proper control of the trust radius d_{TR} around the TS region. In optimization of an energy minimum, the maximum and minimum values of d_{TR} are set to 0.5 and 0.001 \AA , respectively. Between these values, d_{TR} is controlled as follows: When an optimization step increased the function value, d_{TR} is scaled by 0.1, and when the function value is decreased, d_{TR} is scaled by 3.0. Although this step control works well for usual energy minimization, there have been some troubles in the AFIR minimization. Before a barrier, direction of the artificial force and

that of the true potential force are opposite, giving a normal norm of the total force. After passing through the barrier, these two vectors suddenly direct to the same direction, and the norm of the total force becomes too large. This sometimes causes an abnormal jump of atoms involved in the bond rearrangement, making a bump on the potential energy profile along the AFIR path. Hence, when an optimization step gave a structure containing new chemical bonds and when the step changed lengths of the new bonds more than 5%, d_{TR} is reduced to 0.05 \AA , and the optimization step is recalculated, where atoms *i* and *j* with a distance shorter than $1.2(R_i + R_j)$ were considered to be bonded in this study. With this trick, AFIR paths can be calculated very efficiently by the energy minimization method with a relatively large step size used in usual geometry optimization in other regions.

The final point is a special treatment of hydrogen (H) atoms in eq 2. As discussed in Section III, the artificial force accentuates the short-range orbital interactions so that sites with maximal orbital interactions appear as local minima on the AFIR function. Existence of long-range interactions, such as H-bonds, is explicitly seen as local minima even on bare $E(\mathbf{Q})$. In other words, reactive sites for hydrogen-/proton-transfer reactions can be found by energy minimization without the artificial force. Hence, we set the covalent radius parameter *R* in the weight function for H atoms to zero. It should be emphasized that this does not mean that hydrogen-/proton-transfer paths are omitted. Entrances of hydrogen-/proton-transfer paths can be found as local minima even if the weight is small. Once such a H-bond site is located, the artificial force is imposed to H atoms even if $R = 0$, and a hydrogen-/proton-transfer reaction occurs. Without this treatment, AFIR minimization with small γ frequently falls into H-bond sites because of the long-range H-bond interaction. In order to efficiently find reactive sites for bond formations between heavy atoms with small γ , this treatment is recommended.

General Remarks. The present method may be categorized into the bias potential approaches.^{82–85,44} Such methods have been used for accelerating molecular dynamics as well as global optimization. The common idea is to eliminate unwanted local minima using bias potentials constructed by unique mathematical procedures and data collected by extensive PES sampling. However, these methods may not be very effective in the present purpose. Strategies for eliminating flat regions of PESs for weakly interacting reactants should be very different from those for eliminating closed valleys. Hence, the bias employed in the present study is not similar to those for the previous approaches.

What is the best bias potential to eliminate the flat regions of PESs? A variety of functions, such as the harmonic, the Morse, and the Gaussian functions, may be possible candidates. The use of the linear function αr_{AB} in eq 1 was considered by assuming collisions among reactant molecules. In collisions, the driving force to overcome a reaction barrier is the inertia force to keep the translation of reactant molecules toward the TS. Hence, we chose to impose the (model inertia) force using the linear function. The force completely changes the landscape of the PES; the flat asymptote is completely eliminated by the linear function bias, as seen in Figure 1c.

One significant advantage of the present method is that the bias potential is defined uniquely at any given geometry without prior sampling of PES. This advantage arises from a single assumption in the AFIR method that a part of reactant species is connected with chemical bonds in the product. In this sense,

the present method is in part related to previous methods that assume a product structure or a small set of reaction variables.^{37–46} However, the assumption in the AFIR method is much more flexible than those. Consequently, many reaction paths giving unexpected products through unexpected reaction variables are discovered in the AFIR method. In other words, AFIR is essentially a single-end method starting from given reactants as this single assumption does not specify which product will be found. More importantly, no empirical guess of the mechanism is required in the AFIR method.

We acknowledge two methods are especially related to AFIR. One is SCD^{5,6,43} (and its modern form of RGF).^{57–63} If the force is applied between a pair of atoms k and l (when $F(r_{kl}) = E(r_{kl}) + \alpha r_{kl}$ is minimized), then the minimization path should be very similar to the path followed by SCD for r_{kl} . Another is the steered molecular dynamics,⁸⁶ which applies a force to specified atoms to accelerate a particular event in molecular dynamics or to simulate single-molecule experimental techniques, such as atomic force microscopy and optical tweezers. The uniqueness of AFIR is that the direction of the force or the driving coordinate(s) is automatically given by eq 2. Among the numerous possible directions of the force or the huge combinations of the coordinates, eq 2 can choose suitable ones automatically, which enables an automated search. Despite the automatic selection of the driving coordinates, there are few AFIR paths wandering into dead end valleys without any reaction, which is demonstrated in Section IV.

Although we discussed only technical aspects of the AFIR method above, there is another important effect of the artificial force, that is, accentuate the local quantum chemical (frontier orbital) interactions leading to a TS. This helps finding an entrance of reaction paths in the automated search. In other words, the use of an artificial force to push the reactants together is no more than a starting point. After many trails and errors in adjusting the form of the force term, we discovered a special function (AFIR function) on which all reactive sites among multiple reactants can be identified as local minima. The success of the AFIR function is understandable by considering the FMO theory as illustrated in Section III using a very simple reaction $\text{CO}_2 + \text{H} \rightarrow \text{HCO}_2/\text{HOCO}$ as an example.

III. AN EXAMPLE OF LANDSCAPE OF THE AFIR FUNCTION

To illustrate a typical landscape of the AFIR function, we revisit a reaction of $\text{CO}_2 + \text{H} \rightarrow \text{HCO}_2/\text{HOCO}$. This reaction is the first step of $\text{CO}_2 + \text{H} \rightarrow \text{CO} + \text{OH}$ or the final step of the inverse $\text{CO} + \text{OH} \rightarrow \text{CO}_2 + \text{H}$.⁸⁷ Figure 3a shows an interaction potential energy contour map between a CO_2 molecule and a H atom at the B3LYP/cc-pVTZ level, where the positions of the C and O atoms in CO_2 were fixed at those of optimized CO_2 . A similar contour map is reported in ref 87 with a different computation level. There are very shallow minima in the diagonal directions, probably because of the basis set superposition error. These minima are not important in chemical reactions.

Although all directions are almost completely repulsive, many chemists may recognize the reactive sites by looking at the FMOs of CO_2 shown in Figure 3b. The lowest unoccupied molecular orbitals (LUMOs) are degenerate π orbitals with a node on each C=O bond. The highest occupied molecular orbitals (HOMOs) of CO_2 are degenerate π orbitals with a node on the C atom. One reactive site should be the perpendicular direction (see the red arrow in Figure 3a) as LUMOs have the largest lobe toward this

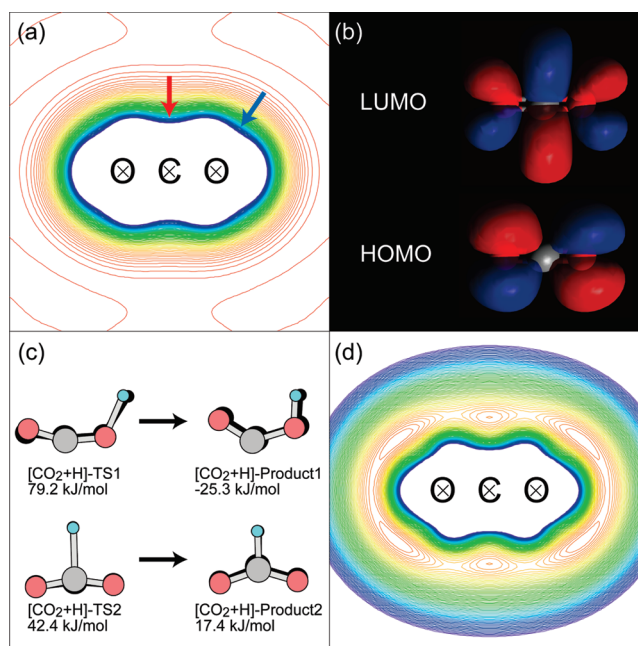


Figure 3. (a) The interaction potential contour map between a CO_2 molecule and a H atom at the B3LYP/cc-pVTZ level, where the contour spacing is 5 kJ/mol and reactive sites are indicated by arrows; (b) the FMOs of a CO_2 molecule at the same computation level; (c) true TS and product structures for the $\text{CO}_2 + \text{H}$ reaction at the same level, where approximate TS and product structures along an AFIR path (see text for details) are shown behind the true structures with blackbody; and (d) the contour plot of the AFIR function with $\gamma = 200$ kJ/mol for the interaction between a CO_2 molecule and a H atom at the B3LYP/cc-pVTZ level.

direction. Another should be the diagonal direction (see the blue arrow in Figure 3a) because of the lobe of HOMOs in this direction. The interaction between these frontier orbitals and the 1s orbital of the H atom lowered the potential energy in these directions, and consequently, dents are seen in these directions in Figure 3a. Actually, the reaction has been found to take place when the H atom collides toward either of these dents (see TSs in Figure 3c).

If a function has local minima only in the reactive directions, then one can find them easily by minimizing the function. The AFIR function does have such a landscape as shown in Figure 3d, which is a contour plot of the AFIR function with $\gamma = 200$ kJ/mol (the second term of eq 2 is added to the plot of Figure 3a). The values of the AFIR function linearly decrease from the long distance to the medium distance because of the artificial force. In the short distance, the strong repulsion between two particles supersedes the artificial force, and the resulting potential curve has a minimum in the medium distance. There we see a clear angular dependence of the depth of the minimum and find local minima at the reactive sites. Once such a reactive site is found as a local minimum of the AFIR function, by increasing the γ value, one can find an approximate reaction path, along which approximate TS and product points can be obtained.

Figure 3c compares approximate AFIR structures with true TSs and true products, where AFIR structures obtained by the algorithm in Figure 2 with $N_{\text{MAX}} = 10$ and $\gamma_{\text{MAX}} = 200$ kJ/mol are shown in black behind the true stationary structures. The AFIR structures, especially the approximate TS structures, are quite

similar to the true ones. Moreover, all AFIR paths with different initial orientations gave one of these two unique reaction paths; there was no miss shot and no useless vibrations or fluctuations in the AFIR search. Before the completion of the algorithm, 564 gradients and 21 Hessian were computed in this application.

As seen in Figure 3a, the frontier (and other important) orbital interactions leading to chemical reactions should have significant effects on the repulsive potential walls. However, orbital interaction works in relatively short ranges compared to H-bond interaction as well as electrostatic and dispersion interactions. Hence, dents indicating entrances of reaction paths are not always seen clearly as local minima. By pressing the reactants to each other by the artificial force to accentuate such dents in the short-range (see Figure 3d), locations of maximal orbital interaction can be visualized as local minima so that the computer can detect them without extensive sampling or scans. Although the FMO theory itself is also very powerful in predictions of reactive sites, it does not work very well when many orbitals are involved at once or when changes in shapes and/or characters of FMOs are significant due to strong interactions with the collision partner. On the other hand, the AFIR method looks at the shape of PES avoiding these difficulties in the FMO theory.

IV. TEST CALCULATIONS

Aldol Reaction. As a case study, the reaction between vinyl alcohol ($\text{H}_2\text{C}=\text{CH}-\text{OH}$) and formaldehyde (H_2CO) was studied by the AFIR method. In this and the following applications, gradients and Hessians were computed at the B3LYP/6-31G level by the Gaussian09 program.⁸⁸ This calculation is expected to give the products of the aldol reaction. In the aldol reaction, vinyl alcohol is assumed to be an important intermediate which reacts with another reactant, an aldehyde molecule.⁸⁹ Hence, the most important point in this application is whether the AFIR method can find the product of aldol reaction or not. At $\gamma_{\text{MAX}} = 100$ kJ/mol, only one path was obtained, which is shown in Figure 4. In Figure 4 and others shown below, six structures are presented for each path: a starting random structure for the corresponding AFIR minimization, a reactant complex structure found by an IRC calculation from the corresponding true TS structure, a true TS structure, a highest energy structure along the corresponding AFIR path (shown in black behind the true TS structure for comparison), a true product structure found by an IRC calculation from the corresponding true TS structure, and a product structure with the corresponding AFIR path (shown in black behind the true product structure). The only path shown in Figure 4 actually corresponds to the final step of aldol reaction. As seen in the TS structure, the approximate structure along the AFIR path is quite similar to the true TS structure. The approximate product structure along the AFIR path is also similar to the true product structure. Thus, the AFIR method worked very well. In this application with $N_{\text{MAX}} = 30$, 33 random structures were considered, and 3490 gradient and 87 Hessian calculations were performed in total.

To find the higher barrier paths systematically, γ_{MAX} was set to a very large value of 1000 kJ/mol. All the obtained paths are listed in Figure 5. The order of the path ID numbers represents the order in which the structure was found in the AFIR search. The most important path leading to the aldol product was found first among the 13 paths. This is probably because the acceptance region of this path is much wider than the others because of the low barrier of this path. All the other paths have barriers higher

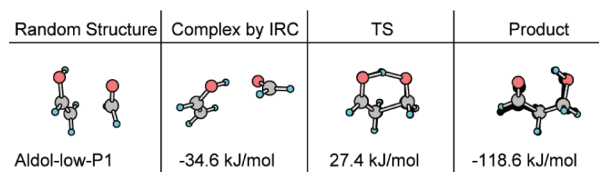


Figure 4. The reaction path between $\text{H}_2\text{C}=\text{CH}-\text{OH}$ and H_2CO obtained by the AFIR search with $\gamma_{\text{MAX}} = 100$ kJ/mol at the B3LYP/6-31G level. Six structures with their energies relative to the isolated reactants are presented: a starting randomly oriented structure as the starting point for the AFIR minimization, a reactant complex structure found by an IRC calculation from the corresponding true TS structure, a true TS structure, a highest energy structure along the corresponding AFIR path (shown in black behind the true TS structure for comparison), a true product structure found by an IRC calculation from the corresponding true TS structure, and a product structure along the corresponding AFIR path (shown in black behind the true product structure).

than 100 kJ/mol. Although deviations between approximate TS and true TS structures are large in some paths, these are still acceptable since in all cases geometry optimization by the RFO method starting from these approximate TS structures converged within 30 optimization steps to the true TS structures. However, in future applications, there may be cases in which optimization of a true TS geometry fails due to a very poor AFIR guess. In such cases, one of the double-end methods^{11,12,37-42} could be applied to the corresponding pair of product and reactant. Besides these paths, many biradical hydrogen-transfer paths were obtained giving $\text{H}_2\text{C}=\text{CH}-\text{O}^\bullet + \text{H}_2^\bullet\text{COH}$, $\text{H}_2\text{C}=\text{C}^\bullet-\text{OH} + \text{H}_2^\bullet\text{COH}$, $\text{H}^\bullet\text{C}=\text{CH}-\text{OH} + \text{H}_2^\bullet\text{COH}$, $\text{H}_2\text{C}=\text{CH}-\text{O}^\bullet + \text{H}_3\text{CO}^\bullet$, $\text{H}_2\text{C}=\text{C}^\bullet-\text{OH} + \text{H}_3\text{CO}^\bullet$, $\text{H}^\bullet\text{C}=\text{CH}-\text{OH} + \text{H}_3\text{CO}^\bullet$, $\text{H}_2\text{C}=\text{C}^\bullet\text{H}\cdots\text{OH}_2 + \text{H}^\bullet\text{CO}$, $\text{H}_2\text{C}-\text{CH}_2-\text{OH} + \text{H}^\bullet\text{CO}$, and $\text{H}_3\text{C}-\text{CH}-\text{OH} + \text{H}^\bullet\text{CO}$. This result demonstrates that the AFIR method, with the combination of the AFIR function in eq 2 and the algorithm in Figure 2, has the ability to find many unknown paths automatically and systematically without prejudice toward assumed mechanisms. In this application with $N_{\text{MAX}} = 50$, 505 random structures were considered, and 59 208 gradient calculations and 1486 Hessian calculations were performed in total. As expected, the cost is much larger in this application than the above application with the small γ_{MAX} due to many high-barrier paths. One may choose to use a small γ_{MAX} if high-barrier paths are not interested.

In the search with $\gamma_{\text{MAX}} = 1000$ kJ/mol, there were five unique AFIR paths that ended without any reaction. For completeness, these paths were further followed with $\gamma = 5000$ kJ/mol starting from these nonreacted end points, and all of these five finally reacted (see Figure S1 in Supporting Information for these additional five paths). In other words, there was no AFIR path that ran into a dead end valley in this particular example. The situation was the same also in the above example for $\text{CO}_2 + \text{H}$. As demonstrated for $\text{CO}_2 + \text{H}$ in Figure 3, all reactive sites emerge as local minima on the AFIR function, which helped to automatically find effective coordinates to be driven by the artificial force also in the case of $\text{H}_2\text{C}=\text{CH}-\text{OH} + \text{H}_2\text{CO}$. One problem in the present algorithm is that each AFIR path is found again and again starting from many random orientations. In the case of the search with $\gamma_{\text{MAX}} = 1000$ kJ/mol, 27 unique paths were obtained from 505 orientations; each product was reached 18.7 times on average. It follows that most paths can be discarded in the course of the minimizations before they reach the products when some

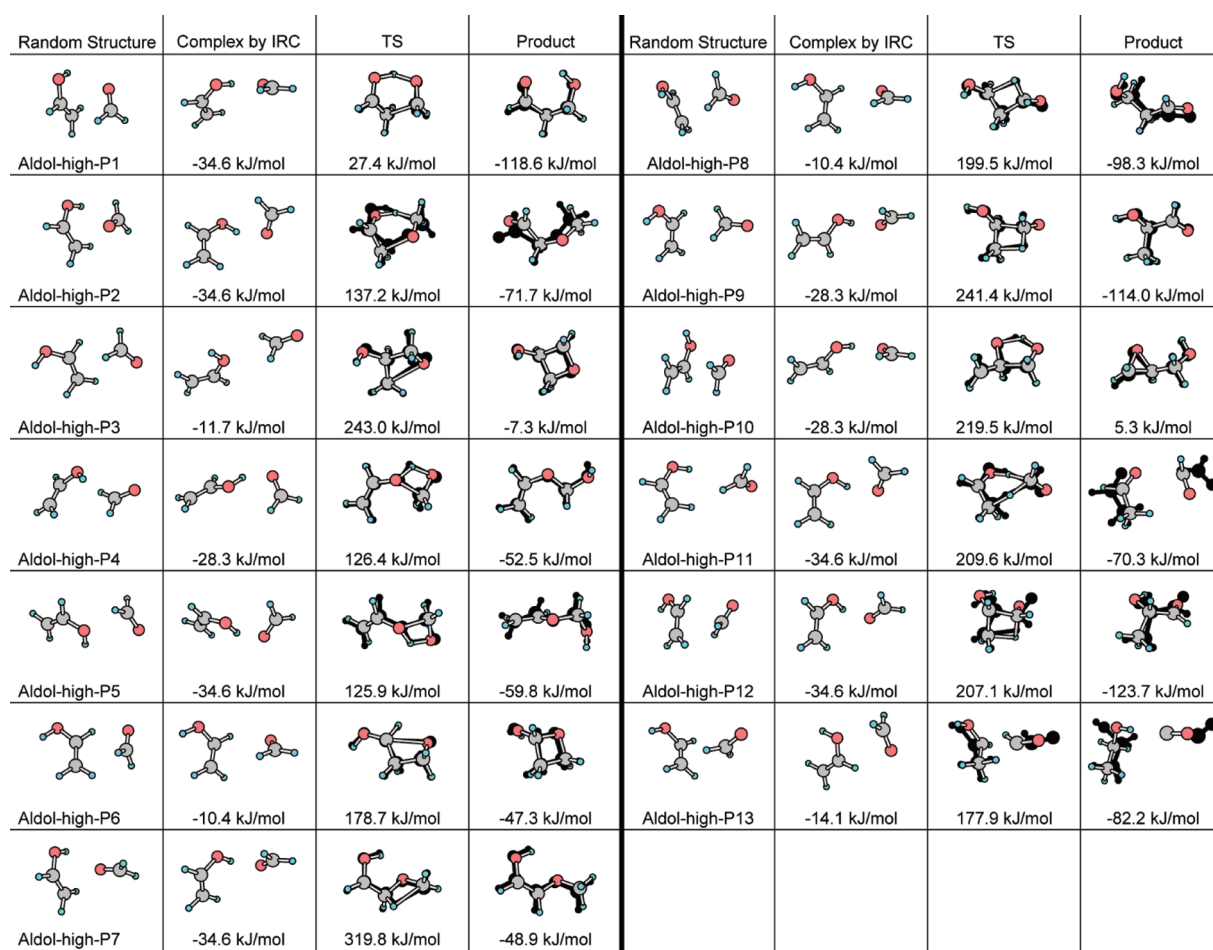


Figure 5. The reaction paths between $\text{H}_2\text{C}=\text{CH}-\text{OH}$ and H_2CO obtained by the AFIR search with $\gamma_{\text{MAX}} = 1000$ kJ/mol at the B3LYP/6-31G level. See the caption of Figure 4 for the meanings of each structure.

similarities are found between the present path and one of previous paths. This trick, once implemented, will reduce the total computation cost dramatically.

Aldol Reaction Catalyzed by a Single Water Molecule. As a case study of three component reactions, one water molecule is reacted together with $\text{H}_2\text{C}=\text{CH}-\text{OH}$ and H_2CO . In this application, in addition to the attraction between $\text{H}_2\text{C}=\text{CH}-\text{OH}$ and H_2CO , two more terms were added to eq 2, one for attraction between $\text{H}_2\text{C}=\text{CH}-\text{OH}$ and H_2O and the other between H_2CO and H_2O . We chose the same α for the three force terms, and thus the γ value is divided three in calculations of α with eq 3. It is possible to apply different α to each term, however, this is not recommended. As discussed above, one important role of the AFIR function is to accentuate intrinsic topography on potential walls of original PESs, as shown in Figure 3d. Not to destroy the anisotropy of the topography, the AFIR function should be as simple as possible.

All paths obtained by a search with $\gamma_{\text{MAX}} = 100$ kJ/mol is presented in Figure 6. There are two aldol paths with different conformations (W1-P2 and W1-P5). With participation of one water molecule, an analog of aldol-high-P4 in Figure 5, i.e., W1-P1 in Figure 6, was found with this small γ_{MAX} . Two new types of paths opened, giving methylene-diol with (W1-P6 and W1-P8) and without (W1-P3, W1-P4, and W1-P7) the enol-keto tautomerization. All the AFIR paths provided good

approximate TS and product geometries also in this three-component system. In this application with $N_{\text{MAX}} = 30$, 97 random structures were considered, and 13 961 gradient and 340 Hessian calculations were performed in total.

Aldol Reaction Catalyzed by Two Water Molecules. In three component reactions among A, B, and C, the AFIR function having three force terms (for A-B, A-C, and B-C attractions) is unique. On the other hand, when four components A, B, C, and D are involved at once, there are six force terms for the A-B, A-C, A-D, B-C, B-D, and C-D attractions. One is the function with all six terms. However, such reactions giving at least six new chemical bonds in one step are highly unlikely. Hence, one needs to consider also AFIR functions with 3-5 force terms, which gives formally 38 possible force combinations. In addition to one six-term function above, there are six (6C_5) five-term possibilities, 15 (6C_4) four-term cases, and 16 (${}^6C_3 - 4$, where the four cases involve an isolated component and are treated as three-component reactions) three-term cases. In a fully systematic search, one should consider all of these 38 possibilities. In the present case study, only one of such cases is studied, where A, B, C, and D are $\text{H}_2\text{C}=\text{CH}-\text{OH}$, H_2CO , H_2O , and H_2O , respectively, and the AFIR function consists of four terms for the A-B, A-C, B-D, and C-D attractions. This is expected to give reaction paths between $\text{H}_2\text{C}=\text{CH}-\text{OH}$ and H_2CO involving two H_2O molecules as proton-transfer agents. If we

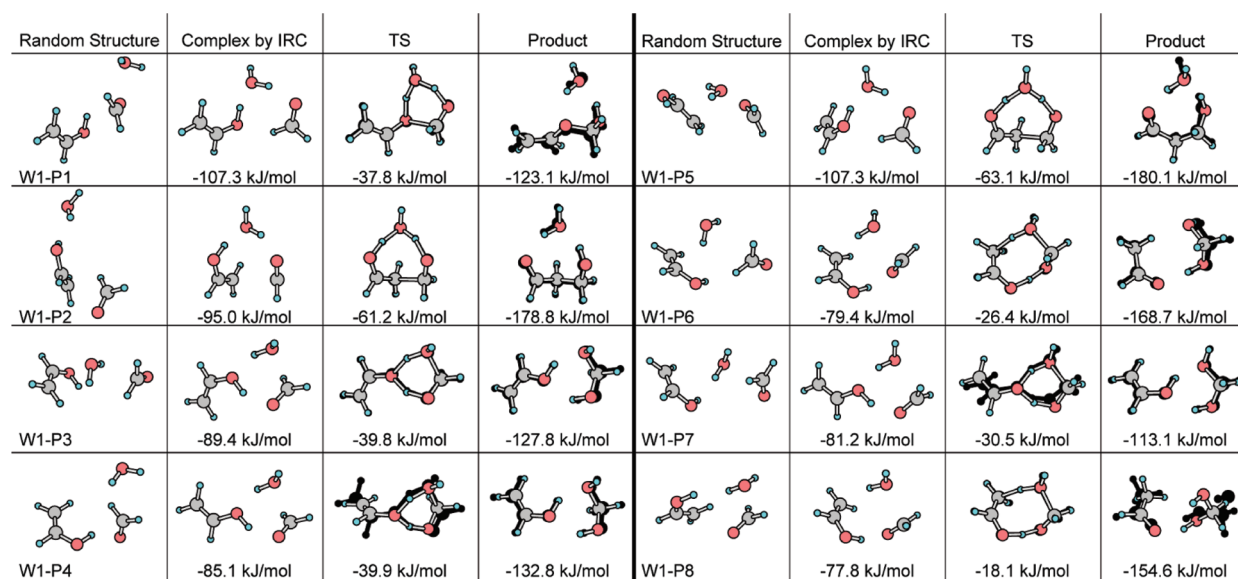


Figure 6. The reaction paths among $\text{H}_2\text{C}=\text{CH}-\text{OH}$, H_2CO , and H_2O obtained by the AFIR search with $\gamma_{\text{MAX}} = 100$ kJ/mol at the B3LYP/6-31G level. See the caption of Figure 4 for the meanings of each structure.

intend to investigate the catalytic role of two solvent molecules in a bimolecular reaction, then this is the unique choice among the 38 possibilities.

All paths obtained by a search with $\gamma_{\text{MAX}} = 100$ kJ/mol are presented in Figure 7. All products in Figure 7 are already seen in Figure 6, although in all paths, the second water molecule is explicitly involved as a proton-transfer agent. The AFIR paths again provided good estimates of TS and product geometries for full optimization. In this application with $N_{\text{MAX}} = 30$, 180 random structures were considered, and 29 508 gradient and 699 Hessian calculations were performed in total.

If Figures 4–7 are compared, then very large catalytic effects of the first and second water molecules are seen. However, the effects are significantly overestimated because of the basis set superposition errors of 6-31G. Figure 8 shows the aldol reaction paths at the B3LYP/cc-pVTZ level, where TSs were optimized starting from those for aldol-high/low-P1, W1–P5, and W2–P4. As shown in Figure 8, when the larger basis sets are employed, the energy lowering in relative potential energies is much smaller. Moreover, if standard-state free energies at room temperature are considered, participation of water molecules increases the barrier height. Hence, in gas phase, catalytic effects of water molecules are not important in this reaction.

V. CONCLUSIONS AND PERSPECTIVES

One goal in development of tools for exploring PESs would be establishing a general approach which can predict the entire sets of chemical reaction mechanisms automatically starting from a given set of reactants and catalysts. Such a method, once realized, will be very useful for exploring unknown chemical worlds by using computers. Toward this unachieved goal, we introduced the AFIR method for finding associative ($A + B \rightarrow X$) paths automatically. AFIR by pressing the reactants to each other with an artificial force can locate the reactive sites of each reactant very efficiently by accentuating the effects of (short-range) orbital interactions. Once such a reactive site is located, further pressing will lead to an approximate TS geometry and eventually to the

product. This procedure can be performed very efficiently just by minimizing a single function called the AFIR function. Approximate TS and product geometries can then easily be reoptimized to corresponding true TS and product structures by geometry optimization without the artificial force.

One of the most successful approaches for recognizing reactive sites should be the FMO theory and the Woodward–Hoffmann rules.^{33–36} Although these theories have been very powerful for explaining chemical reactivities, they have not been applied successfully to systematic predictions of reaction paths in combinations with quantum chemical calculations of PESs. Shapes of FMOs are known to change significantly by interaction with collision partners in some examples,^{90,91} and hence, FMOs for isolated molecules are not always very useful in predicting reactive sites. To our knowledge, for this particular purpose, there has been no systematic theoretical approach beyond these theories. On the other hand, AFIR has the ability to predict the reactive sites with quantum chemical calculations of PESs. Moreover, AFIR can predict TS geometries in courses of the AFIR minimization.

There is another unique feature of AFIR to be noted. In most methods for finding reaction paths without a guess of TS geometry, many cycles of constrained/penalty function optimization or extensive PES sampling are required before obtaining a path. However, in AFIR, just minimization of one single function gives a path. On the AFIR function, there is a local minimum (or minima) at each reactive site. This has been illustrated in Figure 3d and is obvious from many reaction paths shown in Figures 4–7 that were obtained just by minimizing the AFIR functions. Hence, from a given initial orientation, the system is drawn into one of these sites. If a H atom is placed at any direction of Figure 3d, it will automatically fall into one of the reactive sites. This feature makes search for each single path very efficient. Consequently, the automated stochastic search for many paths was possible even with the brute-force algorithm in Figure 2. Improvements of the stochastic part of the algorithm in Figure 2 may increase the efficiency in the total automated searches in the future.

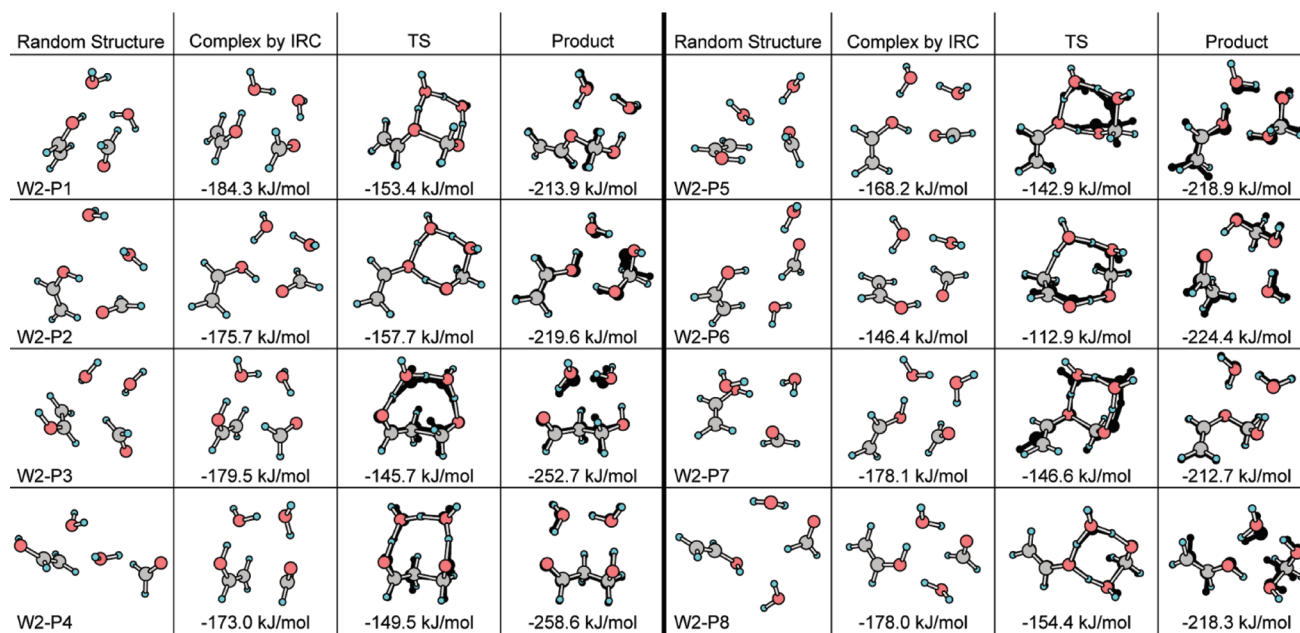


Figure 7. The reaction paths among $\text{H}_2\text{C}=\text{CH}-\text{OH}$, H_2CO , H_2O , and H_2O obtained by the AFIR search with $\gamma_{\text{MAX}} = 100$ kJ/mol at the B3LYP/6-31G level. See the caption of Figure 4 for the meanings of each structure.

We note that the AFIR search should be performed for all important conformers when reactants have several conformers. Small conformational changes may occur during AFIR minimization as seen in P4, P9, and P10 in Figure 5. However, for many reaction paths the conformation does not change. Therefore all the conformations need to be considered for the search. Hence, we suggest to perform conformation sampling for reactants before the AFIR search. Many previous researches are available regarding the expansion of the applicability of reaction path search methods to larger complex systems; see refs 92–96 for most successful examples for the EF and NEB methods. Their strategies may be helpful in future applications of AFIR to complex reactions.

In multistep mechanisms, unimolecular steps may also be involved. It is still under tests how well AFIR can be applied to unimolecular reactions. As discussed in the Introduction, some methods are already available for unimolecular paths. Among them, the ADDF method has been demonstrated to be very powerful for local chemical bond rearrangements.^{64–68} Hence, we believe that the goal of finding total chemical reaction mechanisms automatically starting from given reactants and catalysts will be achieved in the future by a combination of AFIR for $\text{A} + \text{B} \rightarrow \text{X} (+ \text{Y})$ steps and ADDF (or others) for $\text{A} \rightarrow \text{X} (+ \text{Y})$ steps.

There are some useful techniques to make applications of automated reaction path search methods to large flexible systems and to nonadiabatic (photochemical, ion–molecule, and spin-flipping) reactions.^{69,70} Although these techniques were originally introduced in a combination with the ADDF method, in principle they can be combined with any automated reaction path search method including AFIR. One of such techniques is the microiteration technique used in QM/MM geometry optimization.^{97–99} With microiteration and the effective Hessian approach, reaction paths concerned within a given reaction center can be explored automatically.⁶⁹ This technique is expected to greatly expand the applicability of AFIR to larger molecular systems. The development combining AFIR and

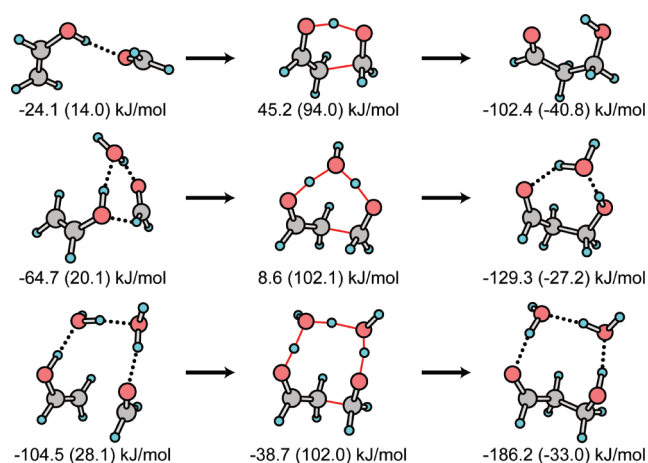


Figure 8. The aldol reaction paths involving zero to two water molecules at the B3LYP/cc-pVTZ level. Values in parentheses show relative standard-state free energies at room temperature.

microiteration within the ONIOM(QM:MM)¹⁰⁰ framework has already been done, and systematic tests are in progress for an organometallic catalytic reaction with flexible ligands. This approach is expected to develop into applications of AFIR to enzymatic reactions. In enzymatic reactions as well as reactions in solution phase, thermal entropy effects play a significant role. To account for the entropy efficiently, several powerful free energy methods are available, such as umbrella sampling,¹⁰¹ metadynamics,⁴⁴ free energy perturbation theory,¹⁰² etc. Development of interfaces with these free energy methods will be a very important subject in the future.

Another such technique is the seam model function (SMF) approach for nonadiabatic reactions. In the SMF approach, many minima on seam of crossing hypersurfaces (MSXs) are searched in two steps: automated exploration of many approximate MSX

structures using a penalty function by an automated reaction path search method, followed by tight MSX optimization using the approximate structures and an accurate MSX optimizer.⁷⁰ Here, “seam of crossing” includes conical intersections between states with the same spin and space symmetry. With SMF and ADFF, photodissociation mechanisms of small atmospheric species have been investigated systematically.^{70,103–106} We already combined SMF with AFIR and applied to exploration of nonadiabatic ignition pathways of unsaturated hydrocarbons.⁷⁵ In this application, we discovered unexpected ignition pathways through low-energy MSXs for aromatic hydrocarbons. Thus, applications of the automatic search method to photochemical, ion–molecule, and spin-flipping reactions are also possible. Interfaces with free energy methods as well as microiteration will be required also in this extension for nonadiabatic reactions in the future.

In these five decades, computational chemistry has become a powerful means for exploration of chemical worlds in collaboration with experiments. In many studies of chemical reactions, it has been a splendid guide to confirm or to judge ideas of computational and/or experimental chemists through the use of sophisticated PES exploration tools. One of the important subjects in the next generation should be predicting reaction paths without presumed mechanisms. This has been achieved at least for reactions of type $A + B \rightarrow X (+ Y)$ in small systems by the present AFIR method. Obviously, further development of automated reaction path search methods will be required in the future to cover reactions of many kinds in a variety of system size.

■ ASSOCIATED CONTENT

S Supporting Information. Additional five pathways for $H_2C=CH-OH+H_2CO$. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: keiji.morokuma@emory.edu; smaeda@fukui.kyoto-u.ac.jp.

■ ACKNOWLEDGMENT

This work is partly supported by a grant from Japan Science and Technology Agency with a Core Research for Evolutional Science and Technology (CREST) in the Area of High Performance Computing for Multiscale and Multiphysics Phenomena at Kyoto University as well as a grant from US AFOSR (grant no. FA9550-10-1-0304) at Emory University. S.M. thanks Prof. Koichi Ohno of Toyota Physical and Chemical Research Institute for helpful discussions and comments as well as his guidance on the reaction path search problem. We thank Prof. Masanobu Uchiyama of RIKEN and The University of Tokyo for helpful discussions and comments in the development. S.M. thanks Dr. Steven K. Burger of McMaster University for helpful discussions about reaction path search methods.

■ REFERENCES

- (1) Glasstone, S.; Laidler, K.; Eyring, H. *The Theory of Rate Processes*; McGraw-Hill: New York, 1941.
- (2) Murrell, J. N.; Laidler, K. J. *Trans. Faraday Soc.* **1968**, *64*, 371–377.
- (3) Brown, A.; Dewar, M. J. S.; Schoeller, W. J. *Am. Chem. Soc.* **1970**, *92*, 5516–5517.

- (4) McIver, J. W., Jr.; Komornicki, A. *J. Am. Chem. Soc.* **1972**, *94*, 2625–2633.
- (5) Hayes, D. M.; Morokuma, K. *Chem. Phys. Lett.* **1972**, *12*, 539–543.
- (6) Jaffe, R. L.; Hayes, D. M.; Morokuma, K. *J. Chem. Phys.* **1974**, *60*, 5108–5109.
- (7) Fukui, K. *J. Phys. Chem.* **1970**, *74*, 4161–4163.
- (8) Ishida, K.; Morokuma, K.; Komornicki, A. *J. Chem. Phys.* **1977**, *66*, 2153–2156.
- (9) Pulay, P. *Mol. Phys.* **1969**, *17*, 197–204.
- (10) Komornicki, A.; Ishida, K.; Morokuma, K.; Ditchfield, R.; Conrad, M. Efficient. *Chem. Phys. Lett.* **1977**, *45*, 595–602.
- (11) Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514–1527.
- (12) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; Wiley: Chichester, U.K., 2007.
- (13) Koga, N.; Morokuma, K. *Chem. Rev.* **1991**, *91*, 823–842.
- (14) Niu, S.; Hall, M. B. *Chem. Rev.* **2000**, *100*, 353–405.
- (15) Ziegler, T.; Autschbach, J. *Chem. Rev.* **2005**, *105*, 2695–2722.
- (16) Houk, K. N.; Cheong, P. H.-Y. *Nature* **2008**, *455*, 309–313.
- (17) Wales, D. J.; Scheraga, H. A. *Science* **1999**, *285*, 1368–1372.
- (18) Floudas, C. A.; Pardalos, P. M. *Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches*, Kluwer Academic Publishers: Dordrecht, The Netherlands, 2000.
- (19) Dobson, C. M.; Šali, A.; Karplus, M. *Angew. Chem., Int. Ed.* **1998**, *37*, 868–893.
- (20) Wales, D. J. *Int. Rev. Phys. Chem.* **2006**, *25*, 237–282.
- (21) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. *Annu. Rev. Biophys.* **2008**, *37*, 289–316.
- (22) Maddox, J. *Nature* **1988**, *335*, 201–201.
- (23) Woodley, S. M.; Catlow, R. *Nat. Mater.* **2008**, *7*, 937–946.
- (24) Müller, K.; Brown, L. D. *Theor. Chim. Acta* **1979**, *53*, 75–93.
- (25) Page, M.; McIver, J. W., Jr. *J. Chem. Phys.* **1988**, *88*, 922–935.
- (26) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154–2161.
- (27) Hratchian, H. P.; Schlegel, H. B. *J. Chem. Theory Comput.* **2005**, *1*, 61–69.
- (28) Schlegel, H. B. *J. Comput. Chem.* **1982**, *3*, 214–218.
- (29) Farkas, Ö.; Schlegel, H. B. *J. Chem. Phys.* **1999**, *111*, 10806–10814.
- (30) Cerjan, C. J.; Miller, W. H. *J. Chem. Phys.* **1981**, *75*, 2800–2801.
- (31) Császár, P.; Pulay, P. *J. Mol. Struct.* **1984**, *114*, 31–34.
- (32) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. J. *Phys. Chem.* **1985**, *89*, 52–57.
- (33) Fukui, K.; Yonezawa, T.; Shingu, H. *J. Chem. Phys.* **1952**, *20*, 722–725.
- (34) Fukui, K. *Acc. Chem. Res.* **1971**, *4*, 57–64.
- (35) Woodward, R. B.; Hoffmann, R. *J. Am. Chem. Soc.* **1965**, *87*, 395–397.
- (36) Woodward, R. B.; Hoffmann, R. *Angew. Chem., Int. Ed.* **1969**, *8*, 781–853.
- (37) Halgren, T. A.; Lipscomb, W. N. *Chem. Phys. Lett.* **1977**, *49*, 225–232.
- (38) Dewar, M. J. S.; Healy, E. F.; Stewart, J. J. P. *J. Chem. Soc., Faraday Trans. 2* **1984**, *80*, 227–233.
- (39) Elber, R.; Karplus, M. *Chem. Phys. Lett.* **1987**, *139*, 375–380.
- (40) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9901–9904.
- (41) E, W.; Ren, W.; Vanden-Eijnden, E. *Phys. Rev. B* **2002**, *66*, 052301/1–4.
- (42) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. *J. Chem. Phys.* **2004**, *120*, 7877–7886.
- (43) Černohorský, M.; Kettou, S.; Koča, J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 705–712.
- (44) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.
- (45) Ensing, B.; De Vivo, M.; Liu, Z.; Moore, P.; Klein, M. L. *Acc. Chem. Res.* **2006**, *39*, 73–81.
- (46) Burger, S. K.; Ayers, P. W. *J. Chem. Theory Comput.* **2010**, *6*, 1490–1497.

- (47) Pancfř, J. *Collect. Czech. Chem. Commun.* **1974**, *40*, 1112–1118.
- (48) Basilevsky, M. V.; Shamov, A. G. *Chem. Phys.* **1981**, *60*, 347–358.
- (49) Basilevsky, M. V. *Chem. Phys.* **1982**, *67*, 337–346.
- (50) Rowe, D. J.; Ryman, A. J. *Math. Phys.* **1982**, *23*, 732–735.
- (51) Hoffman, D. K.; Nord, R. S.; Ruedenberg, K. *Theor. Chim. Acta* **1986**, *69*, 265–279.
- (52) Jørgensen, P.; Jensen, H. J. A.; Helgaker, T. *Theor. Chim. Acta* **1988**, *73*, 55–65.
- (53) Quapp, W. *Theor. Chim. Acta* **1989**, *75*, 447–460.
- (54) Schlegel, H. B. *Theor. Chim. Acta* **1992**, *83*, 15–20.
- (55) Sun, J.-Q.; Ruedenberg, K. *J. Chem. Phys.* **1993**, *98*, 9707–9714.
- (56) Bondensgård, K.; Jensen, F. J. *Chem. Phys.* **1996**, *104*, 8025–8031.
- (57) Quapp, W.; Hirsch, M.; Imig, O.; Heidrich, D. *J. Comput. Chem.* **1998**, *19*, 1087–1100.
- (58) Quapp, W.; Hirsch, M.; Heidrich, D. *Theor. Chem. Acc.* **1998**, *100*, 285–299.
- (59) Bofill, J. M.; Anglada, J. M. *Theor. Chem. Acc.* **2001**, *105*, 463–472.
- (60) Crehuet, R.; Bofill, J. M.; Anglada, J. M. *Theor. Chem. Acc.* **2002**, *107*, 130–139.
- (61) Dallos, M.; Lischka, H.; Monte, E. V. D.; Hirsch, M.; Quapp, W. *J. Comput. Chem.* **2002**, *23*, 576–583.
- (62) Hirsch, M.; Quapp, W. *J. Comput. Chem.* **2002**, *23*, 887–894.
- (63) Hirsch, M.; Quapp, W. *J. Mol. Struct. (Theochem)* **2004**, *683*, 1–13.
- (64) Ohno, K.; Maeda, S. *Chem. Phys. Lett.* **2004**, *384*, 277–282.
- (65) Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2005**, *109*, 5742–5753.
- (66) Ohno, K.; Maeda, S. *J. Phys. Chem. A* **2006**, *110*, 8933–8941.
- (67) Ohno, K.; Maeda, S. *Phys. Scr.* **2008**, *78*, 058122/1–8.
- (68) Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2007**, *111*, 4527–4534.
- (69) Maeda, S.; Ohno, K.; Morokuma, K. *J. Chem. Theory Comput.* **2009**, *5*, 2734–2743.
- (70) Maeda, S.; Ohno, K.; Morokuma, K. *J. Phys. Chem. A* **2009**, *113*, 1704–1710.
- (71) Irikura, K. K.; Johnson, R. D., III. *J. Phys. Chem. A* **2000**, *104*, 2191–2194.
- (72) Müller, E. M.; de Meijere, A.; Grubmüller, H. *J. Chem. Phys.* **2002**, *116*, 897–905.
- (73) Maeda, S.; Morokuma, K. *J. Chem. Phys.* **2010**, *132*, 241102/1–4.
- (74) Maeda, S.; Komagawa, S.; Uchiyama, M.; Morokuma, K. *Angew. Chem., Int. Ed.* **2011**, *50*, 644–649.
- (75) Maeda, S.; Saito, R.; Morokuma, K. *J. Phys. Chem. Lett.* **2011**, *2*, 852–857.
- (76) Collins, M. A. *Theor. Chem. Acc.* **2002**, *108*, 313–324.
- (77) Culot, P.; Dive, G.; Nguyen, V. H.; Ghuysen, J. M. *Theor. Chim. Acta.* **1992**, *82*, 189–205.
- (78) Broyden, C. G. *J. Inst. Math. Appl.* **1970**, *6*, 76–90.
- (79) Fletcher, R. *Comput. J. (Switzerland)* **1970**, *13*, 317–322.
- (80) Goldfarb, D. *Math. Comput.* **1970**, *24*, 23–26.
- (81) Shanno, D. F. *Math. Comput.* **1970**, *24*, 647–656.
- (82) Crippen, G. M.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1969**, *64*, 42–49.
- (83) Piel, L.; Kostrowick, J.; Scheraga, H. A. *J. Phys. Chem.* **1989**, *93*, 3339–3346.
- (84) Huber, T.; Torda, A. E.; van Gunsteren, W. E. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 695–708.
- (85) Grubmüller, H. *Phys. Rev. E* **1995**, *52*, 2893–2906.
- (86) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (87) Schatz, G. C.; Fitzcharles, M. S.; Harding, L. B. *Faraday Discuss. Chem. Soc.* **1987**, *84*, 359–369.
- (88) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.2; Gaussian, Inc.: Wallingford, CT, 2009.
- (89) Kürti, L.; Czako, B. Aldol Reaction. In *Strategic Applications of Named Reactions in Organic Synthesis*; Elsevier B.V.: Amsterdam, The Netherlands, 2005; pp 8–9.
- (90) Fukui, K.; Koga, N.; Fujimoto, H. *J. Am. Chem. Soc.* **1981**, *103*, 196–197.
- (91) Fukui, K. *Science* **1982**, *218*, 747–754.
- (92) Munro, L. J.; Wales, D. J. *Phys. Rev. B* **1999**, *59*, 3969–3980.
- (93) Henkelman, G.; Jónsson, H. *J. Chem. Phys.* **1999**, *111*, 7010–7022.
- (94) Trygubenko, S. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *120*, 2082–2094.
- (95) Carr, J. M.; Trygubenko, S. A.; Wales, D. J. *J. Chem. Phys.* **2005**, *122*, 234903/1–7.
- (96) Sheppard, D.; Terrell, R.; Henkelman, G. *J. Chem. Phys.* **2008**, *128*, 134106/1–10.
- (97) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170–1179.
- (98) Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760–769.
- (99) Vreven, T.; Frisch, M. J.; Kudin, K. N.; Schlegel, H. B.; Morokuma, K. *Mol. Phys.* **2006**, *104*, 701–714.
- (100) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.
- (101) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (102) Hu, H.; Yang, W. *Annu. Rev. Phys. Chem.* **2008**, *59*, 573–601.
- (103) Zhang, P.; Maeda, S.; Morokuma, K.; Braams, B. J. *J. Chem. Phys.* **2009**, *130*, 114304/1–10.
- (104) Maeda, S.; Ohno, K.; Morokuma, K. *J. Phys. Chem. Lett.* **2010**, *1*, 1841–1845.
- (105) Nádasi, R.; Zügner, G. L.; Farkas, M.; Dóbe, S.; Maeda, S.; Morokuma, K. *ChemPhysChem* **2010**, *11*, 3883–3895.
- (106) Xia, H.; Maeda, S.; Morokuma, K. *J. Phys. Chem. Lett.* **2011**, *2*, 934–938.

Early Experiences with Computational Quantum Chemistry

Russell M. Pitzer*

Department of Chemistry, The Ohio State University, Columbus, Ohio 43210, United States

ABSTRACT: A description is given of the progress in computational quantum chemistry in the early 1960s, the time of the early mainframe computers. In particular, the first calculation of the barrier to internal rotation in ethane and the first molecular application of perturbed self-consistent-field equations are described with the accompanying developments, trials, and tribulations.

My Ph.D. research began in the summer of 1960 with W. N. (Colonel) Lipscomb, who had moved to Harvard the previous year. Most of his graduate students were doing crystal structures of boron hydrides, but he had told me that he thought it was now almost possible to do a theoretical calculation of the barrier to internal rotation in ethane.

He said that Michael Barnett and co-workers, working in J. C. Slater's group at MIT, were making good progress in evaluating the needed multicenter integrals over Slater orbitals. Presumably, another reason why he suggested this problem was that he had seen how much computers could help in solving X-ray crystal structures and had anticipated their impact on quantum chemistry as well.

The early theoretical discussions of the ethane internal rotation barrier had been made in the context of the perfect-pairing approximation in valence bond theory, which involved the largely unknown effect of the neglect of many integrals.¹ By 1960, it had been established that the molecular orbital approach was much more computationally efficient.² Thus, the goal of this and many other calculations at this time was a self-consistent-field (SCF) calculation with a minimum basis set of Slater orbitals.

Other students in the Lipscomb group doing theoretical work were Bill Kern (magnetic resonance parameters), Larry Lohr (semiempirical theory), and Roald Hoffman, just returning from a year in Russia and developing extended Hückel theory to apply to carboranes. Bill Kern introduced me to the new Fortran II and to Michael Barnett. The advantages of using a high-level programming language were becoming accepted, but some research groups continued to use assembler languages. Bill also showed me around the main computer center available, an IBM 704 at the regional New England facility at MIT. It was a vacuum-tube computer with 32k 36-bit words of memory and many tape drives. Input was by IBM punched cards. We drove to MIT every day or so to submit jobs and pick up output. The programs for molecular integrals were in various stages of development, and there was no SCF program available for use or adaptation. When the crystallography students found out that we went to MIT often, they asked us to get their liquid nitrogen containers filled up there also.

The Barnett–Coulson method (for Slater orbitals) was to expand all functions in the integrand in spherical harmonics about one of the nuclei (centers) involved.³ The simplest example of this expansion is

$$\frac{e^{-\beta r_b}}{r_b} = \beta \sum_{n=0}^{\infty} (2n + 1) P_n(\cos \theta_a) i_n(\beta r_<) k_n(\beta r_>)$$

where $r_< = \text{minimum}(r_a, \rho_b)$, $r_> = \text{maximum}(r_a, \rho_b)$, and i_n and k_n are spherical Bessel functions, giving an expansion of an exponential function at center b in terms of spherical coordinates at center a . ρ_b is the distance between centers, and center b is on the z axis of the coordinates. Other powers of r_b can be obtained by differentiation with respect to β .

All one- and two-center integrals except two-center exchange resulted in finite expressions. Two-center exchange and all three-center integrals resulted in singly infinite angular momentum expansions. Four-center integrals resulted in triply infinite expansions with triangle conditions on the angular momentum indices. The expansions almost always converged to the desired 0.000001 accuracy in 50 terms or less, but they all had to be checked.

We were in correspondence with Martin Karplus and Shi Shavitt (at Columbia University and IBM Watson Lab) who were using an entirely different approach (Gaussian transforms) to the integral evaluation.⁴ The basic transform is

$$e^{-\beta r} = (\beta/\sqrt{4\pi}) \int_0^{\infty} s^{-3/2} e^{-\beta^2/(4s)} e^{-sr^2} ds$$

which, when used for all four orbitals gives a readily evaluated integral over Gaussian functions, which can then be integrated over the transform variables.

I visited them twice to discuss comparisons on test cases, which was quite helpful in debugging the programs. Col. Lipscomb sent me to the 1962 Gordon Conference on Theoretical Chemistry (forerunner of the present ACTC Conferences), which gave me the opportunity to meet a number of people in the field, including people whose papers I had been reading.

Organizing the integrals for the as yet unwritten SCF program was done by making a list of the symmetry-unique integrals using the D_{3h} and D_{3d} molecular symmetries of eclipsed and staggered ethane. The list I generated was 1653-integrals-long in both cases, with some integrals appearing in both lists. Much later, I discovered that one of these integrals was zero by symmetry, but by that time it was easier just to keep the same list and enter the 0.0 value. The integrals came from several different programs and had to have the orientation of their p orbitals rotated to the molecular axes, so the values were entered into a notebook along with their two or four orbital indices and an "occurrence factor" determined by

Received: May 17, 2011

Published: June 30, 2011

their symmetry and permutation properties.⁵ This stage of hand-processing was the most error-prone part of the work and was rechecked extensively. Each integral and its accompanying information was put on its own punch card.

I read enough about supermatrix formulations of SCF calculations^{6,7} to write an SCF program specific to ethane using these ideas. By late 1962, I had complete sets of integrals and was ready to run the SCF program. By this time, the regional MIT center had an IBM 7090 (transistorized) machine and the Slater group had purchased their own (used) IBM 709 machine, which was set up in a large space under one of MIT's domes. The SCF calculations took comparatively little time, and I soon had a rotational barrier value of 3.3 kcal/mol, reasonably close to the then-quoted experimental value of 2.875 ± 0.125 kcal/mol.⁸ The ethane barrier could not be measured by simple straightforward spectroscopic methods because of its lack of a dipole moment, but more sophisticated techniques later gave a value of 2.90 ± 0.03 kcal/mol.⁹

The credibility of such calculations in 1963 was a serious problem. Calculations at this level (SCF, minimum basis set of Slater orbitals) had been published previously on smaller molecules, with the integrals computed using desk calculators, and some of them were known to contain errors due to their overall difficulty. For example, the April 1960 issue of *Reviews of Modern Physics* contains papers from a theoretical chemistry conference. A paper by B. J. Ransil contains similar calculations, including exponent optimization, for a number of diatomic molecules and were done with no hand processing.¹⁰ I think the results were generally considered to be correct, but of different accuracy for different molecules. On the other hand, there were two similar papers on formaldehyde, and their computed dipole moments had opposite signs.^{11,12} In the case of ethane, a number of similar calculations were subsequently performed within a few years, and all gave similar results, reinforcing the validity of my calculations. It took an additional number of years and the establishment of several complete program systems before disagreement of reported experimental and theoretical results was considered a serious matter requiring further examination of both results. By that time, contracted Gaussian basis sets had been shown to be more efficient computationally than Slater basis sets.

The rotational barrier in ethane turned out to have been a quite fortunate choice of computational project because alkanes do not have the complications of multiple bonds or lone pairs of electrons. One aspect of this is that alkanes have a higher ratio of basis functions to number of electrons than other types of molecules. In comparison, the skew structure of hydrogen peroxide is not obtained until larger basis sets are used.¹³ H. F. Schaefer has summarized this result in private and perhaps in public by saying that minimum basis sets give surprisingly good results for structural properties of alkanes.¹⁴

An additional project during this time involved a younger graduate student, Dick Stevens. Oktay Sinanoglu had given a seminar on his ideas about correlation energy and had spoken to Col. Lipscomb about the possible effects of electron correlation on magnetic properties. This was to be Dick's project. On closer examination, we saw the need to get Hartree–Fock values of such properties first. We worked out the equations for perturbed Hartree–Fock theory using a basis set expansion and notation similar to Roothaan's² for the effects of uniform external magnetic and electric fields. These equations are coupled between the occupied MOs. Other workers had spent considerable effort to find ways to make approximations to uncouple the equations. Our view of the problem at the time was that solving the coupled

equations was not as difficult as doing the integrals, so why not go ahead and solve them with no further approximation? Dick set out to do the programming. He had been given an object deck for a diatomic integral program, but not the source deck. Dick's response was to obtain a core dump of the program and then proceed to translate the numerical op codes into an assembler version for his work, which seemed to me to be a rather daunting task. In due time, he had a program to study the electric and magnetic properties of diatomic molecules and first applied it to LiH.¹⁵ His results helped to establish some of the magnetic properties of this molecule.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pitzer@chemistry.ohio-state.edu.

REFERENCES

- (1) Wilson, E. B., Jr. The Problem of Barriers to Internal Rotation in Molecules. *Adv. Chem. Phys.* **1959**, *2*, 367–393.
- (2) Roothaan, C. C. J. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951**, *23*, 69–89.
- (3) Barnett, M. P.; Coulson, C. A. The Evaluation of Integrals Occurring in the Theory of Molecular Structure. Parts I & II. *Philos. Trans. R. Soc. London* **1951**, *243*, 221–249. Barnett, M. P. The Evaluation of Molecular Integrals by the Zeta-Function Expansion. *Methods Comput. Phys.* **1963**, *2*, 95–153.
- (4) Shavitt, I.; Karplus, M. Multicenter Integrals in Molecular Quantum Mechanics. *J. Chem. Phys.* **1962**, *36*, 550–551. Shavitt, I. The Gaussian Function in Calculations of Statistical Mechanics and Quantum Mechanics. *Methods Comput. Phys.* **1963**, *2*, 1–45.
- (5) Pitzer, R. M. Contribution of Atomic Orbital Integrals to Symmetry Orbital Integrals. *J. Chem. Phys.* **1973**, *58*, 3111–3112.
- (6) Nesbet, R. K. Computer Programs for Electronic Wave-Function Calculations. *Rev. Mod. Phys.* **1963**, *35*, 552–557.
- (7) Roothaan, C. C. J. Self-Consistent Field Theory for Open Shells of Electronic Systems. *Rev. Mod. Phys.* **1960**, *32*, 179–185.
- (8) Pitzer, K. S. Potential Energies for Rotation about Single Bonds. *Discuss. Faraday Soc.* **1951**, *10*, 66–73.
- (9) Hirota, E.; Saito, S.; Endo, Y. Barrier to Internal Rotation in Ethane from the Microwave Spectrum of CH₃CHD₂. *J. Chem. Phys.* **1979**, *71*, 1183–1187.
- (10) Ransil, B. J. Studies in Molecular Structure. II. LCAO-MO-SCF Wave Functions for Selected First-Row Diatomic Molecules. *Rev. Mod. Phys.* **1960**, *32*, 245–254.
- (11) Foster, J. M.; Boys, S. F. Quantum Variational Calculations for HCHO. *Rev. Mod. Phys.* **1960**, *32*, 303–307.
- (12) Goodfriend, P. L.; Birss, F. W.; Duncan, A. B. F. Calculations on the Electronic Structure of the Normal State of Formaldehyde. *Rev. Mod. Phys.* **1960**, *32*, 307–311.
- (13) Dunning, T. H., Jr.; Winter, N. W. Theoretical Determination of the Barriers to Internal Rotation in Hydrogen Peroxide. *J. Chem. Phys.* **1975**, *63*, 1847–1855.
- (14) Schaefer, H. F. *Private communication*.
- (15) Stevens, R. M.; Pitzer, R. M.; Lipscomb, W. N. Perturbed Hartree-Fock Calculations. I. Magnetic Susceptibility and Shielding in the LiH Molecule. *J. Chem. Phys.* **1963**, *38*, 550–560.

Free Energy Profiles along Consensus Normal Modes Provide Insight into HIV-1 Protease Flap Opening

Paulo R. Batista,^{†,§,⊥} Gaurav Pandey,^{§,‡} Pedro G. Pascutti,[†] Paulo M. Bisch,[†] David Perahia,^{*,⊥,§,‡} and Charles H. Robert^{§,||,‡}

[†]Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro, 21941-902, Brasil

[‡]Indian Institute of Technology, Roorkee, 247667, India

[§]CNRS Institut de Biochimie et Biophysique Moléculaire et Cellulaire, Université Paris Sud 11, 91405 Orsay, France

^{||}CNRS Laboratoire de Biochimie Théorique, Institut de Biologie Physico Chimique, Université Paris Diderot, Sorbonne Paris Cité, 75005 Paris, France

[⊥]CNRS BIMoDyM - Laboratoire de Biologie et Pharmacologie Appliquées - École Normale Supérieure de Cachan, 94235 Cachan, France

 Supporting Information

ABSTRACT: Describing biological macromolecular energetics from computer simulations can pose major challenges, and often necessitates enhanced conformational sampling. We describe the calculation of conformational free-energy profiles along carefully chosen collective coordinates: “consensus” normal modes, developed recently as robust alternatives to conventional normal modes. In an application to the HIV-1 protease, we obtain efficient sampling of significant flap opening movements governing inhibitor binding from relatively short simulations, in close correspondence with experimental results.

Predicting protein dynamics remains a challenge in computational biology. The configurational space in which a solvated macromolecule evolves is so large that we cannot reasonably expect to ever explore it fully. Standard molecular dynamics (MD) simulations have their limits: for proteins of normal size, all-atom simulations are typically limited to time scales largely inferior to that of many biologically interesting movements and conformational changes. Although such limits to computation are being pushed back every year (e.g., ref 1), exploration of long-time-scale dynamics by MD alone is still not feasible. The difficulty is especially acute for the identification of new molecules binding flexible biological targets and understanding related allosteric effects.^{2,3} Improved understanding of the dynamics of biological macromolecules can thus facilitate the proposal of new inhibitors or modulators of biological macromolecules in the search for new therapies.^{4,5}

Properly taking into account features of the energy landscape “in the large” allows one to arrive at accurate conclusions concerning the dynamic behavior of macromolecules while using shorter simulations. For example, coarse-grained potentials represent groups of atoms by a single, larger particle, leading to increased efficiency of the simulation.^{6,7} Other approaches are based on the use of biasing techniques to enhance MD sampling along interesting directions in the conformational space, using for example umbrella sampling or metadynamics.^{8–12} In both cases, the goal is essentially to look past noisier, smaller-scale degrees of freedom in order to exploit larger-scale features of the structure and/or the energy surface. Representative collective-coordinate descriptions must be found that are appropriate to the system under study. Principal components obtained from molecular simulations of the macromolecule have been successfully adapted for this purpose.^{13–15} Normal modes analysis of the macromolecular

structure also provides information concerning its most favorable modes of deformation, which correspond to gently sloped directions on the energy landscape.^{16–18} Normal-mode-based coordinates have been correlated with directions of conformational change in known systems^{19–21} and offer productive directions for exploring the mechanisms of conformational change.^{22,23} Consensus normal modes take this one step further through a statistical approach to identifying robust directions from multiple normal mode descriptions, in which the shape of the free energy surface is averaged over several minima via the covariance matrix.²⁴ As well as being more robust to intrinsic variability in the analyzed structure, the resulting consensus modes are more collective than the individual normal modes and provide a more natural means of accommodating solvent effects.²⁴

A particularly challenging case is that of the HIV-1 protease (PR), the major target of anti-AIDS drugs. PR is a homodimer consisting of two chains of 99 amino acid residues in which access to a central substrate binding site is mediated by symmetrically disposed β hairpins called the “flaps”.²⁵ Movement of the flaps has been shown in experimental studies to be critical for inhibitor binding.²⁵ Indeed, 450 crystal or NMR structures of this molecule under various conditions show conformational variability mainly in this region (Figure S1, Supporting Information). Inter-flap distance distributions established by electron paramagnetic resonance (EPR) measurements using spin labeling confirm the importance of flap conformational dynamics in solution.²⁶ Characterizing these dynamics is thus critical in the search for new inhibitors of the protease and for understanding mutations that allow HIV to evade existing therapies.²⁷

Received: April 7, 2011

Published: July 01, 2011

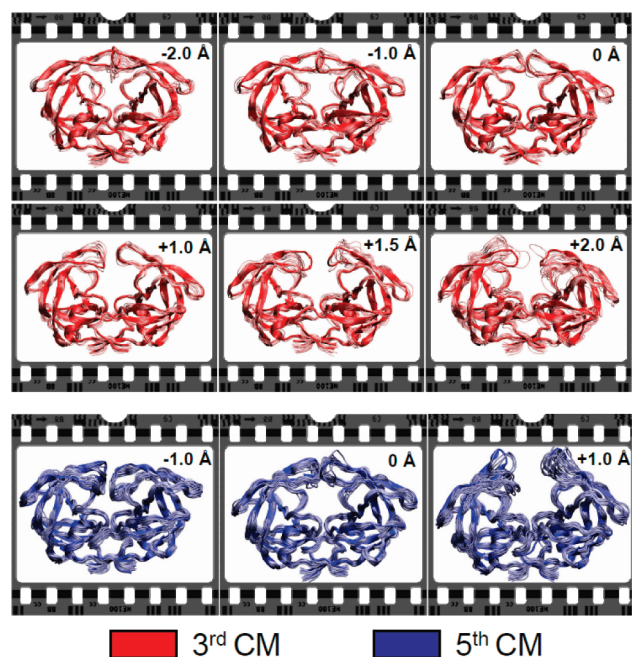


Figure 1. Snapshots of HIV-1 protease taken from MD simulations with umbrella sampling along the consensus mode directions describing flap opening. Each frame contains 20 superposed snapshots sampled at 5 ps intervals from an independent MD trajectory restrained to a particular normal mode value: from -2 to $+2$ Å for the mode 3 and from -1 to $+1$ Å for mode 5.

Nevertheless, the amplitude of the conformational dynamics seen in unbiased MD simulations only accounts for a small part of the conformational variability observed experimentally, due to the long time scales characterizing the motion, which have been estimated to be on the order of several hundred nanoseconds or microseconds.^{28,29} Further, enhanced sampling methods do not always increase exploration of flap opening movements. In a recent study of the binding of a peptide substrate to PR, seven different collective coordinates were biased simultaneously using explicit solvent simulations over $1.6 \mu\text{s}$.³⁰ Only semiopen structures were observed; full flap opening that would be necessary to allow the entrance of the natural polyprotein substrate was not observed.

We show here that umbrella sampling along strongly collective consensus modes of the HIV-1 protease permits calculation of a potential of mean force (PMF) along these reaction coordinates using relatively short MD simulations. The collective coordinates allow statistical sampling of representative conformations of the protein with alternative flap conformations including highly open states and largely account for the conformational dynamics observed in the spin label measurements.

The consensus modes (CM) for the *apo* HIV-1 protease were calculated as described previously from normal modes analyses of 20 protease structures, which were extracted at regular intervals from a 1 ns explicit-solvent MD simulation and quenched.²⁴ An advantage of the consensus modes formalism is its adaptability to practical constraints; here, for example, the consensus modes were calculated for the α carbons alone in order to obtain applicability to mutational studies, yet they reflect a set of PR conformations sampled from the original all-atom, explicit solvent MD simulation as well as the individual hydration layers used in the initial normal mode determinations.²⁴ The third and fifth lowest consensus

modes most involved in flap movement are shown in Figure 1. In the orientation shown, mode 3 is associated with a horizontal separation of the flaps while mode 5 describes vertical movement of the flaps, which expands the active site. A movie of the individual mode movements is provided (see Supporting Information, movies 1 and 2).

Sampling along the consensus mode directions was performed using generalized-Born implicit solvent³¹ MD simulations, using the VMOD command integrated into CHARMM version c35.³² This command adds a term to the energy function which is proportional to the squared deviation of the instantaneous normal coordinate of the protein from a prescribed target value. In this procedure, a given instantaneous structure \mathbf{R} is first superimposed onto a structure \mathbf{R}_0 that served as the reference for the consensus modes calculations. The normal coordinate is then defined as the projection of the mass-weighted coordinate difference vector between \mathbf{R} and the \mathbf{R}_0 onto the mode vector considered. VMOD can be used to bias several collective coordinates simultaneously. The approach has been used along with energy minimization to explore the effects of variability in the shape of the binding pocket in inhibitor docking to proteins^{33,34} and to the study of the energy variations of large conformational changes.^{23,35} In the current study, we used 41 independent MD simulations, each restrained by VMOD to a different consensus normal mode coordinate (from -2 to $+2$ Å, every 0.1 Å). The starting structure was first displaced to the target coordinate by energy minimization using the VMOD restraint; MD heating, equilibration, and production followed, all performed with the restraint in place in order to obtain the desired sampling. PMFs along the consensus mode coordinate were calculated using the weighted histogram (WHAM) procedure as implemented by Grossfield.³⁶ Additional details are given in the Supporting Information and associated Figure S2. We emphasize that exploring a normal mode coordinate using a VMOD restraint is not the same as simply displacing the structure along the normal coordinate. Employment of VMOD restricts only the normal mode coordinate itself; the remaining degrees of freedom are free to adapt to the imposed coordinate restraint.

The PMFs calculated for consensus modes 3 and 5 are shown in Figure 2. The free energy profile along mode 3 (top panel) shows a wide basin for displacements ranging from -1.5 Å to $+1.5$ Å with energy less than $2kT$, while mode 5 (bottom panel) shows a narrower basin between -0.5 and $+0.5$ Å. In both plots, the sign has been chosen so that the negative coordinate corresponds to closure of the flaps and the positive to opening. Outside the basin region, the free energy increases rapidly. We also show in red the distributions of normal mode coordinates calculated for the 450 known PDB structures, as described in the Supporting Information. Most of these structures contain ligands and are thus of the closed form, and correspondingly the bulk of the distribution for each mode is seen to lie to the left of the origin. However, several unbound forms of the protease have been solved as well, and these correspond to the small peaks around 0.5 Å in the distribution of mode 3. For both open and closed forms, the experimentally determined structures are seen to represent low-energy regions of the PMFs, as should be expected. In comparing the PMFs of modes 3 and 5, we see that the profile along mode 5, in which the flaps move vertically in Figure 1, is higher in energy but simpler in form than that of mode 3, associated with horizontal flap separation. As can be seen in Figure 2, mode 3 shows small barriers on the order of $2kT$ which tend to delineate the starting structure from either more-open or more-closed forms.

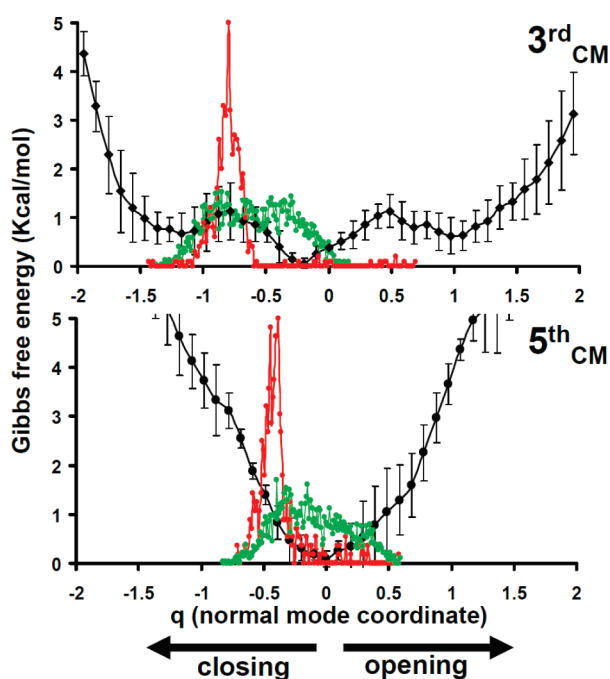


Figure 2. Free energy profiles for PR flap opening motions along CM directions. Two different modes (3rd and 5th CMs) were explored. The black line corresponds to the mean (\pm SD) of the free energies obtained over three independent simulations. The distribution in red shows the normal mode coordinates calculated for the 450 experimental PR structures deposited in the Protein Data Bank. The distribution shown in green corresponds to the normal mode coordinates of structures sampled from an unrestrained 50 ns explicit-solvent MD simulation.

The necessity for umbrella restraints in exploring the flap opening can be seen by examining the normal mode coordinates calculated for the structures obtained from a 50 ns unrestrained MD simulations, shown in the distribution in green (Figure 2). Clearly, the unrestrained MD sampled closed forms of the binding site rather well, producing structures having projections along both modes 3 and 5 that are consistent with the calculated PMFs. But the sampling of open forms along consensus mode 3 was poor. This may be related to the barrier seen at 0.5 Å in Figure 2 (top) for this mode, although this barrier must be of a different nature from that at -0.75 Å of the same height, which did not prevent sampling in the direction of closing. The asymmetry of the sampling seen in the unrestrained MD simulations in this study is consistent with the under-representation of highly open structures seen in over 460 replicated 50 ns explicit solvent MD simulations reported by Sadiq and De Fabritiis,³⁷ in which it was suggested that the time scale for the conformational dynamics accompanying full flap opening is very long and judging from NMR studies may be on the order of microseconds and beyond.^{28,37}

Another way of looking at flap opening is to calculate the distance between the flaps. In EPR²⁶ and MD³⁸ studies of the protease, spin labels attached to residue 55 in each flap allowed determination of the interflap distance distribution in the presence and absence of the protease inhibitor ritonavir. Figure 3 compares the experimental distribution for the *apo*-protease (in red) to the distribution of distances calculated for structures obtained from a 50 ns explicit solvent unrestrained MD simulation (green). It is clear that the unrestrained simulations do not sample structures having the larger interflap distances seen in the EPR results.

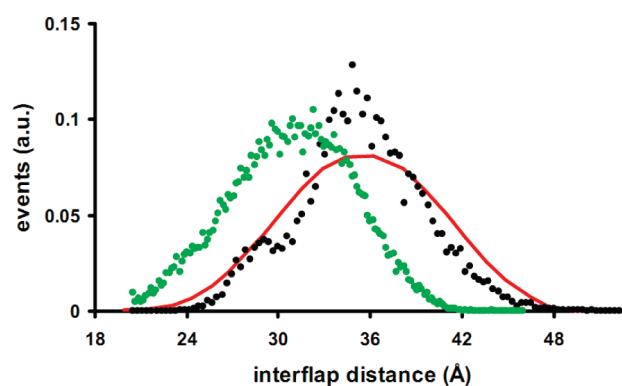


Figure 3. Correspondence between interflap distances (using Lys55 NZ atoms) obtained from MD with umbrella sampling (black) and from EPR of the spin-labeled protease (red). Data for structures extracted from an unrestrained explicit-solvent 50 ns MD are also shown (green).

Figure 3 also shows the distance distribution obtained for structures sampled in the MD simulations employing umbrella sampling along consensus mode 3 (in black), in which each distance has been weighted by the Boltzmann factor corresponding to the PMF shown in Figure 2 in order to allow comparison. The agreement with the experimental curve suggests that both closed and open flap conformations are properly sampled along this reduced coordinate.

Sutto et al., in comparing unbiased and biased exploration of the energy surface for a pentapeptide, concluded that the sampling performance strongly depends on the choice of the collective coordinates.¹⁵ Indeed, other collective coordinates used for the HIV-1 protease did not particularly enhance sampling of large flap opening movements, although they were sufficient to study peptide binding.³⁰ From this point of view, the consensus modes provide valuable collective coordinates for exploring the HIV-1 PR potential energy surface in the direction of flap opening, as will be necessary for characterizing conformational factors playing a role in the binding of the natural substrate.

Several conclusions can be drawn from this study, both for consensus normal mode sampling in general and for the dynamics of the HIV-1 protease. First, the use of umbrella sampling along a well-chosen collective coordinate—here, consensus mode directions—enables one to explore HIV-1 protease flap opening that would require long conventional MD simulations in order to properly account for its conformational and energetic aspects. The free energy increase associated with highly open structures is less than $2kT$ despite their rarity in conventional simulations, underlining the kinetic difficulties that can confront sampling using conventional MD. Also, the distribution of experimental PR structures, the majority containing ligands, on the free energy landscape clearly shows that a population shift mechanism applies,^{39–42} as the binding of the ligand would appear to stabilize closed forms of the apoprotein that are already energetically allowed, in agreement with pulsed EPR experiments.⁴³ It may also be noted that the free energy landscape along the opening/closing mode (Figure 2) is rather flat, with a highly anharmonic nature, while that corresponding to a higher energy mode (5th mode in Figure 2) is closer to a quadratic shape. The PMF approach using the consensus normal mode direction as the collective coordinate allows one to estimate upper bounds for the amplitude of large conformational changes and corresponding energy, and to describe structural aspects that may be very different from those derived from a simple normal-mode analysis itself.

Further, using consensus modes²⁴ allowed us to frame the problem of exploring the conformational space of the HIV-1 protease in terms of the α carbon atoms alone, while the mode determinations themselves still benefit from the high quality of the individual normal mode determinations in which an all-atom representation of the protein was used in addition to a solvation layer.⁴⁴ We note that the protocol used here is entirely adaptable for use on distributed computing platforms, as the simulations corresponding to the individual displacements along the mode coordinate are completely independent. This will enable us to further explore the protease motions to understand the effects of certain mutants such as the drug-resistant MDR769, in which experimental evidence suggests the flap distribution to be shifted toward much larger distances.⁴⁵

Improved free energy evaluations from MD simulations, which as we have shown here can be obtained using enhanced sampling along consensus normal mode directions, offers the possibility of differentiating the behavior of different mutants of HIV-1 in view of obtaining clues about their resistance mechanisms.

■ ASSOCIATED CONTENT

S Supporting Information. Additional data, movies of the individual mode movements, and information regarding methods and protocols are included. This material is available free of charge via the Internet at <http://pubs.acs.org>

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: david.perahia@lbpa.ens-cachan.fr, robert@ibpc.fr

Author Contributions

[#]D.P. and C.H.R. contributed equally to the conception and design of this study.

■ ACKNOWLEDGMENT

We wish to thank the Brazilian agencies CAPES, CNPq, and FAPERJ as well as the French–Brazilian CAPES/COFECUB collaboration project for financial support and the French Ministry of Foreign and European Affairs for supporting the participation of G.P. in this project. The University of Paris Orsay Pluriformation Project DEMAIN provided computing resources used in this study.

■ REFERENCES

- (1) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330* (6002), 341–6.
- (2) Ahmed, A.; Kazemi, S.; Gohlke, H. Protein Flexibility and Mobility in Structure-Based Drug Design. *Front. Drug Des. Discovery* **2007**, *3*, 455–476.
- (3) Costa, M.; Batista, P.; Shida, C.; Robert, C.; Bisch, P.; Pascutti, P. How does heparin prevent the pH inactivation of cathepsin B? Allosteric mechanism elucidated by docking and molecular dynamics. *BMC Genomics* **2010**, *11* (Suppl 5), S5.
- (4) B-Rao, C.; Subramanian, J.; Sharma, S. D. Managing protein flexibility in docking and its applications. *Drug Discovery Today* **2009**, *14* (7–8), 394–400.
- (5) Zwier, M. C.; Chong, L. T. Reaching biological timescales with all-atom molecular dynamics simulations. *Curr. Opin. Pharmacol.* **2010**, *10* (6), 745–752.

- (6) Levitt, M.; Warshel, A. Computer simulation of protein folding. *Nature* **1975**, *253* (5494), 694–8.
- (7) Clementi, C. Coarse-grained models of protein folding: toy models or predictive tools?. *Curr. Opin. Struct. Biol.* **2008**, *18* (1), 10–15.
- (8) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199.
- (9) Dellago, C.; Bolhuis, P. G. Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events. *Adv. Comput. Simulat. Approaches Soft Matter Sci. III* **2009**, *221*, 167–233.
- (10) Spiwok, V.; Lipovova, P.; Kralova, B. Metadynamics in essential coordinates: free energy simulation of conformational changes. *J. Phys. Chem. B* **2007**, *111* (12), 3073–6.
- (11) Tribello, G. A.; Ceriotti, M.; Parrinello, M. A self-learning algorithm for biased molecular dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (41), 17509–17514.
- (12) Autieri, E.; Sega, M.; Pederiva, F.; Guella, G. Puckering free energy of pyranoses: A NMR and metadynamics-umbrella sampling investigation. *J. Chem. Phys.* **2010**, *133* (9), 095104.
- (13) Spiwok, V.; Lipovova, P.; Kralova, B. Metadynamics in essential coordinates: Free energy simulation of conformational changes. *J. Phys. Chem. B* **2007**, *111* (12), 3073–3076.
- (14) Spiwok, V.; Kralova, B.; Tvaroska, I. Continuous metadynamics in essential coordinates as a tool for free energy modelling of conformational changes. *J. Mol. Model.* **2008**, *14* (11), 995–1002.
- (15) Sutto, L.; D'Abramo, M.; Gervasio, F. L. Comparing the Efficiency of Biased and Unbiased Molecular Dynamics in Reconstructing the Free Energy Landscape of Met-Enkephalin. *J. Chem. Theory Comput.* **2010**, *6* (12), 3640–3646.
- (16) Karplus, M.; Kushick, J. N. Method for estimating the configurational entropy of macromolecules. *Macromolecules* **1981**, *14* (2), 325–332.
- (17) Levy, R. M.; Perahia, D.; Karplus, M. Molecular dynamics of an alpha-helical polypeptide: Temperature dependence and deviation from harmonic behavior. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79* (4), 1346–1350.
- (18) Levy, R. M.; Karplus, M.; Kushick, J.; Perahia, D. Evaluation of the configurational entropy for proteins: application to molecular dynamics simulations of an alpha-helix. *Macromolecules* **1984**, *17* (7), 1370–1374.
- (19) Tama, F.; Sanejouand, Y. H. Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **2001**, *14* (1), 1–6.
- (20) Brooks, B.; Karplus, M. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82* (15), 4995–9.
- (21) Yang, Z.; Majek, P.; Bahar, I. Allosteric transitions of supramolecular systems explored by network models: application to chaperonin GroEL. *PLoS Comput. Biol.* **2009**, *5* (4), e1000360.
- (22) Zheng, W.; Brooks, B. R.; Hummer, G. Protein conformational transitions explored by mixed elastic network models. *Proteins* **2007**, *69* (1), 43–57.
- (23) Floquet, N.; Durand, P.; Maigret, B.; Badet, B.; Badet-Denisot, M. A.; Perahia, D. Collective motions in glucosamine-6-phosphate synthase: influence of ligand binding and role in ammonia channelling and opening of the fructose-6-phosphate binding site. *J. Mol. Biol.* **2009**, *385* (2), 653–64.
- (24) Batista, P. R.; Robert, C. H.; Marechal, J. D.; Hamida-Rebai, M. B.; Pascutti, P. G.; Bisch, P. M.; Perahia, D. Consensus modes, a robust description of protein collective motions from multiple-minima normal mode analysis—application to the HIV-1 protease. *Phys. Chem. Chem. Phys.* **2010**, *12* (12), 2850–9.
- (25) Gustchina, A.; Weber, I. T. Comparison of inhibitor binding in HIV-1 protease and in non-viral aspartic proteases: the role of the flap. *FEBS Lett.* **1990**, *269* (1), 269–72.
- (26) Galiano, L.; Bonora, M.; Fanucci, G. E. Interflap distances in HIV-1 protease determined by pulsed EPR measurements. *J. Am. Chem. Soc.* **2007**, *129* (36), 11004–5.

(27) Piana, S.; Carloni, P.; Rothlisberger, U. Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations. *Protein Sci.* **2002**, *11* (10), 2393–402.

(28) Ishima, R.; Freedberg, D. I.; Wang, Y. X.; Louis, J. M.; Torchia, D. A. Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure* **1999**, *7* (9), 1047–55.

(29) Tozzini, V.; Trylska, J.; Chang, C. E.; McCammon, J. A. Flap opening dynamics in HIV-1 protease explored with a coarse-grained model. *J. Struct. Biol.* **2007**, *157* (3), 606–15.

(30) Pietrucci, F.; Marinelli, F.; Carloni, P.; Laio, A. Substrate binding mechanism of HIV-1 protease from explicit-solvent atomistic simulations. *J. Am. Chem. Soc.* **2009**, *131* (33), 11811–8.

(31) Im, W.; Lee, M. S.; Brooks, C. L., III. Generalized born model with a simple smoothing function. *J. Comput. Chem.* **2003**, *24* (14), 1691–702.

(32) Brooks, B. R.; et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.

(33) Floquet, N.; Marechal, J. D.; Badet-Denisot, M. A.; Robert, C. H.; Dauchez, M.; Perahia, D. Normal mode analysis as a prerequisite for drug design: application to matrix metalloproteinases inhibitors. *FEBS Lett.* **2006**, *580* (22), 5130–6.

(34) Sperandio, O.; Mouawad, L.; Pinto, E.; Villoutreix, B.; Perahia, D.; Miteva, M. How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis. *Eur. Biophys. J.* **2010**, *39* (9), 1365–1372.

(35) Floquet, N.; M'Kadmi, C.; Perahia, D.; Gagne, D.; Bergé, G.; Marie, J.; Banères, J.-L.; Galleyrand, J.-C.; Fehrentz, J.-A.; Martinez, J. Activation of the Ghrelin Receptor is Described by a Privileged Collective Motion: A Model for Constitutive and Agonist-induced Activation of a Sub-class A G-Protein Coupled Receptor (GPCR). *J. Mol. Biol.* **2010**, *395* (4), 769–784.

(36) Grossfield, A. WHAM: the weighted histogram analysis method, v.158, <http://membrane.urmc.rochester.edu/content/wham> (accessed July 2011).

(37) Sadiq, S. K.; De Fabritiis, G. Explicit solvent dynamics and energetics of HIV-1 protease flap opening and closing. *Proteins* **2010**, *78* (14), 2873–85.

(38) Ding, F.; Layten, M.; Simmerling, C. Solution structure of HIV-1 protease flaps probed by comparison of molecular dynamics simulation ensembles and EPR experiments. *J. Am. Chem. Soc.* **2008**, *130* (23), 7184–5.

(39) Kumar, S.; Ma, B.; Tsai, C. J.; Sinha, N.; Nussinov, R. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci.* **2000**, *9* (1), 10–9.

(40) Okazaki, K.; Takada, S. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (32), 11182–7.

(41) Kar, G.; Keskin, O.; Gursoy, A.; Nussinov, R. Allosteric and population shift in drug discovery. *Curr. Opin. Pharmacol.* **2010**, *10* (6), 715–22.

(42) Valente, A. P.; Miyamoto, C. A.; Almeida, F. C. Implications of protein conformational diversity for binding and development of new biological active compounds. *Curr. Med. Chem.* **2006**, *13* (30), 3697–703.

(43) Blackburn, M. E.; Veloro, A. M.; Fanucci, G. E. Monitoring inhibitor-induced conformational population shifts in HIV-1 protease by pulsed EPR spectroscopy. *Biochemistry* **2009**, *48* (37), 8765–7.

(44) Ma, J.; Karplus, M. Ligand-induced conformational changes in ras p21: a normal mode and energy minimization analysis. *J. Mol. Biol.* **1997**, *274* (1), 114–31.

(45) Galiano, L.; Ding, F.; Veloro, A. M.; Blackburn, M. E.; Simmerling, C.; Fanucci, G. E. Drug pressure selected mutations in HIV-1 protease alter flap conformations. *J. Am. Chem. Soc.* **2009**, *131* (2), 430–1.

Resolutions of the Coulomb Operator: V. The Long-Range Ewald Operator

Taweetham Limpanuparb* and Peter M. W. Gill

Research School of Chemistry, Australian National University, Canberra ACT 0200, Australia

ABSTRACT: We show that the long-range Ewald operator can be resolved as $\text{erf}(\omega r_{12})/r_{12} = \sum_k \phi_k^*(\mathbf{r}_1) \phi_k(\mathbf{r}_2)$, where ϕ_k is proportional to the product of a spherical Bessel function and a spherical harmonic. We demonstrate the use of this new resolution by calculating the long-range Coulomb energy of the nanodiamond crystallite $\text{C}_{84}\text{H}_{64}$ and the long-range exchange energy of the graphene $\text{C}_{96}\text{H}_{24}$. The resolution appears particularly effective for long-range exchange calculations.

1. INTRODUCTION

We have recently published a series of papers^{1–5} concerned with resolving the Coulomb operator

$$r_{12}^{-1} \equiv |\mathbf{r}_1 - \mathbf{r}_2|^{-1} = \sum_{k=1}^{\infty} |\phi_k\rangle\langle\phi_k| \quad (1)$$

into one-particle functions, where $|\phi_k\rangle$ and $\langle\phi_k|$ are functions of \mathbf{r}_1 and \mathbf{r}_2 , respectively. Such resolutions factorize a Coulomb integral into a sum of products of auxiliary integrals

$$\langle a|r_{12}^{-1}|b\rangle = \sum_{k=1}^{\infty} \langle a|\phi_k\rangle\langle\phi_k|b\rangle \quad (2)$$

and thereby offer the computational benefits of Cholesky decomposition^{6–9} and density fitting,^{10–12} but without the need to solve Cholesky or fitting equations.

In our most recent work,^{4,5} we have shown that the one-particle functions can take the form

$$\phi_k(\mathbf{r}) \equiv \phi_{nlm}(\mathbf{r}) = 2\sqrt{2 - \delta_{n,0}} j_l(nr) Y_{lm}(\mathbf{r}) \quad (3)$$

where j_l is a spherical Bessel function and Y_{lm} is a complex spherical harmonic.¹³ Although this resolution is valid only for $r_1 + r_2 < 2\pi$, we have shown that this weakness can be overcome by a suitable prescaling of the system under study.

There is considerable contemporary interest^{14–36} in partitioning the Coulomb operator as

$$r_{12}^{-1} \equiv S(r_{12}) + L(r_{12}) \quad (4)$$

where S is a singular short-range operator and L is a smooth long-range operator, and then treating the short-range and long-range subproblems separately. Ewald introduced this to chemistry to compute Madelung constants,³⁷ but it can be traced, in the mathematics literature, to Riemann.³⁸

The partition strategy is now employed in many quantum chemical methods. It is particularly prominent in hybrid methodologies, wherein wavefunction-based and density-based approaches are carefully combined to exploit their respective strengths. This has led, for example, to the popular HSE,^{21–24} CAM-B3LYP,²⁵ LC- ω PBE,³⁰ LCgau-BOP,³⁴ and ω B97XD³⁵ methods.

The short-range operator S can be treated efficiently by the use of boxing schemes^{22,39–43} that exploit spatial locality. However,

the long-range operator L is more computationally difficult, and it is natural to ask whether a resolution analogous to eq 1 can be constructed for it.

It turns out that there are many ways to resolve such operators, and we will consider several. Our approaches are general, but in this Letter, we focus on the long-range Ewald operator

$$L(r_{12}) = \frac{\text{erf}(\omega r_{12})}{r_{12}} \quad (5)$$

The partition parameter ω can take any positive value (the limit $\omega \rightarrow \infty$ recovers the Coulomb operator) but, in practice, often lies between 0.1 and 1. We use atomic units throughout.

2. RESOLUTIONS OF THE EWALD OPERATOR

We have investigated five approaches for resolving the Ewald operator: orthonormal expansion, Taylor expansion, Gaussian expansion, Bessel expansion, and Hermite quadrature. The first four are outlined in the Appendix, but we describe the fifth and most promising here.

If we apply $2N$ -point Gauss–Hermite quadrature^{13,44} to the integral representation

$$L(r_{12}) = \frac{2\omega}{\pi} \int_{-\infty}^{\infty} j_0(2\beta\omega r_{12}) \exp(-\beta^2) d\beta \quad (6)$$

we obtain the spherical Bessel expansion:

$$L(r_{12}) = \frac{4\omega}{\pi} \left[\sum_{n=1}^N b_n j_0(2\beta_n \omega r_{12}) + \varepsilon_N(\omega r_{12}) \right] \quad (7)$$

where β_n and b_n are the (positive) Hermite roots and weights.

How accurate are these Bessel expansions? The quadrature error $\varepsilon_N(\omega r_{12})$ for $N = 2, 4, 6, 8,$ and 10 is shown in Figure 1. It is initially tiny, indicating that the expansions are accurate for small ωr_{12} , but eventually breaks away from the axis when the expansion becomes unsatisfactory. (We note, however, that the error is bounded for all ωr_{12} .) It is encouraging to observe that the breakaway point moves rapidly to the right as N is increased,

Published: June 21, 2011

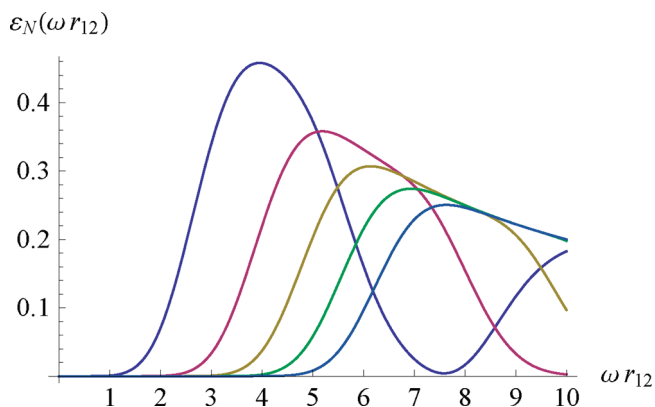


Figure 1. Quadrature error $\varepsilon_N(\omega r_{12})$ in eq 7 for $N = 2$ (left-most), 4, 6, 8, and 10 (right-most).

Table 1. N and N' such that $\varepsilon_N(\omega r_{12}) < \varepsilon$ for $0 \leq \omega r_{12} \leq R$

$-\log_{10} \varepsilon$	$R = 10$		$R = 20$		$R = 30$		$R = 40$		$R = 50$	
	N	N'	N	N'	N	N'	N	N'	N	N'
2	31	8	108	13	235	19	410	25	634	30
3	35	10	117	17	248	25	428	32	658	39
4	38	12	122	21	257	29	441	38	674	47
5	40	14	127	24	264	34	450	43	687	53
6	42	16	132	27	270	37	459	48	697	59
7	44	17	135	29	276	41	467	53	707	65
8	46	19	139	32	281	44	473	57	715	70
9	48	20	142	34	286	48	480	61	723	75
10	50	22	145	36	291	51	486	65	731	80
11	51	23	148	38	295	54	491	69	738	84
12	53	24	151	41	299	57	497	73	744	89

suggesting that even modest values of N yield Bessel expansions that are useful over large domains of ωr_{12} .

In principle, all N terms in eq 7 must be included. However, because $|j_0(x)| \leq 1$ and the Hermite weights b_n decay extremely rapidly, it is possible to truncate eq 7 at $n = N' \ll N$ with a negligible loss of accuracy. The minimum N and N' that guarantee that the quadrature error is below ε over the domain $0 \leq \omega r_{12} \leq R$ are shown in Table 1. This Table reveals that, in a molecule where $\max(r_{12}) \approx 30$ (for example, taxol $C_{47}H_{51}NO_{14}$), an accuracy of 10^{-10} requires only $N' = 51$ terms for $\omega = 1$ or only $N' = 22$ terms for $\omega = 1/3$.

To resolve the j_0 functions in eq 7, we start with the spherical Bessel addition theorem¹³

$$j_0(\lambda r_{12}) = \sum_{l=0}^{\infty} (2l+1) j_l(\lambda r_1) j_l(\lambda r_2) P_l(\cos \theta_{12}) \quad (8)$$

and apply the Legendre addition theorem¹³ to find

$$j_0(\lambda r_{12}) = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^l j_l(\lambda r_1) j_l(\lambda r_2) Y_{lm}^*(\mathbf{r}_1) Y_{lm}(\mathbf{r}_2) \quad (9)$$

Substituting eq 9 into eq 7 then yields our key result—the Ewald resolution:

$$L^{(N,L)}(r_{12}) = \sum_{n=1}^N \sum_{l=0}^L \sum_{m=-l}^l |\phi_{nlm}\rangle \langle \phi_{nlm}| \quad (10a)$$

$$\phi_{nlm}(\mathbf{r}) = 4\sqrt{b_n \omega} j_l(2\beta_n \omega r) Y_{lm}(\mathbf{r}) \quad (10b)$$

3. COMPUTATIONAL CONSIDERATIONS

It is essential to be able to determine *a priori* the minimum values of N and N' that will guarantee that eq 7 is accurate to within ε over the domain of important ωr_{12} values in one's system. By examining the values of N in Table 1, we have devised the simple quadratic estimate

$$N \approx R^2/4 + (\sqrt{-\log_{10} \varepsilon} - 1)R + 3 \quad (11)$$

and it is then easy to show from the asymptotic behavior of the Hermite roots and weights that

$$N' \approx \frac{2}{\pi} \sqrt{-N \ln \varepsilon} \quad (12)$$

To use the Ewald resolution, eq 10, to find long-range energies, we need the auxiliary integrals

$$\langle ab | \phi_{nlm} \rangle = \int a(\mathbf{r}) b(\mathbf{r}) \phi_{nlm}(\mathbf{r}) d\mathbf{r} \quad (13)$$

where we will assume that a and b are Gaussian basis functions centered at \mathbf{A} and \mathbf{B} , respectively. Because the Gaussian product rule allows $a(\mathbf{r})b(\mathbf{r})$ to be expanded as a finite linear combination⁴⁵ of Gaussians with a centroid \mathbf{P} on the line between \mathbf{A} and \mathbf{B} , the problem reduces to finding two-center integrals of the form

$$\langle G_{n'l'm'} | \phi_{nlm} \rangle = \int r'^l \exp(-\zeta r'^2) Y_{l'm'}(\mathbf{r}') \phi_{nlm}(\mathbf{r} + \mathbf{P}) d\mathbf{r}' \quad (14)$$

These can be solved in closed form, and we will discuss elsewhere⁴⁶ an efficient algorithm for Gaussians of arbitrary angular momentum. However, in a basis that contains only s and p functions, the only necessary formulas are

$$\langle G_{000} | \phi_{nlm} \rangle = c_n C_l^{lm00} j_l Y_{lm} \quad (15a)$$

$$\langle G_{200} | \phi_{nlm} \rangle = c_n C_l^{lm00} [3/(2\zeta) - x_n^2] j_l Y_{lm} \quad (15b)$$

$$\langle G_{11m'} | \phi_{nlm} \rangle = c_n x_n [C_{l-1}^{lm1m'} j_{l-1} Y_{l-1,m-m'} - C_{l+1}^{lm1m'} j_{l+1} Y_{l+1,m-m'}] \quad (15c)$$

$$\langle G_{22m'} | \phi_{nlm} \rangle = c_n x_n^2 [C_{l-2}^{lm2m'} j_{l-2} Y_{l-2,m-m'} - C_{l+2}^{lm2m'} j_{l+2} Y_{l+2,m-m'}] \quad (15d)$$

where $x_n \equiv \beta_n \omega / \zeta$, $j_l \equiv j_l(2\beta_n \omega P)$, and $Y_{lm} \equiv Y_{lm}(\mathbf{P})$

$$c_n = 4\sqrt{b_n \omega} (\pi/\zeta)^{3/2} \exp(-\zeta x_n^2) \quad (16)$$

$$C_l^{lm'm'} = (-1)^{m'} \sqrt{\frac{(2l+1)(2l'+1)}{4\pi(2l+1)}} \langle ll'00 | ll'00 \rangle \langle ll'm(-m') | ll'(m-m') \rangle \quad (17)$$

and the final two factors in eq 17 are Clebsch–Gordan coefficients.¹³ We note that $C_l^{lm00} = Y_{00} = 1/(4\pi)^{1/2}$, and thus, eq 15a is analogous to eq 18 of our previous work.⁴

Table 2. (*N,L*) Pairs Required in Long-Range Coulomb and Exchange Calculations

	long-range Coulomb energy of the nanodiamond C ₈₄ H ₆₄			long-range exchange energy of the graphene C ₉₆ H ₂₄		
	$\omega = 0.1$	$\omega = 0.5$	$\omega = 1.0$	$\omega = 0.1$	$\omega = 0.5$	$\omega = 1.0$
$\varepsilon = 10^{-3}$	(3, 0)	(30, 13)	(114, 24)	(1, 4)	(2, 27)	(5, 45)
$\varepsilon = 10^{-6}$	(5, 4)	(49, 23)	(181, 50)	(2, 9)	(4, 44)	(8, 85)
$\varepsilon = 10^{-9}$	(7, 8)	(65, 36)	(241, 68)	(3, 13)	(6, 58)	(12, 99)

We have implemented the Ewald resolution in a stand-alone C program which precomputes the required Hermite roots and weights,⁴⁴ along with the Clebsch–Gordan coefficients.¹³ The j_l and Y_{lm} are calculated recursively, as in our previous work.⁴ We use the relative error

$$\varepsilon = \left| \frac{E^{(N,L)} - E}{E} \right| \quad (18)$$

to measure the accuracy of the approximate energies afforded by eq 10.

4. NUMERICAL RESULTS

The long-range Coulomb energy of a density $\rho(\mathbf{r})$ is

$$E_J = \frac{1}{2} \langle \rho | L(r_{12}) | \rho \rangle \quad (19)$$

and applying the Ewald resolution eq 10 to this yields the approximation

$$E_J^{(N,L)} = \frac{1}{2} \sum_{n=1}^N \sum_{l=0}^L \sum_{m=-l}^l \langle \rho | \phi_{nlm} \rangle^2 \quad (20)$$

We have applied eq 20 to the electron density in the nanodiamond C₈₄H₆₄, which is described in our previous paper.⁴ The (*N,L*) pairs that yield various relative errors ε for various attenuation parameters ω are shown in the middle columns of Table 2.

The long-range exchange energy is

$$E_K = -\frac{1}{2} \sum_{ij}^{\text{occ}} \langle \psi_i \psi_j | L(r_{12}) | \psi_i \psi_j \rangle \quad (21)$$

and applying the Ewald resolution (eq 10) to this yields the approximation

$$E_K^{(N,L)} = -\frac{1}{2} \sum_{n=1}^N \sum_{l=0}^L \sum_{m=-l}^l \sum_{ij}^{\text{occ}} \langle \psi_i \psi_j | \phi_{nlm} \rangle^2 \quad (22)$$

Diamond has a large bandgap, and its exchange interactions decay rapidly with distance. We therefore chose to apply eq 22 to the more interesting π system of the C₉₆H₂₄ graphene,⁴⁷ placing a unit exponent p_π Gaussian on each C atom and using its Hückel orbitals.⁴⁸ The (*N,L*) pairs that yield various relative errors ε for various ω are shown in the final columns of Table 2.

Because the Ewald operator eq 5 is smooth, the (*N,L*) pairs required for the long-range Coulomb energies are much smaller than for the total Coulomb energies.⁴ Moreover, we find that long-range exchange energies require surprisingly small *N* values, reflecting that, even in the highly delocalized graphene system, the exchange interaction decays fairly quickly with distance.^{47,49,50}

5. CONCLUDING REMARKS

There are a number of ways to resolve the long-range Coulomb (Ewald) operator into products of one-particle functions. Our favorite resolution eq 10 employs a spherical Bessel expansion of the Ewald operator and thereby generalizes our earlier quasi-resolution of the Coulomb operator. Numerical results indicate that this Ewald resolution converges rapidly and may be useful in a range of quantum chemical contexts. It looks particularly promising for the efficient calculation of long-range exchange energies. We are implementing the resolution in the Q-Chem package,⁵¹ and we will discuss the efficient evaluation of the auxiliary integrals eq 14 and present timing comparisons elsewhere.⁴⁶

We note finally that the Bessel expansion method is easy to extend to the erf-gau operator^{16,34,52}

$$L_1(r_{12}) = \frac{\text{erf}(\omega r_{12})}{r_{12}} - \frac{2\omega}{\sqrt{\pi}} \exp\left(-\frac{\omega^2 r_{12}^2}{3}\right) \quad (23)$$

Applying Gauss–Hermite quadrature as for the Ewald operator yields

$$\begin{aligned} \frac{2\omega}{\sqrt{\pi}} \exp\left(-\frac{\omega^2 r_{12}^2}{3}\right) &= \frac{4\omega}{\pi} \int_{-\infty}^{\infty} \beta^2 j_0\left(\frac{2}{\sqrt{3}}\beta\omega r_{12}\right) \exp(-\beta^2) d\beta \\ &\approx \frac{8\omega}{\pi} \sum_{n=1}^N b_n \beta_n^2 j_0\left(\frac{2}{\sqrt{3}}\beta_n\omega r_{12}\right) \end{aligned} \quad (24)$$

where β_n and b_n have the same meanings as in eq 7.

APPENDIX

Orthonormal Expansion. One way to resolve $L(r_{12})$ is to find functions f_k that are complete and Ewald-orthonormal, i.e.

$$\langle f_k | L(r_{12}) | f_{k'} \rangle = \delta_{kk'} \quad (25)$$

If these f_k are known, one can show¹ that

$$\phi_k(\mathbf{r}_1) = \int L(r_{12}) f_k(\mathbf{r}_2) d\mathbf{r}_2 \quad (26)$$

If f_k is chosen to be a product of Y_{lm} and a radial function, one eventually obtains

$$\begin{aligned} \phi_k(\mathbf{r}) &= \sqrt{2/\pi} Y_{lm}(r) \int_0^\infty p_n(x) j_l(rx) \hat{L}^{1/2}(x) x dx \\ &= 2\sqrt{2} Y_{lm}(\mathbf{r}) \int_0^\infty p_n(x) j_l(rx) \exp\left(-\frac{x^2}{8\omega^2}\right) dx \end{aligned} \quad (27)$$

where \hat{L} is the Fourier transform of L and the p_n are any functions that form a complete and orthonormal set on $[0, \infty)$. Unfortunately, this approach is thwarted by the difficulty of selecting p_n that yield tractable integrals.

Taylor Expansion. The Taylor expansion of the Ewald operator

$$L(r_{12}) = \frac{2\omega}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-\omega^2 r_{12}^2)^n}{n!(2n+1)} \quad (28)$$

converges for all r_{12} . Because $(r_{12}^2)^n$ expands naturally¹⁸ into a finite sum for any n , it is easy to construct a resolution from eq 28. However, when truncated after $n = N$, the series eq 28 behaves as $(-r_{12}^2)^N$ and is therefore worthless at large r_{12} .

Gaussian Expansion. If we apply $2N$ -point Gauss–Legendre quadrature¹³ to the Ewald integral representation

$$L(r_{12}) = \frac{\omega}{\sqrt{\pi}} \int_{-1}^1 \exp(-\omega^2 \gamma^2 r_{12}^2) d\gamma \quad (29)$$

we obtain the Gaussian expansion¹⁶

$$L(r_{12}) \approx \frac{2\omega}{\sqrt{\pi}} \sum_{n=1}^N g_n \exp(-\omega^2 \gamma_n^2 r_{12}^2) \quad (30)$$

where the γ_n and g_n are the (positive) Legendre roots and weights. The function $\exp(-\lambda r_{12}^2)$ can be partially resolved, using the exponential and Legendre addition theorems¹³ to find

$$\begin{aligned} \frac{\exp(-\lambda r_{12}^2)}{\exp(-\lambda r_1^2 - \lambda r_2^2)} &= \sum_{l=0}^{\infty} (2l+1) i_l(\lambda r_1 r_2) P_l(\cos \theta_{12}) \\ &= 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^l i_l(\lambda r_1 r_2) Y_{lm}^*(\mathbf{r}_1) Y_{lm}(\mathbf{r}_2) \end{aligned} \quad (31)$$

where i_l is a modified spherical Bessel function.¹³ However, this does not mirror the form of eq 1 because we cannot resolve $i_l(\lambda r_1 r_2)$.

Bessel Expansion. The Fourier–Bessel expansion^{13,53}

$$\begin{aligned} L(r_{12}) &= \frac{2}{\pi} \sum_{n=1}^{\infty} j_0(nr_{12}) \int_0^{\pi} L(x) j_0(nx) n^2 x^2 dx \\ &= \frac{2}{\pi} \sum_{n=1}^{\infty} \left[(-1)^{n+1} \operatorname{erf}(\omega\pi) + \exp\left(-\frac{n^2}{4\omega^2}\right) \mathcal{R}\left\{\operatorname{erf}\left(\omega\pi + \frac{n}{2\omega}\right)\right\} \right] j_0(nr_{12}) \\ &= L(\pi) + \frac{2}{\pi} \sum_{n=1}^{\infty} \exp\left(-\frac{n^2}{4\omega^2}\right) \mathcal{R}\left\{\operatorname{erf}\left(\omega\pi + \frac{n}{2\omega}\right)\right\} j_0(nr_{12}) \end{aligned} \quad (32)$$

converges rapidly, but unfortunately, it is valid only on the finite domain $0 \leq r_{12} \leq \pi$. As a consequence, it yields what we have previously termed a “quasi-resolution”,⁴ and to use it in practice, one would need to scale the system to fit within this domain.

AUTHOR INFORMATION

Corresponding Author

*E-mail: taweetham.limpanuparb@anu.edu.au.

ACKNOWLEDGMENT

T.L. thanks the Development and Promotion of Science and Technology Talents Project for a Royal Thai Government Ph.D. scholarship. P.M.W.G. thanks the Australian Research Council for funding (DP0984806 and DP1094170) and the NCI National Facility for supercomputer resources.

REFERENCES

- Varganov, S. A.; Gilbert, A. T. B.; Deplazes, E.; Gill, P. M. W. *J. Chem. Phys.* **2008**, *128*, 201104.
- Gill, P. M. W.; Gilbert, A. T. B. *Chem. Phys.* **2009**, *356*, 86–90.
- Limpanuparb, T.; Gill, P. M. W. *Phys. Chem. Chem. Phys.* **2009**, *11*, 9176–9181.
- Limpanuparb, T.; Gilbert, A. T. B.; Gill, P. M. W. *J. Chem. Theory Comput.* **2011**, *7*, 830–833.
- Dominici, D. E.; Gill, P. M. W.; Limpanuparb, T. A Remarkable Identity Involving Bessel Functions, arXiv: 1103.0058. arXiv.org ePrint archive. <http://arxiv.org/abs/1103.0058> (accessed Mar 1, 2011).

- Beebe, N. H. F.; Linderberg, J. *Int. J. Quantum Chem.* **1977**, *12*, 683–705.
- Koch, H.; Sanchez de Meras, A.; Pedersen, T. B. *J. Chem. Phys.* **2003**, *118*, 9481–9484.
- Aquilante, F.; Lindh, R.; Pedersen, T. B. *J. Chem. Phys.* **2007**, *127*, 114107.
- Weigend, F.; Kattannek, M.; Ahlrichs, R. *J. Chem. Phys.* **2009**, *130*, 164106.
- Vahtras, O.; Almlöf, J.; Feyereisen, M. *Chem. Phys. Lett.* **1993**, *213*, 514–518.
- Jung, Y.; Sodt, A.; Gill, P. M. W.; Head-Gordon, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6692–6697.
- Chinnamsetty, S. R.; Espig, M.; Khoromskij, B. N.; Hackbusch, W.; Flad, H.-J. *J. Chem. Phys.* **2007**, *127*, 084110.
- NIST Handbook of Mathematical Functions*; Olver, F. W. J., Lozier, D. W., Boisvert, R. F., Clark, C. W., Eds.; Cambridge University Press: New York, 2010.
- Savin, A. In *Recent Developments of Modern Density Functional Theory*; Seminario, J. M., Ed.; Elsevier: Amsterdam, 1996.
- Dombroski, J. P.; Taylor, S. W.; Gill, P. M. W. *J. Phys. Chem.* **1996**, *100*, 6272–6276.
- Gill, P. M. W.; Adamson, R. D. *Chem. Phys. Lett.* **1996**, *261*, 105–110.
- Lee, A. M.; Taylor, S. W.; Dombroski, J. P.; Gill, P. M. W. *Phys. Rev. A* **1997**, *55*, 3233–3235.
- Gill, P. M. W. *Chem. Phys. Lett.* **1997**, *270*, 193–195.
- Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.
- Fusti-Molnar, L.; Pulay, P. *J. Chem. Phys.* **2002**, *117*, 7827–7835.
- Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 8207–8215.
- Izmaylov, A. F.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **2006**, *125*, 104103.
- Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 224106.
- Brothers, E. N.; Izmaylov, A. F.; Normand, J. O.; Barone, V.; Scuseria, G. E. *J. Chem. Phys.* **2008**, *129*, 011102.
- Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- Tawada, Y.; Tsuneda, T.; Yanagisawa, S. *J. Chem. Phys.* **2004**, *120*, 8425–8433.
- Baer, R.; Neuhauser, D. *Phys. Rev. Lett.* **2005**, *94*, 043002.
- Gerber, I. C.; Angyan, J. G. *Chem. Phys. Lett.* **2005**, *415*, 100–105.
- Lochan, R. C.; Jung, Y.; Head-Gordon, M. *J. Phys. Chem. A* **2005**, *109*, 7598–7605.
- Vydrov, O. A.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 074106.
- Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.
- Toulouse, J.; Savin, A. *THEOCHEM* **2006**, *762*, 147–150.
- Sato, T.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2007**, *126*, 234114.
- Song, J. W.; Tokura, S.; Sato, T.; Watson, M. A.; Hirao, K. *J. Chem. Phys.* **2007**, *127*, 154109.
- Chai, J. D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.
- Chai, J. D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.
- Ewald, P. P. *Ann. Phys.* **1921**, *64*, 253–287.
- Crandall, R. E. *Experiment. Math.* **1999**, *8*, 367–379.
- Appel, A. W. *SIAM J. Sci. Stat. Comput.* **1985**, *6*, 85–103.
- Rokhlin, V. *J. Comput. Phys.* **1985**, *60*, 187–207.
- Greengard, L. *The Rapid Evaluation of Potential Fields in Particle Systems*; MIT Press: Cambridge, MA, 1987.
- White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1996**, *253*, 268–278.
- Adamson, R. D.; Dombroski, J. P.; Gill, P. M. W. *J. Comput. Chem.* **1999**, *20*, 921–927.
- Takemasa, T. *Comput. Phys. Commun.* **1988**, *48*, 265–270.
- McMurchie, L. E.; Davidson, E. R. *J. Comput. Phys.* **1978**, *26*, 218.

- (46) Limpanuparb, T.; Gill, P. M. W. In preparation.
- (47) Gill, P. M. W.; Lee, A. M.; Nair, N.; Adamson, R. D. *THEO-CHEM* **2000**, *506*, 303–312.
- (48) Heilbronner, E.; Bock, H. *The HMO-Model and Its Applications: Basis and Manipulation*; Verlag Chemie: Berlin, 1976; English translation.
- (49) Kohn, W. *Phys. Rev. Lett.* **1996**, *76*, 3168–3171.
- (50) Prodan, E.; Kohn, W. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 11635–11638.
- (51) Shao, Y.; et al. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.
- (52) Toulouse, J.; Colonna, F.; Savin, A. *Phys. Rev. A* **2004**, *70*, 062505.
- (53) Watson, G. N. *A Treatise on the Theory of Bessel Functions*; Cambridge University Press: Cambridge, U.K., 1995; pp 576–618, reprint of the second (1944) ed.

Path Integral Computation of Quantum Free Energy Differences Due to Alchemical Transformations Involving Mass and Potential

Alejandro Pérez*

Department of Chemistry, New York University, New York, New York 10003, United States, Nano-bio spectroscopy group, Centro Joxe Mari Korta, Avenida de Tolosa, 72, E-20018 Donostia-San Sebastian, Spain, and Institute of Pure and Applied Mathematics, University of California Los Angeles, Los Angeles, California 90095, United States

O. Anatole von Lilienfeld*

Surface and Interface Sciences Department, Sandia National Laboratories, Albuquerque, New Mexico 87185, United States, Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, United States, and Institute of Pure and Applied Mathematics, University of California Los Angeles, Los Angeles, California 90095, United States

ABSTRACT: Thermodynamic integration, perturbation theory, and λ -dynamics methods were applied to path integral molecular dynamics calculations to investigate free energy differences due to “alchemical” transformations. Several estimators were formulated to compute free energy differences in solvable model systems undergoing changes in mass and/or potential. Linear and nonlinear alchemical interpolations were used for the thermodynamic integration. We find improved convergence for the virial estimators, as well as for the thermodynamic integration over nonlinear interpolation paths. Numerical results for the perturbative treatment of changes in mass and electric field strength in model systems are presented. We used thermodynamic integration in *ab initio* path integral molecular dynamics to compute the quantum free energy difference of the isotope transformation in the Zundel cation. The performance of different free energy methods is discussed.

I. INTRODUCTION

The quantum nature of nuclei plays an important role in the prediction of many properties of systems containing light elements such as hydrogen. A successful approach to investigating such effects relies on first principles (*ab initio*) based path-integral molecular dynamics (AI-PIMD).^{1,2} In AI-PIMD, the nuclei are quantized following Feynman’s path integral (PI) formalism,^{3,4} while the many-body potential of the system is computed “on-the-fly” from electronic structure calculations.^{5,6} Proton transfer events are now routinely investigated using AI-PIMD in various hydrogen-bonded systems in gas,^{7–10} condensed phases,^{11,12} or even on surfaces.^{13,14}

Nuclear quantum effects in biological systems have attracted considerable attention in recent years. Various methods have been developed and successfully deployed to investigate nuclear quantum effects in biological systems.^{15,16} For example, using AI-PIMD, we recently identified nuclear quantum effects to be crucial for kinetically stabilizing the canonical Watson–Crick base pair configuration by rendering their rare enol tautomer form metastable.¹⁷

In this study, we explore “alchemical” changes within the path integral formalism. By alchemical changes, we refer to changes of parameters that define a given Hamiltonian, such as atomic numbers, atomic masses, number of particles, or the interaction potential. Studying changes in composition (atomic numbers) holds great promise for future applications in the arena of rational computational materials design, i.e., approaches that attempt to outperform a mere screening of combinatorial sets of materials candidates.^{18–23} The computation of free energy

differences²⁴ between compounds based on thermodynamic integration,²⁵ or free-energy perturbations,²⁶ has only recently been applied using electronic structure methods.^{27–29} For example, a related framework was used to successfully determine the chemical composition of the Earth’s core.³⁰ To the best of our knowledge, similar alchemical changes have not been hitherto investigated within PIMD.

Alchemical transformations involving changes in atomic numbers could one day lead to the engineering of tunneling barriers in proton-conducting materials. For changes involving only the mass, such an approach permits the direct evaluation of kinetic isotope effects. Only very recently, and while we carried out our study, an adaptation of AI-PIMD for the computation of isotope effects in intramolecular hydrogen transfer *via* thermodynamic integration was published.³¹ More recently, Ramirez and Herrero applied thermodynamic integration in path integrals to investigate the isotope effects on the melting temperature of ice Ih.³² Finally, we also note a recent application of the Jarzynski equality within PIMD.³³

In this study, we discuss analytically solvable model systems subjected to changes in mass and potential. Free energy differences were computed numerically using thermodynamic integration and λ -dynamics methods.³⁴ We furthermore consider the perturbative treatment of changes in mass and static electric fields. While ultimately AI-PIMD will be used to investigate such alchemical changes in more realistic systems, our focus here is on

Received: January 24, 2011

Published: June 06, 2011

formulating and assessing the methods for simple model systems. To the best of our knowledge, a path integral scheme to generally estimate free energy changes due to alchemical transformations has yet to be presented.

This paper is organized as follows: In the Theory section, we first recall Feynman's path integral expression for the Helmholtz free energy. Then, we describe thermodynamic integration, perturbation theory, and λ -dynamics methods applied to PIMD. In the subsequent Results section, the investigated model systems and numerical results are discussed. The paper ends with the Conclusions section and the Appendix.

II. THEORY

A. Path Integral Expression for the Helmholtz Free Energy. Feynman's PI formulation^{3,4} of quantum statistical mechanics describes each nucleus by a ring polymer of P (known as the Trotter number) interacting quasi-particles, colloquially referred to as "beads". In the usual formulation, these beads are connected by harmonic springs and experience the effect of an attenuated classical potential.³⁵ Path configurations can then be sampled either stochastically via Monte Carlo methods³⁶ or deterministically using molecular dynamics.³⁷ Thus, path integral molecular dynamics (PIMD) is a straightforward way to include nuclear quantum effects in ordinary molecular dynamics (MD) simulations.

The discrete path integral expression for the canonical quantum partition function of a single particle in 3D is³⁴

$$Z(\lambda) \approx C(\lambda) \int d\mathbf{r}_1 \dots d\mathbf{r}_P e^{-\beta\Phi(\{\mathbf{r}\};\lambda)},$$

with

$$C(\lambda) = \left(\frac{m(\lambda)P}{2\pi\beta\hbar^2} \right)^{3P/2} \quad (1)$$

where P is the Trotter number, i.e., the number of replicas or beads³⁵ representing the quantum particle, and Φ is the path integral effective potential,

$$\Phi(\{\mathbf{r}\};\lambda) = \sum_{s=1}^P \frac{m(\lambda)}{2} \omega_p^2 (\mathbf{r}_s - \mathbf{r}_{s+1})^2 + \sum_{s=1}^P \frac{V(\mathbf{r}_s;\lambda)}{P} \quad (2)$$

where $\omega_p^2 = P/(\beta\hbar^2)$, $\{\mathbf{r}\}$ denotes the bead coordinates, and V is the classical potential. The paths obey the cyclic condition $\mathbf{r}_{P+1} = \mathbf{r}_1$ imposed by the quantum-mechanical trace operation $Z = \text{Tr}[\exp(-\beta\hat{H})]$. Equation 1 shows explicitly the parametric dependence on λ driving the transformation in mass $m(\lambda)$ and potential (see next section).

The canonical average of any observable O at a fixed λ value, $\langle O \rangle_\lambda$, is given by the following integral over the path configurations:

$$\langle O \rangle_\lambda \approx \frac{C(\lambda)}{Z(\lambda)} \int d\mathbf{r}_1 \dots d\mathbf{r}_P O(\mathbf{r}_1, \dots, \mathbf{r}_P; \lambda) e^{-\beta\Phi(\{\mathbf{r}\};\lambda)} \quad (3)$$

In practical path integral calculations, one performs a suitable coordinate transformation of eq 2 to either normal modes or staging coordinates.³⁷ To carry out PIMD, a Hamiltonian is constructed by adding a set of momenta to this transformed effective path integral potential. The dynamical masses of these pseudoparticles are then chosen so as to

bring all free-particle modes to the same time scale. Finally, thermostats are attached to each bead to improve ergodic canonical sampling.³⁷

B. Thermodynamic Integration and Path Integrals. The quantum free energy change of a system due to an alchemical transformation is, by definition,

$$\Delta F = F_f - F_i = -\frac{1}{\beta} \ln \left(\frac{Z_f}{Z_i} \right) \quad (4)$$

where $Z_i(F_i)$ and $Z_f(F_f)$ denote the initial and final canonical quantum partition functions (Helmholtz free energies), respectively. Equation 4 can be evaluated numerically using the thermodynamic integration (TI) method:^{25,38,39}

$$\Delta F = \int_0^1 d\lambda \left(\frac{dF(\lambda)}{d\lambda} \right)_\lambda \quad (5)$$

where λ is a coupling parameter that drives the transformation from an initial state i to a final state f . We remark that while it is well-known that eq 5 can be identified with $\int d\lambda \langle \partial_\lambda V \rangle$ in classical statistical mechanics (V being the potential energy),³⁴ this is not the case for PI (see below).

Consider now a single quantum particle undergoing a change of identity, i.e., a simultaneous change of mass and potential energy function. Since the Helmholtz free energy F is a state function, ΔF is independent of how the two states are interpolated using the λ variable, provided that the two end points are met at $\lambda = 0$ and 1. For simplicity, consider a linear interpolation path and a single λ parameter driving both transformations

$$\begin{aligned} V(\lambda) &= \lambda V_f + (1 - \lambda) V_i \\ m(\lambda) &= \lambda m_f + (1 - \lambda) m_i \end{aligned} \quad (6)$$

with $0 \leq \lambda \leq 1$. Obviously, the intermediate λ states do not correspond to any real system but are only a (perfectly rigorous) way of obtaining the corresponding free energy difference. Although in eq 6 a single parameter λ is used, it may be more efficient to use two independent control parameters, λ_m and λ_v , for the mass and potential transformation, respectively.

As noted above, any interpolation scheme is permissible provided that the end point condition $Z(\lambda = 1) = Z_f$ and $Z(\lambda = 0) = Z_i$ is satisfied. Although a linear interpolation is the simplest choice, it is not necessarily the most efficient. An optimal choice is the scheme that samples many function values in the region of the λ domain where the function to be integrated varies rapidly. In the Results and Discussion section, we explore in detail the relative performance of linear and nonlinear interpolations. Specifically, a simple quartic interpolation in λ will be discussed

$$\begin{aligned} V(\lambda) &= \lambda^4 V_f + (1 - \lambda^4) V_i \\ m(\lambda) &= \lambda^4 m_f + (1 - \lambda^4) m_i \end{aligned} \quad (7)$$

As expected, different interpolation functions lead to different convergence properties for thermodynamic integration using PIMD (TI-PIMD).

C. TI-PIMD: Changes in Mass. The TI-PIMD evaluation of the quantum free energy difference associated with a change of mass enables us to predict isotope effects. For changes being restricted to the mass, i.e., V being independent of λ , the first

derivative of eq 1 with respect to λ yields the average of the free energy derivative:

$$\left(\frac{dF}{d\lambda}\right)^{\text{prim}} = -\frac{1}{\beta} \frac{\partial_{\lambda} Z(\lambda)}{Z(\lambda)} \approx -\left(\frac{\partial_{\lambda} m(\lambda)}{m(\lambda)}\right) \times \left[\frac{3P}{2\beta} - \frac{m}{2} \omega_p^2 \sum_{s=1}^P \langle (\mathbf{r}_s - \mathbf{r}_{s+1})^2 \rangle_{\lambda}\right] \quad (8)$$

Here, the term $3P/(2\beta)$ comes from the normalization prefactor $C(\lambda)$ in eq 1. Equation 8 corresponds to the *primitive* estimator of the first-order derivative of the free energy with respect to mass. This expression is expected to have the same statistical problems as the *primitive* estimator for the quantum kinetic energy.⁴⁰ Better convergence, however, can be achieved using the *virial* relation for bound systems in path integrals:^{34,40}

$$\left\langle \frac{3P}{2\beta} - \frac{m\omega_p^2}{2} \sum_{s=1}^P (\mathbf{r}_s - \mathbf{r}_{s+1})^2 \right\rangle = \left\langle \frac{1}{2P} \sum_{s=1}^P \mathbf{r}_s \cdot \frac{\partial V}{\partial \mathbf{r}_s} \right\rangle \quad (9)$$

Insertion in eq 8 yields the virial estimator for a bound particle

$$\left(\frac{dF}{d\lambda}\right)^{\text{vir}} \approx -\left(\frac{\partial_{\lambda} m(\lambda)}{m(\lambda)}\right) \left\langle \frac{1}{2P} \sum_{s=1}^P \mathbf{r}_s \cdot \frac{\partial V(\mathbf{r}_1 \dots \mathbf{r}_P)}{\partial \mathbf{r}_s} \right\rangle_{\lambda} \quad (10)$$

For a homogeneous system of N particles (all undergoing a simultaneous change in mass), the virial expression becomes

$$\left(\frac{dF}{d\lambda}\right)^{\text{vir}} \approx -\left(\frac{\partial_{\lambda} m(\lambda)}{m(\lambda)}\right) \times \left\langle \frac{3N}{2\beta} + \frac{1}{2P} \sum_{I=1}^N \sum_{s=1}^P (\mathbf{r}_{I,s} - \bar{\mathbf{r}}_I) \cdot \frac{\partial V}{\partial \mathbf{r}_{I,s}} \right\rangle_{\lambda} \quad (11)$$

where $\bar{\mathbf{r}}_I = \sum_{s=1}^P \mathbf{r}_{I,s}/P$ is the centroid of the ring polymer describing particle I .

If only the atoms belonging to a certain element (usually, hydrogen) are undergoing the isotope transformation, then

$$\left(\frac{dF}{d\lambda}\right)^{\text{vir}} \approx -\sum_{I=1}^{N_c} \left(\frac{\partial_{\lambda} m}{m_I}\right) \times \left[\frac{3}{2\beta} + \left\langle \frac{1}{2P} \sum_{s=1}^P (\mathbf{r}_{I,s} - \bar{\mathbf{r}}_I) \cdot \nabla_{\mathbf{r}_I} V(\mathbf{r}_s) \right\rangle_{\lambda} \right] \quad (12)$$

where N_c is the number of atoms in the subset. Similarly, the primitive estimators is

$$\left(\frac{dF}{d\lambda}\right)^{\text{prim}} \approx -\sum_{I=1}^{N_c} \left(\frac{\partial_{\lambda} m}{m_I}\right) \times \left[\frac{3P}{2\beta} - \left\langle \frac{m_I}{2} \omega_p^2 \sum_{s=1}^P (\mathbf{r}_{I,s} - \mathbf{r}_{I,s+1})^2 \right\rangle_{\lambda} \right] \quad (13)$$

It is well-known that *virial* estimators exhibit less fluctuations than their *primitive* counterparts.^{34,40} In the Results and Discussion section, we confirm the favorable statistical properties of the virial estimator.

D. TI-PIMD: Changes in Potential. In the same vein as isotope transformations, changes in the free energy due to varying the underlying classical potential can be also evaluated using

TI-PIMD. The path integral estimator for the free energy derivative with respect to changes in the potential V is given by

$$\frac{dF}{d\lambda} = -\frac{1}{\beta} \left\langle \frac{\partial_{\lambda} Z}{Z} \right\rangle \approx \left\langle \sum_{s=1}^P \frac{\partial_{\lambda} V(\mathbf{r}_s; \lambda)}{P} \right\rangle_{\lambda} \quad (14)$$

In contrast to changes in the mass, there is generally only one version of the estimator for changes in the potential.

As an illustrative example, the free energy dependence on the force constant $\kappa(\lambda)$ of the harmonic potential ($V(x; \lambda) = \kappa(\lambda)x^2/2$) can be evaluated according to

$$\frac{dF(\lambda)}{d\lambda} \approx \left\langle \frac{\partial \kappa(\lambda)}{\partial \lambda} \sum_{s=1}^P \frac{x_s^2}{2P} \right\rangle_{\lambda} \quad (15)$$

The numerical result of this expression can be compared to the analytical free energy derivative with respect to force constant for the harmonic potential,

$$\frac{dF}{d\lambda} = \frac{\hbar}{4} \frac{\partial_{\lambda} \kappa(\lambda)}{\sqrt{m\kappa(\lambda)}} \coth\left(\frac{\beta\hbar}{2} \sqrt{\frac{\kappa(\lambda)}{m}}\right) \quad (16)$$

Numerical results will be presented and discussed below.

E. TI-PIMD: Changes in Mass and Potential. We now consider a simultaneous transformation in mass and potential state function using TI-PIMD. The first derivative of the quantum free energy (see eq 1) with respect to parameter λ is simply the combination of eq 8 and eq 14

$$\frac{dF}{d\lambda} = -\frac{1}{\beta} \left\langle \frac{\partial_{\lambda} Z}{Z} \right\rangle \approx -\frac{1}{\beta} \langle \mathcal{F} \rangle_{\lambda} \quad (17)$$

where the function to be averaged is

$$\mathcal{F} = \frac{3P}{2m(\lambda)} \frac{\partial m(\lambda)}{\partial \lambda} - \beta \sum_{s=1}^P \frac{\partial_{\lambda} V(\{\mathbf{r}_s\}; \lambda)}{P} - \beta \left[\frac{\partial_{\lambda} m(\lambda)}{2} \omega_p^2 \sum_{s=1}^P (\mathbf{r}_s - \mathbf{r}_{s+1})^2 \right]$$

In eq 17, the primitive estimator (eq 8) was used for the contribution of the isotope transformation to the free energy change. Using eq 10, we can also obtain a "virial" expression for eq 17.

The analytical expression for the harmonic oscillator undergoing a simultaneous change in mass $m(\lambda)$ and force constant $\kappa(\lambda)$ is

$$\frac{dF}{d\lambda} = \frac{\hbar}{4\sqrt{\kappa(\lambda)/m(\lambda)}} \left[\frac{m(\lambda) \partial_{\lambda} \kappa(\lambda) - \kappa(\lambda) \partial_{\lambda} m(\lambda)}{m(\lambda)^2} \right] \times \coth\left(\frac{\beta\hbar}{2} \sqrt{\kappa(\lambda)/m(\lambda)}\right) \quad (18)$$

Numerical results and comparison will be presented and discussed below.

F. Perturbative Treatment of Changes in Mass. Path integral calculations are often very expensive computationally. Therefore, it would be desirable to have a method capable of estimating the isotope effect with less or even without any of the intermediate calculations required by TI-PIMD.

If the isotope transformation is not too large, perturbation theory⁴¹ (PT) can be used to estimate the free energy change from a single PIMD simulation. The combination of free energy

perturbation and PIMD is termed here as PT-PIMD. The free energy difference due to a (small) change in mass of all atoms of a certain element (leaving the others unchanged) is

$$\Delta F \approx -\frac{1}{\beta} \ln \left[\left(\frac{m_f}{m_i} \right)^{3N_c P/2} \frac{\int d^{NP} \mathbf{r} e^{-\beta(\Phi_f - \Phi_i)} e^{-\beta\Phi_i}}{\int d^{NP} \mathbf{r} e^{-\beta\Phi_i}} \right]$$

$$= -\frac{1}{\beta} \left[\frac{3N_c P}{2} \ln \left(\frac{m_f}{m_i} \right) + \ln \langle e^{-\beta(m_f - m_i)A} \rangle_i \right],$$

$$\text{with } A = \frac{\omega_p^2}{2} \sum_{l=1}^{N_c} \sum_{s=1}^P (\mathbf{r}_{l,s} - \mathbf{r}_{l,s+1})^2 \quad (19)$$

where N_c is the number of atoms undergoing the isotope transformation, Φ denotes the effective path integral potential similar to eq 2, and the angular brackets denote the canonical sampling over path configurations of a system where the mutating elements have fixed mass m_i .

From eq 19, it is clear that if the difference in mass is very large, there will be very little overlap between states, and the sampling will be very inefficient. However, even in this case, one can split the interval of interest into nw windows and perform PT-PIMD on each individual window. Then, the free energy difference is given by the sum over all PT-PIMD windows,

$$\Delta F = \sum_{i=0}^{nw-1} \Delta F(i\Delta\lambda \rightarrow (i+1)\Delta\lambda) \quad (20)$$

As long as the free energy change in each window is around $\sim 2k_B T$, we can efficiently sample eq 19 using PT-PIMD methods. For further statistics, the $nw - 2$ interior intervals can be sampled forward and backward within *double-wide sampling*.³⁸

G. λ -Dynamics Approach for Quantum Free Energy Differences: λ -PIMD. The free energy difference between two chemical states can be also computed using the so-called λ -dynamics (LD) approach.^{38,42–44} LD follows the same spirit as thermodynamic integration, but instead of running separate simulations for each fixed λ value, this parameter is allowed to vary continuously in a single (albeit long) MD simulation. Here, we propose to combine LD with PIMD. To this end, the λ parameter is endowed with dynamical character according to the following extended PIMD Hamiltonian:

$$H = \sum_{s=1}^P \left[\frac{p_s^2}{2m_s} + \frac{m}{2} \omega_p^2 (\mathbf{r}_s - \mathbf{r}_{s+1})^2 \right] + \frac{p_\lambda^2}{2m_\lambda}$$

$$+ \sum_{s=1}^P [\lambda V_f(\mathbf{r}_s) + (1-\lambda) V_i(\mathbf{r}_s)]/P \quad (21)$$

which allows the continuous interpolation between initial (V_i) and final (V_f) potential energy functions. In LD, the associated mass of the λ particle (m_λ) has to be large enough to allow for proper equilibration of the other degrees of freedom. We term the scheme given by eq 21 as λ -PIMD.

The λ path chosen in eq 21 corresponds to a linear path. The coupling parameter obeys the constraint $\lambda \in [0, 1]$. Usually, this condition is enforced in MD via Lagrange multipliers. However, an elegant alternative (adopted here) is to use the trigonometric parametrization presented in ref 38. Using the change of

variable $\lambda = \cos^2 \theta$, with $\theta \in [0, \pi/2]$, in eq 21 yields

$$H = \sum_{s=1}^P \left[\frac{p_s^2}{2m_s} + \frac{m}{2} \omega_p^2 (\mathbf{r}_s - \mathbf{r}_{s+1})^2 \right] + \frac{p_\theta^2}{2m_\theta}$$

$$+ \sum_{s=1}^P [V_f(\mathbf{r}_s) \cos^2 \theta + V_i(\mathbf{r}_s) \sin^2 \theta]/P \quad (22)$$

The velocity of the λ particle, $v_\theta = p_\theta/m_\theta$, is reversed whenever its position θ hits the boundaries.

The free energy change for this transformation is then obtained from numerical integration of the λ -derivative

$$\Delta F = F_f - F_i = \int_0^1 \langle \Delta V \rangle_\lambda d\lambda$$

$$= \int_0^{\pi/2} \langle \Delta V \rangle_\theta \sin(2\theta) d\theta \quad (23)$$

where $\Delta V(\theta) = (1/P) \sum_{s=1}^P [V_f(\mathbf{r}_s; \theta) - V_i(\mathbf{r}_s; \theta)]$. In eq 23, configurations from the λ -dynamics trajectory are histogrammed according to their $\lambda(\theta)$ value.⁴⁵

We note in passing that in other λ -dynamics methods such as the one proposed in ref 46, λ switching functions are designed so as to induce a free energy barrier between the end points with the goal of enhancing the sampling of these states. Then, the λ particle is kept at a higher temperature than the system to promote barrier crossing along the λ trajectory.⁴⁶ We note that under the physical conditions of our calculations, we did not encounter significant free energy barriers, and the λ particle was thermalized at the same temperature as the physical system.

The λ -PIMD scheme presented here is general, and it could be used for instance to estimate quantum free energy changes for proton transfer, ligand binding affinities, or acid dissociation constants (pK_a). As an illustrative example, in the next section, we present an application of λ -PIMD to study changes in the force constant in the harmonic oscillator. By declaring the mass variable, it could also be used to investigate the isotope effect.

Finally, λ -dynamics could be easily extended by using P independent λ particles. That is, an independent λ particle is associated with each bead of the ring polymer according to the Hamiltonian

$$H = \sum_{s=1}^P \left[\frac{p_s^2}{2m_s} + \frac{p_{\lambda,s}^2}{2m_{\lambda,s}} + \frac{m}{2} \omega_p^2 (\mathbf{r}_s - \mathbf{r}_{s+1})^2 \right]$$

$$+ \sum_{s=1}^P [\lambda_s V_f(\mathbf{r}_s) + (1-\lambda_s) V_i(\mathbf{r}_s)]/P \quad (24)$$

III. RESULTS AND DISCUSSION

Natural units ($\hbar = k_B = m = 1$) are used for the model systems investigated, except for the Zundel cation, hydrogen molecule, and Morse potential, where we used atomic units (au).

A. TI-PIMD: Changes in Mass. 1. *Quantum Harmonic Oscillator.* Using TI-PIMD, the isotope effect was investigated for the harmonic potential $V(x) = m(\lambda)\omega^2 x^2/2$. For this system, the difference in quantum free energy is known analytically

$$\Delta F = F(m_f) - F(m_i) = \frac{1}{\beta} \ln \left[\frac{\sinh(\beta \hbar \omega_f/2)}{\sinh(\beta \hbar \omega_i/2)} \right] \quad (25)$$

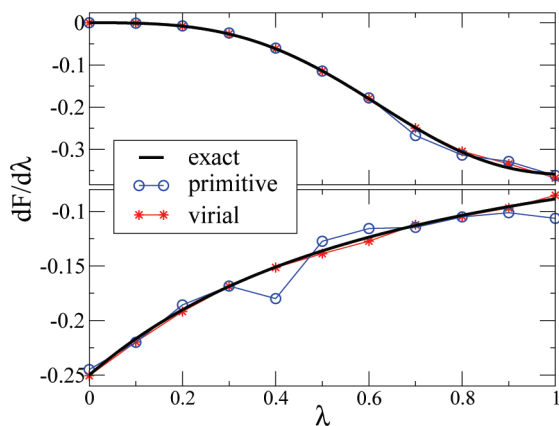


Figure 1. TI-PIMD: First derivative of the quantum free energy with respect to mass for the harmonic potential $V = m(\lambda) \omega^2 x^2/2$. The mass $m(\lambda)$ is a linear (bottom) or quartic (top) interpolation between $m_i = 1 \rightarrow m_f = 2$. Black: exact result (eq 26). Red: virial estimator (eq 10). Blue: primitive estimator (eq 8). Natural units ($\hbar = k_B = 1$) are used. The temperature was set to $\beta\hbar\omega = 10$, and PIMD (with $P = 64$) was conducted for 400 000 steps.

where the oscillator frequencies are $\omega_\lambda^2 = \kappa/m(\lambda)$ and κ is the (fixed) force constant. The derivative of the free energy with respect to λ is also known analytically,

$$\frac{dF}{d\lambda} = -\frac{\hbar}{4} \left(\frac{\partial_\lambda m(\lambda)}{m(\lambda)} \right) \omega_\lambda \coth(\beta\hbar\omega_\lambda/2) \quad (26)$$

Note that the value of eq 26 depends on the interpolation function adopted for $m(\lambda)$.

Figure 1 (bottom) shows the primitive, virial, and exact first-order derivative of the free energy as a function of mass using a linear path. Using the same number of PIMD steps (400 000), the convergence of the primitive estimator (blue) is worse than the virial estimator (red), which is nearly identical to the exact value given by eq 26. The corresponding change of free energy as obtained from thermodynamic integration (eq 5) yields -0.154 and -0.148 for the primitive and virial estimators, respectively. The exact value from eq 25, $\Delta F = -0.147$, compares favorably with these numerical results using eqs 8 and 10. The average error bar for the primitive estimator is 0.0091, which is 5 times larger than the error bar for the virial estimator (0.0018), both for the linear path. This result confirms that the virial expression usually has less statistical error and is preferable for efficient computation of quantum free energy differences.

Figure 1 (top) shows numerical results for the virial and primitive estimators of the free energy derivative for the same system but with the quartic interpolation in mass, $m(\lambda) = \lambda^4 m_f + (1 - \lambda^4) m_i$. Using the same number of normal mode PIMD steps (400 000) as Figure 1 (bottom), the nonlinear interpolation (top) clearly improves the sampling for both estimators. In particular, the statistical error of the primitive estimator (0.0054) is greatly reduced, whereas the virial remains the same: 0.0018. Both estimators yield results nearly indistinguishable (and therefore converged) from the exact values, eq 26. In this case, the change of free energy integrates to -0.151 for the primitive and to -0.150 for the virial estimator, the exact value being -0.147 .

2. Double Quantum Harmonic Oscillator. We investigated a three-body system that resembles the highly relevant proton

transfer scenario in many naturally occurring systems, such as in ref 17. This model system consists of three collinear masses $S_1 - S_2 \cdots S_3$, where S_1 and S_3 sites represent identically electronegative “heavy” atoms (e.g., oxygens) with equal masses ($m_1 = m_3$). The extension to the asymmetric, heterogeneous ($m_1 \neq m_3$) case is straightforward. The intermediate atom S_2 is supposed to represent the light atom that will undergo a change in mass (from hydrogen to deuterium). The particles are assumed to interact harmonically. The Hamiltonian operator of this system is

$$\hat{\mathcal{H}} = -\frac{\hbar^2}{2} \sum_{i=1}^3 \frac{1}{m_i} \frac{\partial^2}{\partial x_i^2} + \frac{\kappa}{2} [(x_2 - x_1 - x_{\text{eq}})^2 + (x_3 - x_2 - x_{\text{eq}})^2] \quad (27)$$

where κ is the force constant and x_{eq} is the equilibrium distance. This many-body system can be solved exactly using Jacobi coordinates

$$\begin{aligned} x_s &= \frac{1}{\sqrt{2}}(x_3 - x_1 - 2x_{\text{eq}}) \\ x_a &= \frac{1}{\sqrt{2}}(x_1 + x_3 - 2x_2) \\ X &= [m_1(x_1 + x_3) + m_2x_2]/(2m_1 + m_2) \end{aligned}$$

where x_a and x_s represent the asymmetric and symmetric stretching modes, respectively, and X is the center of mass position. In terms of these Jacobi coordinates, the Hamiltonian becomes separable

$$\begin{aligned} \hat{\mathcal{H}} &= \left[-\frac{\hbar^2}{2} \frac{\partial_X^2}{m_X} \right] + \left[-\frac{\hbar^2}{2} \frac{\partial_s^2}{m_s} + \frac{m_s}{2} \omega_s^2 x_s^2 \right] \\ &+ \left[-\frac{\hbar^2}{2} \frac{\partial_a^2}{m_a} + \frac{m_a}{2} \omega_a^2 x_a^2 \right] \end{aligned} \quad (28)$$

where the effective masses $m_a = m_1 m_2 / (2m_1 + m_2)$ and $m_s = m_1$ and corresponding frequencies $\omega_n = (\kappa/m_n)^{1/2}$, $n = a, s$, have been introduced. In eq 28, the first term represents the center of mass motion, the second term is the symmetric vibrational mode, and the third term is the asymmetric stretch. Note that for this simple system, x_a is the only mode that involves motion of the light atom S_2 , thereby exclusively contributing to the isotope effect. The symmetric mode x_s (only the end point masses move in opposite directions) does not contribute to the isotope effect.

The first derivative of the free energy with respect to the mass of S_2 can be solved analytically,

$$\frac{dF}{d\lambda} = -\frac{\hbar\omega_a}{4} \left(\frac{\partial_\lambda m_a}{m_a} \right) \coth\left(\frac{\beta\hbar\omega_a}{2}\right) - \frac{1}{2\beta} \left(\frac{\partial_\lambda m_2}{M} \right) \quad (29)$$

where $M = 2m_1 + m_2$ is the total mass of the system and $\partial_\lambda m_a$ and $\partial_\lambda m_2$ are the interpolation-function-dependent partial derivatives of the effective asymmetric and transforming masses, respectively. The first and second terms in eq 29 are the contributions from the asymmetric stretching mode and from the free motion of the center of mass, respectively.

For this model system, the numerical and analytic results are presented in Figure 2 for the quartic interpolation of the mass in λ , $m_2(\lambda) = \lambda^4 m_D + (1 - \lambda^4) m_H$. The partial derivatives for this quartic interpolation path are $\partial_\lambda m_a(\lambda) = 2m_1^2 m_2(\lambda) / (2m_1 + m_2(\lambda))^2$ and $\partial_\lambda m_2(\lambda) = 4\lambda^3$. The initial hydrogen mass $m_H = 1$ was changed to the mass of deuterium $m_D = 2$. The equilibrium

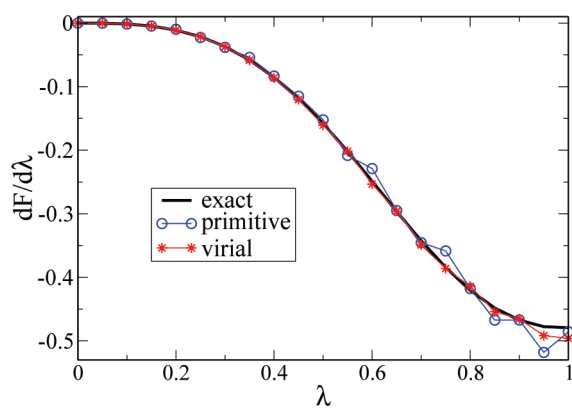


Figure 2. TI-PIMD: First order derivative of the quantum free energy with respect to mass of the central atom in the double oscillator model (eq 27). A quartic nonlinear interpolation path $m(\lambda) = \lambda^4 m_D + (1 - \lambda^4) m_H$ between the initial mass of hydrogen ($m_H = 1$) and final mass of deuterium ($m_D = 2$) was used. Natural units are used. Black: exact result (eq 29). Red: virial estimator (eq 10). Blue: primitive estimator (eq 8).

distance x_{eq} in eq 27 was set to 1. The inverse temperature $\beta = 10$ and Trotter number $P = 64$ were used. The numerical estimators were computed using 100 000 normal mode PIMD steps per window ($\Delta t = 0.01$) for a total of 21 trajectories.

Figure 2 shows again the superior convergence behavior of the virial estimator. The corresponding integrated quantum free energy changes are -0.204 for the primitive estimator (eq 8) and -0.205 for the virial estimator (eq 10). These values are in excellent agreement with the exact result (-0.204). The average error bar for the primitive estimator is 0.012, which is 5 times larger than the error for the virial estimator (0.002). We note that the overall shape of this derivative is similar to the corresponding single harmonic oscillator in Figure 1, suggesting that the first term in eq 29 is dominating.

In more general situations (heterogeneous system A–B···C and/or anharmonic potentials), all vibrational modes are coupled and are therefore expected to contribute to the isotope effect. Such scenarios could be modeled using an antiparallel double Morse or Lennard-Jones potential.

3. Isotope Effect on the Zundel Cation from First Principles. The Zundel cation (H_5O_2^+), which consists of a proton shared by two water molecules (see inset of Figure 3), is a model for strongly hydrogen-bonded systems and is of paramount importance in condensed phases and in enzymatic reactions. Using a combination of centroid molecular dynamics (CMD) and empirical valence bond (EVB) theory, Schmitt and Voth found that the deuterated Zundel complex in liquid water exhibits a free energy barrier increase of about 0.2 kcal/mol compared to the all-hydrogen complex, leading to a decrease of 1.4 times in the hopping rate of proton transfer.⁴⁷ Here, we investigated the stability of H_5O_2^+ versus its deuterated counterpart (D_5O_2^+) with *ab initio* path-integral molecular dynamics. The virial estimator of eq 13 was implemented in the density functional theory code CPMD (version 3.13.2)⁴⁸ for the PIMD thermodynamic integration. The details of the *ab initio* PIMD simulation are as follows. The exchange and correlation were approximated with the BLYP^{49–51} functional. Numerical norm-conserving Trouillier–Martins pseudopotentials⁵² were used to represent core electronic orbitals. These pseudopotentials were augmented with a dispersion energy correction as described in

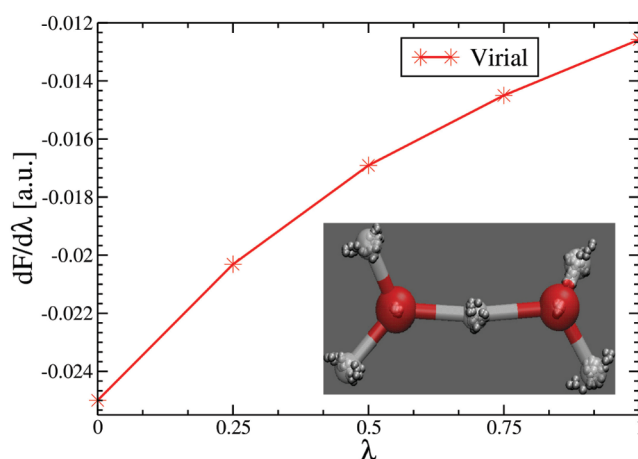


Figure 3. Results for the virial estimator for the *ab initio* TI-PIMD simulation ($P = 32$) of the Zundel cation at 300 K. The predicted change in free energy for the all-H to all-D transformation is $\Delta F = -0.01755$ au (-11.02 kcal/mol). The inset shows a typical snapshot (created with the visualization program VMD^{59,60}) of the Zundel cation from our PIMD simulations.

refs 53 and 54. A kinetic energy cutoff of 100 Ry was employed for the plane wave basis set expansion of the electronic valence orbitals. The Zundel cation was placed in a cubic box of volume 12^3 \AA^3 . Isolated molecule boundary conditions were imposed using the algorithm of Martyna and Tuckerman.⁵⁵ The Car–Parrinello (CP) algorithm⁵ was used with a fictitious electronic mass of 340 au and a time step of 2 au (≈ 0.0484) fs. These parameters were carefully chosen to ensure the stability of the simulations. With these parameters, the adiabaticity of the CP scheme was preserved, and the fictitious electronic kinetic energy was stable and always less than 0.003 au. Normal mode variables and 32 beads were used for the discrete path integral. Massive Nosé–Hoover chain thermostats⁵⁶ of length 3 and a characteristic frequency of 4400 cm^{-1} were employed to ensure adequate canonical sampling at 300 K.⁵⁷ A total of five independent PIMD trajectories, each of 3.7 ps, were simulated corresponding to ($\lambda = 0, 0.25, 0.5, 0.75, 1.0$) values in the thermodynamic integration. A linear path was used for the mass interpolation between hydrogens and deuteriums. We checked the correctness of our CPMD implementation by first simulating the hydrogen molecule at 300 K using a setup similar to that described above. An estimate of the free energy change of the isotopic transformation $\text{H}_2 \rightarrow \text{D}_2$ is available by other quantum chemistry methods (≈ -2 kcal/mol) and was reproduced numerically (see details in the Appendix).

Figure 3 shows the virial TI-PIMD results for the change in mass of the hydrogen atoms to deuterium in the Zundel cation at 300 K. The system is expected to be more stable upon deuteration mostly due to the suppression in the zero point energy. The predicted change in quantum free energy, $\Delta F = -0.01755$ au (-11.02 kcal/mol), was obtained from the numerical integration of the TI-PIMD curve in Figure 3 using Simpson's rule. This value agrees well with the thermochemistry result (-10.16 kcal/mol) computed using the quantum chemistry code Firefly⁵⁸ at the BLYP/6-311++G** level under the harmonic and rigid rotor approximation. We remark that converging error bars below 1 kcal/mol (our average standard deviation was 1.36 kcal/mol) would require very long PIMD simulations and extremely well equilibrated path integral configurations. Finally, we mention that

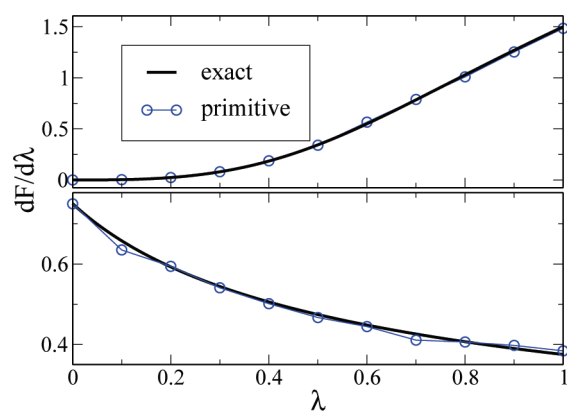


Figure 4. TI-PIMD: Free energy derivative with respect to the force constant ($\kappa_i = 1 \rightarrow \kappa_f = 4$) in the harmonic oscillator. The inverse temperature was $\beta = 10$. Natural units are used. PIMD (with $P = 64$) was conducted for 400 000 steps. Bottom: linear path $\kappa(\lambda) = \lambda\kappa_f + (1 - \lambda)\kappa_i$. Top: nonlinear path $\kappa(\lambda) = \kappa_f\lambda^4 + (1 - \lambda^4)\kappa_i$.

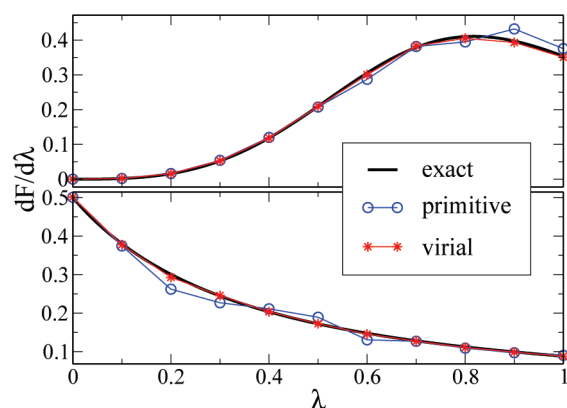


Figure 5. TI-PIMD: First order free energy derivative with respect to simultaneous changes in force constant and mass in the harmonic oscillator: $m_i = 1, \kappa_i = 1 \rightarrow m_f = 2, \kappa_f = 4$. Bottom: linear interpolation, eq 6. Top: nonlinear path, quartic order in λ . A total of 400 000 PIMD (64 beads) steps were simulated at inverse temperature $\beta = 10$.

computing changes in quantum kinetic energy due to isotope transformation may be useful to compare with deep inelastic neutron scattering experiments.

B. TI-PIMD: Changes in Potential. Free energy differences due to variations in the force constant κ of the harmonic potential $V(x) = \kappa x^2/2$ were computed using TI-PIMD. Figure 4 compares numerical values from eq 16 with analytical results according to eq 15 for transformation $\kappa_i = 1 \rightarrow \kappa_f = 4$ in the force constant. Two interpolation paths have been investigated, first an interpolation of the force constant that is linear in λ (bottom), then an interpolation that is quartic in λ , $\kappa(\lambda) = \kappa_f\lambda^4 + (1 - \lambda^4)\kappa_i$ (top). In analogy to the results for changing masses, the results displayed in Figure 4 suggest that also for changing potentials a quartic interpolation path leads to superior convergence behavior. The typical error bars for the primitive estimator are 0.0039 (quartic) and 0.0044 (linear), while for the virial they are 0.0022 (linear) and 0.0022 (quartic). However, as in the previous example and not without irony, due to statistical errors, the numerical integrated result of the less converged linear interpolation (0.5028) compares

Table 1. Morse Potential $V(x) = D[1 - e^{-a(x-x_{\text{eq}})}]^2$ Parameters Describing the Diatomic Molecules^a

Morse parameters	H ₂	HCl
r_{eq}	1.4014	2.4086
D	0.1745	0.1676
a	1.0213	0.9684
μ	918.0584	1785.6425
ω	0.0199	0.013268

^a Also shown is the reduced mass μ and the harmonic frequency $\omega = a((2D/\mu))^{1/2}$. All values are in atomic units. The thermal energy at 300 K is $k_B T = 0.00095$ au.

fortuitously better to the exact value (0.5) than the result of the more converged quartic interpolation (0.5214).

C. TI-PIMD: Simultaneous Changes in Mass and Potential. Using TI-PIMD, we have investigated simultaneous changes in mass and potential not only for the quantum harmonic oscillator, but also for transforming H₂ into HCl.

1. Quantum Harmonic Oscillator. Figure 5 shows the TI-PIMD results for virial and primitive estimators of the quantum free energy of the harmonic oscillator shown together with exact results. Linear (bottom) and quartic (top) interpolation functions in λ (eq 7) were considered. As for the previous cases, the latter exhibits better convergence with the exception of the last two windows. After TI, the change in free energy according to the linear interpolation yields 0.211 for the primitive and 0.215 for the virial. The nonlinear interpolation slightly outperforms the linear interpolation when compared to the exact value given by eq 18 ($\Delta F = 0.207$): The primitive estimator predicts 0.206, whereas the virial leads to 0.203. The typical error bar for the primitive estimator is 0.0088, which is nearly 4 times larger than the error bar for the virial estimator (0.0022), both for the linear path. Similarly, for the quartic path, the average error bar for the primitive estimator is 0.066, while the error for the virial estimator is 0.0097.

We note that both derivatives bear more resemblance with the example discussed in the previous section where only the potential was changed, and not the mass. Furthermore, the effect of changing the mass in addition to changing the potential clearly introduces further curvature to the free energy derivative profile of the quartic interpolation. This raises the question of whether the increase in curvature leads to finer resolution requirements for the thermodynamic integration, thereby removing the advantage gained from better convergence of the sampled estimator.

2. Transmutation of a Diatomic Molecule. In this section, the alchemical transformation of molecular hydrogen into hydrogen chloride at 300 K is investigated using TI-PIMD. We approximate the internuclear potential of the diatomic molecules by a Morse potential:

$$V(x) = D[1 - e^{-a(x - x_{\text{eq}})}]^2$$

where x_{eq} is the internuclear equilibrium distance, D is the dissociation energy measured from the potential minimum, and a is related to the curvature of the potential at the equilibrium position. Table 1 displays all of the Morse parameters employed for these calculations. The reduced mass and all of the parameters (a, D, x_{eq}) were linearly interpolated in λ between initial (H₂) and final (HCl) states. The free energy difference was obtained using normal mode TI-PIMD ($P = 64$) and 4 million steps ($dt = 0.1$ au). Since at 300 K H₂ and HCl are mostly ground-state-dominated, we can approximate the

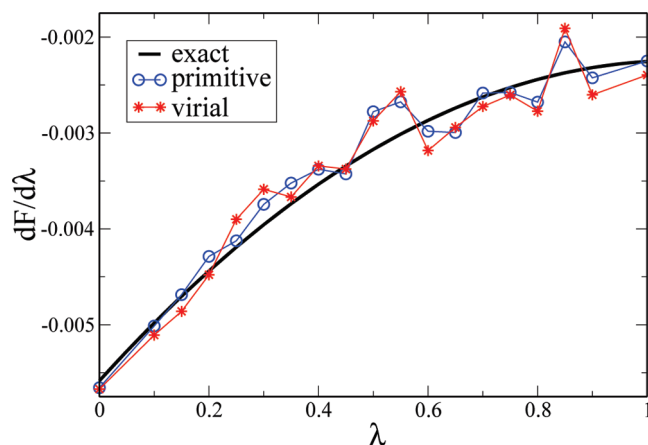


Figure 6. TI-PIMD: First-order free energy derivative with respect to simultaneous changes in Morse parameters and mass corresponding to the alchemical transformation of $\text{H}_2 \rightarrow \text{HCl}$ at $T = 300$ K. Black: Exact analytical estimate using a harmonic approximation to the Morse potential. Blue: Primitive estimator. Red: Virial estimator. Atomic units are used.

Morse potential harmonically to get an estimate of the change in free energy,

$$\begin{aligned} \Delta F &= F(\text{HCl}) - F(\text{H}_2) \approx k_B T \ln \left[\frac{\sinh(\beta \hbar \omega_{\text{HCl}}/2)}{\sinh(\beta \hbar \omega_{\text{H}_2}/2)} \right] \\ &= -0.003324 \text{ au} \end{aligned} \quad (30)$$

TI-PIMD results (eq 17) for the primitive and virial estimators of the free energy derivative with respect to transforming the Morse potential of H_2 into HCl are presented in Figure 6, together with the exact result for the harmonically approximated Morse potential. In contrast to all the previous examples, and despite the large number of PIMD steps taken, the remaining statistical error (on the order of 0.00091 au) is still significant. This behavior is exacerbated as approaching the end point ($\lambda = 1$) when the new physical parameters instantly appear where other interactions were previously present. Such numerical instabilities are commonly ameliorated in classical simulations by softening the underlying potential and making it λ -dependent. Nonetheless, the resulting integrated free energy differences, -0.003326 au and -0.003320 au, for the primitive and virial estimators, respectively, both compare very favorably to the analytical estimate in eq 30 (-0.003324 au).

The shape of Figure 6 is reminiscent of Figure 1 (bottom), suggesting that the linear transformation of H_2 into HCl is dominated by the change in mass, rather than by the change in potential.

D. Mass Transformation in the Harmonic Oscillator via PT-PIMD. In this section, we use PT-PIMD to recompute the free energy difference of the harmonic oscillator, which was already investigated with TI-PIMD in section III.A.1. Figure 7 illustrates the convergence of the running average of the free energy difference with the number of PT-PIMD steps. Two cases are considered: a “large” change in mass, $m_i = 1 \rightarrow m_f = 2$ (top), and a “small” one, $m_i = 1 \rightarrow m_f = 1.5$ (bottom). Unlike TI-PIMD, these results were obtained using a single, albeit substantially longer, PT-PIMD simulation. As one would expect, the smaller change in mass (bottom) converges more rapidly to the analytical

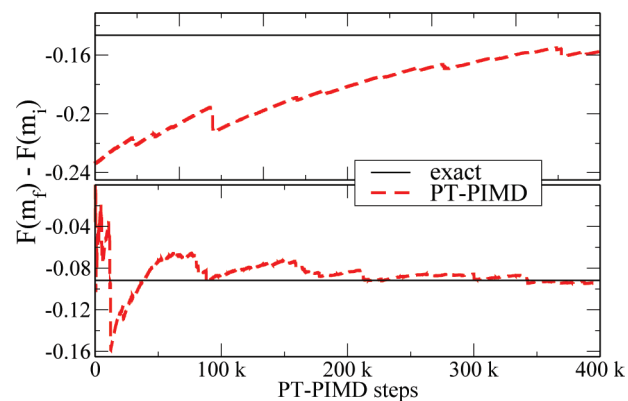


Figure 7. PT-PIMD: Running average of the free energy difference for an isotope transformation in the harmonic potential computed using perturbation theory, eq 19. Top: $m_i = 1 \rightarrow m_f = 2$. Bottom: $m_i = 1 \rightarrow m_f = 1.5$. Natural units are used. Red dashes: PT-PIMD using eq 19. Black solid: Exact result. The inverse temperature is $\beta = 10$, and the Trotter number is $P = 64$.

reference result than the larger (top) for which the number of steps (400 000) is clearly insufficient. The error bar for the latter (0.0424) is more than 5 times larger than the error for the “small” transformation (0.0075), indicating that the calculation suffers from slow convergence. Clearly, for the large transformation, the end point states are too disparate (there is little overlap) and PT methods do not converge. A remedy for such a situation would be to perform a series of PT calculations on mutually overlapping intermediate states between the end points.

The sawtooth pattern in Figure 7 is a common feature when averaging exponentials.⁶¹ The spikes indicate that the averages are dominated by rare events. As mentioned before, a possible remedy to dampen these sudden fluctuations in ΔF consists of further reducing the interval, i.e., to use a multistage PT-PIMD approach. Despite the abrupt initial transient, the running average for the small mass transformation converges nicely toward the exact value as the simulation goes on, Figure 7 (bottom).

E. PT-PIMD: Effect of a Uniform Static Electric Field. Perturbation theory can also be used to compute electric polarizabilities, which are very important in understanding not only weakly bonded intermolecular and long-range interactions⁶² but also the ferro/paraelectric transition in some hydrogen-bonded materials. In this context, Sebastiani and Srinivasan recently investigated the phase transition in potassium dihydrogen phosphate (KDP), which is known to exhibit strong nuclear quantum effects.⁶³ In this subsection, we suggest an estimator that could be used to further characterize these phase transitions. The change in free energy of a single particle of charge q due to a uniform static electric field of strength $E = |\mathbf{E}|$ is

$$\Delta F = F(E) - F(0) = -\frac{1}{\beta} \ln \langle \exp[-\beta(q\mathbf{E} \cdot \bar{\mathbf{r}})] \rangle_{E=0} \quad (31)$$

where $\bar{\mathbf{r}} = \sum_{s=1}^P \mathbf{r}_s / P$ is the centroid of the ring polymer. For bulk systems, eq 31 must be modified to account for periodicity.

Figure 8 shows the numerical results for the running average of the free energy difference in the harmonic oscillator with and without an homogeneous static electric field of varying strength. When the magnitude of the electric field is not too large (bottom), the converged estimate of the corresponding change

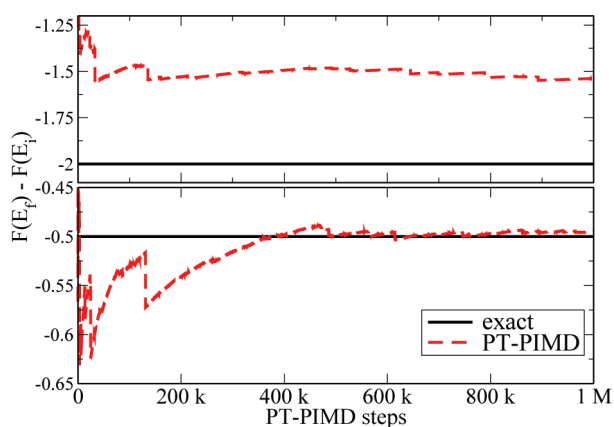


Figure 8. PT-PIMD: Running averages of the free energy difference of the harmonic potential $m\omega x^2/2$ due changes in the static external electric field. Top: $E_i = 0 \rightarrow E_f = 2$. Bottom: $E_i = 0 \rightarrow E_f = 1$ (natural units are used). Red dashes: PT-PIMD using eq 31. Black solid: Exact. The inverse temperature is $\beta = 10$, and the Trotter number is $P = 64$.

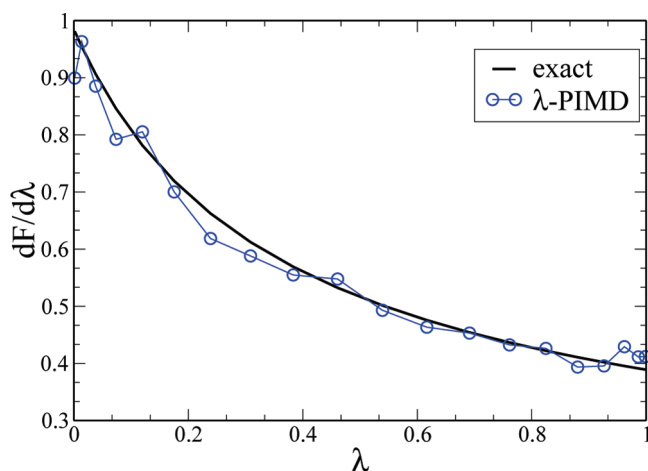


Figure 9. λ -PIMD: First order free energy derivative versus λ for harmonic potential $\kappa(\lambda)x^2/2$ at $T = 0.5$. Black solid: Exact, eq 16. Blue solid: Numerical using λ -PIMD, i.e. $\langle \Delta V \rangle_\lambda$. One million PIMD steps were accumulated with $P = 64$ and $dt = 0.01$ at the inverse temperature $\beta = 10$ (natural units are used).

in free energy (-0.4947) is in very good agreement with the exact value (-0.5), as given by $\Delta F = -((q^2 E^2)/(2m\omega^2))$, where $q = 1$ is the charge and $m = \omega = 1$. However, when the applied electric field is twice as large (top), the estimator eq 31 converges to the wrong value (-1.5411), which deviates significantly from the exact value of -1.1250 . We note that in both cases the error bar is similar (0.042 for large and 0.031 for small transformation), so it is not a reliable indicator of the accuracy of the final result. This example, as with the mass transformation in the previous section, highlights the limitations of the PT-PIMD approach.

F. Changes in Potential of the Quantum Harmonic Oscillator Revisited: λ -PIMD. In this final section, we revisit changes of the force constant in the quantum harmonic oscillator in analogy to section III.B. Instead of TI-PIMD, however, we used λ -PIMD. Specifically, we numerically tested eq 22 for the same linear change in force constant as in section III.B, i.e., $V(x) = \kappa(\lambda)x^2/2$, with $\kappa(\lambda) = \lambda\kappa_f + (1 - \lambda)\kappa_i$ and $\kappa_i = 1$ and $\kappa_f = 4$.

The corresponding free energy change for this transformation is obtained from numerical integration of the λ -derivative over λ ,

$$\Delta F = F_f - F_i = \int_0^1 d\lambda \langle \Delta V \rangle_\lambda = \int_0^{\pi/2} d\theta \langle \Delta V \rangle_\theta \sin(2\theta) \quad (32)$$

where $\Delta V = (1/P)(\kappa_f - \kappa_i)\sum_{s=1}^P x_s^2/2$ for the harmonic oscillator. Numerical results for the first-order derivative of the free energy with respect to λ at $T = 0.5$ are shown in Figure 9. The integrated predicted free energy difference amounts to 0.555 (average error bar 0.009) after 1 million PIMD steps ($P = 64$ beads). This value is in good agreement with the exact value for the harmonic oscillator (0.5635).

IV. CONCLUSIONS

New path integral estimators have been formulated for the calculation of free energy changes due to transformations in the mass and/or potential and changes in the external electric field. Perturbation theory and λ -dynamics are found to be useful alternatives to thermodynamic integration for the prediction of the free energy differences of alchemical transformations from a single simulation. All free energy methods were implemented within path integral molecular dynamics (TI-PIMD, PT-PIMD, λ -PIMD) and applied to various solvable model systems, including the single and double harmonic oscillator and the Morse potential. The methods introduced are rather general and valid for any force-field-based PIMD calculation of an alchemical change. The estimators were also applied to investigate the isotope transformation in the Zundel cation using *ab initio* path integral molecular dynamics.

Future work will deal with the extension of these ideas to *ab initio* PIMD for the simulation of more realistic systems. For example, it will be interesting to investigate a complete change of identity of a molecule from first principles. While this is feasible in force-field-based PIMD calculations, it is more challenging in *ab initio* PIMD due to the use of pseudopotentials, or atomic basis functions. As an example, we would first investigate the iso-electronic transformation of the Zundel cation to the protonated water–ammonia complex cation. Finally, it will also be interesting to devise a way to predict changes in the quantum free energy barrier and not just on traced quantities such as the overall free energy.

APPENDIX A. ISOTOPE EFFECT ON INTERNAL ENERGY

The formalism presented so far can be easily extended to compute changes of any thermodynamic quantity due to alchemical changes. In this section, we derive an estimator for the change of internal energy corresponding to the isotope effect.

The path integral expression for the internal energy according to the *primitive* form is

$$\langle E \rangle = \lim_{P \rightarrow \infty} \frac{C}{Z} \int d\mathbf{r}_1 \dots d\mathbf{r}_P \varepsilon_{\text{prim}}(\lambda) e^{-\beta\Phi(\{\mathbf{r}\};\lambda)} \quad (A1)$$

where $C(\lambda) = ((m(\lambda)P)/(2\pi\beta\hbar^2))^{3P/2}$ is a normalization constant,

$$\Phi(\{\mathbf{r}\};\lambda) = \sum_{s=1}^P \left[\frac{m(\lambda)}{2} \omega_p^2 (\mathbf{r}_s - \mathbf{r}_{s+1})^2 + \frac{V(\mathbf{r}_s)}{P} \right]$$

is the effective path integral potential and

$$\varepsilon_{\text{prim}}(\lambda) = \frac{3P}{2\beta} + \sum_{s=1}^P \left[\frac{V(\mathbf{r}_s)}{P} - \frac{m(\lambda)}{2} \omega_p^2 (\mathbf{r}_s - \mathbf{r}_{s+1})^2 \right]$$

is the so-called primitive energy estimator.

Taking the derivative of eq A1 with respect to the parameter λ gives

$$\frac{dE(\lambda)}{d\lambda} = \langle \partial_\lambda \varepsilon_{\text{prim}} \rangle + \beta [\langle \varepsilon_{\text{prim}} \rangle \langle \partial_\lambda \Phi \rangle - \langle (\partial_\lambda \Phi) \varepsilon_{\text{prim}} \rangle] \quad (\text{A2})$$

where we note that $\partial_\lambda \varepsilon_{\text{prim}} = -\partial_\lambda \Phi$. Rearranging into an expression more amenable for numerical computation gives

$$\frac{dE(\lambda)}{d\lambda} = \left(\frac{\partial_\lambda m}{m} \right) \{ \langle \mathcal{J} \rangle + \beta [\langle \mathcal{J} \varepsilon_{\text{prim}} \rangle - \langle \varepsilon_{\text{prim}} \rangle \langle \mathcal{J} \rangle] \} \quad (\text{A3})$$

where $\mathcal{J} = -(m/2) \omega_p^2 \sum_{s=1}^P (\mathbf{r}_s - \mathbf{r}_{s+1})^2$.

As a test, we can compare $E(m_f) - E(m_i) = \int_0^1 d\lambda (dE^{\text{prim}}/d\lambda)_\lambda$ to the exact expression for a model system. The analytical expression for the internal energy of the harmonic oscillator is $U(\beta) = (\hbar\omega/2) \coth(\beta\hbar\omega/2)$. Thus, the analytic change of internal energy is

$$\Delta E(\beta) = \frac{\hbar}{2} [\omega_f \coth(\beta\hbar\omega_f/2) - \omega_i \coth(\beta\hbar\omega_i/2)] \quad (\text{A4})$$

where the oscillator frequencies are $\omega_{i(f)} = (\kappa/m_{i(f)})^{1/2}$. The exact analytical expression for the first derivative of the total energy with respect to λ is for the harmonic oscillator is

$$\frac{dE(\lambda)}{d\lambda} = \frac{\hbar\omega\partial_\lambda m}{2m} \left[\frac{(\beta\hbar\omega/2)}{\sinh^2(\beta\hbar\omega/2)} - \frac{\coth(\beta\hbar\omega/2)}{2} \right] \quad (\text{A5})$$

Finally, the derivative of the entropy with respect to λ is simply $\partial_\lambda S = (\partial_\lambda (U - F))/T$. Thus, the isotope effect on the entropy can also be estimated.

APPENDIX B. SECOND-ORDER DERIVATIVES

In 1994, Smith and van Gunsteren published a paper on predictions of classical free energy differences using only the information of free energy derivatives computed at the initial state.⁶⁴ On the basis of that work, we have investigated extrapolations of the free energy from the initial state within PIMD.

From Figure 1, it is clear that the change in free energy is not linear. Using only the first-order derivative information at the initial state ($\lambda = 0$) leads to inaccurate predictions of the overall free energy difference. Thus, it is necessary to include second- and higher-order derivatives. Expanding the free energy difference in a Taylor series to second order,

$$\Delta F_{\text{if}} \approx (\lambda_f - \lambda_i) \left(\frac{dF}{d\lambda} \right)_{\lambda_i} + \frac{1}{2!} (\lambda_f - \lambda_i)^2 \left(\frac{d^2F}{d\lambda^2} \right)_{\lambda_i} \quad (\text{B1})$$

we could in principle estimate the difference in free energies using only the information from the initial state. To accomplish this, higher derivatives of the free energy with respect to parameter λ are needed. The primitive estimator for the second

derivative of the free energy with respect to mass reads

$$\begin{aligned} \frac{d^2F}{d\lambda^2} = & \left\langle (\partial_\lambda m + \partial_\lambda^2 m) \frac{\omega_p^2}{2} \mathcal{J} - \frac{3P}{2\beta} \left(\frac{\partial_\lambda m}{m} + \frac{\partial_\lambda^2 m}{m} + \frac{(\partial_\lambda m)^2}{m^2} \right) \right\rangle \\ & + \frac{1}{\beta} \left\langle \frac{P\partial_\lambda m}{2m} - \beta \frac{\partial_\lambda^2 m}{2} \omega_p^2 \mathcal{J} \right\rangle \quad (\text{B2}) \end{aligned}$$

where the sum \mathcal{J} denotes $\sum_{s=1}^P (\mathbf{r}_s - \mathbf{r}_{s+1})^2$ with $\mathbf{r}_{P+1} = \mathbf{r}_1$. Note that the last term in this expression vanishes if we assume a linear interpolation path for the mass ($\partial_\lambda^2 m = 0$). This expression is problematic because it involves a significant cancellation of terms. The numerical results can be checked with the analytical expression for the second-order derivative of the quantum harmonic oscillator with respect to changes in mass

$$\begin{aligned} \left(\frac{d^2F}{d\lambda^2} \right) = & - \left(\frac{1}{2} \right) \left(\frac{\partial_\lambda m}{m} \right)^2 \frac{\hbar\omega}{2} \\ & \times \left[\frac{\beta\hbar\omega/2}{2 \sinh^2(\beta\hbar\omega/2)} - \frac{3}{2 \tanh(\beta\hbar\omega/2)} \right] \\ & - \left(\frac{\hbar\omega}{4} \right) \left(\frac{\partial_\lambda^2 m}{m} \right) \coth(\beta\hbar\omega/2) \quad (\text{B3}) \end{aligned}$$

Our numerical tests (not shown) suggest, however, that convergence of higher order derivatives is difficult to achieve.

Fortunately, an empirical way to circumvent this nonlinearity issue was recently presented in the context of potential energy differences by one of us.⁶⁵ In that work, higher-order terms in the Taylor expansion were absorbed into a correction factor accompanying the first derivative

$$\begin{aligned} \Delta F_{\text{if}} = & (\lambda_f - \lambda_i) \left(\frac{dF}{d\lambda} \right)_{\lambda_i} + \frac{1}{2!} (\lambda_f - \lambda_i)^2 \left(\frac{d^2F}{d\lambda^2} \right)_{\lambda_i} + \dots \\ \approx & (\lambda_f - \lambda_i) \mathcal{G}(\lambda_i) \left(\frac{dF}{d\lambda} \right)_{\lambda_i} \quad (\text{B4}) \end{aligned}$$

where \mathcal{G} is a renormalization factor that accounts for the nonlinear terms. The hope is to calculate \mathcal{G} numerically for a reference system (say \mathcal{G}^{ref}) and then use this coefficient (without modification) to estimate the isotope free energy change in other systems. This approach is admittedly somewhat *ad hoc* because it assumes transferability of \mathcal{G} to other systems. In other words, the effect of higher-order terms is similar in the reference system and in the system of interest. Nonetheless, there might be some justification to this method if the leading term in the Taylor expansion is the first derivative (and the higher terms are small and cancel among each other). Work is underway to find numerically stable *virial* expressions for the second-order derivatives.

APPENDIX C. CONVERGENCE WITH THE NUMBER OF STEPS

We have carried out a convergence test on the free energy change versus the number of steps for the isotope transformation in the harmonic oscillator. Figure 10 shows that the convergence is achieved well within the first 400 000 steps for this toy model.

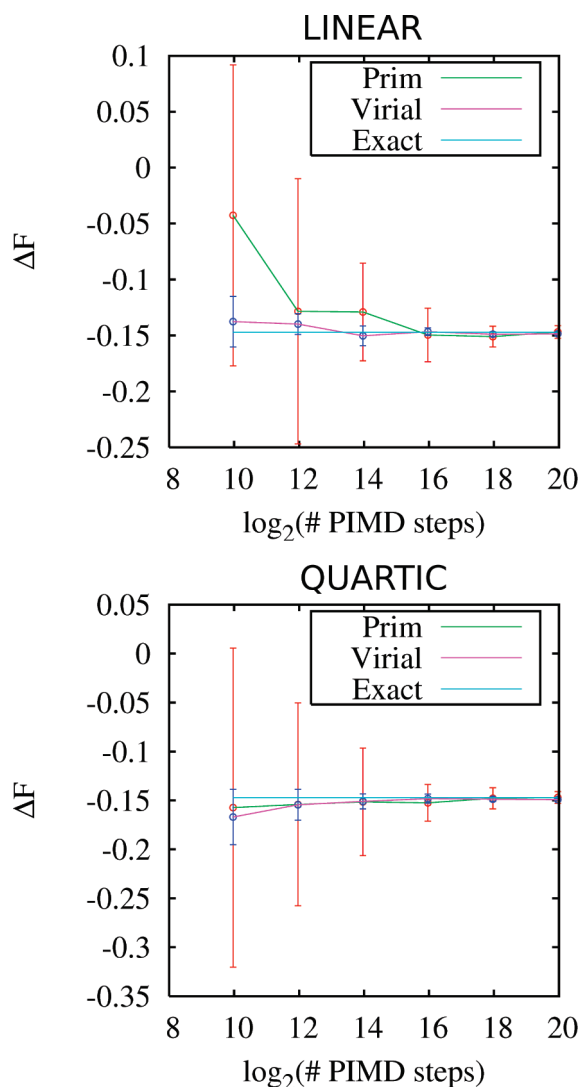


Figure 10. Convergence of the free energy difference (in natural units) with the number of PIMD steps for the mass transformation in the harmonic oscillator at $\beta\hbar\omega = 10$. Top: linear path. Bottom: quartic path.

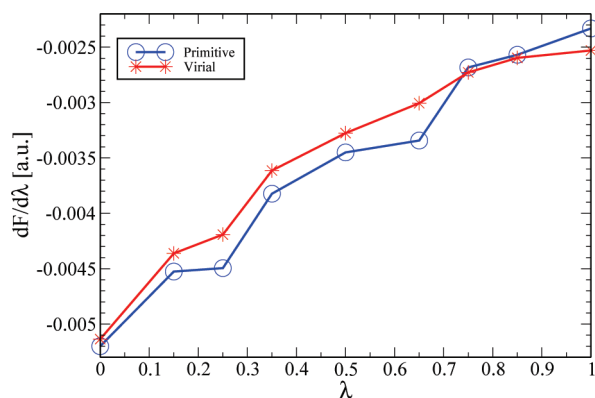


Figure 11. Numerical results (in au) from the *ab initio* TI-PIMD simulation ($P = 8$) of the hydrogen molecule at 300 K.

The figure also confirms that the quartic path leads to a faster convergence than the linear path.

APPENDIX D. ISOTOPE TRANSFORMATION ON THE HYDROGEN MOLECULE: *AB INITIO* APPROACH

To test our numerical implementation in CPMD, we simulated the $\text{H}_2 \rightarrow \text{D}_2$ isotope transformation. The hydrogen molecule was placed in a cubic box of volume 9^3 \AA^3 . The kinetic energy cutoff for the plane wave expansion of electronic orbitals was 75 Ry. Eight beads were used for the path integral discretization. The rest of the parameters were identical to the ones employed for the Zundel ion. The results of this test are shown in Figure 11. The predicted change in free energy is $\Delta F = -0.00358 \text{ au}$ (-2.25 kcal/mol) for the primitive and $\Delta F = -0.00345 \text{ au}$ (-2.17 kcal/mol) for the virial estimator. These values are in good agreement with the values from other quantum chemistry methods ($\approx -2 \text{ kcal/mol}$). The average standard deviation was 1.72 kcal/mol for the primitive and 0.56 kcal/mol for the virial estimator. The perturbation theory approach of eq 19 yields a value of -0.00375 au (-2.36 kcal/mol), which agrees very well with TI-PIMD results.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: ap1484@nyu.edu (A.P.), anatole@alcf.anl.gov (O.A.v.L.).
Website: <http://www.alcf.anl.gov/~anatole> (O.A.v.L.).

ACKNOWLEDGMENT

We gratefully acknowledge stimulating discussions with Mark E. Tuckerman. We thank the hospitality of UCLA/IPAM where this work was completed and the City University of New York (CUNY) High Performance Computing Center (HPCC) at College of Staten Island for computational resources. A.P. is grateful for support from the Horizon fellowship at New York University, the SNL's LDRD project No. 117866, and the SNL's summer student internship program at the Computer Science Research Institute. O.A.v.L. acknowledges SNL's LDRD Truman project No. 120209. SNL is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. This research also used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357.

REFERENCES

- (1) Marx, D.; Parrinello, M. *J. Chem. Phys.* **1996**, *104*, 4077–4082.
- (2) Tuckerman, M. E.; Marx, D.; Klein, M. L.; Parrinello, M. *J. Chem. Phys.* **1996**, *104*, 5579–5588.
- (3) Feynman, R. P. *Rev. Mod. Phys.* **1948**, *20*, 367–387.
- (4) Feynman, R. P.; Hibbs, A. R. *Quantum Mechanics and Path Integrals*; McGraw-Hill: New-York, 1965.
- (5) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
- (6) Iftimie, R.; Minary, P.; Tuckerman, M. E. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6654–6659.
- (7) Tuckerman, M. E.; Marx, D. *Phys. Rev. Lett.* **2001**, *86*, 4946–4949.
- (8) Miura, S.; Tuckerman, M. E.; Klein, M. L. *J. Chem. Phys.* **1998**, *109*, 5290–5299.
- (9) Morrone, J. A.; Car, R. *Phys. Rev. Lett.* **2008**, *101*, 017801.
- (10) Tuckerman, M. E.; Marx, D.; Klein, M. L.; Parrinello, M. *Science* **1997**, *275*, 817–820.

- (11) Marx, D.; Tuckerman, M. E.; Hutter, J.; Parrinello, M. *Nature* **1999**, *397*, 601–604.
- (12) Tuckerman, M. E.; Marx, D.; Parrinello, M. *Nature* **2002**, *417*, 925–929.
- (13) Li, X.-Z.; Probert, M. I. J.; Alavi, A.; Michaelides, A. *Phys. Rev. Lett.* **2010**, *104*, 066102.
- (14) Davidson, E. R. M.; Alavi, A.; Michaelides, A. *Phys. Rev. B* **2010**, *81*, 153410.
- (15) Pu, J.; Gao, J.; Truhlar, D. G. *Chem. Rev.* **2006**, *106*, 3140–3169.
- (16) Hwang, J. K.; Warshel, A. *J. Am. Chem. Soc.* **1996**, *118*, 11745.
- (17) Pérez, A.; Tuckerman, M. E.; Hjalmarsen, H. P.; von Lilienfeld, O. A. *J. Am. Chem. Soc.* **2010**, *132*, 11510–11515.
- (18) von Lilienfeld, O. A.; Lins, R.; Rothlisberger, U. *Phys. Rev. Lett.* **2005**, *95*, 153002.
- (19) Wang, M.; Hu, X.; Beratan, D. N.; Yang, W. *J. Am. Chem. Soc.* **2006**, *128*, 3228.
- (20) Marcon, V.; von Lilienfeld, O. A.; Andrienko, D. *J. Chem. Phys.* **2007**, *127*, 064305.
- (21) Rinderspacher, B. C.; Andzelm, J.; Rawlett, A.; Dougherty, J.; Beratan, D. N.; Yang, W. *J. Chem. Theory Comput.* **2009**, *5*, 3321.
- (22) Froemming, N. S.; Henkelman, G. *J. Chem. Phys.* **2009**, *131*, 234103.
- (23) Sheppard, D.; Henkelman, G.; von Lilienfeld, O. A. *J. Chem. Phys.* **2010**, *133*, 084104.
- (24) van Gunsteren, W. F.; Daura, X.; Mark, A. E. *Helv. Chim. Acta* **2002**, *85*, 3113.
- (25) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300–313.
- (26) Oostenbrink, C. *J. Comput. Chem.* **2009**, *30*, 212.
- (27) von Lilienfeld, O. A.; Tuckerman, M. E. *J. Chem. Phys.* **2006**, *125*, 154104.
- (28) von Lilienfeld, O. A.; Tuckerman, M. E. *J. Chem. Theory Comput.* **2007**, *3*, 1083.
- (29) Leung, K.; Rempe, S. B.; von Lilienfeld, O. A. *J. Chem. Phys.* **2009**, *130*, 204507.
- (30) Alfé, D.; Gillan, M. J.; Price, G. D. *Nature* **2000**, *405*, 172.
- (31) Zimmermann, T.; Vanicek, J. *J. Chem. Phys.* **2009**, *131*, 024111.
- (32) Ramírez, R.; Herrero, C. P. *J. Chem. Phys.* **2010**, *133*, 144511.
- (33) de la Peña, L. H.; Peshherbe, G. H. *J. Phys. Chem. B* **2010**, *114* (16), 5404–5411.
- (34) Tuckerman, M. E. *Statistical mechanics: Theory and molecular simulation*; Oxford University Press: Oxford, U. K., 2010.
- (35) Chandler, D.; Wolynes, P. G. *J. Chem. Phys.* **1981**, *74*, 4078–4095.
- (36) Zamalin, V. M.; Norman, G. E. *USSR Comp. Math. Math. Phys.* **1973**, *13* (2), 408–420.
- (37) Tuckerman, M.; Berne, B.; Martyna, G.; Klein, M. *J. Chem. Phys.* **1993**, *99*, 2796–2808.
- (38) *Free Energy Calculations: Theory and Applications in Chemistry and Biology*; Chipot, C., Pohorille, A., Eds.; Springer: New York, 2007; Springer Series in Chemical Physics, Vol. 86.
- (39) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*; Academic Press: New York, 2002.
- (40) Herman, M. F.; Bruskin, E. J.; Berne, B. *J. Chem. Phys.* **1982**, *76*, 5150.
- (41) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.
- (42) Liu, Z.; Berne, B. *J. Chem. Phys.* **1993**, *99*, 6071–6077.
- (43) Kong, X.; Brooks, C. L., III. *J. Chem. Phys.* **1996**, *105*, 2414–2423.
- (44) Guo, Z.; Brooks, C. L.; Kong, X. *J. Phys. Chem. B* **1998**, *102*, 2032–2036.
- (45) Bitetti-Putzer, R.; Yang, W.; Karplus, M. *Chem. Phys. Lett.* **2003**, *377*, 633–641.
- (46) Abrams, J. B.; Rosso, L.; Tuckerman, M. E. *J. Chem. Phys.* **2006**, *125*, 074115.
- (47) Schmitt, U. W.; Voth, G. A. *Chem. Phys. Lett.* **2000**, *329*, 36–41.
- (48) CPMD, version 3.13.2; IBM Corp: Armonk, New York, 1990–2008; MPI für Festkörperforschung Stuttgart: Stuttgart, Germany, 1997–2001. <http://www.cpmid.org/> (accessed June 2011).
- (49) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (50) Colle, R.; Salvetti, D. *Theor. Chim. Acta* **1988**, *37*, 329–334.
- (51) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (52) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
- (53) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004.
- (54) Lin, I.-C.; Coutinho-Neto, M. D.; Felsenheimer, C.; von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U. *Phys. Rev. B* **2007**, *75*, 205131.
- (55) Martyna, G. J.; Tuckerman, M. E. *J. Chem. Phys.* **1999**, *110*, 2810.
- (56) Martyna, G. J.; Klein, M. L.; Tuckerman, M. E. *J. Chem. Phys.* **1992**, *97*, 2635–2643.
- (57) *Quantum Simulations of Complex Many-Body Systems: from theory to algorithms*; Grotendorst, J., Marx, D., Muramatsu, A., Eds.; John von Neumann Institute for Computing: Princeton, NJ, 2002.
- (58) Granovsky, A. A. Firefly, version 7.1.G. <http://classic.chem.msu.su/gran/firefly/index.html> (accessed June 2011).
- (59) Kohlmeyer, A. VMD scripts for visualization of CPMD data. <http://klein-group.icms.temple.edu/cpmd-vmd/files.html> (accessed June 2011).
- (60) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (61) Pohorille, A.; Jarzynski, C.; Chipot, C. *J. Phys. Chem. B* **2010**, *114*, 10235–10253.
- (62) French, R. H.; Parsegian, V. A.; Podgornik, R.; Rajter, R. F.; et al. *Rev. Mod. Phys.* **2010**, *82*, 1887.
- (63) Srinivasan, V.; Sebastiani, D. *J. Phys. Chem. C* **2011**, *115*, 12631.
- (64) Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1994**, *100*, 577–585.
- (65) von Lilienfeld, O. A. *J. Chem. Phys.* **2009**, *131*, 164102.

Are Peptides Good Two-State Folders?

Alexander M. Berezhkovskii,[†] Florentina Tofoleanu,^{‡,§} and Nicolae-Viorel Buchete^{*,†,§}

[†]Mathematical and Statistical Computing Laboratory, Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892, United States

[‡]School of Physics and [§]Complex and Adaptive Systems Laboratory, University College Dublin, Belfield, Dublin 4, Ireland

ABSTRACT: The folding kinetics of proteins is frequently single-exponential, as basins of folded and unfolded conformations are well separated by a high barrier. However, for relatively short peptides, a two-state character of folding is rather the exception than the rule. In this work, we use a Zwanzig-type model of protein conformational dynamics to study the dependence of folding kinetics on the protein chain length, M . The analysis is focused on the gap in the eigenvalue spectrum of the rate matrix that describes the protein's conformational dynamics. When there is a large gap between the two smallest in magnitude nonzero eigenvalues, the corresponding relaxation times have qualitatively different physical interpretations. The longest of these two times characterizes the interbasin equilibration (i.e., folding), whereas the second time characterizes the intrabasin relaxation. We derive approximate analytical solutions for the two eigenvalues that show how they depend on M . From these solutions, we infer that there is a large gap between the two, and thus, the kinetics is essentially single-exponential when M is large enough such that 2^{M+1} is much larger than M^2 .

INTRODUCTION

The kinetics of many long proteins is single-exponential. This implies that folding of such proteins is a two-state process that involves transitions between the folded and unfolded states of the protein, which are separated by a high barrier. However, this is not necessarily the case for short peptides. The goal of the present work is to study how the character of the folding kinetics (i.e., whether the kinetics is single-exponential or not) depends on the protein chain length. We analyzed this question within the framework of a simple model of the protein conformational dynamics that is similar to the model suggested by Zwanzig, Szabo, and Bagchi to study Levinthal's paradox.^{1–4} Being simple, the model allows for an analytical solution, which is used to establish the relation between the character of the folding kinetics and the protein chain length. Our results are of particular interest to those researchers who study the folding of short peptides, either experimentally or computationally, with atomic-level detail or using simplified models,^{5–24} as their results are frequently interpreted within the framework of simple two-state models. Many theoretical and experimental studies use small homoproteins such as dialanine (Ala2) or pentaalanine (Ala5, the smallest peptide that forms one full helical loop) as test systems.^{25–37}

Our interest in this problem was initiated by a molecular dynamics (MD) study of the folding kinetics of pentaalanine in explicit solvent,^{35,36} in which it was shown that the folding kinetics is better approximated by a four-state rather than by a two-state model. Figure 1 illustrates some of the most populated unfolded conformational states of Ala5, together with the all-helical folded state. As shown herein, a possible cause for the deviation from a single-exponential relaxation might be the relatively small number of unfolded conformational states available to small peptides. Note that a complex, non-two-state, length-dependent folding kinetics has been reported in experimental studies.^{6–8,12,38–41}

To analyze the character of the folding kinetics, we derived approximate expressions for the first two nonzero eigenvalues of

the rate matrix that describes the conformational dynamics of our model protein. These eigenvalues are important because a large gap between them is a fingerprint of the two-state character of the kinetics. With analytical expressions for the eigenvalues in hand, we analyzed how the gap depends on the chain length of the protein. We checked the accuracy of our analytical results by comparison with the exact eigenvalues of the rate matrix found numerically. The comparison showed that the two agree well if the peptide is not too short. For our model, we found that the kinetics is essentially single-exponential when the chain length M of the protein satisfies $2^{M+1} \gg M^2$. To be more precise, our model protein must contain nine or more residues to be a good two-state folder.

MODEL AND METHODS

Consider two-state protein folding described by the kinetic scheme



where U and F denote the unfolded (denatured) and folded (native) states of the protein, respectively, and k_F and k_U are the corresponding folding and unfolding rate constants. To understand the mechanism of protein folding, one has to establish a relation between eq 1 and the underlying protein dynamics.

In our model, the protein is a homopolymer that contains M identical residues, each of which can be in either a folded (f) or an unfolded (u) state. To characterize the state of the residue, we introduce an "indicator", s , and assign $s = 0$ and 1 to the residue in states f and u, respectively. A protein conformation is completely characterized by a set $\{s\} = (s_1, s_2, \dots, s_M)$, where s_i indicates the state of residue i , $i = 1, 2, \dots, M$. The number of residues in state u

Received: April 22, 2011

Published: June 30, 2011

in a protein conformation $\{s\}$, $m(\{s\})$, is given by

$$m(\{s\}) = \sum_{i=1}^M s_i \quad (2)$$

Thus, $0 \leq m \leq M$. The conformation with no unfolded residues (i.e., $s_i = 0$ for all i , and hence, $m = 0$) represents the folded state of the protein. In all other conformations, $m > 0$, and the protein contains m residues in the u state and $M - m$ residues in the f state. The number of conformations with a fixed value m of unfolded residues is given by the binomial coefficient

$$\binom{M}{m} = \frac{M!}{m!(M-m)!}$$

and the total number of the conformations of our protein is

$$\sum_{m=0}^M \binom{M}{m} = 2^M$$

Let $P_m(t)$ be the probability of finding the protein in one of the conformations with m unfolded residues at time t , $\sum_{m=0}^M P_m(t) = 1$. We assume that the evolution of $P_m(t)$ is due to

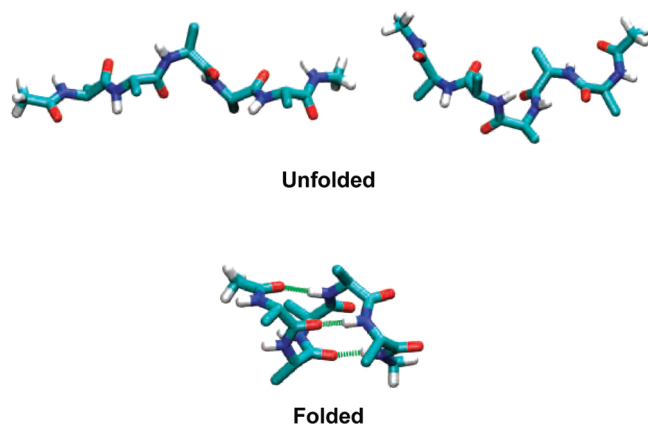


Figure 1. Folded and unfolded conformations of pentaalanine.

independent transitions between f and u states of individual residues. We take these transitions to be described by the kinetic scheme



where k is the rate constant, which is assumed to be the same for transitions in both directions. To introduce a collective behavior, responsible for the folding of our protein, we assume that the unfolding of the first and second residues occurs with the modified rate constants $\varepsilon_0 k$ and $\varepsilon_1 k$, respectively, with $\varepsilon_0, \varepsilon_1 \ll 1$. The precise choices of ε_0 and ε_1 are specified later. Figure 2A illustrates the connectivity network of the conformational dynamics of our peptide of length $M = 5$.

The evolution of the distribution function $P_m(t)$ is described by the master equation

$$\begin{aligned} \dot{P}_0(t) &= -\varepsilon_0 M P_0(t) + P_1(t) \\ \dot{P}_1(t) &= \varepsilon_0 M P_0(t) - [1 + \varepsilon_1(M-1)]P_1(t) + 2P_2(t) \\ \dot{P}_2(t) &= \varepsilon_1(M-1)P_1(t) - M P_2(t) + 3P_3(t) \\ &\vdots \\ \dot{P}_m(t) &= (M-m+1)P_{m-1}(t) - M P_m(t) \\ &\quad + (1 - \delta_{mM})(m+1)P_{m+1}(t), \text{ for } 3 \leq m \leq M \end{aligned} \quad (4)$$

where the time t is expressed in units of $1/k$ and δ_{mn} is the Kronecker delta. This set of equations describes the relaxation of $P_m(t)$ to the equilibrium distribution P_m^{eq} given by

$$P_m^{eq} = \frac{[\varepsilon_0 + (1 - \varepsilon_0)\delta_{m0}][\varepsilon_1 + (1 - \varepsilon_1)\delta_{m1}]}{[1 - (1 - \varepsilon_1)\delta_{m0}][1 + \varepsilon_0(1 - \varepsilon_1)M + \varepsilon_0\varepsilon_1(2^M - 1)]} \binom{M}{m} \quad (5)$$

One can check that P_m^{eq} satisfies eq 4 and the normalization condition, $\sum_{m=0}^M P_m^{eq} = 1$.

To determine the parameters ε_0 and ε_1 , we impose two requirements: (i) The probability of finding the protein in the native state ($m = 0$) is 0.5 (i.e., the calculations are done at the protein melting point), and (ii) the protein in conformations with one unfolded residue ($m = 1$) makes forward (i.e., $1 \rightarrow 2$) and backward (i.e., $1 \rightarrow 0$)

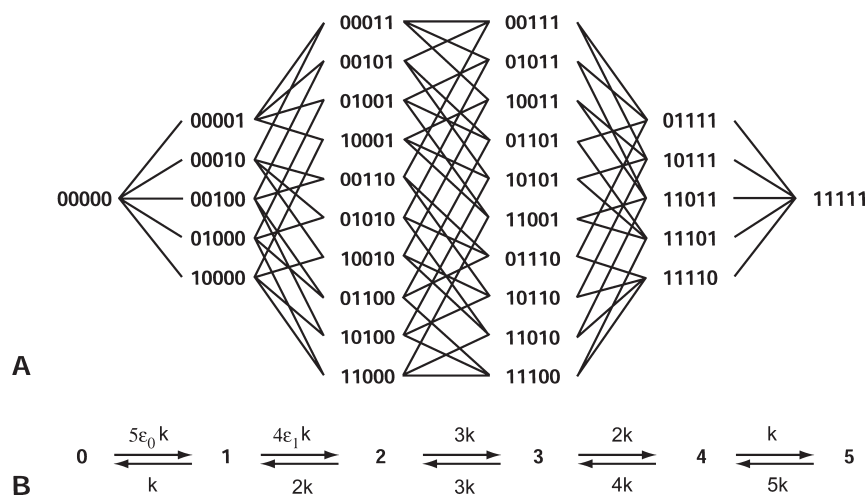


Figure 2. Schematic representation of conformational dynamics of our model peptide for length $M = 5$. (A) Connectivity network showing that only transitions between nearest neighbors are permitted. Conformations of the peptide are described using a binary notation for each residue, with 0 and 1 denoting folded and unfolded states of the residue. (B) Effective conformational dynamics along the discrete one-dimensional reaction coordinate, m .

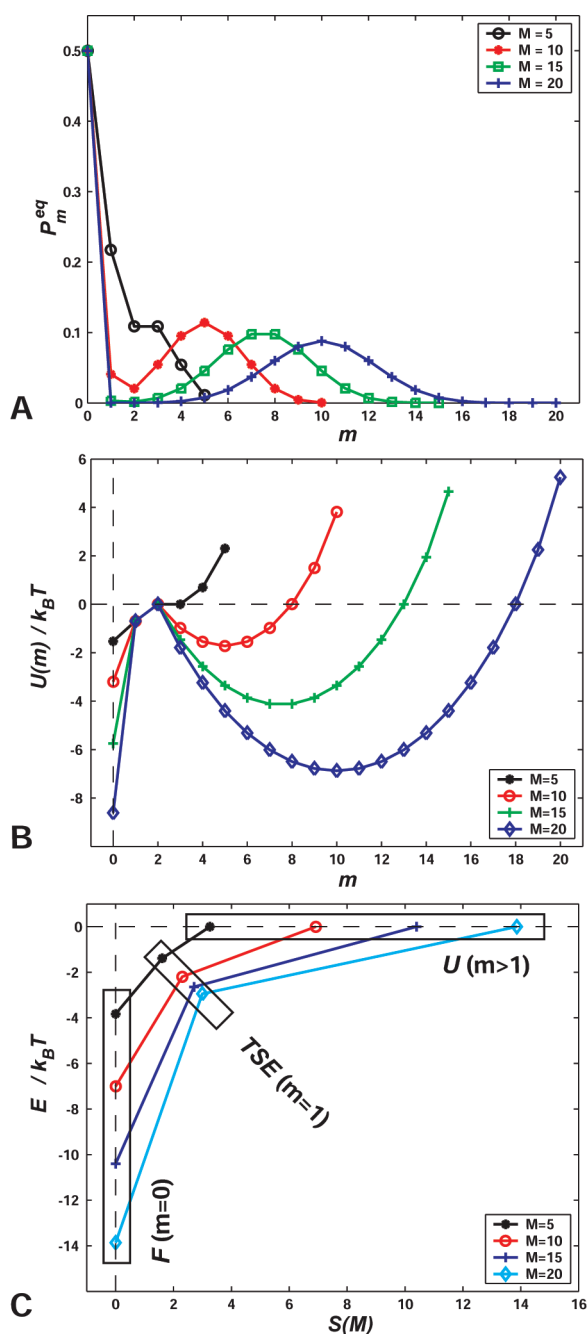


Figure 3. (A) Equilibrium distribution P_m^{eq} and (B) potential of mean force $U(m)$ as functions of the reaction coordinate m . (C) Folding funnel.

transitions with equal probability. In other words, conformations with $m = 1$ can be considered as transition states between the folded and unfolded conformations of the protein. These requirements lead to the following values of the parameters

$$\varepsilon_0 = \frac{M-1}{2^M + M^2 - 2M - 1}, \quad \varepsilon_1 = \frac{1}{M-1} \quad (6)$$

As a consequence, (i) the equilibrium distribution in eq 5 simplifies and takes the form

$$P_m^{eq} = \frac{[\varepsilon_0 + (1-\varepsilon_0)\delta_{m0}][\varepsilon_1 + (1-\varepsilon_1)\delta_{m1}]}{2[1 - (1-\varepsilon_1)\delta_{m0}]} \binom{M}{m} \quad (7)$$

and (ii) the product $\varepsilon_1(M-1)$ in eq 4 becomes unity.

Figure 3 illustrates some properties of the proteins of different chain lengths M . Figure 3A shows the equilibrium distribution P_m^{eq} . In our model, the equilibrium probability of the folded state is always 0.5. Interestingly, the second peak (for the unfolded ensemble) appears only when $M > 5$. Using eq 7, one can check that P_m^{eq} has a minimum at $m = 2$ for all $M > 5$. As the size M of the peptide chain increases, the folded and unfolded basins of the peptide's chain become better separated, and the transition state conformations corresponding to $m = 1$ become less populated. In addition, the distribution of unfolded conformations becomes more and more Gaussian.

The value of m characterizes the “distance” of the conformation from the native state. We chose the number of unfolded residues, m , as a discrete one-dimensional reaction coordinate. The effective conformational dynamics along this coordinate is illustrated in Figure 2B. Note that m is similar to the fraction of native contacts, Q , which has been used as a reaction coordinate in numerous studies of protein folding.^{42–45}

We use the equilibrium distribution, P_m^{eq} , to introduce the potential of mean force, $U(m)$, along the reaction coordinate

$$U(m) = -k_B T \ln \frac{P_m^{eq}}{P_2^{eq}} \quad (8)$$

where k_B is the Boltzmann constant, T is the absolute temperature, and we choose $U(2)$ as the reference such that $U(2) = 0$ for all peptides. Figure 3B shows $U(m)$ for proteins of different lengths M . Note that (i) the potential of mean force has a barrier only when $M > 5$, and this barrier is always at $m = 2$; (ii) the transition state conformations with $m = 1$ have a lower energy that is independent of the chain length; and (iii) the barrier height increases quickly with M .

In our model, all unfolded (U) conformations with $m \geq 2$ have the same energy. This is a consequence of the identity of the forward and backward rate constants in eq 3. The energy of the conformations with $m = 1$, which form the transition state ensemble (TSE), is $k_B T \ln(1/\varepsilon_1) = k_B T \ln(M-1)$ lower than the energy of the unfolded conformations. Finally, the energy of the folded (F) conformation is $k_B T \ln[1/(\varepsilon_0\varepsilon_1)] = k_B T \ln(2^M + M^2 - 2M - 1)$ lower than that of the unfolded conformations. The dimensionless entropy of each of the three groups of conformations is defined as the natural logarithm of the number of conformations in the group

$$S_F = 0, \quad S_{TSE} = \ln M, \quad S_U = \ln(2^M - M - 1) \quad (9)$$

We use these relations to draw the folding funnels^{42,43} for our model proteins of different chain lengths M , shown in Figure 3C, where we chose the energy of the unfolded conformations as zero. Note that the TSE conformations with $m = 1$ are in the funnel. As the chain length increases, the slope of the TSE-U part of the funnel decreases, whereas the slope of the TSE-F part increases.

RESULTS AND DISCUSSION

To analyze the character of the relaxation kinetics, one needs to know the eigenvalues of the evolution operator. In vector–matrix notation, the master equation in eq 4 takes the form

$$\dot{P}(t) = KP(t) \quad (10)$$

Here, the evolution operator \mathbf{K} is the $(M + 1) \times (M + 1)$ three-diagonal rate matrix

$$\mathbf{K} = \begin{pmatrix} -\alpha_0 & \beta_1 & 0 & \cdots & 0 & 0 & 0 \\ \alpha_0 & -\gamma_1 & \beta_2 & \cdots & 0 & 0 & 0 \\ 0 & \alpha_1 & -\gamma_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\gamma_{M-2} & \beta_{M-1} & 0 \\ 0 & 0 & 0 & \cdots & \alpha_{M-2} & -\gamma_{M-1} & \beta_M \\ 0 & 0 & 0 & \cdots & 0 & \alpha_{M-1} & -\beta_M \end{pmatrix} \quad (11)$$

with matrix elements given by

$$\begin{aligned} \alpha_0 &= \varepsilon_0 M, & \alpha_1 &= 1, & \alpha_m &= M - m, & \text{for } m = 2, 3, \dots, M - 1 \\ \beta_m &= m, & \text{for } m &= 1, 2, \dots, M \\ \gamma_1 &= 2, & \gamma_m &= \alpha_m + \beta_m = M, & \text{for } m &= 2, \dots, M - 1 \end{aligned} \quad (12)$$

Eigenvalues, $-\lambda_i$, of the rate matrix and the corresponding eigenvectors, φ_i , are solutions of the eigenvalue problem

$$\mathbf{K}\varphi_i = -\lambda_i\varphi_i, \quad i = 1, 2, \dots, M + 1 \quad (13)$$

The eigenvalue with the smallest magnitude is equal to zero, $\lambda_1 = 0$, because the master equation, eq 10, describes the relaxation of the initial distribution to equilibrium. We assume that the eigenvalues are ordered in increasing values of their magnitudes, $\lambda_1 = 0 < \lambda_2 \leq \lambda_3 \leq \dots \leq \lambda_{M+1}$.

When the system consists of two basins separated by a high barrier, intrabasin relaxation occurs much faster than interbasin exchange. As a consequence, the exchange is a memory-less Markov process, and hence, equilibration of basin's populations is single-exponential. The eigenvalues of the evolution operator of such a system have a simple physical interpretation. The smallest nonzero eigenvalue, $-\lambda_2$, describes interbasin equilibration of the populations, whereas the other nonzero eigenvalues describe intrabasin relaxation to local equilibria in the two basins. The fact that the former process is much slower than the latter leads to the inequality $\lambda_2 \ll \lambda_3$, which implies that there is a large gap in the eigenvalue spectrum of the evolution operator.

For our simple model, one can find approximate solutions for λ_2 and λ_3 (see Appendix A)

$$\lambda_2 \approx \varepsilon_0 M = \frac{M(M-1)}{2^M + M^2 - 2M - 1}, \quad \lambda_3 \approx 2 \quad (14)$$

This allows us to establish the condition of applicability of the two-state description of folding

$$\lambda_3 \gg \lambda_2 \Rightarrow 2^{M+1} \gg M^2 \quad (15)$$

which is the main result of the present study. We show below that the relaxation is single-exponential to a good approximation for $M \geq 9$.

In our model, $k_F = k_U = \lambda_2/2$, and according to eq 14, at large M , we have $\ln \tau_F \propto M$, where $\tau_F = 1/k_F$ is the folding time. Note that linear scaling of the logarithm of the folding time with the protein size has been reported in the literature.^{46,47} A weaker scaling with the chain length, $\ln \tau_F \propto M^\gamma$, where $\gamma < 1$, has also been discussed.⁴⁷⁻⁵²

Finally, we compare the values of λ_2 and λ_3 given in eq 14 to the eigenvalues found numerically by diagonalization of the rate matrix (eq 11). The ratios of the eigenvalues for proteins of different lengths are presented in Table 1. One can see that eq 14

Table 1. Ratios of the Approximate Results in eq 14, $\lambda_{2,3}^{\text{app}}$, to Their Counterparts Obtained by Numerical Diagonalization of the Rate Matrix in eq 11, $\lambda_{2,3}^{\text{diag}}$

	M					
	5	10	20	30	40	50
$\lambda_2^{\text{app}}/\lambda_2^{\text{diag}}$	0.86	1.13	1.07	1.04	1.03	1.01
$\lambda_3^{\text{app}}/\lambda_3^{\text{diag}}$	1.02	1.24	1.09	1.04	1.03	1.02

provides good estimations for λ_2 and λ_3 when M is large enough. As expected, the agreement gets better as M increases. In Table 2, the ratios λ_3/λ_2 obtained numerically are compared with $2^{M+1}/M^2$, which is the large- M estimation of the ratio that follows from eq 14. Comparison shows that the simple expression $2^{M+1}/M^2$ appears to be a good approximation for the ratio of the eigenvalues.

One can interpret λ_2^{-1} and λ_3^{-1} as slow and fast relaxation times of the system, $\tau_{\text{slow}} = 1/\lambda_2$, $\tau_{\text{fast}} = 1/\lambda_3$. Figure 4 shows the ratio $(\tau_{\text{slow}} - \tau_{\text{fast}})/\tau_{\text{fast}}$ which characterizes the gap between the relaxation times, as a function of the chain length. One can see that the gap quickly increases with M and exceeds 10 for $M \geq 9$. Written in terms of λ_2 and λ_3 , the ratio takes the form $(\tau_{\text{slow}} - \tau_{\text{fast}})/\tau_{\text{fast}} = \lambda_3/\lambda_2 - 1$. When M is large, $\lambda_3/\lambda_2 \approx 2^{M+1}/M^2$, so that $(\tau_{\text{slow}} - \tau_{\text{fast}})/\tau_{\text{fast}} \approx 2^{M+1}/M^2$. Figure 4 shows that $2^{M+1}/M^2$ is close to the exact values of the ratio obtained numerically even at not-so-large values of M .

In summary, a simple model of the conformational dynamics has been used to analyze how the character of the folding kinetics depends on the chain length of the protein. We found that, to a good approximation, a protein can be considered as a two-state folder when the number of its residues is nine or larger. In view of this estimation, it is not surprising that the folding kinetics of pentaalanine deviates from single-exponential behavior.^{35,36}

APPENDIX A

When M is large, the potential of mean force $U(m)$ has a well-pronounced double-well structure with a high barrier that separates folded and unfolded conformations (Figure 3B). We assume that equilibration in the basin of unfolded conformations, $m = 2, 3, \dots, M$, is a fast process, $\lambda_4, \lambda_5, \dots, \lambda_{M+1} \gg 1$, so that this basin is always in local equilibrium and

$$P_m(t) = \frac{P_m^{\text{eq}}}{P_2^{\text{eq}}} P_2(t) = \frac{2(M-2)!}{m!(M-m)!} P_2(t), \quad m = 2, 3, \dots, M \quad (16)$$

The probability of finding the protein in the unfolded basin at time t , $P_U(t)$, is

$$P_U(t) = \sum_{m=2}^M P_m(t) = \frac{2(2^M - M - 1)}{M(M-1)} P_2(t) \quad (17)$$

Using this expression and the fact that $P_U(t) = 1 - P_0(t) - P_1(t)$, we can write $P_2(t)$ in terms of $P_0(t)$ and $P_1(t)$

$$\begin{aligned} P_2(t) &= \frac{M(M-1)}{2(2^M - M - 1)} P_U(t) \\ &= \frac{M(M-1)}{2(2^M - M - 1)} [1 - P_0(t) - P_1(t)] \end{aligned} \quad (18)$$

Table 2. Ratio of the First Two Nonzero Eigenvalues Obtained by Numerical Diagonalization of the Rate Matrix in eq 11, $\lambda_3^{\text{diag}}/\lambda_2^{\text{diag}}$, and the Large- M Estimation of This Ratio, $2^{M+1}/M^2$, as Functions of the Chain Length

	M					
	5	10	20	30	40	50
$\lambda_3^{\text{diag}}/\lambda_2^{\text{diag}}$	3.89	22.21	5.43×10^3	2.46×10^6	1.41×10^9	9.08×10^{11}
$2^{M+1}/M^2$	2.56	20.48	5.24×10^3	2.39×10^6	1.37×10^9	9.01×10^{11}

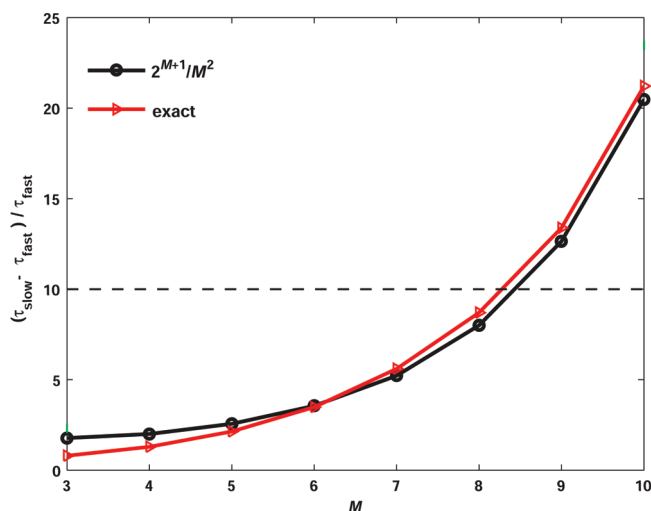


Figure 4. Ratio of the relaxation times $(\tau_{\text{slow}} - \tau_{\text{fast}})/\tau_{\text{fast}}$ and its large- M estimation, $2^{M+1}/M^2$, as functions of the chain length. The exact values (triangles) were obtained by diagonalization of the rate matrix.

Substituting this expression into the second equation of the set in eq 4, we find that the first two equations of this set form closed evolution equations for the probabilities $P_0(t)$ and $P_1(t)$

$$\begin{aligned} \dot{P}_0(t) &= -\varepsilon_0 M P_0(t) + P_1(t) \\ \dot{P}_1(t) &= \varepsilon_0 M P_0(t) - 2P_1(t) + \frac{M(M-1)}{2^M - M - 1} [1 - P_0(t) - P_1(t)] \end{aligned} \quad (19)$$

where we have used the relation $\varepsilon_1(M-1) = 1$ (eq 6).

Equations 19 describe relaxation of $P_0(t)$ and $P_1(t)$ to their equilibrium values (eq 7)

$$P_0^{\text{eq}} = 1/2, \quad P_1^{\text{eq}} = \varepsilon_0 M P_0^{\text{eq}} = \frac{M(M-1)}{2(2^M + M^2 - 2M - 1)} \quad (20)$$

Denoting the deviation of $P_i(t)$ from P_i^{eq} by $\Delta P_i(t)$, $\Delta P_i(t) = P_i(t) - P_i^{\text{eq}}$, $i = 0, 1$, and using eq 19, we find that the deviations satisfy

$$\begin{aligned} \Delta \dot{P}_0(t) &= -\varepsilon_0 M \Delta P_0(t) + \Delta P_1(t) \\ \Delta \dot{P}_1(t) &= \left[\varepsilon_0 M - \frac{M(M-1)}{2^M - M - 1} \right] \Delta P_0(t) \\ &\quad - \left[2 + \frac{M(M-1)}{2^M - M - 1} \right] \Delta P_1(t) \end{aligned} \quad (21)$$

The second equation can be simplified at large M , because the first term on the right-hand side can be neglected as the factor in

front of $\Delta P_0(t)$ is proportional to $\varepsilon_0^2 \ll 1$, whereas the factor in front of $\Delta P_1(t)$ is close to 2. As a result, the second equation takes the form

$$\Delta \dot{P}_1(t) = -2\Delta P_1(t) \quad (22)$$

From eqs 21 and 22, one can see that the relaxation to equilibrium is biexponential and that λ_2 and λ_3 are given by the expressions in eq 14.

AUTHOR INFORMATION

Corresponding Author

*E-mail: buchete@ucd.ie.

ACKNOWLEDGMENT

We thank Gerhard Hummer and Attila Szabo for many helpful and stimulating discussions. This study was supported by the Intramural Research Program of the National Institutes of Health (NIH), Center for Information Technology, and it used the computational resources of the Irish Centre for High-End Computing (ICHEC). F.T. and N.V.B gratefully acknowledge financial support from the Irish Research Council for Science, Engineering & Technology (IRCSET).

REFERENCES

- (1) Zwanzig, R.; Szabo, A.; Bagchi, B. Levinthal's Paradox. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 20.
- (2) Zwanzig, R. Simple Model of Protein Folding Kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9801.
- (3) Zwanzig, R. Two-State Models of Protein Folding Kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 148.
- (4) Buchete, N. V.; Straub, J. E. Mean First-Passage Time Calculations for the Coil-to-Helix Transition: The Active Helix Ising Model. *J. Phys. Chem. B* **2001**, *105*, 6684.
- (5) Go, M.; Go, N.; Scheraga, H. A. Molecular Theory of the Helix-Coil Transition in Polyamino Acids. 3. Evaluation and Analysis of S and Sigma for Polyglycine and Poly-L-Alanine in Water. *J. Chem. Phys.* **1971**, *54*, 4489.
- (6) Huang, C. Y.; Klemke, J. W.; Getahun, Z.; DeGrado, W. F.; Gai, F. Temperature-Dependent Helix-Coil Transition of an Alanine Based Peptide. *J. Am. Chem. Soc.* **2001**, *123*, 9235.
- (7) Wang, T.; Du, D. G.; Gai, F. Helix-Coil Kinetics of Two 14-Residue Peptides. *Chem. Phys. Lett.* **2003**, *370*, 842.
- (8) Wang, T.; Zhu, Y.; Getahun, Z.; Du, D.; Huang, C.-Y.; DeGrado, W. F.; Gai, F. Length Dependent Helix-Coil Transition Kinetics of Nine Alanine-Based Peptides. *J. Phys. Chem. B* **2004**, *108*, 15301.
- (9) Buchete, N. V.; Straub, J. E.; Thirumalai, D. Orientation-Dependent Coarse-Grained Potentials Derived by Statistical Analysis of Molecular Structural Databases. *Polymer* **2004**, *45*, 597.
- (10) Buchete, N. V.; Straub, J. E.; Thirumalai, D. Continuous Anisotropic Representation of Coarse-Grained Potentials for Proteins by Spherical Harmonics Synthesis. *J. Mol. Graph.* **2004**, *22*, 441.
- (11) Van Giessen, A. E.; Straub, J. E. Coarse-Grained Model of Coil-to-Helix Kinetics Demonstrates the Importance of Multiple Nucleation Sites in Helix Folding. *J. Chem. Theory Comput.* **2006**, *2*, 674.
- (12) Mukherjee, S.; Chowdhury, P.; Bunagan, M. R.; Gai, F. Folding Kinetics of a Naturally Occurring Helical Peptide: Implication of the Folding Speed Limit of Helical Proteins. *J. Phys. Chem. B* **2008**, *112*, 9146.
- (13) Buchete, N. V.; Straub, J. E.; Thirumalai, D. Dissecting Contact Potentials for Proteins: Relative Contributions of Individual Amino Acids. *Proteins* **2008**, *70*, 119.

- (14) Rosta, E.; Buchete, N. V.; Hummer, G. Thermostat Artifacts in Replica Exchange Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2009**, *5*, 1393.
- (15) Freddolino, P. L.; Harrison, C. B.; Liu, Y. X.; Schulten, K. Challenges in Protein-Folding Simulations. *Nat. Phys.* **2010**, *6*, 751.
- (16) Li, M. S.; Klimov, D. K.; Thirumalai, D. Finite Size Effects on Calorimetric Cooperativity of Two-State Proteins. *Physica A* **2005**, *350*, 38.
- (17) Berezhkovskii, A.; Szabo, A. Perturbation Theory of Φ -Value Analysis of Two-State Protein Folding: Relation between p_{fold} and Φ Values. *J. Chem. Phys.* **2006**, *125*, 104902.
- (18) Muñoz, V.; Eaton, W. A. A Simple Model for Calculating the Kinetics of Protein Folding from Three-Dimensional Structures. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11311.
- (19) Muñoz, V. Thermodynamics and Kinetics of Downhill Protein Folding Investigated with a Simple Statistical Mechanical Model. *Int. J. Quantum Chem.* **2002**, *90*, 1522.
- (20) Bruscolini, P.; Pelizzola, A. Exact Solution of the Muñoz-Eaton Model for Protein Folding. *Phys. Rev. Lett.* **2002**, *88*, 258101.
- (21) Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. Contact Order Revisited: Influence of Protein Size on the Folding Rate. *Protein Sci.* **2003**, *12*, 2057.
- (22) Weikl, T. R.; Palassini, M.; Dill, K. A. Cooperativity in Two-State Protein Folding Kinetics. *Protein Sci.* **2004**, *13*, 822.
- (23) Ferguson, A.; Liu, Z.; Chan, H. S. Desolvation Barrier Effects Are a Likely Contributor to the Remarkable Diversity in the Folding Rates of Small Proteins. *J. Mol. Biol.* **2009**, *389*, 619.
- (24) Murza, A.; Kubelka, J. Beyond the Nearest-Neighbor Zimm–Bragg Model for Helix–Coil Transition in Peptides. *Biopolymers* **2009**, *91*, 120.
- (25) Brooks, C. L.; Case, D. A. Simulations of Peptide Conformational Dynamics and Thermodynamics. *Chem. Rev.* **1993**, *93*, 2487.
- (26) Brooks, C. L. Helix–Coil Kinetics: Folding Time Scales for Helical Peptides from a Sequential Kinetic Model. *J. Phys. Chem.* **1996**, *100*, 2546.
- (27) Hummer, G.; Garcia, A. E.; Garde, S. Helix Nucleation Kinetics from Molecular Simulations in Explicit Solvent. *Proteins* **2001**, *42*, 77.
- (28) Margulis, C. J.; Stern, H. A.; Berne, B. J. Helix Unfolding and Intramolecular Hydrogen Bond Dynamics in Small α -Helices in Explicit Solvent. *J. Phys. Chem. B* **2002**, *106*, 10748.
- (29) Hummer, G.; Kevrekidis, I. G. Coarse Molecular Dynamics of a Peptide Fragment: Free Energy, Kinetics, and Long-Time Dynamics Computations. *J. Chem. Phys.* **2003**, *118*, 10762.
- (30) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. Structure and Dynamics of the Homologous Series of Alanine Peptides: A Joint Molecular Dynamics/NMR Study. *J. Am. Chem. Soc.* **2007**, *129*, 1179.
- (31) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic Discovery of Metastable States for the Construction of Markov Models of Macromolecular Conformational Dynamics. *J. Chem. Phys.* **2007**, *126*, 155101.
- (32) Wickstrom, L.; Okur, A.; Simmerling, C. Evaluating the Performance of the ff99SB Force Field Based on NMR Scalar Coupling Data. *Biophys. J.* **2009**, *97*, 853.
- (33) Ruzhytska, S.; Jacobi, M. N.; Jensen, C. H.; Nerukh, D. Identification of Metastable States in Peptide's Dynamics. *J. Chem. Phys.* **2010**, *133*, 164102.
- (34) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins* **2010**, *78*, 1950.
- (35) Buchete, N. V.; Hummer, G. Coarse Master Equations for Peptide Folding Dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057.
- (36) Buchete, N. V.; Hummer, G. Peptide Folding Kinetics from Replica Exchange Molecular Dynamics. *Phys. Rev. E* **2008**, *77*, 030902.
- (37) Buchner, G. S.; Murphy, R. D.; Buchete, N. V.; Kubelka, J. Dynamics of Protein Folding: Probing the Kinetic Network of Folding–Unfolding Transitions with Experiment and Theory. *Biochim. Biophys. Acta* **2011**, *1814*, 1001.
- (38) Thompson, P. A.; Eaton, W. A.; Hofrichter, J. Laser Temperature Jump Study of the Helix–Coil Kinetics of an Alanine Peptide Interpreted with a 'Kinetic Zipper' Model. *Biochemistry* **1997**, *36*, 9200.
- (39) Yang, W. Y.; Gruebele, M. Folding at the Speed Limit. *Nature* **2003**, *423*, 193.
- (40) Kubelka, J.; Hofrichter, J.; Eaton, W. A. The Protein Folding 'Speed Limit'. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76.
- (41) Amunson, K. E.; Ackels, L.; Kubelka, J. Site-Specific Unfolding Thermodynamics of a Helix-Turn-Helix Protein. *J. Am. Chem. Soc.* **2008**, *130*, 8146.
- (42) Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. Navigating the Folding Routes. *Science* **1995**, *267*, 1619.
- (43) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. Diffusive Dynamics of the Reaction Coordinate for Protein Folding Funnel. *J. Chem. Phys.* **1996**, *104*, 5860.
- (44) Best, R. B.; Hummer, G. Reaction Coordinates and Rates from Transition Paths. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6732.
- (45) Best, R. B.; Hummer, G. Coordinate-Dependent Diffusion in Protein Folding. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 1088.
- (46) Plotkin, S. S.; Wang, J.; Wolynes, P. G. Statistical Mechanics of a Correlated Energy Landscape Model for Protein Folding Funnel. *J. Chem. Phys.* **1997**, *106*, 2932.
- (47) Gutin, A. M.; Abkevich, V. I.; Shakhnovich, E. I. Chain Length Scaling of Protein Folding Time. *Phys. Rev. Lett.* **1996**, *77*, 5433.
- (48) Thirumalai, D. From Minimal Models to Real Proteins: Time Scales for Protein Folding Kinetics. *J. Phys. I* **1995**, *5*, 1457.
- (49) Finkelstein, A. V.; Badretdinov, A. Y. Rate of Protein Folding near the Point of Thermodynamic Equilibrium between the Coil and the Most Stable Chain Fold. *Fold. Des.* **1997**, *2*, 115.
- (50) Finkelstein, A. V.; Badretdinov, A. Y. Influence of Chain Knotting on the Rate of Folding. *Fold. Des.* **1998**, *3*, 67.
- (51) Wolynes, P. G. Folding Funnel and Energy Landscapes of Larger Proteins within the Capillarity Approximation. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 6170.
- (52) Naganathan, A. N.; Muñoz, V. Scaling of Folding Times with Protein Size. *J. Am. Chem. Soc.* **2005**, *127*, 480.

Assessing Excited State Methods by Adiabatic Excitation Energies

Robert Send,[†] Michael Kühn,[†] and Filipp Furche^{*,†}[†]Institut für Physikalische Chemie, Karlsruher Institut für Technologie, Kaiserstrasse 12, 76131 Karlsruhe, Germany[‡]Department of Chemistry, University of California—Irvine, 1102 Natural Sciences II, Irvine, California 92697-2025, United States Supporting Information

ABSTRACT: We compile a 109-membered benchmark set of adiabatic excitation energies (AEEs) from high-resolution gas-phase experiments. Our data set includes a variety of organic chromophores with up to 46 atoms, radicals, and inorganic transition metal compounds. Many of the 91 molecules in our set are relevant to atmospheric chemistry, photovoltaics, photochemistry, and biology. The set samples valence, Rydberg, and ionic states of various spin multiplicities. As opposed to vertical excitation energies, AEEs are rigorously defined by energy differences of vibronic states, directly observable, and insensitive to errors in equilibrium structures. We supply optimized ground state and excited state structures, which allows fast and convenient evaluation of AEEs with two single-point energy calculations per system. We apply our benchmark set to assess the performance of time-dependent density functional theory using common semilocal functionals and the configuration interaction singles method. Hybrid functionals such as B3LYP and PBE0 yield the best results, with mean absolute errors around 0.3 eV. We also investigate basis set convergence and correlations between different methods and between the magnitude of the excited state relaxation energy and the AEE error. A smaller, 15-membered subset of AEEs is introduced and used to assess the correlated wave function methods CC2 and ADC(2). These methods improve upon hybrid TDDFT for systems with single-reference ground states but perform less well for radicals and small-gap transition metal compounds. None of the investigated methods reaches “chemical accuracy” of 0.05 eV in AEEs.

1. INTRODUCTION

In view of the enormous interest in electronically excited states in analytical, photochemical, and material science applications, the amount of reliable data available to assess and calibrate theoretical excited state treatments is surprisingly small. The most common benchmarks are vertical excitation energies (VEEs) obtained either from higher level theory or from experimental absorption maxima (AM).^{1–22} An extensive VEE benchmark set including a review of previous VEE benchmarks has recently been published by Jacquemin and co-workers.¹⁷

While VEEs are readily obtained in quantum chemistry, they are not observable experimentally. The comparison of computed VEEs to experimental AM is based on three main assumptions: (i) The transition probability is maximized if ground and excited state structures are identical (Franck–Condon approximation). (ii) The transition probability is maximized at the ground state minimum. This is true only if the electronic transition probability is nonzero and does not increase upon geometry distortion, i.e., due to symmetry breaking vibrations. (iii) The absorption maximum is not affected by vibronic or rotational effects. As Herzberg already pointed out,²³ these conditions are not satisfied even in apparently simple cases such as the transition from the ground state to the 1^1B_2 state in CS_2 and the transition from the ground state to the $1^1A''$ state in HCN. The error introduced by comparing VEEs and AM can reach the order of typical method errors, as Dierksen and Grimme illustrated using simulated vibronic spectra: For the $S_0 \rightarrow S_1$ excitations of anthracene, pyrene, and pentacene, the difference between VEEs and AM ranges from 0.25 to 0.33 eV.²⁴

The past decade has seen a number of approaches to improving the accuracy of VEE benchmarks. Thiel and co-workers introduced a large benchmark set based on VEEs with experimental AM

replaced by *best estimates*.^{11,12,21} The latter are obtained from correlated ab initio calculations with large basis sets where available. Otherwise, triple- ζ complete active space second-order perturbation theory (CASPT2) results are used for singlet and triple- ζ approximate third-order coupled cluster (CC3)²⁵ results for triplet excited states. While this method requires only one single-point calculation per system, it is limited to small systems and lacks experimental validation. Jacquemin and co-workers addressed the inherent error of the VEE-AM benchmark by comparing VEEs to Thiel’s *best estimates* and experimental adiabatic excitation energies (AEEs).¹⁷ However, the conclusions are limited since excited state relaxation is neglected in the computed and unknown in the experimental AEEs.

A comprehensive comparison of spectroscopic and computational data is possible by investigating simulated vibronic spectra. Today, such calculations are feasible for fairly large molecules, and the number of vibronic simulations has grown in recent years.^{24,26–29} Simulated vibronic spectra reveal the position of AM, VEEs, and AEEs. These simulations may be systematically improved by including anharmonic corrections, Dushinsky rotation effects, Herzberg–Teller corrections, and hot bands.³⁰ However, such calculations are computationally expensive and demand a high level of expertise. The validity of most approaches is limited to harmonic potential energy surfaces and the existence of an excited state minimum. Currently, a statistically significant assessment of excited state methods using a large number of systems and states is hardly feasible at this level.

Received: February 25, 2011

Published: July 05, 2011

Benchmarks using AEEs represent a compromise between simple VEE-based benchmarks and elaborate vibronic simulations. The AEE is defined as the energy difference between the lowest vibrational level of the excited and ground electronic states. As opposed to VEEs, AEEs are directly observable experimentally, e.g., from 0-0 transitions in high-resolution absorption and fluorescence spectra, and can be measured with high accuracy. On the theoretical side, AEEs require optimized geometries of both the ground and the excited electronic states. Thus, the size and chemical diversity of existing AEE benchmark sets is limited.^{16,31–35} However, with progress in analytical excited state gradient theory,^{31,32} excited state structure optimization is no longer an obstacle, even for systems containing 100 or more atoms. Another significant advantage of AEEs over VEEs is their quadratic dependence on errors in the ground and excited state structures, making them less sensitive to the quality of the structures used. Finally, it is possible to refine computed AEEs to very high accuracy by adding zero-point vibrational energy (ZPVE), relativistic, and diagonal Born–Oppenheimer corrections.³⁶

In this paper, we compile a set of 109 accurate experimental AEEs for benchmark purposes. We include a variety of different systems and states. The benchmark selection procedure is explained in section 2. To the best of our knowledge, this is the largest and most diverse AEE benchmark available so far. Using statistical analysis, we investigate basis set errors (section 4.1) and the performance of time-dependent density functional theory (TDDFT)³⁷ using common hybrid and nonhybrid exchange–correlation (XC) functionals (section 4.2). Configuration interaction singles (CIS)^{38,39} is also considered. We discuss the implications of systematic corrections (section 4.3), correlations between methods (section 4.4), and the influence of large excited state relaxation energies on hybrid TDDFT (section 4.6). Difficult cases for common functionals are discussed in section 5. Section 6 summarizes structural changes observed in the excited state geometry optimizations. A smaller, 15-membered subset of AEEs is introduced in section 7 and used to benchmark the wavefunction-based approximate second-order coupled cluster (CC2)⁴⁰ and the second-order algebraic diagrammatic construction approximation (ADC(2))⁴¹ methods. We present our conclusions in section 8.

2. BENCHMARK SELECTION PROCEDURE

It is impossible to construct a benchmark set that is representative of the entire chemical compound space. Thus, any selection of benchmark data is necessarily biased. It is important to spell out this bias clearly because it intrinsically limits the significance of any benchmark set. Without a well-motivated selection procedure, it must be suspected that a benchmark set has been chosen to favor or disfavor a particular method. This is not our intention.

We chose to bias the present benchmark set toward systems that are expected to be important for a large number of applications. In part, this was accomplished by using experimental reference data only. Experiments are generally more expensive and time-consuming than calculations and are likely more relevant than calculations for systems and states that may not be directly observable. Also, highly accurate calculations of excitation energies are presently feasible for small molecules with 10 or less atoms only. An important selection criterion was the quality of the experimental data: Only vibrationally resolved gas-phase data were included; questionable or controversial results were avoided. Our data set samples important aliphatic, aromatic, and

Table 1. Composition of the Benchmark Set

characteristic	count
molecules	91
excited states	109
organic molecules	64
inorganic molecules	27
transition metal compounds	6
aliphatic molecules	20
aromatic molecules	43
antiaromatic molecules	1
organic radicals	4
singlet excited states	86
triplet excited states	12
spin-unrestricted excited states	11

antiaromatic chromophores; it includes radicals and inorganic main group and selected transition metal compounds (see Table 1). Examples of special interest for applications include coumarin 153, a merocyanine dye, a chromophore model of the photoactive yellow protein, push–pull stilbenes, fluorene, and azaindole. The size of the molecules in our set ranges from 2 to 46 atoms (terylene). Most states are singlets, doublets, or triplets, but a few states of higher spin multiplicity are included as well. The majority of states are of valence type, some have considerable Rydberg character, such as the 3^1A state of aminomethane. Gas-phase results for long-range charge-transfer excitations were not available to us.

3. COMPUTATIONAL DETAILS

All ground and excited state structures were optimized using Becke's three-parameter hybrid functional with Lee–Yang–Parr correlation (B3LYP).⁴² Earlier benchmarks have shown that the excited state minimum structures are similar in accuracy to minimum structures obtained using ground-state density functional theory (DFT) if the Kohn–Sham reference is stable.³¹ The sensitivity of the results to the level of theory used to compute equilibrium structures is further investigated in section 7.3. Excited-state optimizations were started from the ground-state minimum structures. The geometry optimizations employed triple- ζ valence basis sets with one set of polarization functions (def2-TZVP).⁴³ Ground-state energies and density matrices were converged to 10^{-7} au. Fine quadrature grids of size m4⁴⁴ were used in the DFT and TDDFT calculations. The optimized structures were confirmed to be local energy minima by calculating vibrational frequencies. For excited states, second derivatives of the energy were determined by numerical differentiation of analytical gradients using central differences. Energy and density convergence criteria of 10^{-9} au were used for the numerical second derivatives.

For the excitation energy calculations, basis sets of split valence quality with polarization functions except on hydrogen atoms def2-SV(P)⁴³ and of triple- ζ quality (def2-TZVP) were employed. Diffuse-augmented def2-SVPD and def2-TZVPD bases from the newly developed hierarchy of property-optimized basis sets were also investigated.⁴⁵ For Sn, small core pseudopotentials (ECP-28)⁴⁶ were used. AEEs were computed from the total energy difference of ground and excited states at the optimized B3LYP/TZVP structures. The resulting energy differences were corrected by the ZPVEs obtained from vibrational frequency

Table 2. AEEs of All Excited States Contained in the Benchmark Set in eV^a

molecule	state	experiment	B3LYP	reference
1,6-epoxy-10-annulene	1 ¹ B ₁	2.98	3.32	58
2-chloropyrimidine	2 ¹ A'	3.98	3.73	59
2-cyclopenten-1-one	2 ¹ A	3.37	3.28	60
7-azaindole	2 ¹ A	4.29	4.09	61
β -dinaphthyleneoxide	2 ¹ A	3.63	3.25	62
acetaldehyde	2 ¹ A	3.69	3.81	63
acetone	2 ¹ A	3.77	3.87	63
acrolein	1 ¹ A''	3.21	3.02	64
acrolein	1 ³ A''	3.01	2.44	65
aminobenzonitrile	1 ¹ B ₁	4.15	4.25	66
aminoethane	2 ¹ A'	5.21	5.30	67
aminomethane	2 ¹ A'	5.18	5.23	67
aminomethane	3 ¹ A	6.22	6.36	67
aniline	2 ¹ A	4.22	4.34	68
anisole	2 ¹ A	4.51	4.73	69
anthracene	1 ¹ B _{2u}	3.43	2.90	70
AsF	1 ³ Π	3.19	3.10	71
azulene	1 ¹ B ₁	1.77	1.98	72
BeH	1 ² Π	2.48	2.58	71
benzaldehyde	2 ¹ A'	4.36	4.37	73
benzaldehyde	1 ¹ A''	3.34	3.27	74
benzaldehyde	1 ³ A''	3.12	2.71	74
benzene	1 ¹ B _{1u}	4.72	5.14	75
benzonitrile	1 ¹ B ₁	4.53	4.82	76
benzophenone ketyl radical	2 ² A	2.29	2.31	77
BeO	1 ¹ Π	1.17	1.21	71
BF	1 ¹ Π	6.34	6.14	71
BH	1 ¹ Π	2.87	2.70	71
biphenyl	1 ¹ B ₁	4.37	4.24	78
biphenylene	1 ¹ B _{3u}	3.50	3.63	79
C ₂ H ₂	2 ¹ A	5.23	4.72	80
CCl ₂	1 ¹ B ₂	2.14	1.89	81
CH ₂ O	2 ¹ A	3.49	3.62	82
CH ₂ O	1 ³ A	3.12	2.73	82
CH ₂ S	1 ¹ A ₂	2.03	2.07	82
CH ₂ S	1 ³ A	1.80	1.44	82
cinnoline	1 ¹ A''	2.82	2.43	83
CO	1 ¹ Π	8.07	7.99	71
CO	1 ³ Π	6.04	5.58	71
CrF	2 ⁶ Π	1.01	1.24	84
CrH	2 ⁶ Σ^+	1.62	1.78	85
CS ₂	3 ³ A	3.25	3.25	80
Cu ₂	1 ¹ Σ_u^+	2.53	2.75	86
Cu ₂	1 ¹ Π_u	2.71	2.63	86
CuH	2 ¹ Σ^+	2.91	3.06	71
cynoacetylene	2 ¹ A'	4.77	4.74	23
cyclohexadienyl radical	2 ² A	2.26	2.49	87
DBH ⁸⁸	2 ¹ A	3.66	3.45	89
DCS ⁹⁰	2 ¹ A	3.36	2.95	91
dimethylaminobenzonitrile	1 ¹ B	4.02	4.04	66
DMPD ⁹²	2 ¹ A	3.64	3.68	93
fluorene	2 ¹ A'	4.19	4.19	94
glyoxal	1 ¹ A _u	2.72	2.43	80

Table 2. Continued

molecule	state	experiment	B3LYP	reference
HCN	1 ¹ A''	6.48	5.97	80
HCOOH	2 ¹ A	4.64	4.78	95
HCP	2 ¹ A'	4.31	4.31	97
hexatriene	1 ¹ B _u	4.93	4.32	71
hydroquinone	1 ¹ B _u	4.15	4.28	100
indole	2 ¹ A'	4.37	4.31	101
Li ₂	1 ¹ Σ_u^+	1.74	1.99	71
merocyanine dye ⁹⁹	2 ¹ A	2.58	2.53	100
methyl-4-hydroxycinnamate	2 ¹ A	4.08	3.96	101
Mg ₂	1 ¹ Σ_u^+	3.23	3.44	71
N ₂	1 ¹ Δ_u	8.94	8.32	71
N ₂	1 ¹ Π_g	8.59	8.57	71
N ₂	1 ³ Π_g	7.39	6.94	71
naphthalene	1 ¹ B _{2u}	3.97	3.96	102
NH	1 ³ Π	3.70	3.92	71
NH ₃	1 ¹ A'' ₂	5.73	5.73	103
NH ₃	3 ¹ A	7.34	7.59	103
NO ₃	1 ² E'	1.87	2.13	104
octatetraene	1 ¹ B _u	4.41	3.72	105
o-cyanobenzyl radical	2 ² A	2.50	2.65	106
oxalylfluoride	1 ¹ A _u	4.02	3.72	107
P ₂	1 ¹ Π_g	4.27	4.12	71
p-benzoquinone	2 ¹ A	2.48	2.28	108
p-benzoquinone	1 ¹ B _{1g}	4.07	3.34	109
p-benzoquinone	1 ¹ B _{2g}	2.49	2.26	108
p-diethynylbenzene	1 ¹ B _{2u}	4.25	4.17	110
p-phenylenediamine	2 ¹ A	3.70	3.69	93
PH ₂	2 ² B ₂	2.27	2.37	80
phenol	2 ¹ A	4.51	4.78	111
porphyrin	1 ¹ B _{1u}	2.02	2.20	80
propynal	2 ¹ A	3.24	3.25	112
pyrene	2 ¹ A	3.34	3.40	113
pyridine	2 ¹ A	4.31	4.14	114
pyridone lactam	2 ¹ A	3.70	3.70	115
pyridone lactim	2 ¹ A	4.48	4.60	115
pyrimidine	2 ¹ A	3.85	3.67	116
pyrimidine	4 ¹ A	5.00	5.16	114
quinoline	2 ¹ A	3.99	3.51	116
quinoline	1 ³ A'	2.79	2.26	116
quinoxaline	2 ¹ A ₁	3.97	4.25	116
quinoxaline	2 ¹ A	3.36	3.02	83
quinoxaline	1 ³ A	2.68	2.19	117
ScO	1 ² Π	2.04	2.14	118
SiF ₂	1 ¹ B ₂	5.34	5.35	80
SiO	1 ¹ Π	5.31	5.20	80
SnF ₂	1 ¹ B ₂	5.05	4.65	119
styrene	2 ¹ A	4.31	4.40	120
syn-coumarin 153	2 ¹ A	3.21	2.96	121
terylene	2 ¹ A'	2.39	1.90	122
tetrazine	2 ¹ A	2.25	1.99	123
thioacetone	2 ¹ A	2.33	2.35	124
thioacetone	1 ³ A	2.14	1.85	124
toluene	1 ¹ A''	4.65	5.01	125
t-stilbene	2 ¹ A	4.00	3.51	126

Table 2. Continued

molecule	state	experiment	B3LYP	reference
vinyl radical	2 ² A	2.48	2.56	127
VO	1 ⁴ Π	1.56	1.41	118

^aThe calculated zero-point vibrational energy corrected values are compared to experimental values. All theoretical values were obtained using the B3LYP functional and def2-TZVP basis sets.

calculations at the B3LYP/TZVP level. Excited state relaxation energies were computed from the difference of calculated VEEs and AEEs. The fluorescence energies are vertical de-excitation energies computed at the optimized B3LYP/TZVP excited state structure. A complete list of fluorescence and relaxation energies is given as Supporting Information.

The following density functionals were investigated: The Perdew–Wang parametrization of the local spin-density approximation (LSDA);⁴⁷ the generalized gradient approximation (GGA) functionals of Becke and Perdew (BP86)^{48,49} and of Perdew, Burke, and Ernzerhof (PBE);⁵⁰ the meta-GGA functional of Tao, Perdew, Staroverov, and Scuseria (TPSS);⁵¹ and the B3LYP⁴² and PBE0⁵² hybrid GGAs. Wave function methods included CIS,⁵³ the approximate coupled cluster singles and doubles method CC2,⁴⁰ and the second-order algebraic diagrammatic construction approximation (ADC(2)).⁴¹ In the CC2 and ADC(2) calculations, the resolution-of-the-identity approximation is employed along with def2-TZVP basis and auxiliary basis sets.^{54–56} To obtain excited state energies, the ADC(2) excitation energies are added to ground state MP2 energies. All calculations were carried out using the TURBOMOLE program suite.⁵⁷

A correct assignment is vital for comparison of experimental excited state data with theoretical results. We adopted the following procedure to ensure that the computed AEEs correspond to the experimentally observed states:

1. Vertical excitation energies were computed at the optimized ground state minimum, including all excitations with energies close to the experimental AEE.
2. A set of excited states was preselected on the basis of oscillator strength if absorption properties are available from the experiment.
3. The selected excited states were optimized, and the experimental states were assigned using oscillator strengths and further experimental evidence, such as rotational intensities and rotational constants.
4. In diatomic molecules, symmetry information from rovibronic spectra was required to be consistent with the excited state term symbol.

Cases where a clear-cut assignment was not possible were discarded from the benchmark set.

Covariances Σ were computed according to

$$\Sigma(X, Y) = \frac{1}{N-1} \sum_{i=1}^N (E_i^{\text{calc}}(X) - E_i^{\text{exp}}(X))(E_i^{\text{calc}}(Y) - E_i^{\text{exp}}(Y)) \quad (1)$$

where E_i^{calc} and E_i^{exp} are the computed and experimental AEEs and N is the number of excited states in our benchmark set. X and Y denote data sets. Sample standard deviations σ were obtained from Σ according to

$$\sigma(X) = \sqrt{\Sigma(X, X)} \quad (2)$$

Histograms and normalized Gaussians with a mean and SD identical to those of the sample were used to visualize the results.

Table 3. Performance of B3LYP AEEs on the 109 Set Using Different Basis Sets^a

basis	MAE	ME	SD	MaxAE
def2-SV(P)	0.25	−0.07	0.30	0.69
def2-SVPD	0.22	−0.13	0.29	0.78
def2-TZVP	0.21	−0.08	0.28	0.73
def2-TZVPD	0.22	−0.11	0.29	0.74

^aMAE denotes the mean absolute error, ME the mean error, SD the standard deviation, and MaxAE the maximum absolute error. All values are in eV.

Finally, correlation coefficients ρ were computed according to

$$\rho(X, Y) = \frac{\Sigma(X, Y)}{\sigma(X) \sigma(Y)} \quad (3)$$

4. RESULTS

4.1. Basis Set Convergence. The B3LYP AEEs using def2-TZVP basis sets are compared to experimental reference values in Table 2. A statistical analysis of the basis set convergence is presented in Table 3 and illustrated in Figure 1; a complete list of the results is provided as Supporting Information. The basis set convergence of other density functionals and CIS is similar. Higher l -quantum numbers and more diffuse functions are necessary for CC2 and ADC(2), but a systematic assessment of the basis set convergence of these methods is beyond our present scope. Overall, the B3LYP results change very little when going from the smaller def2-SV(P) to the larger def2-TZVP basis set or adding diffuse augmentation. Table 3 suggests that the residual basis set error of the B3LYP/TZVP results is 0.03 eV or less, roughly an order of magnitude less than the method error. Diffuse or triple- ζ basis sets are important for π – π^* excitations as in styrene and for π – σ^* excitations as in aminoethane and diatomic molecules where electrons are excited into diffuse orbitals. Further examples are discussed in section 5.

4.2. Exchange-Correlation Functionals. The performance of different functionals was assessed using def2-TZVP basis sets. A statistical analysis is given in Table 4, and a complete list of results is available as Supporting Information. The hybrid functionals B3LYP and PBE0 have the smallest SD (0.28 eV and 0.30 eV) and a mean error (ME) close to zero (−0.08 eV and 0.01 eV), followed by the TPSS meta-GGA with a SD of 0.41 eV and a ME of −0.20 eV. The GGA functionals PBE and BP86 feature a larger SD of 0.49 eV and a ME of −0.33 eV. LSDA performs just as well as the GGA functionals with a SD of 0.49 eV and a ME of −0.31 eV. CIS strongly overestimates the experimental results as reflected by its large positive ME, and the SD of 1.16 eV is almost 4 times that of B3LYP and PBE0.

The histograms in Figure 2 illustrate that all density functionals except PBE0 underestimate experimental AEEs. The maximum of the distribution lies between zero deviation and the ME. An underestimation by 0.5 eV or more is fairly common for nonhybrid functionals. CIS shows a broad error distribution and seems to have little predictive power for AEEs. The maximum positive and negative deviations for each level are listed in Table 5. The table shows that maximum deviations increase from hybrid GGAs to (meta-)GGAs and to LSDA. The total range of errors increases with the maximum deviations. Again, CIS

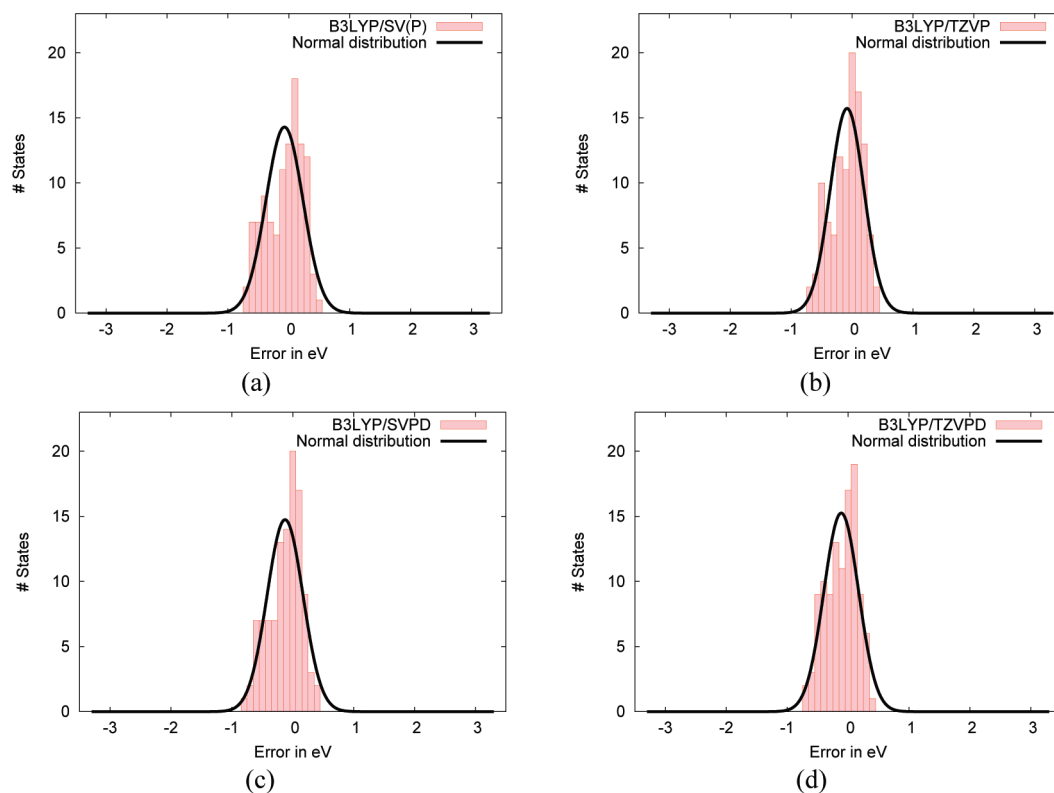


Figure 1. Error histograms of B3LYP AEEs obtained using different basis sets (see insets). Also shown are normal distributions with the computed ME and SD of the sample.

Table 4. Performance of AEEs Obtained from Various Exchange-Correlation Functionals Using def2-TZVP Basis Sets^a

method	MAE	ME	SD	MaxAE
B3LYP	0.21	-0.08	0.28	0.73
PBE0	0.25	0.01	0.30	0.66
TPSS	0.32	-0.20	0.41	1.05
BP86	0.39	-0.32	0.49	1.14
PBE	0.40	-0.33	0.49	1.14
LSDA	0.39	-0.31	0.49	1.19
CIS	0.98	0.90	1.16	3.04

^a See Table 3 for further explanation. All values are in eV.

performs more weakly than the density functional methods, with a total range of errors of almost 4 eV.

4.3. A Posteriori Corrections. Jacquemin and co-workers proposed to use linear regression to correct TDDFT excitation energies in solution and reduce the ME and SD.³⁴ This procedure is useful for samples of similar compounds. In Table 6, we show how the statistical parameters are affected when AEEs are shifted by a constant that makes the ME vanish. CIS benefits most from such a shift, which partially corrects the systematic overestimation of AEEs. LSDA and (meta-)GGAs, which underestimate excitation energies, are considerably improved, with reductions of ~ 0.1 eV in the MAE and SD. As expected from their small MEs, the hybrid results change little: B3LYP and PBE0 still outperform all other functionals, even after application of the shift. Thus, our

gas-phase data support a posteriori corrections for nonhybrid functionals only.

4.4. Correlations. We compare the similarity of the AEEs obtained with two functionals by calculating correlation coefficients and the maximum absolute deviation between two calculated AEEs (Table 7). The thus obtained correlation coefficients deviate from unity in the second to fourth digit. The maximum absolute deviations range from 0.2 to 1.2 eV.

The correlation analysis shows marked differences between LSDA, GGA, meta-GGA, and hybrid functionals. Within each rung of functionals, correlations are higher; e.g., the BP86 and PBE functionals perform very similarly, as indicated by a correlation coefficient of 0.9997 and a maximum deviation of 0.19 eV. Table 7 also shows that LSDA and GGA AEEs are highly correlated, in line with the empirical observation that GGAs tend to improve little upon LSDA for excitation energies. The TPSS results are quite similar to the GGA results, confirming that the improvement from GGAs to meta-GGAs is quite small for excitation energies.¹²⁸ On the other hand, the hybrid functionals PBE0 and B3LYP show lower correlations to other methods. This underlines the fact that hybrid exchange systematically improves over LSDA and GGAs for AEEs, consistent with the results of section 4.2.

4.5. Zero Point Vibrational Energies. The ZPVE corrections to the AEEs in our benchmark set range from -0.30 eV to $+0.26$ eV; they are positive in only eight cases. A complete list of the ZPVE corrections is available as Supporting Information. The largest ZPVE corrections occur when amino groups become planar in the excited state. Examples are aminomethane, aminoethane, and ammonia.

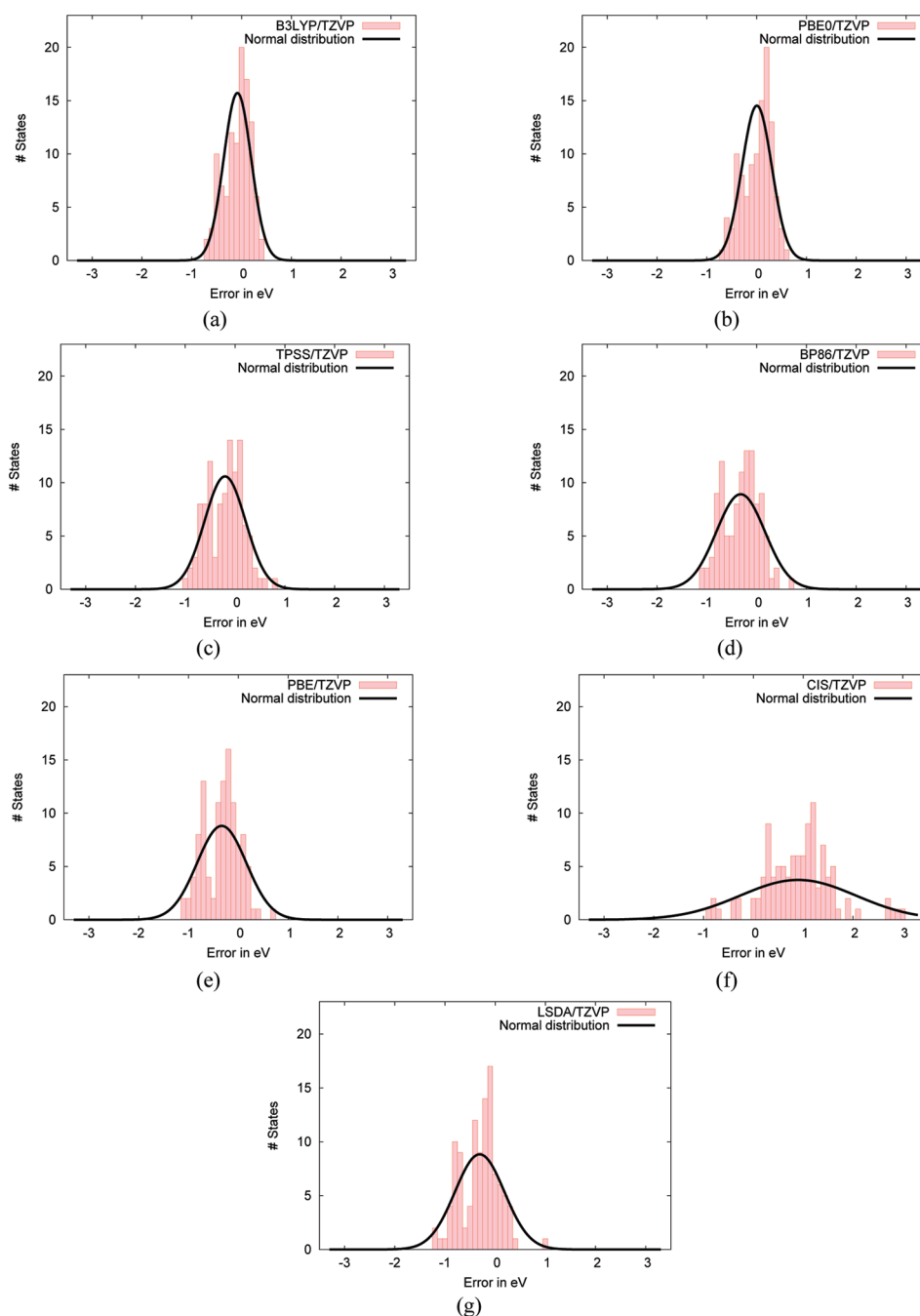


Figure 2. Error histograms of AEEs obtained with different functionals (see insets). Also shown are normal distributions with the computed ME and SD of the sample.

On our data set, the inclusion of ZPVE corrections does not significantly improve the overall accuracy of the B3LYP/TZVP AEEs, see Table 4, but it does not worsen them either. ZPVE corrections are a byproduct of molecular Hessian calculations, which are needed to identify minima. This is particularly important for excited states, where symmetry breaking and saddle point structures are common.

4.6. Excited State Relaxation. Do B3LYP AEEs deteriorate if the molecular structure changes strongly upon excitation? To address this important question, we computed the correlation coefficient between excited state relaxation energies and errors in the corresponding AEEs. If such a correlation existed, it could be

used as a diagnostic for AEEs. However, on our present data set, the correlation coefficient between excited state relaxation energies and AEE errors is -0.29 ; i.e., there is essentially no correlation. This result suggests that the incorrect dissociation behavior of spin-restricted TDDFT potential energy surfaces¹²⁹ has little effect on the AEEs in our benchmark set.

5. DIFFICULT AND NOTEWORTHY CASES

5.1. Aminomethane: 3^1A . Diffuse functions are necessary for an adequate description of the 3^1A state of aminomethane: The B3LYP AEE changes from 6.37 eV to 6.22 eV when the basis set

Table 5. Excited States with Maximum Negative Deviation (MND) and Maximum Positive Deviation (MPD) from the Experimental AEE for Each Level of Theory^a

level	MND	MPD
B3LYP/SV(P)	p-benzoquinone: 1^1B_{1g} −0.69	benzene: 1^1B_{1u} 0.46
B3LYP/SVPD	p-benzoquinone: 1^1B_{1g} −0.78	benzene: 1^1B_{1u} 0.37
B3LYP/TZVP	p-benzoquinone: 1^1B_{1g} −0.73	benzene: 1^1B_{1u} 0.42
B3LYP/TZVPD	p-benzoquinone: 1^1B_{1g} −0.74	benzene: 1^1B_{1u} 0.40
PBE0/TZVP	quinoline: $1^3A'$ −0.66	CuH: $2^1\Sigma^+$ 0.63
TPSS/TZVP	p-benzoquinone: 1^1B_{1g} −1.05	NH: $1^3\Pi$ 0.79
BP86/TZVP	p-benzoquinone: 1^1B_{1g} −1.14	NO ₃ : $1^2E'$ 0.71
PBE/TZVP	p-benzoquinone: 1^1B_{1g} −1.14	NO ₃ : $1^2E'$ 0.74
LSDA/TZVP	quinoline: 2^1A −1.19	NO ₃ : $1^2E'$ 1.00
CIS/TZVP	N ₂ : $1^1\Delta_u$ −0.85	Cu ₂ : $1^1\Pi_u$ 3.04

^a All values are in eV.

Table 6. Performance of AEEs Obtained from Various Exchange-Correlation Functionals and CIS Using def2-TZVP Basis Sets^a

method	MAE	SD	MaxAE
B3LYP	0.22	0.26	0.64
PBE0	0.25	0.30	0.66
TPSS	0.30	0.36	0.99
BP86	0.30	0.37	1.03
PBE	0.29	0.36	1.07
LSDA	0.30	0.38	1.31
CIS	0.55	0.73	2.14

^a For each method, a constant shift has been applied to make the ME vanish. See Table 3 for further explanation. All values are in eV.

size is increased from the def2-SV(P) basis set to def2-SVPD and from 6.36 eV to 5.97 eV when enlarging the def2-TZVP basis set to def2-TZVPD. This is consistent with a Rydberg $\pi \rightarrow s$ excitation.¹³⁰ The agreement of the nonaugmented basis set result with the experimental value of 6.22 eV⁸⁷ is fortuitous.

5.2. Benzene: 1^1B_{1u} . The excited state potential energy surfaces of benzene are a challenge for excited state methods.^{75,131,132} The PBE0 AEE of 5.26 eV and the CC2 AEE of 4.92 eV⁷⁵ are larger than the experimental AEE of 4.72 eV. The corresponding CASPT2 result of 4.37 eV is too small and might approach the CC2 result when the modified zeroth-order Hamiltonian is used.^{132–134} B3LYP, CC2, and CASPT2 excited state bond lengths agree up to 1 pm.

5.3. CuH: $2^1\Sigma^+$. The ground state of CuH has $1^1\Sigma^+$ symmetry and is dominated by a $2s(sd^{10})$ configuration of the copper atom.^{135,136} For the $2^1\Sigma^+$ state, multireference configuration interaction calculations show a mixed d^9-d^{10} character at the equilibrium bond lengths and strong spin-orbit coupling to the $1^3\Pi$ state.¹³⁵ Within TDDFT, all functionals yield AEEs within 0.1 eV of the experiment. An exception is PBE0 with an AEE too large by 0.63 eV. Diffuse augmentation is necessary but deteriorates the agreement with experimental results. Upon excitation, the dipole moment increases by ~ 2.4 D in the B3LYP calculation. The charge-transfer character of the excitation, the multiconfiguration character of the ground state, and the large spin-orbit coupling make this system difficult for response methods.^{135,137}

Table 7. Correlation Coefficients ρ and Maximum Absolute Deviations $|\Delta_{\max}|$ (in eV) between AEEs Computed with Different Density Functionals

(a) correlation matrix						
	B3LYP	PBE0	TPSS	BP86	PBE	LSDA
B3LYP	1	0.9983	0.9909	0.9888	0.9881	0.9832
PBE0	0.9983	1	0.9880	0.9849	0.9836	0.9781
TPSS	0.9909	0.9880	1	0.9979	0.9971	0.9912
BP86	0.9888	0.9849	0.9979	1	0.9997	0.9964
PBE	0.9881	0.9836	0.9971	0.9997	1	0.9975
LSDA	0.9832	0.9781	0.9912	0.9964	0.9975	1
(b) maximum absolute deviations matrix						
	B3LYP	PBE0	TPSS	BP86	PBE	LSDA
B3LYP	0	0.48	0.69	0.93	0.94	0.91
PBE0	0.48	0	0.95	1.16	1.17	1.14
TPSS	0.69	0.95	0	0.77	0.79	0.81
BP86	0.93	1.16	0.77	0	0.19	0.40
PBE	0.94	1.17	0.79	0.19	0	0.36
LSDA	0.91	1.14	0.81	0.40	0.36	0

5.4. Cu₂: $1^1\Pi_u$. For the $1^1\Pi_u$ state of Cu₂, large basis sets are essential to reproduce the experimental result of 2.71 eV.⁸⁶ The AEEs change from 2.02 eV to 2.15 eV and to 2.63 eV when going from def2-SV(P) to def2-SVPD¹³⁸ and to def2-TZVP. There is little further change when the basis set is increased to def2-TZVPD (2.62 eV) or QZVP⁴³ (2.60 eV). The TDDFT excitation vectors are dominated by one single-electron excitation, whereas the CIS excitation vector shows a multiconfigurational state 3.04 eV above the experimental AEE. The Hartree-Fock and hybrid functional ground state place the bonding $4s \sigma_g$ orbital above all antibonding $3d$ orbitals.¹³⁹ The results are stable with respect to tightened ground state convergence criteria.

5.5. p-Benzoquinone: 1^1B_{1g} . The AEE of p-benzoquinone is underestimated by more than 0.6 eV at the TDDFT level and overestimated by 0.8 eV at the CIS level. While the optimized ground state has C–C bond lengths between 134 and 148 pm, the optimized excited state has C–C bond lengths between 141 and 144 pm. This indicates a reduction of the quinoidal character upon excitation, and an excitation into a hydroquinone like state, in agreement with CASPT2 results of Weber and co-workers.¹⁴⁰

5.6. Quinoline: $1^3A'$. The ground state of quinoline shows a near triplet instability¹⁴¹ at the B3LYP level. This is reflected in the weak performance of the density functional methods for the lowest triplet state $1^3A'$. The quality of the singlet excitations appears to be unaffected by nearby triplet instabilities.

5.7. NO₃: $1^2E'$. The AEE predicted by CIS is lower than the experimental value of 1.87 eV¹⁰⁴ by 0.76 eV. Since CIS usually overestimates excitation energies, this underestimation indicates a multiconfigurational ground state. The underestimation of the CIS AEE is in line with the strong vibronic coupling between the ground and $1^2E'$ excited state.¹⁴² Inclusion of double and higher excitations will mostly lower the ground state energy and increase the AEE. Multireference contributions are noticeable but not dominant in the excited state.¹⁴³ Nonhybrid functionals overestimate the AEE by 0.71–1.00 eV, indicating that NO₃ is too electron-rich for a reliable description by semilocal functionals.

Table 8. Composition of the 15-Membered Subset and Computed AEEs in eV

molecule	state	experiment	CC2/TZVPD	B3LYP/TZVP	reference
acetaldehyde	2 ¹ A	3.69	3.70	3.81	63
anisole	2 ¹ A	4.51	4.60	4.73	69
benzene	1 ¹ B _{1u}	4.72	4.94	5.14	75
benzophenone ketyl radical	2 ² A	2.29	2.84	2.31	77
BF	1 ¹ Π	6.34	6.39	6.14	71
C ₂ H ₂	2 ¹ A	5.23	5.36	4.72	80
cinnoline	1 ¹ A''	2.82	2.77	2.43	83
CO	1 ¹ Π	8.07	7.99	7.99	71
glyoxal	1 ¹ A _u	2.72	2.72	2.43	80
p-diethynylbenzene	1 ¹ B _{2u}	4.25	4.59	4.17	110
propynal	2 ¹ A	3.24	3.27	3.25	112
pyridone lactim	2 ¹ A	4.48	4.70	4.60	115
quinoline	2 ¹ A	3.99	3.66	3.51	116
ScO	1 ² Π	2.04	2.02	2.14	118
VO	1 ⁴ Π	1.56	1.96	1.41	118

5.8. Tetrazine: 2¹A. Hybrid functionals underestimate the AEE by ~0.26 eV and slightly outperform the CASPT2 result of 1.94 eV of Schütz et al.¹⁴⁴ The TDDFT excitation vectors tend to converge to solutions dominated by many different single excitations with default convergence settings. Excitation vectors dominated by one single excitation in agreement with experimental results are obtained with tighter energy and density convergence criteria in the ground state calculation (10⁻⁸ au).

6. STRUCTURAL CHANGES

The main symmetry breaking structural changes encountered upon excited state relaxation fall into three groups:

- (1) out-of-plane bending of a C=O or C=S group
- (2) planarization of an amino group
- (3) others

In group 1 systems, a $\pi \rightarrow \pi^*$ excitation of a C=O or C=S chromophore leads to a $sp^2 \rightarrow sp^3$ rehybridization of the carbon atom. In group 2 molecules, planarization is caused by $s \rightarrow p$ promotion of an electron occupying a nitrogen lone pair orbital. The nonplanarity of excited ketyl groups and the planarity of excited ammonia is discussed by Herzberg.²³ The four other cases are the twisted intermolecular charge-transfer state of dimethylaminobenzonitrile,³² the bend excited states of HCN and C₂H₂,²³ and biphenyl. In biphenyl, the ground state is nonplanar due to the steric repulsion of hydrogen atoms in the adjacent rings. The excitation to the 1¹B₁ state takes place from an orbital that is antibonding between the rings to an orbital that is bonding between the rings. This causes the phenyl–phenyl bond to shorten by 7 pm to 141 pm upon excitation and a twist of the rings to a nearly planar structure.¹⁴⁵

7. REDUCED DATA SET

7.1. Composition. Out of the original 109-membered set, we generated a reduced set containing only 15 states. The reduced set was required to approximately reproduce the B3LYP/TZVP ME, MAE, and SD of the original set and cover a variety of different chromophores and states but was randomly chosen otherwise. The resulting reduced data set is shown in Table 8,

Table 9. Performance of Computed AEEs on the Reduced 15-Membered Subset^a

(a) full subset			
method	MAE	ME	SD
B3LYP/TZVP	0.21	-0.08	0.28
CC2/TZVP	0.21	0.16	0.29
CC2/TZVPD	0.17	0.10	0.24
ADC(2)/TZVP	0.23	0.14	0.34
(b) subset after exclusion of VO and benzophenone ketyl radical			
method	MAE	ME	SD
B3LYP/TZVP	0.23	-0.08	0.30
CC2/TZVP	0.16	0.10	0.20
CC2/TZVPD	0.12	0.05	0.17
ADC(2)/TZVP	0.15	0.04	0.20

^a See Table 3 for further explanation. All values are in eV.

and results for density functionals other than B3LYP and CIS are available as Supporting Information.

Due to its small sample size, the reduced set cannot replace the original 109 set. The aim of the reduced set is to provide a convenient and fast initial screening test for new methods, which is desirable for computationally demanding theories.

7.2. Performance of CC2 and ADC(2). To illustrate the use of the reduced data set, we investigate the performance of the correlated wave function methods CC2 and ADC(2) on the reduced set (Table 9). CC2 and ADC(2) behave similarly. On average, CC2 overestimates excitation energies, as reflected in a positive ME. However, the addition of diffuse functions considerably lowers the CC2 ME, MAE, and SD,^{22,146} and CC2/TZVPD outperforms all other methods on the reduced data set.

Scrutiny of the results in Table 8 shows that the largest CC2 errors occur for the benzophenone ketyl radical and for the 1⁴Π state of VO. It can be argued that CC2 and ADC(2) are not expected to work for these open-shell systems whose ground states show some amount of multireference character. Diagnostic criteria to identify such cases are available,^{118,147} and thus it can

Table 10. Comparison of Errors in AEEs Obtained from Various Exchange-Correlation Functionals and CIS Using def2-TZVP Basis Sets on the 15-Membered Subset for B3LYP and PBE0 Equilibrium Structures^a

	Structure	B3LYP	PBE0	TPSS	BP86	PBE	LSDA	CIS
ME	B3LYP	-0.08	-0.01	-0.21	-0.34	-0.35	-0.36	0.97
ME	PBE0	-0.08	-0.01	-0.20	-0.33	-0.34	-0.35	1.00
MAE	B3LYP	0.21	0.22	0.26	0.38	0.38	0.42	1.07
MAE	PBE0	0.21	0.22	0.26	0.37	0.38	0.41	1.09
SD	B3LYP	0.28	0.28	0.38	0.50	0.51	0.57	1.24
SD	PBE0	0.28	0.28	0.38	0.49	0.51	0.56	1.24

^a See Table 3 for further explanation. All values are in eV.

be justified to consider a “single-reference” subset which does not contain VO and benzophenone ketyl. Whereas the B3LYP errors are essentially unchanged, CC2/TZVPD performs very well on the single-reference subset, yielding a MAE of 0.12 eV.

7.3. Sensitivity to Equilibrium Structures. To evaluate the sensitivity of our results to errors in the B3LYP ground and excited state structures, the reduced data set was evaluated using fully optimized PBE0 structures. Table 10 compares the mean errors, mean absolute errors, and standard deviations for B3LYP and PBE0 structures on the reduced data set. The results are virtually identical, with maximum deviations of 0.03 eV for CIS and 0.01 eV for density functional methods. This relative insensitivity to the molecular structure reflects the quadratic dependence of adiabatic excitation energies on nuclear displacements, because both the ground and the excited state are in a minimum for adiabatic excitations. Thus, errors in the underlying equilibrium structures are less of a concern here than for vertical excitation energies, which depend linearly on nuclear displacements.

8. CONCLUSION

“There are lies, damned lies, and statistics.” – It is possible to construct excited state “benchmarks” that (dis)favor a particular method or conclusion. For example, a data set containing mostly Rydberg and charge-transfer excitations will make common density functionals look bad, while single-reference wave function methods will perform poorly for transition metal compounds and radicals with multireference ground states. This is consistent with the work of Jacquemin and co-workers,¹⁷ who observe a significant dependence of their statistical error measures on the subset of molecules or states considered. We have biased the present benchmark set toward systems and states where accurate spectroscopic data are available. A central new aspect is the focus on adiabatic rather than vertical excitation energies which are rigorously defined by differences of observable vibronic energy levels. The main remaining sources of inaccuracies on the theoretical side are the quality of the B3LYP/TZVP structures used and the neglect of relativistic and beyond-Born–Oppenheimer corrections. We estimate^{148,149} that these effects are 0.05 eV or less on average.

None of the methods tested here comes close to the desirable “chemical accuracy” of ~ 0.05 eV in excitation energies. Clearly, there is ample room for improvement. We hope that the present work will catalyze the development of excited state methods by providing a critical, accurate, and easy to use benchmark.

■ ASSOCIATED CONTENT

S Supporting Information. Tables of AEEs calculated with different functionals and basis sets. Table of the ZPVE for the ground and excited states including the ZPVE correction. Table with VEEs calculated at the ground state minimum, fluorescence energies, and excited state relaxation energies. Statistical parameters of the reduced data set. Figure with the structures of all molecules (Lewis notation). Coordinates of all ground and excited state structures. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: filipp.furche@uci.edu.

■ ACKNOWLEDGMENT

We thank Jefferson E. Bates for providing us with his implementation of TPSS excitation energies. This work was supported by the Center for Functional Nanostructures (CFN) of the Deutsche Forschungsgemeinschaft (DFG) within project C3.9, and by the Center of Chemical Innovation “Chemistry at the Space-Time Limit” (CaSTL) of the NSF, grant no. CHE-0802913. F.F. acknowledges additional support by the AirUCI Environmental Molecular Sciences Institute funded by the NSF, grant no. CHE-0909227.

■ REFERENCES

- (1) Adamo, C.; Scuseria, G. E.; Barone, V. *J. Chem. Phys.* **1999**, *111*, 2889–2899.
- (2) Tozer, D. J.; Amos, R. D.; Handy, N. C.; Roos, B. O.; Serrano-Andres, L. *Mol. Phys.* **1999**, *97*, 859–868.
- (3) Jaramillo, J.; Scuseria, G. E. *Theor. Chem. Acc.* **2000**, *105*, 62–67.
- (4) Parac, M.; Grimme, S. *J. Phys. Chem. A* **2002**, *106*, 6844–6850.
- (5) Fabian, J.; Diaz, L. A.; Seifert, G.; Niehaus, T. *THEOCHEM* **2002**, *594*, 41–53.
- (6) van Faassen, M.; de Boeij, P. L. *J. Chem. Phys.* **2004**, *120*, 8353–8364.
- (7) Grimme, S.; Neese, F. *J. Chem. Phys.* **2007**, *127*, 154116.
- (8) Jacquemin, D.; Perpète, E. A.; Vydrov, O. A.; Scuseria, G. E.; Adamo, C. *J. Chem. Phys.* **2007**, *127*, 094102.
- (9) Jacquemin, D.; Perpète, E. A.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105.
- (10) Jacquemin, D.; Perpète, E. A.; Scuseria, G. E.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2008**, *4*, 123–135.
- (11) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.
- (12) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103.
- (13) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.
- (14) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.
- (15) Gilbert, A. T. B.; Besley, N. A.; Gill, P. M. W. *J. Phys. Chem. A* **2008**, *112*, 13164–13171.
- (16) Goerigk, L.; Moellmann, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4611–4620.
- (17) Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420–2435.
- (18) Fabian, J. *Dyes Pigm.* **2010**, *84*, 36–53.
- (19) Caricato, M.; Trucks, G. W.; Frisch, M. J.; Wiberg, K. B. *J. Chem. Theory Comput.* **2010**, *6*, 370–383.
- (20) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2010**, *6*, 1532–1537.

- (21) Silva-Junior, M. R.; Thiel, W. *J. Chem. Theory Comput.* **2010**, *6*, 1546–1564.
- (22) Silva-Junior, M. R.; Sauer, S. P. A.; Schreiber, M.; Thiel, W. *Mol. Phys.* **2010**, *108*, 453–465.
- (23) Herzberg, G. *Molecular Spectra and Molecular Structure*; Van Nostrand and Reinhold, New York, 1966; , Vol. III, pp 172–173.
- (24) Dierksen, M.; Grimme, S. *J. Chem. Phys.* **2004**, *120*, 3544–3554.
- (25) Christiansen, O.; Koch, H.; Jørgensen, P. *J. Chem. Phys.* **1995**, *103*, 7429–7441.
- (26) Perić, M.; Marian, C. M.; Peyerimhoff, S. D. *J. Chem. Phys.* **2001**, *114*, 6086–6099.
- (27) Neugebauer, J.; Baerends, E. J.; Nooijen, M. *J. Phys. Chem. A* **2005**, *109*, 1168–1179.
- (28) Jankowiak, H.; Stuber, J. L.; Berger, R. *J. Chem. Phys.* **2007**, *127*, 234101.
- (29) Diehl, F. P.; Roos, C.; Jankowiak, H.; Berger, R.; Köhn, A.; Diezemann, G.; Basché, T. *J. Phys. Chem. B* **2010**, *114*, 1638–1647.
- (30) Berger, R.; Fischer, C.; Klessinger, M. *J. Phys. Chem. A* **1998**, *102*, 7157–7167.
- (31) Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2002**, *117*, 7433–7448.
- (32) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2005**, *122*, 064105.
- (33) Grimme, S.; Izgorodina, E. I. *J. Chem. Phys.* **2004**, *305*, 223–230.
- (34) Jacquemin, D.; Perpète, E. A.; Scalmani, G.; Frisch, M. J.; Assfeld, X.; Ciofini, I.; Adamo, C. *J. Chem. Phys.* **2006**, *125*, 164324.
- (35) Hellweg, A.; Grün, S. A.; Hättig, C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4119–4127.
- (36) Helgaker, T.; Klopper, W.; Tew, D. P. *Mol. Phys.* **2008**, *106*, 2107–2143.
- (37) Marques, M. A. L.; Ullrich, C. A.; Nogueira, F.; Rubio, A.; Burke, K.; Gross, E. K. U. *Time-Dependent Density Functional Theory: Lecture notes in physics*; Springer: Heidelberg, Germany, 2006.
- (38) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6807.
- (39) Maurice, D.; Head-Gordon, M. *Mol. Phys.* **1999**, *96*, 1533–1641.
- (40) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409–418.
- (41) Schirmer, J. *Phys. Rev. A* **1982**, *26*, 2395–2416.
- (42) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (43) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.
- (44) Treutler, O.; Ahlrichs, R. *J. Chem. Phys.* **1995**, *102*, 346–354.
- (45) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2010**, *133*, 134105.
- (46) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–124.
- (47) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249.
- (48) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (49) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- (50) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (51) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (52) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.
- (53) Foresman, J. B.; Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *J. Phys. Chem.* **1992**, *96*, 135–149.
- (54) Hättig, C.; Weigend, F. *J. Phys. Chem.* **2000**, *113*, 5154–5162.
- (55) Köhn, A.; Hättig, C. *J. Chem. Phys.* **2003**, *119*, 5021–5036.
- (56) Hättig, C. *Adv. Quantum Chem.* **2005**, *50*, 37–60.
- (57) TURBOMOLE 6.2; TURBOMOLE GmbH: Karlsruhe, Germany, 2010. <http://www.turbomole.com> (accessed July 2011).
- (58) Giugni, A.; Eramo, R.; Cavalieri, S.; Pietraperzia, G.; Becucci, M.; Gellini, C.; Moroni, L.; Salvi, P. R. *Chem. Phys. Lett.* **2000**, *330*, 315–324.
- (59) He, Y.; Wu, C.; Kong, W. *Chem. Phys. Lett.* **2004**, *391*, 38–43.
- (60) Cheatham, C. M.; Laane, J. *J. Chem. Phys.* **1991**, *94*, 7734–7743.
- (61) Huang, Y.; Arnold, S.; Sulkes, M. *J. Phys. Chem.* **1996**, *100*, 4734–4738.
- (62) Borisevich, N. A.; Povedailo, V. A.; Tolkachev, V. A.; Yakovlev, D. L. *Dokl. Phys.* **2006**, *51*, 17–21.
- (63) Baba, M.; Hanazaki, I.; Nagashima, U. *J. Chem. Phys.* **1985**, *82*, 3938–3947.
- (64) Hollas, J. M. *Spectrochim. Acta* **1963**, *19*, 1425–1441.
- (65) Alves, A. C. P.; Christoffersen, J.; Hollas, J. M. *Mol. Phys.* **1971**, *20*, 625–644.
- (66) Ramos, R. C.; Fujiwara, T.; Zgierski, M. Z.; Lim, E. C. *J. Phys. Chem. A* **2005**, *109*, 7121–7126.
- (67) Hubin-Franskin, M.-J.; Delwiche, J.; Giuliani, A.; Ska, M.-P.; Motte-Tollet, F.; Walker, I. C.; Mason, N. J.; Gingell, J. M.; Jones, N. C. *J. Chem. Phys.* **2002**, *116*, 9261–9268.
- (68) Sinclair, W. E.; Pratt, D. W. *J. Chem. Phys.* **1996**, *105*, 7942–7956.
- (69) Eisenhardt, C. G.; Pietraperzia, G.; Becucci, M. *Phys. Chem. Chem. Phys.* **2001**, *3*, 1407–1410.
- (70) Staicu, A.; Rouille, G.; Sukhorukov, O.; Henning, T.; Huisken, F. *Mol. Phys.* **2004**, *102*, 1777–1783.
- (71) Huber, K. P.; Herzberg, G. *Molecular Spectra and Molecular Structure*; Van Nostrand and Reinhold: New York, 1979; Vol. IV.
- (72) Ruth, A. A.; Kim, E.; Hese, A. *Phys. Chem. Chem. Phys.* **1999**, *1*, 5121–5128.
- (73) Silva, C. R.; Reilly, J. P. *J. Phys. Chem.* **1996**, *100*, 17111–17123.
- (74) Ohmori, N.; Suzuki, T.; Ito, M. *J. Phys. Chem.* **1988**, *92*, 1086–1093.
- (75) Christiansen, O.; Stanton, J. F.; Gauss, J. *J. Chem. Phys.* **1998**, *108*, 3987–4001.
- (76) Yamamoto, R.; Ishikawa, S.; Ebata, T.; Mikami, N. *J. Raman Spectrosc.* **2000**, *31*, 295–304.
- (77) Hamatani, S.; Tsuji, K.; Kawai, A.; Shibuya, K. *Phys. Chem. Chem. Phys.* **2003**, *5*, 1370–1375.
- (78) Im, H.-S.; Bernstein, E. R. *J. Chem. Phys.* **1988**, *88*, 7337–7347.
- (79) Hochstrasser, R. M. *Can. J. Chem.* **1961**, *39*, 765–772.
- (80) Herzberg, G. *Molecular Spectra and Molecular Structure*; Van Nostrand and Reinhold: New York, 1966; Vol. II.
- (81) Clouthier, D. J.; Karolczak, J. *J. Chem. Phys.* **1991**, *94*, 1–10.
- (82) Clouthier, D. J.; Ramsay, D. A. *Annu. Rev. Phys. Chem.* **1983**, *34*, 31–58.
- (83) Glass, R. W.; Robertson, L. C.; Merritt, J. A. *J. Chem. Phys.* **1970**, *53*, 3857–3864.
- (84) Wallin, S.; Koivisto, R.; Launila, O. *J. Chem. Phys.* **1996**, *105*, 388–397.
- (85) Chowdhury, P. H.; Merer, A. J.; Rixon, S. J.; Bernath, P. F.; Ram, R. S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 822–826.
- (86) Page, R. H.; Gudeman, S. S. *J. Chem. Phys.* **1991**, *94*, 39–51.
- (87) Nakajima, M.; Schmidt, T. W.; Sumiyoshi, Y.; Endo, Y. *Chem. Phys. Lett.* **2007**, *449*, 57–62.
- (88) 2,3-Diazabicyclo[2,2,1]hept-2-ene.
- (89) Steel, C.; Thomas, T. F. *Chem. Commun. (London)* **1966**, 900–902.
- (90) 4-(Dimethylamino)-4'-cyanostilbene.
- (91) Rijkenberg, R. A.; Bebelaar, D.; Buma, W. J.; Hofstraat, J. W. *J. Phys. Chem. A* **2002**, *106*, 2446–2456.
- (92) N,N-Dimethyl-p-phenylenediamine.
- (93) Ozeki, H.; Okuyama, K.; Takahashi, M.; Kimura, K. *J. Phys. Chem.* **1991**, *95*, 4308–4313.
- (94) Kauffman, J. F.; Côté, M. J.; Smith, P. G.; McDonald, J. D. *J. Chem. Phys.* **1989**, *90*, 2874–2891.
- (95) Ioannoni, F.; Moule, D. C.; Clouthier, D. J. *J. Phys. Chem.* **1990**, *94*, 2290–2300.
- (96) Leopold, D. G.; Pendley, R. D.; Roebber, J. L.; Hemley, R. J.; Vaida, V. *J. Chem. Phys.* **1984**, *81*, 4218–4229.
- (97) Humphrey, S. J.; Pratt, D. W. *J. Chem. Phys.* **1993**, *99*, 5078–5086.
- (98) Hollas, J. M. *Spectrochim. Acta* **1963**, *19*, 753–767.
- (99) 5-[2-(3-Ethyl-2-benzothiazolylidene)-ethylidene]-3-ethyl-2-thioxo-4-thiazolidi.
- (100) Takayanagi, M. *Chem. Phys. Lett.* **2002**, *366*, 525–530.
- (101) de Groot, M.; Gromov, E. V.; Köppel, H.; Buma, W. J. *J. Phys. Chem. B* **2008**, *112*, 4427–4434.
- (102) Reyle, C.; Brechignac, P. *Eur. Phys. J. D* **2000**, *8*, 205–210.
- (103) Glowina, J. H.; Riley, S. J.; Colson, S. D.; Nieman, G. C. *J. Chem. Phys.* **1980**, *73*, 4296–4310.

- (104) Nelson, H. H.; Pasternack, L.; McDonald, J. R. *J. Phys. Chem.* **1983**, *87*, 1286–1288.
- (105) Heimbrook, L. A.; Kohler, B. E.; Levy, I. J. *J. Chem. Phys.* **1984**, *81*, 1592–1598.
- (106) Lee, S. K.; Lee, G. W. *Chem. Phys. Lett.* **2005**, *410*, 6–10.
- (107) Liverman, M. G.; Beck, S. M.; Monts, D. L.; Smalley, R. E. *J. Chem. Phys.* **1979**, *70*, 192–198.
- (108) Horst, G. T.; Kommandeur, J. *Chem. Phys.* **1979**, *44*, 287–293.
- (109) Trommsdorff, H. P.; Kahane-Paillous, J. *Spectrochim. Acta* **1967**, *23A*, 1661–1670.
- (110) Stearns, J. A.; Zwier, T. S. *J. Phys. Chem. A* **2003**, *107*, 10717–10724.
- (111) Martinez, S. J., III; Alfano, J. C.; Levy, D. H. *J. Mol. Spectrosc.* **1992**, *152*, 80–88.
- (112) Brand, J. C. D.; Chan, W. H.; Liu, D. S.; Callomon, J. H.; Watson, J. K. G. *J. Mol. Spectrosc.* **1974**, *50*, 304–309.
- (113) Salvi, P. R.; Foggi, P.; Castellucci, E. *Chem. Phys. Lett.* **1983**, *98*, 206–211.
- (114) Bolovinos, A.; Tsekeris, P.; Philis, J.; Pantos, E.; Andritsopoulos, G. *J. Mol. Spectrosc.* **1984**, *103*, 240–256.
- (115) Nimlos, M. R.; Kelley, D. F.; Bernstein, E. R. *J. Phys. Chem.* **1989**, *93*, 643–651.
- (116) Innes, K. K.; Ross, I. G.; Moomaw, W. R. *J. Mol. Spectrosc.* **1988**, *132*, 492–544.
- (117) Dewey, H. J.; Hadley, S. G. *Chem. Phys. Lett.* **1972**, *17*, 574–577.
- (118) Bauschlicher, C.; Langhoff, S. *J. Chem. Phys.* **1986**, *85*, 5936–5942.
- (119) Hauge, R. H.; Hastie, J. W.; Margrave, J. L. *J. Phys. Chem.* **1968**, *72*, 3510–3511.
- (120) Syage, J. A.; Adel, F. A.; Zewail, A. H. *Chem. Phys. Lett.* **1983**, *103*, 15–22.
- (121) Pryor, B. A.; Palmer, P. M.; Chen, Y.; Topp, M. R. *Chem. Phys. Lett.* **1999**, *299*, 536–544.
- (122) Hoheisel, G.; Hese, A. *J. Mol. Spectrosc.* **2006**, *235*, 211–217.
- (123) Smalley, R. E.; Wharton, L.; Levy, D. H.; Chandler, D. W. *J. Mol. Spectrosc.* **1977**, *66*, 375–388.
- (124) Judge, R. H.; Moule, D. C.; Bruno, A. E.; Steer, R. P. *Chem. Phys. Lett.* **1983**, *102*, 385–389.
- (125) Bolovinos, A.; Philis, J.; Pantos, E.; Tsekeris, P.; Andritsopoulos, G. *J. Mol. Spectrosc.* **1982**, *94*, 55–68.
- (126) Syage, J. A.; Lambert, W. R.; Felker, P. M.; Zewail, A. H.; Hochstrasser, R. M. *Chem. Phys. Lett.* **1982**, *88*, 266–270.
- (127) Pushkarsky, M. B.; Mann, A. M.; Yeston, J. S.; Moore, C. B. *J. Chem. Phys.* **2001**, *115*, 10738–10744.
- (128) Tao, J.; Tretiak, S.; Zhu, J. X. *J. Chem. Phys.* **2008**, *128*, 084110.
- (129) Giesbertz, K. J. H.; Baerends, E. J. *Chem. Phys. Lett.* **2008**, *461*, 338–342.
- (130) Dunn, K. M.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 123–129.
- (131) Penfold, T. J.; Worth, G. A. *J. Chem. Phys.* **2009**, *131*, 064303.
- (132) Bernhardsson, A.; Forsberg, N.; Malmqvist, P.; Roos, B. O.; Serrano-Andrés, L. *J. Chem. Phys.* **2000**, *112*, 2798–2809.
- (133) Ghigo, G.; Roos, B. O.; Malmqvist, P. *Chem. Phys. Lett.* **2004**, *396*, 142–149.
- (134) Send, R.; Valsson, O.; Filippi, C. *J. Chem. Theory Comput.* **2011**, *7*, 444–455.
- (135) Marian, C. M. *J. Chem. Phys.* **1991**, *94*, 5574–5585.
- (136) Harrison, J. F. *Chem. Rev.* **2000**, *100*, 679–716.
- (137) Furche, F.; Perdew, J. P. *J. Chem. Phys.* **2006**, *124*, 44103.
- (138) Wang, X.; Wan, X.; Zhou, H.; Takami, S.; Kubo, M.; Miyamoto, A. *THEOCHEM* **2002**, *579*, 221–227.
- (139) Barden, C. J.; Rienstra-Kiracofe, J. C.; Schaefer, H. F., III. *J. Chem. Phys.* **2000**, *113*, 690–700.
- (140) Weber, J.; Malsch, K.; Hohlneicher, G. *Chem. Phys.* **2001**, *264*, 275–318.
- (141) Bauernschmitt, R.; Ahlrichs, R. *J. Chem. Phys.* **1996**, *104*, 9047–9052.
- (142) Stanton, J. F. *Mol. Phys.* **2009**, *107*, 1059–1075.
- (143) Eisfeld, W.; Morokuma, K. *J. Chem. Phys.* **2001**, *114*, 9430–9441.
- (144) Schütz, M.; Hutter, J.; Lüthi, H. P. *J. Chem. Phys.* **1995**, *103*, 7048–7057.
- (145) Mank, D.; Raytchev, M.; Amthor, S.; Lambert, C.; Fiebig, T. *Chem. Phys. Lett.* **2003**, *376*, 201–206.
- (146) Lehtonen, O.; Sundholm, D.; Send, R.; Johansson, M. P. *J. Chem. Phys.* **2009**, *131*, 024301.
- (147) Lee, T. J. *Chem. Phys. Lett.* **2003**, *372*, 362–367.
- (148) Eliav, E.; Kaldor, U. *Phys. Rev. A* **1994**, *50*, 1121–1128.
- (149) Handy, N. C.; Yamaguchi, Y.; Schaefer, H. F. *J. Chem. Phys.* **1985**, *84*, 4481–4484.

Determining the Numerical Stability of Quantum Chemistry Algorithms

Gerald Knizia,^{*,†} Wenbin Li,[‡] Sven Simon,[‡] and Hans-Joachim Werner[†]

[†]Institut für Theoretische Chemie, Universität Stuttgart, Pfaffenwaldring 55, D-70569 Stuttgart, Germany

[‡]Institut für Parallele und Verteilte Systeme, Universität Stuttgart, Universitätsstraße 38, D-70569 Stuttgart, Germany

ABSTRACT: We present a simple, broadly applicable method for determining the numerical properties of quantum chemistry algorithms. The method deliberately introduces random numerical noise into computations, which is of the same order of magnitude as the floating point precision. Accordingly, repeated runs of an algorithm give slightly different results, which can be analyzed statistically to obtain precise estimates of its numerical stability. This noise is produced by automatic code injection into regular compiler output, so that no substantial programming effort is required, only a recompilation of the affected program sections. The method is applied to investigate: (i) the numerical stability of the three-center Obara–Saika integral evaluation scheme for high angular momenta, (ii) if coupled cluster perturbative triples can be evaluated with single precision arithmetic, (iii) how to implement the density fitting approximation in Møller–Plesset perturbation theory (MP2) most accurately, and (iv) which parts of density fitted MP2 can be safely evaluated with single precision arithmetic. In the integral case, we find a numerical instability in an equation that is used in almost all integral programs. Due to the results of (ii) and (iv), we conjecture that single precision arithmetic can be applied whenever a calculation is done in an orthogonal basis set and excessively long linear sums are avoided.

1. INTRODUCTION

Numerical computations are done with number representations of limited accuracy. In special circumstances, an unfortunate amplification of rounding errors can occur, such that the final calculation result (for example, a total molecular energy) is much less accurate than expected or desired. Algorithms with this behavior are called numerically unstable. Unfortunately, in general it is very hard to determine if a concrete implementation of a quantum chemistry (QC) method suffers from numerical instabilities, and if it does, where the sources lie. In this article we present a numerical analysis method based on random rounding. This method is easy to apply and can answer both questions with minimal programming effort.

Various accuracy pitfalls arise from the properties of the ubiquitous floating point (FP) arithmetic, which violates common mathematical identities.¹ For example, the associativity law of addition does not hold for FP numbers a , b , c : In general, $(a + b) + c \neq a + (b + c)$, and which summation order is taken can have stark impact on the calculation result. For example, if we consider single precision FP numbers with about seven decimal digits of precision ($\epsilon \approx 6 \times 10^{-8}$) and set

$$a = 1.4142136 \quad b = -1.4142136$$

$$c = 1.2345678 \times 10^{-5}$$

then we get for $a + b + c$, depending on the summation order:

$$\begin{aligned} (a + b) + c &= 1.2345678 \times 10^{-5} \quad \text{or} \\ a + (b + c) &= 1.2397766 \times 10^{-5} \end{aligned} \quad (1)$$

In the second variant the last five digits are completely bogus; the result's relative accuracy is only 5×10^{-4} , despite the fact that a single precision variable could store results accurate to 6×10^{-8} . Under unfortunate conditions, such rounding errors can accumulate and amplify in nontransparent ways and induce an inaccurate final calculation result.

QC programs usually do not state how many digits of a calculation result can be trusted—because this information is not readily available and obtaining it used to require lots of work. For the same reason a detailed numerical analysis of a program is usually not done unless obvious numerical problems show up, and even then typically the problems and/or solutions are only published as side remarks, if at all (e.g., refs 2–9). While one can expect that time-proven programs are “accurate enough” in practice, this does not necessarily hold for newly developed programs. In particular, algorithms which rely on deep recurrence relations are prone to numerical problems (e.g., molecular integral evaluation for Gaussian basis sets),¹⁰ as are molecular dynamics algorithms which depend on time-reversibility (e.g., transition path sampling),¹¹ and algorithms employing any kind of large, semi-redundant basis set (e.g., density fitting basis sets in Hartree–Fock^{12,13} and MP2 theories,^{14,15} complementary auxiliary basis sets in F12 theory^{16–18} or nonredundant orbital rotations,^{19,20} or configuration spaces^{20,21} in multireference theories).

As another aspect, new computational hardware systems, like general purpose graphics processing units (GPUs), field-programmable gate arrays (FPGAs), or Cell processors, are seeing a rapid increase in popularity. For example, several groups have implemented parts of QC programs, like molecular integrals,^{22–25} Hartree–Fock or density functional theory,^{23,24,26–29} or density-fitted MP2,^{30–32} on GPUs or similar platforms.^{33,34} Unlike on central processing units (CPUs), where until very recently single precision arithmetic offered only minor processing speed benefits, arithmetic in single precision is much faster than in double precision on such novel hardware. This is the case because in this context double precision support is typically reduced to a (economical) minimum, because low-precision arithmetic can lead to distinct advantages in the hardware costs associated with

Received: April 7, 2011

Published: July 14, 2011

chip area, clock period, latency, and power consumption. For example, the hardware resources required for an addition circuit scale linearly with word length, while for a multiplication circuit they even scale quadratically.³⁵ In light of these hardware developments, it is worthwhile to investigate whether (and how) expensive parts of calculations, for example, the perturbative (T) triples correction³⁶ in coupled cluster theory, can be evaluated with sufficient accuracy also with single precision arithmetic. Some of the aforementioned articles^{24,26,32} also dealt with these issues; additionally ref 37 explicitly investigated whether parts of Cholesky-decomposition MP2 (CD-MP2) can be done in single precision arithmetic.

For these reasons we believe that better tools for numerical error analysis are required. In this article we describe one novel, easily applied approach, which can be executed without extensive source code modification.

So let us assume that we compute a FP number E_{dbl} with some program, and we want to find out how many digits of E_{dbl} can be trusted. Then traditionally one of the following techniques is used:

Comparison with a High Precision Reference. If the program can be changed to employ high-precision or multiprecision FP arithmetic instead of the standard double precision arithmetic, then we can in this way calculate a more accurate $E = E_{\text{hp}}$ in high precision again and see if it agrees with the double precision result E_{dbl} .

The main problem here is that the source code of the calculation program usually has to be modified, and this can imply a huge workload. (It should be noted that some recent compilers (e.g., Intel Fortran) can redefine intrinsic floating point data types to quadruple precision, thus potentially greatly simplifying the calculation of E_{hp} reference values. However, since this implies a change of the memory layout of the program, addressing, I/O operations, data compression, interfaces within the program to external programs and libraries, etc. may still need to be adjusted in nonstraightforward ways.) Additionally, high-precision arithmetic must be emulated and can become very slow. Furthermore, since only one E_{dbl} is available, and this result might by chance be very bad or very good, the precise size of the expected rounding errors is hard to estimate. Also, in some cases the convergence of the result with increasing precision can be misleading; an extreme example was constructed by Rump:^{38,39}

$$f = (333.75 - a^2)b^6 + a^2(11a^2b^2 - 121b^4 - 2) + 5.5b^8 + \frac{a}{2b}$$

where $a = 77617$ and $b = 33096$. When the computations are performed in single, double, and quadruple precision, the following results are obtained

$$\text{Single precision : } f_s = 1.172604$$

$$\text{Double precision : } f_d = 1.17260394005318$$

$$\text{Quad precision : } f_q = 1.1726039400531786318588 \dots$$

$$\text{Exact result : } f = -0.827396059946821368141 \dots$$

By looking at the first three results, the single precision result seems to be accurate, yet the correct answer is not even found with quadruple precision (it requires 122 mantissa bits to get the leading term correct). Thus, in pathological cases the comparison with a high-precision reference can lead to erratic results, unless

one can guarantee that the high-precision reference is really precise enough. The method we will present does not have this defect.

Interval Arithmetic (IA). IA is a well-known method of numerical stability analysis.⁴⁰ In IA, a single standard FP number is replaced by a pair of FP numbers representing its strict upper and lower bound; the basic arithmetic operations are defined on those intervals and incorporate rounding errors into the bounds. IA generates strict bounds on rounding errors, but it is less useful to obtain information on the practical numerical stability of algorithms. The strict bounds tend to exaggerate the actual accuracy loss due to rounding by orders of magnitude, since error cancellation cannot possibly occur.⁴¹ Additionally, IA is difficult to implement into an existing program, requiring large-scale changes to the entire source code, and IA is much slower than straight FP arithmetic.

Since both high-precision reference and IA computations require large modifications of the affected programs, they are not commonly applied in QC. The only actually applied “stability analysis” method we are aware of is to try if the same calculation gives different results if other compilers or other numerical libraries (BLAS etc.) are used. Obviously, the reliability of this approach is limited.

Instead of the sketched methods, we propose to use a modification of a little-known probabilistic error analysis technique by Vignes.^{42,43} The original method has been termed *contrôle et estimation stochastique des arrondis de calculs* (CESTAC) or in a slight modification *discrete stochastic arithmetic* (DSA).⁴⁴ These methods work by estimating rounding errors during the run time of a program in a statistical manner and thus attempt to model correct rounding error propagation. They have been used successfully for numerical analysis in a few previous cases.^{45–47} The original implementations of the DSA method are based on operator overloading, and thus are restricted to programming languages supporting this feature (e.g., not Fortran 77 or C, two popular languages for numerical software), and furthermore, they incur the same (or even more) program adjustment overhead as IA or high-precision reference calculations.

In contrast to CESTAC/DSA, our method is noninvasive (does not require source-code modification) and focuses on obtaining rounding error estimates of final calculation results, not of every single intermediate number. With our approach, accurate numerical stability estimates can be obtained quickly and with little programming effort. Additionally, our method works for most popular programming languages, as long as they are compiled.

2. METHOD

The goal of a numerical stability analysis is: (i) to assess how the tiny intermediate rounding errors of FP arithmetic combine into the uncertainty of the final calculation result and (ii) to quantify how uncertainties are introduced or amplified by the various subtasks of a calculation. Both questions can be answered *experimentally* by introducing random numerical noise into the performed FP arithmetic. Concretely, if we want to judge the numerical impact of a subtask S (say, the calculation of molecular integrals) on the final result E (say, a total molecular energy), we proceed as follows:

- For each FP operation in S , we add numerical noise on the same order of magnitude as the precision of the individual

FP numbers occurring. For a double precision number $f \cdot 2^m$, where f is the mantissa and m is the exponent, this magnitude is $\pm 2^{-53}$ ($\approx 10^{-15.95}$, in the 16th decimal digit).

- Because the noise is so small, under optimal conditions it does not affect the final calculation E noticeably. If, however, some amplification of the rounding errors takes place, then our introduced numerical noise is amplified in exactly the same way as the inherent finite precision rounding error (whose impact is what we actually want to quantify).
- If we now run the entire calculation of E multiple times, we will observe some numerical scattering in its result, which is caused by the noise introduced by S . By statistically analyzing the individual results E_i ($i = 1, \dots, n$) obtained in n runs of the program, we can estimate how many digits of accuracy in E are lost due to S . A first indication is given by the order of magnitude of the root-mean-square deviation (rmsd):

$$\text{rmsd}(E) = \sqrt{\frac{1}{n} \sum_i (E_i - \langle E_i \rangle)^2} \quad (2)$$

which approximates the standard deviation of the scattering of the calculation results. Due to the central limit theorem,⁴⁹ in most cases one can expect the scattering in E to follow a normal distribution. In this case a statistically sound reliability analysis can be obtained even for small n (say, 3, ..., 5) by weighting with Student's t -distribution.⁴⁹ If the scattering exactly follows a normal distribution around the exact result, then $2 \cdot \text{rmsd}(E)$ is a good estimate of the 95% confidence interval of the numerical error. But since the actual distributions often have wider tails, errors of $3 \cdot \text{rmsd}(E)$ or even $4 \cdot \text{rmsd}(E)$ can occasionally be seen.

- The original methods^{42–44,50} now go at great length to estimate the number of trustworthy digits from this deviation. We omit this part because: (i) it provides no additional information to the rmsd and (ii) there are cases when the obtained results are wrong (see below).

Until here we closely followed the CESTAC/DSA prescriptions in refs 42–44. But the indiscriminate nature of the method—every FP operation acquires noise—makes it possible to implement it not by modifying the source code of a program but rather the compilation process transcribing it. This is very important when the program in question consists of several million lines of code, like Molpro,⁵¹ which we will use as our test subject. As most compilers can produce an intermediate representation of a compiled program in terms of assembler code, and as the logical context of the FP operations is not required, this modification can even be accomplished without changing the compiler program itself. Rather than that, the intermediate assembly representation of the program being compiled is altered to include the noise production. (The assembly representation of a program contains a list of mnemonics of machine instructions with operands. It corresponds closely to the final binary program code. All instructions and accesses to registers and memory are explicit there.)

We implemented the concept for the Fortran 95/x86–64 and C++/x86–64 combination of platforms. Concretely, we developed a Python script which pretends to be either the g95 Fortran compiler or the GNU C/C++ compiler, respectively. The script accepts the same command line arguments as the compiler and can thus be seamlessly substituted for it when compiling the program to be analyzed. The script first calls the real g95 or g++

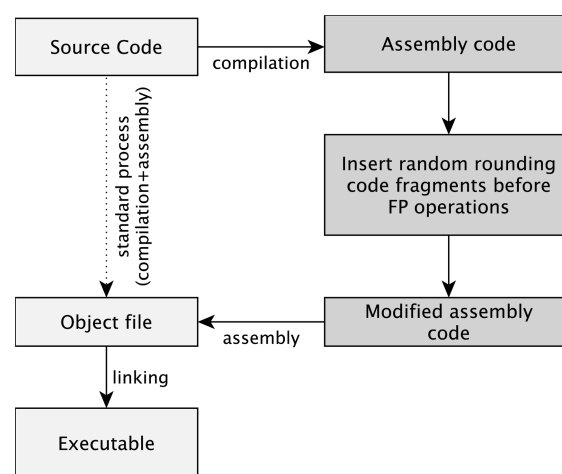


Figure 1. Modified compilation process for random rounding analysis. The script causes the compiler to generate assembly code, modifies the assembly to include the random rounding code, and then again uses the real compiler for assembly and linking.

compiler in order to compile input source code to assembler files. It then scans generated assembly code, and whenever it encounters a FP instruction from a predefined list, it injects code for changing the x87 rounding mode in a pseudorandom manner (see below). It then runs the real compiler again to assemble and/or link the resulting modified assembly files (see Figure 1). The script and support files will be made freely available on the Web at http://www.theochem.uni-stuttgart.de/random_rounding/.

The rounding mode of the FPU determines if intermediate FP computations round nonprecise digits toward the next representable number (“nearest”), toward $+\infty$ (“up”), or toward $-\infty$ (“down”), with “nearest” being the default choice for obvious reasons. But by changing this rounding mode randomly to up/down before the operations, computations are influenced only on the order of magnitude of the machine epsilon, which is exactly what we want for introducing noise. (On hardware platforms not supporting x87-style rounding modes, this behavior can be emulated by multiplying FP operands with $1 \pm \varepsilon$ where $\varepsilon \approx 2^{-53}$ before executing the original FP operations, as Vignes originally suggested.⁴³)

The technical details of how the random rounding mode is implemented are described in the Appendix. Here we just note that the implementation allows for switching between the rounding modes “nearest, 64bit”, “random up/down, 64bit”, “nearest, 32bit”, and “random up/down, 32bit” during the runtime of the program by simple function calls. To apply the proposed analysis method it is thus sufficient to recompile the parts of the program which are to be analyzed (or the whole program, if this is desired) with the noise-injecting meta-compiler and to call the functions for switching between the rounding modes at the appropriate places. Then the program is executed multiple times, and the random scattering of the numerical results is used to assess its numerical stability. The modified program is significantly slower than the original program, typically by a factor of 5–50. But as we will demonstrate later on, meaningful calculations can still be performed easily, since the bad scaling of QC algorithms allows one to offset this cost by choosing a somewhat smaller molecule for numerical testing. Note also that the slowdown is specific to the x87 pipeline hardware architecture and can be greatly reduced on other architectures.⁵²

2.1. Pitfalls of Random Rounding Analysis. In calculations with FP numbers, by far the most important and most transparent source of numerical problems is the elimination of significant digits due to subtraction, because this process leads to an amplification of relative accuracy loss (sketched around eq 1. In general we always eliminate relative accuracy when subtracting FP numbers of similar magnitude: If $a \approx -b$, then $a + b$ is inaccurate unless both a and b are exact.). This is called amplification error in this section. This process is very accurately modeled by the random rounding approach described above and is hard to describe in other ways. However, random rounding may lead to problems in two situations sketched in this section.

2.1.1. Iterative Processes. Although we have not personally seen this, it is possible that the analysis of iterative processes is jeopardized by their self-adjusting nature. However, provided that the same sequence of floating point operations is executed in each iteration, a numerical analysis can still be carried out easily. For this it is sufficient to reseed the pseudorandom number generator used for the rounding mode generation to the same seed value at the beginning of each iteration. In this way the same rounding sequences are used in each iteration during a single run of the program, and the results obtained in one run of the program are equivalent to a sequence one could get if the entire program was run on another machine (which would lead to different rounding errors, but in always the same way in each iteration). Additionally it might be advantageous to fix the number of iterations to some specified value because, otherwise, the number of iterations could differ between different runs in a sample if convergence thresholds are only closely met or missed.

2.1.2. Summing Up Large Amounts of Very Small Numbers. Apart from the elimination of significant digits due to subtraction, there is also another, qualitatively different kind numerical accuracy loss: the simple, nonamplified accumulation of less-than-machine precision (ϵ) sized individual errors which just happen to be done often enough to still cause significant deviations in final results. This case can occur in some quantum chemistry algorithms due to the immense amount of data handled, and it requires special consideration (think of the assembly of correlation energies from integrals and wave function amplitudes via four- or six-dimensional sums).

For an extreme example, let us assume that we work with single precision FP numbers (which have approximately 7.2 decimal digits of precision) and that we want to calculate the sum:

$$s = (\dots \underbrace{((1.0 + 1e-8) + 1e-8) + \dots}_{10^8 \text{ times}}) + 1e-8. \quad (3)$$

The exact value of the sum is $s_{\text{exact}} = 2.0$. But with default single precision FP arithmetic (i.e., using round to nearest) we would get $s_{\text{FP}} = 1.0$, because in each single operation $s := s + 1e-8$ the change in s is smaller than 0.5 units in the last place (ulp), and it remains at $s = 1.0$. Note that there is no amplification of rounding errors at all—every single operation has an accurate result. And yet the entire process combined results in a relative error of 100%. If we would execute this operation with random rounding instead of standard FP arithmetic, then we would get a different result. For each operation $s := s + a$, the random rounding process introduces a numerical scattering on the order of ± 0.5 ulp on s , the direction of which depends only on the sign of $(s + a) - s$ (here: always positive, since $a = 1e-8$). As described, typically this is exactly what one wants to have, but in the current case, this does not lead to a noticeable scattering in the sum s but rather to a

systematic bias in its probabilistic mean value. The random rounding result obtained is $s \approx 62.736$ with standard deviation $\delta s \approx 0.016$, instead of $s = 2.0$ with $\delta s \approx 1$ as one might expect. And in fact, one would get the same result of $s \approx 62.7$ when 10^8 times summing up $1e-9$, $1e-15$, or $1e-30$ to 1.0. In the latter cases the default round-to-nearest results of $s_{\text{FP}} = 1.0$ would be almost exact by accident.

We have seen that the mass truncation error is not properly reflected in the scattering produced by random rounding. However, the method is still useful for this case because a systematic bias is easily measured (by comparing the random-rounding mean value to the default round-to-nearest result), and its presence indicates that the investigated algorithm may suffer from mass truncation errors. While the possible truncation error can be greatly exaggerated by the difference between mean and round-to-nearest results (unlike the amplification error, the straight truncation error is reflected more in a worst-case than average way), the absence of such a bias is a very strong indication of the absence of truncation errors. Thus, if an algorithm shows neither sizable bias in the mean values nor sizable scattering, it is most likely numerically stable.

Note that a mass truncation error of this kind is: (i) rare and easily identified in the source code, due to the immense amount of data required to produce nonentirely negligible effects and (ii) easily fixed by employing a compensated summation algorithm^{53,54} (five lines of code!) or by performing the sum in higher precision.

Note also that an artificial version of the bias problem will arise when treating molecules with higher point group symmetry but using only the D_{2h} subgroups of this symmetry explicitly. In this case there are many almost zero values in the calculation (which would be exactly zero if wave functions were exactly converged). These values have no influence on default round-to-nearest results but may create systematic shifts in random rounding results. We recommend to avoid this problem by choosing suitable test molecules for analyzing algorithms.

3. APPLICATIONS

In order to demonstrate the practical feasibility of the method, we apply it to some numerical questions which arose during our own development activities on the Molpro program package.⁵¹ Concretely, these are: (i) Which parts of correlation calculations can safely be done in single precision floating point arithmetic, now that hardware is gaining momentum which actually profits from that?, (ii) How numerically accurate are molecular integrals for high angular momenta?, and (iii) How do various ways of implementing the density fitting equations perform in relation to one another in terms of accuracy? These questions are answered in the following sections.

3.1. Perturbative (T) Correction of CCSD(T). We investigated whether the perturbative (T) correction³⁶ of the coupled cluster singles and doubles (CCSD) method can be evaluated accurately with single precision FP arithmetic. This correction is a prime candidate for acceleration on new high-performance platforms, because it is very expensive and its working equations are simple and well-conditioned if evaluated in an orthogonal basis. Concretely, for a converged Hartree–Fock reference function the closed-shell working equations are^{55,56}

$$E_{(T)} = \frac{1}{3} \sum_{ijk}^{\text{occ vir}} \sum_{abc} (4W_{abc}^{ijk} + W_{bca}^{ijk} + W_{cab}^{ijk}) \times (V_{abc}^{ijk} - V_{cba}^{ijk}) / D_{abc}^{ijk} \quad (4)$$

Table 1. Numerical Stability of the Single Precision (T) Triples Correction^a

	SO ₂ /AVSZ	naphthalene/VDZ
Double Precision, Round to Nearest		
RHF	−547.319 262 404 907	−383.382 869 789 653
CCSD	−0.720 154 348 436	−1.347 263 987 000
(T)	−0.040 519 128 711	−0.063 745 284 695
Single Precision, Random Rounding		
(T), #1	−0.040 519 128 614	−0.063 745 284 643
(T), #2	−0.040 519 128 781	−0.063 745 284 703
(T), #3	−0.040 519 128 515	−0.063 745 284 682
(T), #4	−0.040 519 128 717	−0.063 745 284 712
(T), #5	−0.040 519 128 709	−0.063 745 284 708
(T), #6	−0.040 519 128 883	−0.063 745 284 648
(T), #7	−0.040 519 128 690	−0.063 745 284 757
(T), #8	−0.040 519 128 673	−0.063 745 284 659
mean	−0.040 519 128 698	−0.063 745 284 689
rmsd	0.000 000 000 102	0.000 000 000 036

^aThe first three lines give energies (in au) of standard double precision runs, and the next eight lines give individual values obtained with single precision and random rounding.

$$W_{abc}^{ijk} = P_{abc}^{ijk}[(bd|ai)T_{cd}^{kj} - (ck|jl)T_{ab}^{il}] \quad (5)$$

$$V_{abc}^{ijk} = W_{abc}^{ijk} + (bj|ck)t_a^i + (ai|ck)t_b^j + (ai|bj)t_c^k \quad (6)$$

$$D_{abc}^{ijk} = \varepsilon_i + \varepsilon_j + \varepsilon_k - \varepsilon_a - \varepsilon_b - \varepsilon_c \quad (7)$$

where i, j, k, l and a, b, c, d run over canonical occupied and virtual orbitals, respectively, t_a^i and T_{ab}^{ij} are the converged CCSD cluster amplitudes, ε_r are the canonical orbital energies, and

$$P_{abc}^{ijk}[X_{abc}^{ijk}] = X_{abc}^{ijk} + X_{acb}^{ikj} + X_{cab}^{kij} + X_{cba}^{kji} + X_{bca}^{jki} + X_{bac}^{jik} \quad (8)$$

is a permutation operator.

In order to quantify the rounding errors introduced by single precision arithmetic in the (T) correction, we calculated Hartree–Fock reference functions, CCSD wave functions, and transformed four-index integrals in standard double precision and then made multiple calculations of the (T) correction with single precision random rounding. The numerical stability was estimated from the numerical scattering produced in eight runs of the (T) correction as explained before. For the test cases, we chose SO₂/AVSZ,^{57,58} a small molecule with a large basis set, and a naphthalene/VDZ,⁵⁹ a larger molecule with a small basis set. The SO₂ molecular geometry was optimized on MP2/AVTZ, and naphthalene was optimized on DF-KS(PBE)⁶⁰/def2-TZVPP.^{61,62}

In these calculations it turned out to be essential that the final correlation energy (eq 4) is summed up in double precision when using large basis sets, because otherwise the previously discussed mass truncation error occurs and therefore the results become rather inaccurate (or, alternatively, a single precision compensated summation algorithm should be used^{53,54}). This summation is a $O(o^3v^3)$ process, with o/v the number of occupied/virtual orbitals, respectively, and thus its computational expense is insignificant compared to the $O(o^3v^4)$ and $O(o^4v^3)$ steps (eq 5) in the (T) correction. Similarly, for the case of CD-MP2

correlation energies, Vysotskiy et al. found that the final summation should be done in double precision.³⁷

The results obtained are given in Table 1. The numbers were obtained in the default (T) triples implementation⁶³ of Molpro, which does not take any special measures to preserve numerical accuracy; apart from switching back to double precision for the energy summation, no further adjustments were done. Nevertheless, even so the obtained (T) triples energies from single precision calculations were far more accurate than the about ≈ 0.1 kJ/mol ($40 \mu E_h$) which would be sufficient for single point calculations. In fact, we have seen not an amplification of rounding errors but rather a cancellation: The final results obtained are *more* accurate than the machine epsilon for single precision arithmetic ($\approx 6e-8$), since all the individual energy contributions in the $O(o^3v^3)$ sum are tiny and well-behaved. These are the entries the single precision arithmetic applies to because, as described, the final summation is done in double precision. This high accuracy is certainly unexpected. We cross-checked the result by performing additional calculations of (T) contributions in which all standard matrix multiplications have been replaced by a routine which explicitly converts its input to single precision FP numbers, calls sgemm (single precision BLAS matrix multiplication), and then converts the results back to double precision. These tests showed that the random rounding results are correct: deviations between final (T) energies with single and double precision matrix multiplications were in the 10th or 11th digit (e.g., for SO₂/AVQZ we obtained $-0.039\ 463\ 660\ 530$ au and $-0.039\ 463\ 660\ 633$ au for single and double precision matrix multiplications, respectively).

The results obtained here clearly show that the $O(N^7)$ steps of (T) calculations should be done with single precision arithmetic if that leads to large computational savings and that it is unlikely that this can cause problems in any situation, except maybe for the calculation of higher numerical derivatives.

3.2. Three-Index Molecular Integrals with High Angular Momenta. One of the most fundamental processes in QC is the calculation of integrals over Gaussian-type orbital (GTO) basis functions. The calculation of such integrals is based on deep recurrence relations, in which intermediate results for derivatives of integrals over lower angular momenta (AM) are linearly recombined into integrals over higher AM.⁶⁴ Integral evaluation for high AM (≥ 6) is thus prone to numerical problems, unless designed carefully.

Here we analyze the numerical properties of a very efficient method for evaluating three-center two-electron integrals $(ab|c)$, namely the Obara–Saika scheme^{65–67} in Ahlrichs’ three-center solid harmonic modification,⁶⁸ as implemented in Molpro’s AIC integral core.⁶⁹ Such integrals occur in both conventional and explicitly correlated QC methods involving the density fitting approximation (see below). For high-accuracy benchmark calculations it would be desirable to be able to calculate integrals with angular momenta of up to $(ki|l)$ (i.e., $l = 7, 6$, and 8), as they would occur in explicitly correlated QC methods applied with sextuple- ζ orbital basis sets. An interest in such highly accurate calculations was indicated in some previous articles.^{70–74}

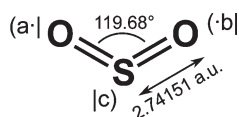
While it is straightforward to write code for calculating these integrals, the numerical properties of the resulting quantities are completely unknown, and furthermore, in this regard there exists no prior experience to draw upon. We thus applied our numerical analysis method to this integration algorithm in order to check its numerical properties. The concrete equations are noted in the Appendix, where the process is also described in more detail.

Table 2. Order of Magnitude (v) and Instability (r) of Integrals ($ab|c$) with $\alpha = \beta$ and $\gamma = 0.1$ on the SO_2 Geometry (a, b on O, c on S)^a

$\alpha = \beta = \dots$	(ss s)	(pp p)	(dd d)	(ff f)	(gg g)	(hh h)	(ii i)	(kk k)	(ll l)
$v, \alpha = \beta = 10.0$	1.17e-48	1.39e-47	1.40e-46	1.02e-45	5.57e-45	2.37e-44	8.16e-44	2.34e-43	5.71e-43
$r, \alpha = \beta = 10.0$	1.39e-13	1.42e-13	1.20e-13	1.61e-13	1.65e-13	3.39e-13	4.17e-13	2.14e-12	7.01e-12
$v, \alpha = \beta = 1.0$	9.65e-05	1.07e-04	1.03e-04	8.10e-05	5.51e-05	3.39e-05	1.97e-05	1.11e-05	6.19e-06
$r, \alpha = \beta = 1.0$	4.98e-15	3.70e-15	4.45e-15	2.71e-14	9.21e-14	6.69e-13	1.93e-12	1.12e-11	5.93e-11
$v, \alpha = \beta = 0.1$	2.03e+00	2.66e-01	8.38e-02	3.49e-02	1.98e-02	1.14e-02	7.81e-03	5.18e-03	3.99e-03
$r, \alpha = \beta = 0.1$	4.38e-15	3.28e-15	5.38e-15	3.47e-14	7.39e-14	3.37e-13	8.73e-13	2.15e-12	7.88e-12
$v, \alpha = \beta = 0.01$	2.88e+00	9.56e-02	2.46e-02	4.99e-03	1.47e-03	4.71e-04	1.47e-04	5.51e-05	1.82e-05
$r, \alpha = \beta = 0.01$	1.85e-15	1.90e-15	4.79e-15	1.08e-14	1.60e-14	4.71e-14	1.41e-13	1.25e-12	2.73e-12
$v, \alpha = \beta = 0.001$	1.10e+00	4.52e-03	1.07e-03	2.46e-05	6.60e-06	2.46e-07	6.23e-08	3.05e-09	7.07e-10
$r, \alpha = \beta = 0.001$	2.13e-15	1.65e-15	4.42e-15	1.21e-14	1.83e-14	6.43e-14	2.63e-13	4.81e-12	9.35e-12

^aWhere v is the rms batch average of the integral values (see text) and r the relative numerical instability (theoretical optimum: $1e-16$).

We tested the integration accuracy by calculating batches of primitive ($ab|c$) integrals, for a fixed geometry and varying exponents α, β, γ and angular momenta l_a, l_b, l_c (where $l_a \geq l_b$, see Appendix). (An integral batch is the collection of integrals for all combinations of solid harmonic components of a, b and c . E.g., in a (pd|f) batch there are $3 \cdot 5 \cdot 7 = 105$ entries.) The functions a and b were centered on the oxygen atoms of a SO_2 molecule, and the function c was centered on the sulfur atom:



This combination of distributing a, b , and c turned out to be the most challenging one because there is significant overlap between the ($ab|$) and the ($|c$) charge distributions, and yet there is nontrivial angular momentum transfer. The SO_2 geometry was chosen because we previously experienced numerical problems with SO compounds during early stages of the development of F12 methods; SO_2 , particularly, has short bond lengths and high exponents in the basis functions. For each test we calculated an integral batch 16 times with random rounding turned on and stored the calculated final integrals. From the 16 runs we then calculated root-mean-square (rms) deviations from the mean for each individual component of the integral batch. As a summary we report the value $v = \text{rms values}$ of the integrals (to give an indication of average order of magnitude of the integrals in the batch) and the value $r = \text{rms of the rms deviations}$ for the components, divided by v (r is thus some measure of the relative instability of the calculated values). The best possibly achievable rates are around $1e-16$, which is the machine precision. Any value of r below $1e-10$ is to be considered good enough for practical use.

Our analysis shows two main results: First, integrals of type ($as|c$) (i.e., where b is a s function) and general integrals ($ab|c$) where $\alpha \geq \beta$ can be computed accurately to high angular momenta, even if relatively low exponents are used on either center. For example, Table 2 gives the integral instabilities for $\gamma = 0.1$ and varying $\alpha = \beta$. We can see that, although there does seem to be an exponential increase of the relative instability with angular momentum, the actual values of r are still small ($<1e-10$) even for very high angular momenta ($l_a = l_b = l_c = 8$), for tiny absolute integral values ($v \approx 1e-45$ with $\alpha = \beta = 10$) and even for very

Table 3. Order of Magnitude (v) and Relative Instability (r) of Integrals ($gg|s$) with $\gamma = 0.5$ and Varying α, β on the SO_2 Geometry^a

(gg s), $\gamma = 0.5$	$\beta = 100.0$	$\beta = 10.0$	$\beta = 1.0$	$\beta = 0.1$
$v, \alpha = 100.0$	0.00e+00	6.95e-83	2.54e-10	1.37e-07
$r, \alpha = 100.0$		2.82e-13	4.23e-14	2.40e-11
$v, \alpha = 10.0$	6.95e-83	1.58e-42	4.25e-07	6.95e-05
$r, \alpha = 10.0$	4.25e-11	1.78e-13	2.61e-14	1.98e-13
$v, \alpha = 1.0$	2.54e-10	4.25e-07	8.72e-03	1.58e-02
$r, \alpha = 1.0$	4.93e-07	1.17e-10	9.86e-14	5.33e-14
$v, \alpha = 0.1$	1.37e-07	6.95e-05	1.58e-02	1.42e-01
$r, \alpha = 0.1$	7.81e-04	1.24e-07	2.25e-11	5.28e-14

^aNote that the table would be symmetric under optimal conditions.

diffuse exponents like $\alpha = \beta = 0.001$ (which are, in fact, so small that they are already far from reasonable).

The second main result, however, is disturbing. If the exponent β is much larger than the exponent α , then the integration becomes numerically unstable even for moderate angular momenta (for this it has to be noted that in the integration procedure, as is usually done, angular momentum is first accumulated on center A (where $l_a \geq l_b$) and then transferred to B in the final step, see Appendix). Table 3 shows the average values and instabilities of ($gg|s$) integrals with varying values of α and β and fixed γ . Note that integrals are supposed to be symmetric regarding the exchange of a and b . We can see that there is a major difference in the stabilities, depending on whether β/α is large or small, and that unacceptably large relative instabilities of almost $1e-3$ are already reached in semirealistic cases [$\beta = 100, \alpha = 0.1, (gg|s)$ integrals]. While this problem is unlikely to show up in standard calculations of valence correlation energies with conventional QC methods (due to the nonpresence of high-AM functions with steep exponents), there are other cases in which it does occur. For example, in F12 treatments, high-exponent high-AM functions are required for the resolution of the identity (RI) approximation, and these are affected by this issue; basis sets for all-electron calculations of higher elements include affected functions (e.g., ANO-RCC for the Rn atom⁷⁵ includes f

primitives with exponent 689.88), and calculations involving bond centered functions (often used for calculations of potential energy surfaces⁷⁶) are affected because distance between basis function centers is decreased, and thus, the unstable integrals have a higher relative impact on the energy. The problems get much worse with increasing ratio β/α and with increasing angular momentum l_a and l_b (the l_c dependence, however, is weak). In fact, for angular momenta k and l we have seen cases where the results were completely bogus, and the integral values suggested the presence of non-negligible integrals of size $1e-7$, where the actual (negligible) integral values were in the range of $1e-60$.

The reason for this asymmetric numerical instability turns out to lie in the contracted transfer equation eq 21, which is applied as final step in almost every efficient integral core in use (in both Rys quadrature^{2,77-80} based and Obara–Saika⁶⁵⁻⁶⁷ based cores). An analysis of the issue and its possible remedies are given in the Appendix. Our current understanding is that this problem cannot easily be solved without both additional computational cost and significant changes to the existing integral programs. An extensive literature search showed that this issue was discovered before,⁸¹ but it does not appear to be well-known.

3.3. Density Fitting Equations in DF-MP2. Density fitting (DF)⁸²⁻⁸⁵ is an approximation technique widely used to accelerate the evaluation of transformed four-index molecular integrals in a large variety of contexts.⁸⁶⁻⁹⁶ For example, in DF-MP2,⁸⁷ the two-electron exchange integrals $(ia|jb)$ are approximated as

$$(ia|jb) \approx \sum_A D_{ia}^A(A|jb) \quad (9)$$

where i, j denote occupied orbitals, a, b denote virtual orbitals, and A, B are basis functions from an auxiliary fitting basis set. The fitting coefficient D_{ia}^A approximates the $|ia\rangle$ orbital product density such that $|ia\rangle \approx \sum_A D_{ia}^A|A\rangle$, and it is formally determined by^{92,97}

$$D_{ia}^A = \sum_B [J^{-1}]_{AB}(B|ia) \quad (10)$$

$$J_{AB} = (A|B) \quad (11)$$

There are multiple ways to evaluate eqs 9 and 10 (in parentheses are the central LAPACK operations involved):

- (i) An inverse matrix J^{-1} is explicitly constructed from J 's LU decomposition (`dgetrf` and `dgetri`), and eqs 9 and 10 are evaluated as written.
- (ii) Equation 10 is evaluated as a system of linear equations using J 's LU decomposition and pivoted Gaussian elimination (`dgetrf` and `dgetrs`).
- (iii) A matrix $J^{-1/2}$ is evaluated from J 's spectral decomposition (`dsyev`), and eqs 9 and 10 are evaluated in the symmetric form:

$$(ia|jb) \approx \sum_C \tilde{D}_{ia}^C \tilde{D}_{jb}^C \quad (12)$$

$$\tilde{D}_{ia}^C = \sum_A (ia|A)[J^{-1/2}]_{AC} \quad (13)$$

- (iv) A matrix $J^{-1/2}$ is evaluated as in (iii), and eq 10 is calculated as

$$D_{ia}^A = \sum_B [J^{-1/2}]_{AB} \left(\sum_C [J^{-1/2}]_{BC}(C|ia) \right) \quad (14)$$

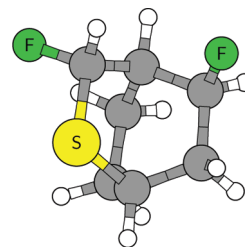


Figure 2. Cage-like molecule used for numerical testing of the DF-MP2 methods. The structure was chosen to be tightly packed yet realistic, in order to create a large redundancy when using diffuse basis sets.

(i.e., there are two transformations with $J^{-1/2}$ instead of one with J^{-1} as done in (i); apart from that, (i) and (iv) are identical).

The strategies (i–iii) have been used in various programs (e.g., (i) in refs 98 and 99 (ii) in refs 3 and 100–104 (iii) in refs 86 and 105), with (ii) and (iii) being the most common choices and (iv) is investigated out of curiosity. In any case, the closed-shell DF-MP2 correlation energy is given by

$$E_{\text{MP2}} = \sum_{ijab} \frac{[2(ia|jb) - (ib|ja)](ia|jb)}{\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j} \quad (15)$$

where the ϵ_r are the canonical orbital eigenvalues obtained in the Hartree–Fock calculation.

That (i) works poorly has been found previously, but to our knowledge, a detailed numerical stability study comparing the various methods of calculating $(ia|jb)$ has never been performed. With our statistical analysis method, we can easily determine the impact of the various choices on the final calculation result E_{MP2} .

To that end, we performed random rounding DF-MP2 calculations on an artificial cage-like molecule (Figure 2) using the aug-cc-pVQZ orbital basis set⁵⁷ (1264 basis functions) and the associated AVQZ/MP2FIT basis set of Weigend et al.¹⁰⁶ (2506 basis functions). As reference we used a tightly converged DF-RHF wave function, obtained with VQZ/JKFIT⁹⁸ as fitting set. This combination of molecule and orbital basis set was chosen because it presents a formidable challenge due to the high redundancy of the basis sets in the closely packed environment (e.g., the ratio between the highest and the lowest eigenvalue of the J_{AB} matrix is $1.2e+11$). For each of the four described ways of evaluating the DF equation systems, we performed eight calculations employing double precision random rounding throughout the entire DF-MP2 procedure and measured the final correlation energies obtained in the runs.

The results are shown in Table 4. Here we can see that, not surprisingly, the pivoted Gaussian elimination (ii) is the most accurate way of solving the DF equations, while method (i), based on the explicit calculation of J^{-1} , is the least accurate one. The numerical scattering obtained for method (i) is already large enough to cause significant problems when calculating numerical derivatives (for example, in calculations of vibrational frequencies or molecular properties). Method (i) is thus best avoided. The two methods based on the $J^{-1/2}$ decomposition, (iii) and (iv), are similarly stable, and although their accuracy is worse than that of the pivoted elimination, it is not much worse, and the errors introduced through them are still acceptable. This is perhaps most surprising for the asymmetric formula (iv) which, after all, is just a re-expression of method (i) with two $J^{-1/2}$ multiplications instead of one with J^{-1} . While this method is obviously not something one should use for new programs

Table 4. Numerical Stability of the Different Variants of Solving the Density Fitting Equations in DF-MP2^a

DF-MP2	(i) expl. J^{-1}	(ii) LU/pivoted elim.	(iii) sym. $J^{-1/2}$	(iv) asym. $J^{-1/2} \cdot J^{-1/2}$
#1	-2.004 579 669 344 919	-2.004 579 715 663 295	-2.004 579 715 665 265	-2.004 579 715 673 580
#2	-2.004 579 690 862 029	-2.004 579 715 663 242	-2.004 579 715 669 322	-2.004 579 715 679 456
#3	-2.004 579 777 141 105	-2.004 579 715 662 883	-2.004 579 715 699 197	-2.004 579 715 675 233
#4	-2.004 579 738 407 475	-2.004 579 715 662 652	-2.004 579 715 675 525	-2.004 579 715 669 635
#5	-2.004 579 722 504 768	-2.004 579 715 662 659	-2.004 579 715 661 420	-2.004 579 715 678 294
#6	-2.004 579 728 945 659	-2.004 579 715 663 033	-2.004 579 715 667 403	-2.004 579 715 670 761
#7	-2.004 579 756 469 682	-2.004 579 715 662 888	-2.004 579 715 660 053	-2.004 579 715 684 286
#8	-2.004 579 658 373 851	-2.004 579 715 661 850	-2.004 579 715 669 734	-2.004 579 715 663 029
mean	-2.004 579 717 756 186	-2.004 579 715 662 813	-2.004 579 715 670 990	-2.004 579 715 674 284
rmsd	0.000 000 039 039 628	0.000 000 000 000 426	0.000 000 000 011 605	0.000 000 000 006 174

^a The rows contain individual energy values (in au) obtained with random rounding for the cage molecule in Figure 2 with the aug-cc-pVQZ basis set. The standard values obtained with an unmodified code are $E(\text{DF} - \text{HF}) = -867.520895102141$ au and $E(\text{DF} - \text{MP2}) = -2.004579708652200$ au.

Table 5. Results Obtained for the DF-MP2 Correlation Energy of the Ethanol Molecule Using Regular (left) and Augmented (right) Basis Sets of Double- ζ to Quadruple- ζ Cardinalities^a

SP steps	card.	cc-pVnZ		aug-cc-pVnZ	
		mean	rmsd	mean	rmsd
int, DF, asm	DZ	-0.480 233 851	0.000 000 658	-0.507 113 084	0.000 001 830
DF, asm	DZ	-0.480 233 484	0.000 000 592	-0.507 109 818	0.000 001 928
asm	DZ	-0.480 233 835	0.000 000 005	-0.507 105 174	0.000 000 003
(none)	DZ	-0.480 233 824		-0.507 105 163	
int, DF, asm	TZ	-0.601 515 520	0.000 001 835	-0.647 189 353	0.009 018 198
DF, asm	TZ	-0.601 516 124	0.000 001 726	-0.636 752 456	0.009 256 710
asm	TZ	-0.601 514 922	0.000 000 002	-0.612 762 154	0.000 000 001
(none)	TZ	-0.601 514 915		-0.612 762 147	
int, DF, asm	QZ	-0.643 779 496	0.000 040 769	-84.512 702 083	14.999 787 051
DF, asm	QZ	-0.643 715 620	0.000 022 033	-60.558 243 374	5.506 838 535
asm	QZ	-0.643 521 307	0.000 000 001	-0.648 550 399	0.000 000 001
(none)	QZ	-0.643 521 307		-0.648 550 395	

^a Int (three-index integrals), DF (density fitting), and Asm (assembly) refer to the steps in the DF-MP2 calculation to which single precision random rounding was applied (the other steps used double precision round-to-nearest).

[it is both slower and less accurate than (ii)], it might be used as a quick fix to older programs employing the J^{-1} construction or, generally, any other case outside the context of density fitting where projection operators or inverse matrices are employed.

A comparison of the random rounding mean values with the results of an unmodified code (see caption of Table 4) shows a small bias of $\approx 1e-8$ H in all cases. As explained previously, this can indicate the presence of a mass truncation error, most likely in the energy summation. As this effect is reflected in a worst case way and still produces negligible errors, we can assume that DF-MP2 is generally numerically very stable as long as the explicit inversion of J_{AB} is avoided.

3.4. Speeding Up DF-MP2 with Mixed Precision Arithmetic. Vysostoskiy and Cederbaum³⁷ recently performed a numerical analysis of Choleski-decomposition MP2, where they found that CD-MP2 energies can still be evaluated accurately if the CD-analog of the DF assembly step eq 9 is performed in single precision arithmetic. We here perform a similar analysis for DF-MP2 but extend the coverage also to other steps of the total calculation.

The total DF-MP2 method consists of three steps with different formal scaling properties:

- (i) The evaluation of two- and three-index molecular integrals [an $O(N^3)$ process, where N is some measure of the molecular size].
- (ii) The calculation of the fitting coefficients D_{ia}^A according to eq 10 [$O(N^4)$].
- (iii) The assembly of the final exchange integrals according to eq 9 [$O(N^5)$].

While for small molecules integral evaluation (i) dominates the computational cost, for large molecules, eventually the assembly step (iii) becomes dominant (usually if more than ≈ 1000 – 2000 basis functions are involved).

We thus investigated if mixed precision schemes can be employed to evaluate the DF-MP2 energy accurately. Concretely, we tested the resulting numerical accuracy of the following combinations:

- (i) The total DF-MP2 method except for the energy summation (eq 15) and the calculation of $J^{-1/2}$ is evaluated in

single precision. That means that three-center integrals, fitting coefficients, and the final exchange integrals are all evaluated in single precision arithmetic.

- (ii) Unlike in (i), the integrals are evaluated in standard double precision arithmetic.
- (iii) Unlike in (i) and (ii), also the \tilde{D}_{ia}^A fitting coefficients (eq 13) are evaluated in double precision arithmetic. Only the final assembly of exchange integrals is done in single precision random rounding.

In any case, the final summation of the energies (eq 15) is performed in double precision arithmetic, as suggested in ref 37. Additionally, the matrix $[J]_{AB}$ and its inverse square root are calculated in double precision. The symmetric density fitting formula eq 12 is used. As a test subject we use the ethanol molecule. Again we perform eight calculations with random rounding for each of the three computational schemes (i–iii), and as basis sets, we employ VDZ–VQZ⁵⁹ and AVDZ–AVQZ⁵⁷ with the same fitting sets as before.

The results are noted in Table 5. We can see that for small, nonaugmented basis sets it appear as if reasonable accuracy can be obtained also when using single precision arithmetic globally. However, it is obvious that the accuracy decreases quickly with increasing size of the basis sets unless both the integrals and the density fitting equations are evaluated in double precision (this effect is even more pronounced for larger systems). When augmented basis sets are employed, the errors due to single precision arithmetic in integrals or density fitting equations are amplified enormously. Already at AVTZ level the results are completely bogus, and at AVQZ level the numerical errors exceed the actual calculation result by orders of magnitude.

If, however, only the assembly step is performed in single precision, then in all cases the results are very accurate. The main difference between the assembly step and the DF equations step is that the assembly step is done in an orthogonal basis set. This is the same effect we have already seen previously for the (T) correction. In fact, this leads us to the conjecture that likely in all parts of calculations which are done in an orthogonal basis, performing the actual core operations in single precision is numerically fine and produces acceptable numerical errors, as long as excessively long linear sums (and thus the mass truncation error) can be avoided. The reason for this is that in an orthogonal basis, the numerical size of the quantities in question carries information about their importance, while in a nonorthogonal basis, the relative importance may still be magnified or diminished by the same order of magnitude as the condition number of the space's overlap matrix.

4. CONCLUSIONS

Previously the means of numerical accuracy analysis for quantum chemistry programs were rather limited, and for this reason, such an analysis was rarely done in practice. However, as our discovery of the instability in the widely used angular momentum transfer equation showed, even established techniques can have serious defects in unusual circumstances. For this reason, better tools, especially simpler tools, and a more frequent application of them were in dire need.

Our applications clearly demonstrate that the proposed method is a practicable way of analyzing complex numerical software packages. Additionally, unlike the previous techniques, it is simple enough to be routinely applicable to newly developed software. We thus believe that our approach is a valuable tool in

the development of new QC algorithms and for the identification of persistent problems in algorithms already existing.

■ APPENDIX

A. Technical Details. On the technical side, the injected code changes the rounding mode by cycling through a table of 2^{16} x87 floating point control words (FPCW), one of which is loaded to the FPU before each arithmetic FP operation. A FPCW¹⁰⁷ is a 16 bit integer whose bits control different aspects of the operating mode of the FPU (e.g., two bits control the rounding mode, and two other bits control the calculation precision of the FPU). Since this table can be modified at run time, one can easily switch between normal calculations (all table entries are “64 bit precision, round to nearest”), random rounding (table contains “64 bit precision, round up” or “64 bit precision, round down” randomly distributed) or reduced precision (e.g., all entries “32 bit precision, round to nearest”), such that a selective investigation of sub-algorithms of the program in question can be done.

As noted previously, we generate the numerical noise by changing the FPU rounding modes randomly before each relevant floating point operation. This is realized by injecting the following x86-64 assembler code before each relevant FP operation. (These are the addition/subtraction operations `fadd`, `faddp`, `fiadd`, `fsub`, `fsubp`, `fisub`, `fsubr`, `fsubrp`, `fisubr` and the multiplication/division operations `fmul`, `fmulp`, `fdiv`, `fdivp`, `fdivr`, `fdivrp`, `fprem`, `fprem1`, `fmul`, `fdiv`, `fdivr`). We make sure that the compiler generates only x87 FPU code and does not perform the FP operations by some other means (e.g., SSE instructions):

```
xchg rax, [LC_CWINDEX]
lea ax, [rax+1]
fldcw word ptr [LC_CWDATA+2*rax]
xchg rax, [LC_CWINDEX]
```

The variable `LC_CWDATA` is the start address of an array with 2^{16} entries of FPU control words. The variable `LC_CWINDEX` is a counter which contains the current index in this table. In each execution of the injected code, the following happens:

- (i) The `xchg` instruction exchanges the current values of the general purpose register `rax` and of the memory location at `LC_CWINDEX`. This allows us to modify the table index (subsequently in `rax`) and to keep a copy of the original register value.
- (ii) The load effective address (`lea`) instruction increments `ax` (the lower 16 bits of `rax`) by one and resets it to zero if it overflows 2^{16} . The effect of the instruction is equal to “`add ax, 1`”, apart from `lea` not affecting CPU flags. Due to the implicit reset to zero on overflow, we can directly use `rax` as index for the 2^{16} entry rounding mode table.
- (iii) The load floating point control word (`fldcw`) instruction reads a FPCW from our table and activates its use by the FPU. Thus, all following FPU operations will be performed with a precision and rounding mode as specified by the FPCW, until another FPCW is loaded.
- (iv) Finally, the second `xchg` instruction resets the `rax` register back to its original value and at the same time writes back our updated table index.

All of the injected instructions preserve the CPU flags register, and thus can be inserted without affecting any of the surrounding code, apart from the FPU control word.

The entries of the table LC_CWDATA can be changed at runtime by the program itself. For example, when switching into “random rounding” mode, a function is called which distributes “round up” and “round down” control words randomly in the table using the KISS pseudorandom number generator¹⁰⁸ in David Jones JKISS modification.¹⁰⁹ In order to investigate an isolated sub-task of the program, the table is set to “random rounding” before beginning the sub-task and is reset to “default rounding” afterwards. This way all FP operations between the two calls are subject to the random rounding behavior, and all other FP operations proceed as normal.

In practice it may be sufficient to inject code for FPU random rounding mode changes only before additions and subtractions, since these operations are most prone to causing accuracy loss. Because the rounding mode has a “state nature” and remains active until set to something different, the other instructions would still be affected by random rounding, since they occur interleaved with addition/subtraction operations. Here, however, the random rounding was also applied to multiplications and divisions. Apart from that, we reset the rounding mode to “64 bit, nearest” before calling mathematical library functions, like `exp` or `log`. The reason for this is that the implementation of these functions may depend on exact rounding properties of intermediate values, and thus the random rounding might misrepresent the accuracy loss caused by them.

B. Equations for the Molecular Integrals. A Cartesian GTO basis function has the functional form

$$a_j(\mathbf{r}) = (r_x - A_x)^{a_x} (r_y - A_y)^{a_y} (r_z - A_z)^{a_z} \times \sum_{\alpha} \Gamma_{j\alpha} \exp(-\alpha \|\mathbf{r} - \mathbf{A}\|^2) \quad (16)$$

where \mathbf{A} is the center of the basis function, α runs over primitive exponents, j runs over contractions (with contraction coefficients $\Gamma_{j\alpha}$), and $\mathbf{a} = (a_x, a_y, a_z)$ indexes the Cartesian powers of the polynomial. In the actual basis functions used, the polynomial prefactors are transformed into solid harmonics $S_{m_a}^l(\mathbf{r} - \mathbf{A})$.⁶⁴ Similar definitions hold for $b_k(\mathbf{r})$ and $c_l(\mathbf{r})$, with their own centers and exponents. Using the notation of Ahlrichs,^{67,68} the three-center primitive integral over the scalar integral kernel $K(t)$:

$$(\mathbf{ab}|\mathbf{c}) = \iint a(\mathbf{r}_1) b(\mathbf{r}_1) K(\|\mathbf{r}_1 - \mathbf{r}_2\|) c(\mathbf{r}_2) d^3 r_1 d^3 r_2 \quad (17)$$

is then calculated by the following recurrence relations:

$$(\mathbf{00}|\mathbf{0})^m = \left(\frac{\pi}{\xi + \gamma}\right)^{3/2} \exp\left(-\frac{\alpha\beta}{\alpha + \beta} \|\mathbf{A} - \mathbf{B}\|^2\right) G_m(\rho, T) \quad (18)$$

$$\begin{aligned} (\mathbf{a0}|\mathbf{0})^m &= (P_i - A_i) ((\mathbf{a} - \mathbf{1}_i) \mathbf{0}|\mathbf{0})^m + \frac{a_i - 1}{2\xi} ((\mathbf{a} - \mathbf{2}_i) \mathbf{0}|\mathbf{0})^m \\ &\quad - \frac{\rho}{\xi} (P_i - C_i) ((\mathbf{a} - \mathbf{1}_i) \mathbf{0}|\mathbf{0})^{m+1} \\ &\quad - \frac{\rho(a_i - 1)}{\xi^2} ((\mathbf{a} - \mathbf{2}_i) \mathbf{0}|\mathbf{0})^{m+1} \end{aligned} \quad (19)$$

$$(\mathbf{a0}|\mathbf{c})^m = \frac{\rho}{\gamma} (P_i - C_i) (\mathbf{a0}|\mathbf{c} - \mathbf{1}_i)^{m+1} + \frac{a_i}{2(\eta + \gamma)} ((\mathbf{a} - \mathbf{1}_i) \mathbf{0}|\mathbf{c} - \mathbf{1}_i)^{m+1} \quad (20)$$

$$(\mathbf{ab}|\mathbf{c}) = ((\mathbf{a} + \mathbf{1}_i)(\mathbf{b} - \mathbf{1}_i)|\mathbf{c}) - (B_i - A_i)(\mathbf{a}(\mathbf{b} - \mathbf{1}_i)|\mathbf{c}) \quad (21)$$

where $i = x, y, \text{ or } z$ (no summation), $G_m(\rho, T)$ is a family of scalar functions depending only on the integral kernel K ,⁶⁷ and the other occurring intermediates are $\zeta = \alpha + \beta$ and

$$\rho = \frac{\xi\gamma}{\xi + \gamma} \quad \mathbf{P} = \frac{\alpha\mathbf{A} + \beta\mathbf{B}}{\alpha + \beta} \quad T = \rho \|\mathbf{P} - \mathbf{C}\|^2. \quad (22)$$

The final integrals are $(\mathbf{ab}|\mathbf{c}) = (\mathbf{ab}|\mathbf{c})^{m=0}$. Equation 21 is called the transfer equation, because it transfers angular momentum from a to b . Note that it does not contain any exponents and thus can be applied to contracted integrals instead of primitives.

In our implementation, the eqs 18–20, followed by the solid harmonic transformation of c , are evaluated for each combination of primitive integrals. The primitive integrals $(\mathbf{a0}|\mathbf{c})$ (with total AM from l_a to $l_a + l_b$ in the first label) are then accumulated into their respective general contractions. In eq 19 we evaluate $(\mathbf{P} - \mathbf{A})$ as $(\beta / (\alpha + \beta))(\mathbf{B} - \mathbf{A})$, as this increases the numerical stability when α/β is large. Finally, for each already contracted intermediate $(\mathbf{a0}|\mathbf{c})$, first the label a is transformed to solid harmonics, then the transfer equation eq 21 is applied to form $(\mathbf{ab}|\mathbf{c})$, and then label b is transformed to solid harmonics.

C. On the Asymmetric Instability of the Contracted Transfer Equation. As noted previously, the discovered asymmetric numerical instability is caused by the contracted transfer equation (eq 21), which is used as final step in not only the described integral core but also in almost every other one. This equation is popular because it can be applied to already contracted integral intermediates, as already pointed out, and can thus lead to significant time savings.

In the one-dimensional case, the numeric instability can be described as follows: We are calculating the desired integrals

$$\begin{aligned} (l_a, l_b | l_c) &= \iint (x_1 - A_x)^{l_a} e^{-\alpha(x_1 - A_x)^2} \cdot \\ &\quad (x_1 - B_x)^{l_b} e^{-\beta(x_1 - B_x)^2} \dots d^3 x_1 d^3 x_2 \end{aligned} \quad (23)$$

by first calculating $(l_a + l_b, 0 | l_c)$, that is

$$\iint (x_1 - A_x)^{l_a + l_b} e^{-\alpha(x_1 - A_x)^2} e^{-\beta(x_1 - B_x)^2} \dots \quad (24)$$

and then re-expressing some of the $(x_1 - A_x)$ as $(x_1 - B_x) + (B_x - A_x)$ (this is what this contracted transfer relation is doing). The problem now is this: if the exponent on B is much larger than that on A , then in the integral only such x_1 will have contributions where x_1 is very close to B . This means that, on average, $(x_1 - B_x)$ is a small quantity, which, we express as a difference of two large quantities, $(x_1 - A_x)$ and $(B_x - A_x)$; this situation inevitably leads to accuracy loss. The exponents do not explicitly occur in these transfer formulas, but the integral intermediates already contain them implicitly; namely encoded as the difference between the various components of the $l_a + l_b$ batch.

Since the reverse direction $\alpha > \beta$ works fine (see above) and since we have the integral identity $(ab|c) = (ba|c)$, one could think that this problem can simply be solved by always accumulating the angular momentum on the center with the large

exponent first (i.e., exchanging a and b if α is much smaller than β). However, this solution leads to various technical problems, because: (i) the transfer equation is generally implemented only for $l_a \geq l_b$ (since this is much more efficient than $l_a < l_b$), and thus if $l_a \neq l_b$, the order of a and b is already fixed by their angular momenta, and more importantly, (ii) in general one deals with contracted basis functions, in which primitive integrals with several exponents are set into some fixed linear combination. Thus both $\alpha < \beta$ and $\beta > \alpha$ will occur at the same time for a single contracted integral (and the transfer equation is applied to contracted intermediates).

A preliminary investigation suggests that the quantity:

$$r = \left(\frac{\|P - B\|}{\|A - B\|} \right)^{l_b} = \left(\frac{\alpha}{\alpha + \beta} \right)^{l_b}$$

is a good indicator of the relative accuracy loss due to the angular momentum transfer. Thus, for the moment the best workaround to the problem would be to estimate the accuracy loss with this relation for each primitive combination and, if it turns out to be unacceptable for some combination to do the computationally expensive angular momentum transfer from b to a for these primitives, additionally to the a to b transfer for the other primitives in the contraction.

Our current understanding is that this issue applies to almost every efficient integral core in use and that it cannot easily be solved without additional computational cost. We might revisit this issue in the future.

AUTHOR INFORMATION

Corresponding Author

*E-mail: knizia@theochem.uni-stuttgart.de.

ACKNOWLEDGMENT

The authors thank Toru Shiozaki for helpful discussions on integral evaluation and Fabian Giesen for helpful discussions on random number generators. This work was funded by the DFG within the SRC SimTech at Stuttgart University.

REFERENCES

- (1) Knuth, D. E. *The Art of Computer Programming. Vol. 2: Semi-numerical Algorithms*, 2nd ed.; Addison-Wesley: Upper Saddle River, NJ, 1980; pp 213–223.
- (2) Lindh, R.; Ryu, U.; Liu, B. *J. Chem. Phys.* **1991**, *95*, 5889.
- (3) Werner, H.-J.; Adler, T. B.; Manby, F. R. *J. Chem. Phys.* **2007**, *126*, 164102.
- (4) Knizia, G.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054104.
- (5) Rauhut, G.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054105.
- (6) Huang, X.; Valeev, E. F.; Lee, T. J. *J. Chem. Phys.* **2010**, *133*, 244108.
- (7) Bromley, M. W. J.; Mitroy, J. *Int. J. Quantum Chem.* **2007**, *107*, 1150.
- (8) Ekström, U.; Visscher, L.; Bast, R.; Thorvaldsen, A. J.; Ruud, K. *J. Chem. Theory Comput.* **2010**, *6*, 1971.
- (9) Takashima, H.; Amisaki, T.; Kitamura, K.; Nagashima, U. *Comput. Phys. Commun.* **2002**, *148*, 182.
- (10) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic Structure Theory*; Wiley: Chichester, U.K., 2000; pp 336–397.
- (11) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291.
- (12) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285.
- (13) Weigend, F. *J. Comput. Chem.* **2007**, *29*, 167.
- (14) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175.
- (15) Hättig, C. *Phys. Chem. Chem. Phys.* **2005**, *7*, 59.
- (16) Valeev, E. F. *Chem. Phys. Lett.* **2004**, *395*, 190.
- (17) Yousaf, K. E.; Peterson, K. A. *J. Chem. Phys.* **2008**, *129*, 184108.
- (18) Yousaf, K. E.; Peterson, K. A. *Chem. Phys. Lett.* **2009**, *476*, 303.
- (19) Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1985**, *82*, 5053.
- (20) Knowles, P. J.; Werner, H.-J. *Chem. Phys. Lett.* **1985**, *115*, 259.
- (21) Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1988**, *89*, 5803.
- (22) Asadchev, A.; Allada, V.; Felder, J.; Bode, B. M.; Gordon, M. S.; Windus, T. L. *J. Chem. Theory Comput.* **2010**, *6*, 696.
- (23) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2008**, *4*, 222.
- (24) Ufimtsev, I. S.; Martinez, T. J. *Comput. Sci. Eng.* **2008**, *10*, 26.
- (25) Luehr, N.; Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2011**, *7*, 949.
- (26) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 1004.
- (27) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 2619.
- (28) Yasuda, K. *J. Comput. Chem.* **2007**, *29*, 334.
- (29) Yasuda, K. *J. Chem. Theory Comput.* **2008**, *4*, 1230.
- (30) Vogt, L.; Olivares-Amaya, R.; Kermes, S.; Shao, Y.; Amador-edolla, C.; Aspuru-Guzik, A. *J. Phys. Chem. A* **2008**, *112*, 2049.
- (31) Olivares-Amaya, R.; Watson, M. A.; Edgar, R. G.; Vogt, L.; Shao, Y.; Aspuru-Guzik, A. *J. Chem. Theory Comput.* **2010**, *6*, 135.
- (32) Watson, M.; Olivares Amaya, R.; Edgar, R. G.; Aspuru-Guzik, A. *Comput. Sci. Eng.* **2010**, *12*, 40.
- (33) Brown, P.; Woods, C.; McIntosh-Smith, S.; Manby, F. R. *J. Chem. Theory Comput.* **2008**, *4*, 1620.
- (34) Brown, P.; Woods, C. J.; McIntosh-Smith, S.; Manby, F. R. *J. Comput. Chem.* **2010**, *31*, 2008.
- (35) Muller, J.-M.; Brisebarre, N.; de Dinechin, F.; Jeannerod, C.-P.; Lefèvre, V.; Melquiond, G.; Revol, N.; Stehlé, D.; Torres, S. *Handbook of Floating-Point Arithmetic*; Birkhäuser: Boston, 2010; pp 281–282 and 288–302.
- (36) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (37) Vysotskiy, V. P.; Cederbaum, L. S. *J. Chem. Theory Comput.* **2011**, *7*, 320.
- (38) Rump, M. *Algorithms for verified inclusions: Theory and Practice. Reliability in Computing: The Role of Interval Arithmetic in Methods in Scientific Computing*; Moore, R. E., Ed.; Academic Press: Boston, 1988; pp 109–126.
- (39) Loh, E.; Walster, G. W. *Reliab. Comput.* **2002**, *8*, 245.
- (40) Moore, R. E.; Bierbaum, F. *Methods and applications of interval analysis*; SIAM: Philadelphia, PA, 1979; pp 9–16.
- (41) Albertsen, N. C.; Chesneau, J.-M.; Christiansen, S.; Wirgin, A. *Math. Comput. Simul.* **1999**, *48*, 307.
- (42) Vignes, J.; Porte, M. L. *Proc. IFIP Congr.* **1974**, 610.
- (43) Vignes, J. *Math. Comput. Simul.* **1993**, *35*, 233.
- (44) Vignes, J. *Numer. Algor.* **2004**, *37*, 377.
- (45) Scott, N.; Jezequel, F.; Denis, C.; Chesneaux, J.-M. *Comput. Phys. Commun.* **2007**, *176*, 507.
- (46) Guilain, S.; Vignes, J. *Math. Comput. Simul.* **1994**, *37*, 73.
- (47) Abbasbandy, S.; Araghi, M. A. *Appl. Numer. Math.* **2004**, *50*, 279.
- (48) Jezequel, F.; Chesneaux, J.-M. *Comput. Phys. Commun.* **2008**, *178*, 933.
- (49) Barlow, R. J. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*; Wiley: Chichester, U.K., 1989; pp 49–51.
- (50) Li, W.; Simon, S. *Numerical Error Analysis for Statistical Software on Multi-Core Systems*. In Proceedings of Compstat'2010, Paris, France, August 22–27, 2010; Lechevallier, Y.; Saporta, G., Eds.; Physica: Heidelberg, Germany, 2010; pp 1287–1294.

- (51) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Lindh, R.; Mitrushenkov, A.; Rauhut, G.; Shamasundar, K. R.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hesselmann, A.; Hetzer, G.; Hrenar, T.; Jansen, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; O'Neill, D. P.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO*, development version 2010.2, a package of ab initio programs; University College Cardiff Consultants Limited: Wales, U.K., 2010; <http://www.molpro.net>.
- (52) Li, W.; Simon, S. On the Numerical Sensitivity of Computer Simulations on Hybrid and Parallel Computing Systems. In Proceedings of 2011 International Conference on High Performance Computing Simulation, Istanbul, Turkey, July 4–8, 2011, IEEE: New York, 2011; accepted.
- (53) Kahan, W. *Commun. ACM* **1965**, *8*, 40.
- (54) Dekker, T. J. *Numer. Math.* **1971**, *18*, 224.
- (55) Rendell, A. P.; Lee, T. J.; Komornicki, A. *Chem. Phys. Lett.* **1991**, *178*, 462.
- (56) Constans, P.; Ayala, P. Y.; Scuseria, G. E. *J. Chem. Phys.* **2000**, *113*, 10451.
- (57) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (58) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1993**, *98*, 1358.
- (59) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (60) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (61) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
- (62) Weigend, F. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057.
- (63) Deegan, M. J. O.; Knowles, P. J. *Chem. Phys. Lett.* **1994**, *227*, 321.
- (64) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic Structure Theory*; Wiley: Chichester, U.K., 2000; pp 207–218.
- (65) Obara, S.; Saika, A. *J. Chem. Phys.* **1986**, *84*, 3963.
- (66) Obara, S.; Saika, A. *J. Chem. Phys.* **1988**, *89*, 1540.
- (67) Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3072.
- (68) Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2004**, *6*, 5119.
- (69) Werner, H.-J.; Knizia, G.; Manby, F. R. *Mol. Phys.* **2010**, *109*, 407.
- (70) Feller, D.; Peterson, K. A.; Hill, J. G. *J. Chem. Phys.* **2010**, *133*, 184102.
- (71) Hill, J. G.; Peterson, K. A.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2009**, *131*, 194105.
- (72) Hill, J. G.; Mazumder, S.; Peterson, K. A. *J. Chem. Phys.* **2010**, *132*, 054108.
- (73) Lique, F.; Klos, J.; Hochlaf, M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 15672.
- (74) Barnes, E. C.; Petersson, G. A. *J. Chem. Phys.* **2010**, *132*, 114111.
- (75) Roos, B. O.; Lindh, R.; Malmqvist, P.-A.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2005**, *108*, 2851.
- (76) Cybulski, S. M.; Toczyłowski, R. R. *J. Chem. Phys.* **1999**, *111*, 10520.
- (77) Rys, J.; Dupuis, M.; King, H. F. *J. Comput. Chem.* **1983**, *4*, 154.
- (78) Ten-no, S. *Chem. Phys. Lett.* **2004**, *398*, 56.
- (79) Shiozaki, T. *Chem. Phys. Lett.* **2009**, *479*, 160.
- (80) Flocke, N.; Lotrich, V. *J. Comput. Chem.* **2008**, *29*, 2722.
- (81) Ishida, K. *J. Chem. Phys.* **1993**, *98*, 2176.
- (82) Boys, S. F.; Shavitt, I. *Report WIS-AF-13*; University of Wisconsin: Madison, WI, 1959.
- (83) Whitten, J. L. *J. Chem. Phys.* **1973**, *58*, 4496.
- (84) Baerends, E. J.; Ellis, D. E.; Ros, P. *Chem. Phys.* **1973**, *2*, 41.
- (85) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. *J. Chem. Phys.* **1979**, *71*, 3396.
- (86) Weigend, F.; Häser, M. *Theor. Chim. Acta* **1997**, *97*, 331.
- (87) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143.
- (88) Schütz, M.; Hetzer, G.; Werner, H.-J. *J. Chem. Phys.* **1999**, *111*, 5691.
- (89) Schütz, M. *J. Chem. Phys.* **2000**, *113*, 9986.
- (90) Schütz, M.; Werner, H.-J. *J. Chem. Phys.* **2001**, *114*, 661.
- (91) Schütz, M. *Phys. Chem. Chem. Phys.* **2002**, *4*, 3941.
- (92) Manby, F. R. *J. Chem. Phys.* **2003**, *119*, 4607.
- (93) Ten-no, S.; Manby, F. R. *J. Chem. Phys.* **2003**, *119*, 5358.
- (94) Neese, F.; Schwabe, T.; Grimme, S. *J. Chem. Phys.* **2007**, *126*, 124115.
- (95) Neese, F.; Wennmohs, F.; Hansen, A. *J. Chem. Phys.* **2009**, *130*, 114108.
- (96) Neese, F.; Hansen, A.; Liakos, D. G. *J. Chem. Phys.* **2009**, *131*, 064103.
- (97) Vahtras, O.; Almlöf, J.; Feyereisen, M. W. *Chem. Phys. Lett.* **1993**, *213*, 514.
- (98) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285.
- (99) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154.
- (100) Schütz, M.; Manby, F. R. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3349.
- (101) Polly, R.; Werner, H.-J.; Manby, F. R.; Knowles, P. J. *Mol. Phys.* **2004**, *102*, 2311.
- (102) Schütz, M.; Werner, H.-J.; Lindh, R.; Manby, F. R. *J. Chem. Phys.* **2004**, *121*, 737.
- (103) Werner, H.-J.; Manby, F. R. *J. Chem. Phys.* **2006**, *124*, 054114.
- (104) Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 154103.
- (105) Köhn, A.; Hättig, C. *J. Chem. Phys.* **2003**, *119*, 5021.
- (106) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175.
- (107) *Intel 64 and IA-32 Architectures Software Developer's Manual*; Volume 1: Basic Architecture. Intel, 2009; ch. 8.1.5 "x87 FPU Control Word"; Intel: Santa Clara, CA; <http://www.intel.com/products/processor/manuals/>. Accessed March 22, 2011.
- (108) Marsaglia, G. sci.math newsgroup post, message-id <I Sidna-cF9d6F8ajXTWcgg@comcast.com>. 1999; <http://groups.google.com/group/sci.math/msg/9959175f66dd138f>. Accessed March 22, 2011).
- (109) Jones, D. *Good Practice in (Pseudo) Random Number Generation for Bioinformatics Applications*. 2010; www.cs.ucl.ac.uk/staff/d.jones/GoodPracticeRNG.pdf. Accessed March 22, 2011).

Aurophilic Interactions from Wave Function, Symmetry-Adapted Perturbation Theory, and Rangehybrid Approaches

Ru-Fen Liu,[†] Christina A. Franzese,[‡] Ryan Malek,[‡] Piotr S. Żuchowski,[§] János G. Ángyán,^{*,†} Małgorzata M. Szczyński,^{*,‡} and Grzegorz Chałasiński^{‡,||}

[†]Crystallographie, Résonance Magnétique et Modélisations, Institut Jean Barriol, Nancy University and CNRS, F-54506 Vandoeuvre-lès-Nancy, France

[‡]Department of Chemistry, Oakland University, Rochester, Michigan 48309, United States

[§]School of Chemistry, The University of Nottingham, Nottingham NG7 2RD, Great Britain

^{||}Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warszawa, Poland

ABSTRACT: The aurophilic interaction is examined in three model systems $\text{Au}_2(^3\Sigma_g^+)$, $(\text{AuH})_2$, and $(\text{HAuPH}_3)_2$ which contain interactions of pairs of the Au centers in the oxidation state (I). Several methods are employed ranging from wave function theory-based (WFT) approaches to symmetry-adapted perturbation theory (SAPT) and range-separated hybrid (RSH) density functional theory (DFT) methods. The most promising and accurate approach consists of a combination of the DFT and WFT approaches in the RSH framework. In this combination the short-range DFT handles the slow convergence of the correlation cusp, whereas the long-range WFT is best suited for the long-range correlation. Of the three tested RSH DFT methods, the one which uses a short-range exchange functional based on the Ernzerhof–Perdew exchange hole model with a range-separation parameter of 0.4 bohr^{-1} seems to be the best candidate for treatment of gold. In combination with the long-range coupled cluster singles, doubles, and noniterative triples [CCSD(T)] treatment it places the strength of aurophilic bonding in $(\text{HAuPH}_3)_2$ at 5.7 kcal/mol at $R = 3.09 \text{ \AA}$. This value is somewhat larger than our best purely WFT result based on CCSD(T), 4.95 kcal/mol ($R = 3.1 \text{ \AA}$), and considerably smaller than the Hartree–Fock+dispersion value of 7.4 kcal/mol ($R = 2.9 \text{ \AA}$). The 5.7 kcal/mol estimate fits reasonably well within the prediction of the empirical relationship proposed by Schwerdtfeger et al. (*J. Am. Chem. Soc.* **1998**, *120*, 6587). A direct computation of dispersion energy, including exchange corrections, results in values of ca. -9 kcal/mol for $\text{Au}_2(^3\Sigma_g^+)$ and $(\text{AuH})_2$ and -13 kcal/mol for $(\text{HAuPH}_3)_2$ at the distance of a typical aurophilic bond, $R = 3.0 \text{ \AA}$.

I. INTRODUCTION

Experimental studies of two-coordinate gold compounds reveal structural motifs where Au(I) centers are in direct contact with each other but are not chemically bound. This provides the evidence for intermolecular bonding between seemingly closed-shell ($5d$)¹⁰ Au(I) cores known as aurophilic interactions.¹ The d^{10} subshells are spherically symmetric and exhibit valence repulsion; thus, the attractive interactions must originate, or so the explanation goes, from the dispersion effect, i.e., a nonlocal, dynamic correlation.^{2,3} Aurophilic interactions are comparable in strength to hydrogen bonding and play an important competing role in the assembly of gold compounds.¹ Other reasons for interest in these interactions include their relevance to the electronic structure of gold nanoparticles, where Au(I) cores provide the confining potential for a collective behavior of jellium electrons.⁴ Therefore, study of these interactions has wider implications for the coordination chemistry of gold, gold clusters, and gold nanoparticle–ligand interactions.

Previous computational studies of the aurophilic effect employed wave function theory (WFT) and density functional theory (DFT) based methodologies, as surveyed recently in exhaustive review papers.^{2,3} To our knowledge, no publication appeared on the use of recent rangehybrid methods for description of aurophilicity. Except for a very recent publication,⁵ aurophilic interactions have not been systematically studied by symmetry-adapted

perturbation theory (SAPT), which is a method intrinsically designed for intermolecular forces.

To summarize some highlights of previous studies, one should start from the work of Pyykkö and Zhao, who proposed a clever model for studying the aurophilic effect, $(\text{XAuPH}_3)_2$ ($X = \text{ligand}$) (in a gauche orientation), and first determined that the interaction between the Au(I) centers originated from electron correlation effects.⁶ An important finding of Pyykkö et al. was that in the Møller–Plesset (MP2), CCSD, and CCSD(T) series the correlation shows an oscillatory behavior. Therefore, the use of MP2 may overestimate interaction energies by as much as a factor of 2 compared with CCSD(T).⁷ The work of Magnko et al.⁸ on the same model $(\text{XAuPH}_3)_2$ ($X = \text{H, Cl}$) employed supermolecular, local MP2 (LMP2) calculations. From the analysis of LMP2 excitations the authors concluded that for the complex with $X = \text{H}$ about one-half of the post-Hartree–Fock (HF) attraction originated from dispersion-type excitations while the rest came from ionic contributions. Both post-HF contributions displayed distinctly different distance dependence. More recently, Pyykkö and Zaleski-Ejgierd estimated the basis set limit for the MP2 interaction⁹ in the same system. O’Grady and Kaltsoyannis¹⁰ carefully compared several WFT treatments

Received: April 8, 2011

Published: July 01, 2011

and made the first attempt to test DFT with the BP86 functional in the context of aurophilic (and metallophilic) interactions. They were surprised to observe that “DFT values are not significantly out of step with the ab initio results” although “the reasons why are not clear”. Assadollahzadeh and Schwerdtfeger¹¹ presented further comparisons of the BP86 DFT and WFT approaches for other ligands in the studies of the nonadditivity of aurophilic interactions.

Density functional theory in the Kohn–Sham formulation and with standard exchange–correlation functionals is able to provide a reliable description of short-range dynamic correlation effects, and in special circumstances it is able to mimic some amount of static correlation as well.¹² However, strongly non-local dynamic correlation effects, responsible for London dispersion forces, still remain a formidable challenge for DFT¹³ despite vigorous efforts to design new DFT-based methods, which are suited for this purpose. For aurophilic interactions, the words of Pyykkö, “No proof has been given that any supramolecular DFT treatment would reproduce the aurophilic attraction for a good physical reason”,³ remain certainly valid for all conventional local and semilocal functionals used in a simple Kohn–Sham framework. However, it might be worthwhile to revisit some aspects of Pyykkö’s statement in light of recent works aimed at including the London dispersion forces in DFT calculations. Extensively parametrized semiempirical functionals, like M05-2X¹⁴ and M06-2X¹⁵ of the Truhlar group, which proved to be extremely successful for a wide range of phenomena, including binding in van der Waals complexes around their equilibrium structures, fail to grasp physical origins of dispersion forces as reflected by the fact that the asymptotic $1/R^6$ decay of the potential energy is completely missed. Therefore, from our point of view, they are of rather limited interest. Promising new developments include the nonlocal vdW functionals,^{16,17} the a posteriori correction of the DFT total energy with C_6/R^6 or higher-order atom–atom type corrections, exemplified by the work of Grimme,¹⁸ which in its more recent variant adjusts the atomic dispersion coefficient by an appropriate algorithm to the local bonding environment.¹⁹ Following different routes, the atomic dispersion coefficients can be made functionals of the density, as in the approach of Tkatchenko and Scheffler²⁰ or in the model of Becke and Johnson of the local properties of the exchange hole.²¹ Another type of recent development ensures the presence of dynamic correlation effects, responsible for London dispersion forces, by a generalization of the concept of hybrid functionals. In this category, one should mention the double-hybrid functionals with an explicit, usually empirically weighted MP2 contribution, such as B2LYP²² or XYG3,²³ and finally the combination of short-range DFT with long-range WFT treatment of the dynamical correlation in a range-separated hybrid framework.^{24–27} It is worth mentioning that several of these rangehybrid approaches have been used for a systematic study of the group 11 hydrides and halides.²⁸ Some of these different categories of functionals designed for dispersion forces have recently been compared and tested by Sherrill and co-workers.²⁹ We think that it is timely to explore the impact of some of these new advances on the understanding of aurophilic interactions.

In all previous works on aurophilic interactions (except for a very recent contribution⁵) their nature was inferred from supermolecular calculations, i.e., from such observations as the repulsive HF potential and the R^{-6} asymptotic behavior of the supermolecular interaction energy upon inclusion of correlation.³

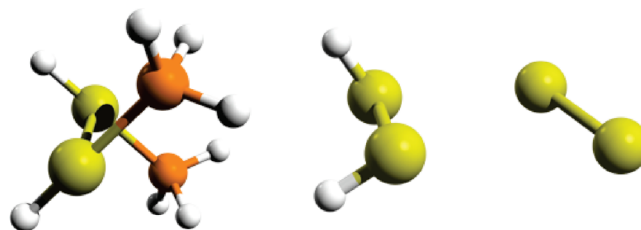


Figure 1. Geometrical configurations of three model aurophilic compounds.

One of the goals of the present paper is to determine the nature of the interaction between the two Au(I) centers embedded in a molecular framework by direct application of the perturbation theory approach. The Au(I)—Au(I) interacting unit will be included in three model systems of decreasing complexity. We begin with the same model system as that used in the work of Pyykkö and Zhao⁶—a dimer of the two-coordinate Au species: HAuPH₃ in a gauche configuration (i.e., with a dihedral angle of 90°, which allows for minimization of electrostatic forces by ensuring a vanishing dipole–dipole interaction). Next, the two PH₃ ligands are removed and one has (AuH)₂ in a gauche conformation. A thallium analog of this model was used by Schwerdtfeger in 1991³⁰ as a model system for Tl(I)—Tl(I) thalophilic interactions to show that the attraction is essentially a correlation effect. Next, the H atoms are removed, leaving Au₂ in the $^3\Sigma_u^+$ state (see Figure 1).

A broader objective of the present study is to demonstrate that the challenges of aurophilic interactions, which are indeed considerable, can be met using the methodologies based on the concept of range separation that combine the advantages of DFT for short-range electron–electron interactions with the long-range capabilities of WFT methods. The main advantage of WFT methods is that they allow for a clear-cut identification of dispersion forces as a post-HF correlation effect. While one can be sure that at the HF level the London dispersion forces cannot be present, such clear guidance is absent in Kohn–Sham calculations. On one hand, it has been observed that numerous exchange–correlation (xc) functionals lead to sometimes quite reasonable attractive wells on potential energy curves even for the simplest, fully dispersion-bound systems, such as rare gas binary complexes. Typically, LDA and GGA functionals, usually based on a Perdew-type exchange functional (e.g., PBE, PW91), which satisfy the Lieb–Oxford bound³¹ locally, fall in this category. On the other hand, in functionals based on an exchange component as developed by Becke, no binding is observed in the above-mentioned category of dispersion-bound complexes; on the contrary, the potential curves are often significantly more repulsive than the HF ones.³²

How does one choose a functional which provides reasonable reference energy for explicit, physically sound dispersion corrections? One strategy consists in tailoring an xc functional to reproduce as well as possible the genuinely dispersion-free HF potential curve. Several groups have shown recently the feasibility of this approach.^{33,34} An alternative strategy is based on the use of mutually polarized Kohn–Sham densities of the monomers and an explicit calculation of the intermolecular Coulomb and exchange effects.³⁵ The spurious overpolarization of the monomers has been avoided by a Pauli–blockade method,³⁶ leading to dispersion-free total energy. Finally, similar results can be obtained in the rangehybrid scheme, where long-range

(predominantly) intermolecular electron–electron (e–e) interactions are treated at the HF level while short-range (intramolecular) e–e interactions are handled by a short-range xc functional. Such hybrid calculations lead to essentially dispersion-free total energies, as first shown by the group of Hirao,³⁷ constituting a convenient starting point for long-range WFT correlation treatments of dispersion effects.

II. METHOD AND COMPUTATIONAL DETAILS

The gold atoms are described using the 19-electron, small-core (1s–4f), relativistic effective core potential (ECP) of Figgen et al.³⁸ ECP is combined with the augmented correlation-consistent basis set of triple- ζ quality (unless stated otherwise) recently optimized by Peterson and Puzzarini (aug-cc-pVTZ-PP), which includes orbitals up to the g symmetry.³⁹ The remaining atoms are described by the aug-cc-pVTZ basis set.⁴⁰

The supermolecular calculations are performed at several post-HF levels of theory up to CCSD(T)⁴¹ defining the interaction energy as

$$E_{\text{int}}^{\text{X}} = E_{\text{dimer}}^{\text{X}} - 2E_{\text{monomer}}^{\text{X}}, \text{ where } X = \text{HF}, \\ \text{HF} + \text{MP2}, \text{HF} + \text{CCSD}, \text{HF} + \text{CCSD}(\text{T})$$

where the monomer contributions are calculated in the basis set of the whole dimer. For direct comparison with the dispersion energy we also use the quantity $E_{\text{int}}^{\text{CORR}}$ which is the correlation contribution to the interaction energy at a given level and defined as the difference between $E_{\text{int}}^{\text{X}}$ and $E_{\text{int}}^{\text{HF}}$. In the correlated calculations the outer-core $5s^25p^6$ electrons were active.

The calculations of dispersion energy and its exchange counterpart employ two variants of symmetry adapted perturbation theory. For $^3\text{Au}_2$ WFT open-shell symmetry adapted perturbation theory [SAPT(WFT)] is used.⁴² The dispersion energy is obtained from the time-dependent (TD) coupled HF theory as described in ref 43 and henceforth denoted $E_{\text{disp}}^{(2)}(\text{CHF})$, and its exchange counterpart is $E_{\text{exdisp}}^{(2)}(\text{CHF})$. For the remaining dimers, which are closed shell, SAPT theory based on the DFT description of monomers⁴⁴ is used as implemented in MOLPRO.⁴⁵ According to SAPT(DFT), the total interaction energy through the second-order perturbation theory, DSAPT[2], is expressed as the following sum of the electrostatic (es), induction (ind), and dispersion (disp) terms as well as their respective exchange counterparts

$$\text{DSAPT}[2] = E_{\text{es}}^{(1)} + E_{\text{exch}}^{(1)} + E_{\text{ind}}^{(2)} + E_{\text{exind}}^{(2)} + E_{\text{disp}}^{(2)} + E_{\text{exdisp}}^{(2)} \quad (1)$$

where the perturbation terms are obtained from Kohn–Sham (KS) orbitals. The induction and dispersion terms as well as their exchange counterparts are obtained from the coupled KS (CKS) approach (see refs 46 and 47 for details). The dispersion and exchange–dispersion terms will henceforth be denoted $E_{\text{disp}}^{(2)}(\text{CKS})$ and $E_{\text{exdisp}}^{(2)}(\text{CKS})$, respectively. Additionally, a residual HF term, δ_{HF} , is defined as a supplement to eq 1 with higher order terms and other residual effects taken from the HF level

$$\delta_{\text{HF}} = E_{\text{int}}^{\text{HF}} - E_{\text{es}}^{(1)} - E_{\text{exch}}^{(1)} - E_{\text{ind}}^{(2)}(\text{CHF}) - E_{\text{exind}}^{(2)}(\text{CHF}) \quad (2)$$

where the perturbation terms subtracted from the supermolecular HF interaction energy are all evaluated from the HF orbitals. The KS orbitals for the SAPT(DFT) calculations are obtained using the asymptotically corrected⁴⁸ PBE0 functional.⁴⁹

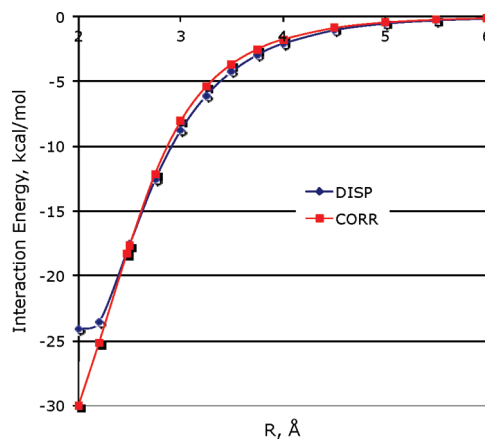


Figure 2. Comparison between the total dispersion energy $E_{\text{disp}}^{(2)}(\text{CHF}) + E_{\text{exdisp}}^{(2)}(\text{CHF})$ (denoted DISP) of $^3\text{Au}_2$ and the correlation contribution to the interaction energy at the CCSD(T) level of theory (CORR).

Range-separated hybrid (RSH) calculations have been performed with a range-separation parameter μ of 0.5 and 0.4 bohr⁻¹ for sr-LDA and sr-PBE functionals, respectively. Several independent studies indicated these values as optimal.^{50–52} The short-range LDA and two flavors of short-range PBE xc functionals have been explored in the present work. The sr-LDA functional is composed of the short-range LDA exchange⁵³ and the complementary short-range correlation derived from Quantum Monte Carlo simulations on the long-range interacting homogeneous electron gas.⁵⁴ The sr-PBE functionals differ only in their exchange components. One of the variants is based on the form suggested by Toulouse et al.⁵⁵ and extended by Goll et al.^{56,57} and keeps the same form as the original PBE functional but introduces a μ dependence of the parameters. The second type of sr exchange functional is constructed from an analytical model of the PBE exchange hole, developed by Ernzerhof and Perdew⁵⁸ (referred to here as PBE-EP), which is integrated with the short-range interaction function, $\text{erfc}(\mu r)/r$ (see also ref 59). This approach is analogous to the method followed in the construction of the HSE functional,⁶⁰ the main difference being that the role of the short and long range is inverted. The sr-PBE correlation functional used here has been described in ref 57. The applied rangehybrid methods will be designated by the following acronyms: sr-{LDA,PBE,PBE-EP}+lr-{MP2,CCSD,CCSD(T)}. The basis set is the same augmented valence triple- ζ quality as for the full-range correlated WFT and SAPT calculations, with the only difference being that it is not augmented on the phosphine H atoms.

The geometrical parameters are as follows. In gauche (AuH)₂ complex $r(\text{Au–H}) = 1.524 \text{ \AA}$.⁶¹ In (HAuPH_3)₂ we use the same dimer configuration as in ref 8. In the WFT and SAPT calculations the intramonomer HAuPH_3 geometrical parameters were taken from ref 8 ($r(\text{Au–P}) = 2.385 \text{ \AA}$, $r(\text{Au–H}) = 1.606 \text{ \AA}$, $r(\text{P–H}) = 1.416 \text{ \AA}$, $\angle(\text{H–P–Au}) = 118.5^\circ$), whereas in the rangehybrid calculations these parameters were optimized at the CCSD(T) level of theory ($r(\text{Au–P}) = 2.325 \text{ \AA}$, $r(\text{Au–H}) = 1.598 \text{ \AA}$, $r(\text{P–H}) = 1.407 \text{ \AA}$, $\angle(\text{H–P–Au}) = 119^\circ$).⁶² The basis sets in these geometry optimizations were ECP-VTZ for Au and VTZ (without augmentation functions) on the P and H atoms. The coordinate $R(\text{Au–Au})$ was varied keeping the monomer geometries unchanged.

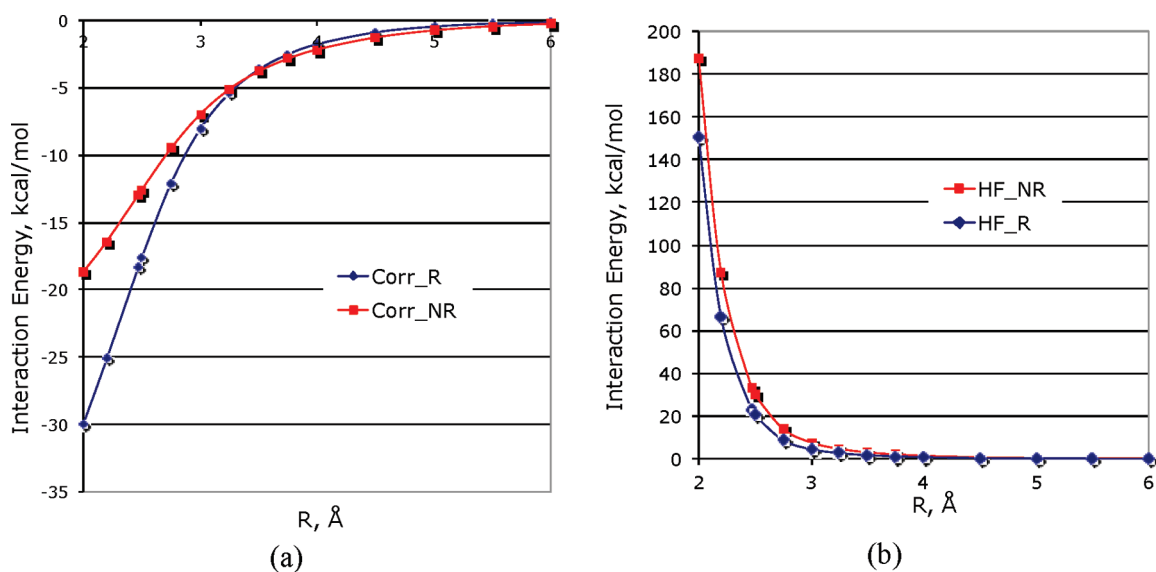


Figure 3. Comparison of “nonrelativistic” (NR) and “relativistic” (R) treatment of the interaction energy in $^3\text{Au}_2$: (a) correlation CCSD(T) contribution to the interaction energy and (b) uncorrelated HF component. For definitions see the text.

III. RESULTS AND DISCUSSION

1. WFT and SAPT Methods. *a. Dispersion Energy in $\text{Au}_2(^3\Sigma_u^+)$.* It is convenient to begin our discussion with the interaction between Au atoms in the $\text{Au}_2(^3\Sigma_u^+)$ state. The open-shell calculations in this section are performed using spin-restricted coupled cluster formalism and spin-unrestricted formalism in the SAPT dispersion calculations (unless stated otherwise).

In the ground state Au_2 is covalently bound by a single σ bond formed by two 6s electrons. In $^3\Sigma_u^+$ state the molecule is bound by long-range forces. The key difference between the two states is the exchange energy, which distinguishes the singlet from the triplet state. The ground state $\text{Au}_2(^1\Sigma_g^+)$ has been carefully benchmarked by Peterson and Puzzarini³⁹ using CCSD(T). Their calculation, which included relativistic effects, basis set limit estimation, and core–valence correlation, yielded a well depth of 53.2 kcal/mol at $R_e = 2.47$ Å. Our present treatment places the well depth of the ground state some 7% above their value. For the $\text{Au}_2(^3\Sigma_u^+)$ state we obtain a well depth of 3.698 kcal/mol at the roughly optimized R_e of 2.924 Å at the UCCSD(T). A recent, more saturated CCSD(T) result for D_e for this complex is reported to be 4.3 kcal/mol.⁶³

In our calculations of the CHF dispersion energy, which is known to be more basis-set dependent than the CKS variants,⁶⁴ we employ a slightly larger basis set (aug-cc-pVQZ-PP from which the more compact g and h orbitals were removed) augmented by a bond function (3s, 3p, 3d, 3f).

A comparison between the sum of $E_{\text{disp}}^{(2)}(\text{CHF}) + E_{\text{exdisp}}^{(2)}(\text{CHF})$ of $^3\text{Au}_2$ and the correlation contribution to this interaction $E_{\text{int}}^{\text{CCSD(T)}}$ is presented in Figure 2.

One can see that the two quantities agree extremely well in the wide range of distances. Two values of dispersion energy are particularly noteworthy. Around the van der Waals minimum of $^3\text{Au}_2$ ($R = 3$ Å) the dispersion effect amounts to -8.8 kcal/mol with $E_{\text{disp}}^{(2)}(\text{CHF}) = -12.36$ kcal/mol and $E_{\text{exdisp}}^{(2)}(\text{CHF}) = 3.57$ kcal/mol. Around the chemically bound minimum $^1\text{Au}_2$ ($R = 2.47$ Å) the total dispersion effect amounts to -18.22 kcal/mol out of which the dispersion is -29.75 kcal/mol and the exchange

dispersion is 11.53 kcal/mol. The C_6 dispersion coefficient fit to the long-range tail of the RHF-UCCSD(T) correlation interaction energy of $^3\text{Au}_2$ amounts to 345 au. This value increases by 4.5% upon freezing the 5s, 5p electrons in the calculations.

Next, we examine the effect of the scalar relativity on the dispersion interaction and more precisely on the correlation contribution to the $^3\text{Au}_2$ bonding. The relativistic effects lower the static dipole polarizability of Au from 64 to about 36 au.⁶⁵ They also lead to the doubling of the electron affinity of Au (from 1.283 to 2.295 eV) and to the increase of the ionization potential by some 2 eV (see Table 3 in ref 3). In gold–ligand interactions Au– PH_3 these effects result in a “relativistic bond” as described by Granatier et al.⁶⁶ Our present comparison is indirect and qualitative in nature and involves the following strategy: For the “nonrelativistic” treatment we perform all-electron calculations, without the Douglas–Kroll option, and with the fully nonrelativistic basis set named Hy-PolX.⁶⁷ The “relativistic” calculation employs the above-mentioned relativistic ECP with aug-cc-pVTZ-PP, which contains a similar number and type of polarization functions as Hy-PolX. To ensure that the same number of electrons is correlated in both calculations, the former keeps 68 electrons on each Au in the core while the latter keeps 8 electrons on each Au in the core. The result of this comparison is shown in Figure 3.

It appears that in the long range the “nonrelativistic” curve is slightly more attractive, consistent with the larger nonrelativistic static polarizability, if London’s model of dispersion is to be followed. In the short range the “relativistic” curve is more attractive, reflecting the larger ionization potential at the relativistic level. It is also interesting that the curves cross each other around 3.2 Å. Thus, at shorter distances, relativistic effects slightly strengthen and at longer distances slightly weaken the correlation interaction in $^3\text{Au}_2$. However, the most significant changes between the two approaches appear (not unexpectedly) at the HF level of theory. As seen in Figure 3b, the relativistic effects shift the repulsive wall toward shorter distances, enabling the atoms to approach closer and thus benefitting from a stronger dispersion attraction.

Table 1. Three-Body (3-b) Contributions in ${}^4\text{Au}_3$ from Restricted Open-Shell CCSD(T) Calculations (in kcal/mol)^a

geometry	$E_{\text{int}}(\text{abc})$	$E_{3\text{-b}}(\text{HF})$	$E_{3\text{-b}}(\text{corr})$
D_{3h}	-11.49	-2.33	0.51
D_{coll}	-6.30	0.41	-0.21

^aThe three-body term is partitioned into the HF and correlation contributions.

Table 2. Comparison of ${}^3\text{Au}_2$, $(\text{AuH})_2$, and $(\text{HAuPH}_3)_2$ Interaction Energies at $R = 3 \text{ \AA}$ (in kcal/mol)^a

term	${}^3\text{Au}_2$	$(\text{AuH})_2$	$(\text{HAuPH}_3)_2$
HF	4.43	3.21	5.63
HF+CCSD(T)	-3.64	-2.93	-4.83
HF+CCSD	-2.12	-1.75	-2.82
HF+MP2	-5.05	-5.73	-8.92
HF+Disp	-4.36	-5.94	-7.40
$E_{\text{disp+ex}}^{(2)}(\text{CKS})$	-8.80 ^b	-9.06	-13.02 (-12.28)
$E_{\text{disp+ex}}^{(2)}(\text{UCKS})$		-19.07	-27.01 (-17.47)

^aThe values in parentheses for $(\text{HAuPH}_3)_2$ correspond to $E_{\text{disp+ex}}^{(2)}$ from CHF and UCHF calculations, respectively. ^bfrom the CHF calculations.

Finally, it is also possible to evaluate the magnitude and sign of the three-body terms in the interaction of three Au atoms. To this end, we choose the high-spin, quartet Au_3 for which the restricted open-shell CCSD(T) calculations are performed. The results are shown in Table 1.

$E_{\text{int}}(\text{abc})$ denotes the entire trimer interaction energy (i.e., both two-body and three-body terms). Assuming that the correlation contribution to the interaction energy consists mainly of the dispersion energy (as shown above in the two-body case), we observe that the three-body dispersion interaction is positive in the equilateral triangle configuration and slightly negative in the linear configuration, which is consistent with the Axilrod–Teller–Muto angular dependence of the dispersion interactions as well as with our previous diagrammatic analysis of the n -body contributions to the supermolecular MP perturbation theory interaction energies.⁶⁸ One can also notice that the nonadditivity of the HF interaction energy has an opposite sign as it results from the first-order exchange three-body term (see a similar behavior for other metals from Ia and Ib groups⁶⁹).

b. $(\text{AuH})_2$. A comparison between the results for ${}^3\text{Au}_2$, $(\text{AuH})_2$, and $(\text{HAuPH}_3)_2$ is shown in Table 2. Let us focus on the first two, leaving $(\text{HAuPH}_3)_2$ for later discussion. The calculations for $(\text{AuH})_2$ are performed for the orientation of the monomers shown in Figure 1 (C_2 point group). The supermolecular dimer CCSD(T) calculations indicate that there may be some multi-configurational character in the reference function, namely, rather large T1 (0.03) and D1 (0.1) diagnostic^{70,71} values have been found. Indeed, the EOM-CCSD calculations show that there are two low-lying singly excited states ${}^1\text{B}$ and ${}^1\text{A}$ at 2.97 and 3.13 eV, respectively, at $R = 3.0 \text{ \AA}$. Nevertheless, the CCSD(T) calculations provide results consistent with those for ${}^3\text{Au}_2$ and yield a CCSD(T) interaction curve with the minimum at $R(\text{Au}–\text{Au}) = 3.09 \text{ \AA}$ with a well depth of 3.03 kcal/mol.

Both ${}^3\text{Au}_2$ and $(\text{AuH})_2$ have purely repulsive HF potential curves, and they are bound only upon inclusion of correlation effects. Both complexes have further similar characteristics: they are relatively weakly bound by about -3.6 and -2.93 kcal/mol,

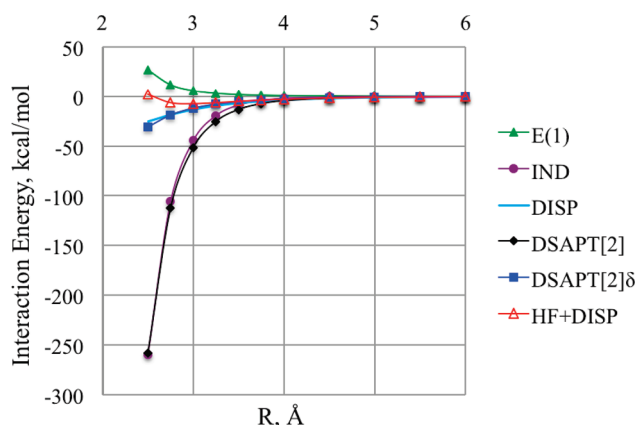


Figure 4. Radial dependence of the interaction energy terms in $(\text{HAuPH}_3)_2$ from SAPT(DFT). $E(1)$ denotes the sum of the electrostatic and exchange energy, IND denotes the sum of the CKS induction and exchange-induction terms, DISP denotes the sum of the CKS dispersion and exchange-dispersion terms, DSAPT[2] denotes the sum of the SAPT(DFT) terms through the second order, see eq 1, and DSAPT[2] δ includes also the residual δ_{HF} term of eq 2.

respectively, and their total dispersion effect (including exchange) is very similar, -8.8 vs -9.1 kcal/mol, respectively. It is particularly noteworthy that uncoupled KS dispersion energy is severely overestimated in magnitude. Compared to HF+CCSD(T), HF+MP2 leads to a stronger binding. Also shown is the HF+Disp approximation, which combines HF with the coupled dispersion and exchange–dispersion terms. In ${}^3\text{Au}_2$ this approximation gives values in reasonable agreement with HF+CCSD(T) as discussed in the previous section. In the case of $(\text{AuH})_2$ it exceeds HF+CCSD(T) by a factor of 2.

c. $(\text{HAuPH}_3)_2$. The supermolecular HF+CCSD(T) calculations for this dimer, in the geometry shown in Figure 1, yield a well depth of 4.95 kcal/mol at around $R(\text{Au}–\text{Au}) = 3.1 \text{ \AA}$. In this value the effects of outer-core correlation (i.e., 5s and 5p electrons in Au) contribute only 0.2 kcal/mol toward the stabilization. Incidentally, both types of diagnostics, T1 and D1, are in the normal range, indicating that the single-reference character improves upon addition of the PH_3 ligands. However, this result obtained in aug-cc-pVTZ is probably far from basis set saturation. The HF interaction energy results in a purely repulsive potential curve. SAPT(DFT) values of dispersion energy terms $E_{\text{disp}}^{(2)}(\text{CKS})$ and $E_{\text{exdisp}}^{(2)}(\text{CKS})$ are also shown. In addition, we also list corresponding values of dispersion energy calculated using the CHF and UCHF approach, $E_{\text{disp+ex}}^{(2)}(\text{CHF})$ and $E_{\text{disp+ex}}^{(2)}(\text{UCHF})$, respectively. The other SAPT(DFT) components are not shown because the theory appears to diverge for aurophilic interactions.

Figure 4 shows the R dependence of the SAPT(DFT) contributions for $(\text{HAuPH}_3)_2$. DSAPT[2], the sum of all the terms through the second order (see eq 1), is dominated by the induction effects and falls precipitously for shorter distances without any minimum. Adding the residual HF term δ_{HF} (eq 2) as suggested in the instances of using ECP (see refs 72 and 73) also fails to produce the minimum (see DSAPT[2] δ curve). One should add that δ_{HF} , which is strongly repulsive, eludes physical justification. By contrast, HF+Disp displays a minimum with reasonable position and depth (see Figure 4).

The causes of the SAPT divergence are related to the overestimated induction energy (-131 kcal/mol at $R = 3.0 \text{ \AA}$!), which cannot be properly constrained by the exchange effects

(see Figure 4). The approximate treatment of the exchange interactions with the partner's core as a result of using the ECP (see ref 72) is a contributing factor but not the cause of it. Broadly speaking, it is the fundamental inseparability of the induction and exchange, which is the root of this problem.⁷⁴ It should be stressed that the calculated values of the dispersion and exchange dispersion terms are sound.

The equilibrium results obtained at various levels of theory are shown in Table 2, where they can be compared with both $(\text{AuH})_2$ and ${}^3\text{Au}_2$. The $(\text{HAuPH}_3)_2$ dimer is more strongly bound than the other Au-containing systems, but the pattern is similar: the dimer is unbound at the HF level of theory; HF+MP2 overestimates the binding compared to HF+CCSD(T) to an even greater degree. The same is true of the HF+Disp approximation. The strengthening of the aurophilic bonding upon addition of the PH_3 group to AuH can be explained by a subtle balance between the first-order repulsion and dispersion attraction. For example, at $R = 3.0 \text{ \AA}$ addition of PH_3 results in the 2.4 kcal/mol net gain in the first-order interaction ($E_{\text{es}}^{(1)} + E_{\text{exch}}^{(1)}$), which almost exactly accounts for the change in the HF term (Table 2). Simultaneously, adding the ligands leads to enhanced net dispersion ($E_{\text{disp}}^{(2)} + E_{\text{exdisp}}^{(2)}$) stabilization of -4 kcal/mol .

In aurophilic interactions one is faced with a number of unpleasant realities: (i) the cluster expansion converges slowly as evidenced by the large contribution from the triples; (ii) the HF+Disp approximation overestimates the interaction as compared to HF+CCSD(T). Furthermore, as seen in Table 2, the UCHF dispersion energy is too large in magnitude compared to both CHF and CKS (although not as outrageously so as UCKS). Since the UCHF type of dispersion resides implicitly in the supermolecular MP2 interaction contribution,^{75,76} HF+MP2 should also lead to overbinding. Finally, it is not possible to substitute the HF interaction for another "dispersionless" approach based on SAPT(DFT) because the latter is divergent and the quality of its components cannot be ascertained.

2. Range-Separated DFT + WFT Approaches for Aurophilic Interactions. The combined sr-DFT + lr-WFT approaches have proven very successful in such challenging systems as rare gas dimers^{56,77} and alkaline earth dimers.^{78–80} Therefore, it is interesting to test them in the circumstance of the aurophilic bond. Such approaches are predicated on the range separation of the e–e interactions,^{24,53,78,81–83} where the short range is described within the DFT theory and the long range, nonlocal correlation is obtained from the WFT methodology. The calculations in this section are performed in a slightly different monomer geometry of HAuPH_3 (see section II), which has a negligible effect on the interaction energy.⁶²

The first emerging question is how to select the sr-DFT containing only the short-range correlation effects, i.e., which is expected to be essentially dispersion free. It is difficult to predict how the dispersion-free potential should appear for $\text{Au(I)}-\text{Au(I)}$ and by what criteria it should be judged. Unlike for rare-gas dimers where one is certain that such potential should be repulsive there is no such certitude in the present case, especially in light of the findings of refs 8 and 84 that have identified the nondispersion, post-HF ionic contributions to the aurophilic bond.

Figure 5a and 5b presents the sr-DFT potentials for $(\text{AuH})_2$ (Figure 5a) and $(\text{HAuPH}_3)_2$ (Figure 5b) obtained with the following range-separated functionals described in section II: sr-LDA ($\mu = 0.5$), sr-PBE ($\mu = 0.4$), and sr-PBE-EP ($\mu = 0.4$).

The results show that sr-PBE-EP is the least attractive of the three sr-DFT options for both systems, although it is much less

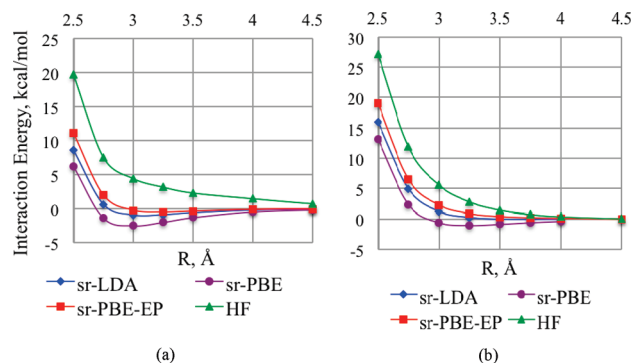


Figure 5. Comparison of sr-LDA, sr-PBE, and sr-PBE-EP potentials for (a) $(\text{AuH})_2$ and (b) $(\text{HAuPH}_3)_2$.

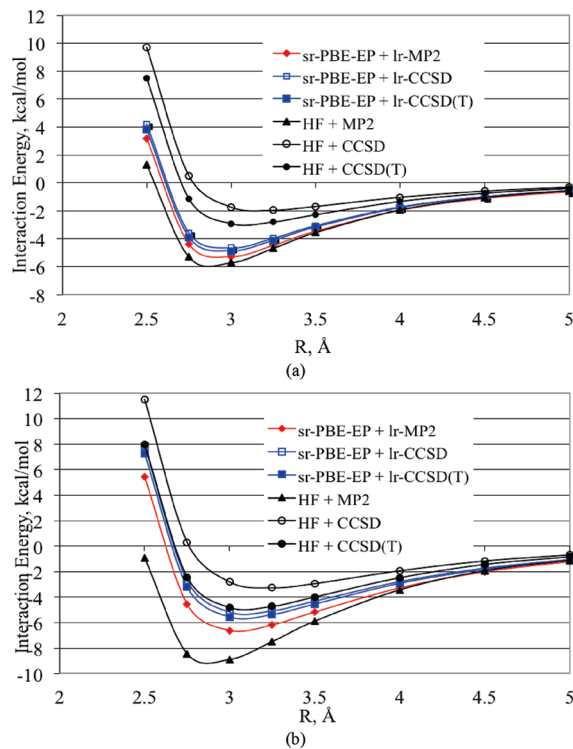


Figure 6. Comparison between sr-DFT + lr-WFT and full-range WFT approaches for (a) $(\text{AuH})_2$ and (b) $(\text{HAuPH}_3)_2$.

so than HF. sr-PBE is the most attractive of the three, whereas sr-LDA closely follows sr-PBE-EP. One can notice that even sr-PBE-EP and sr-LDA are slightly attractive for $(\text{AuH})_2$. However, the interaction between two AuH has some multiconfigurational character (see section b), which could manifest itself as short-range correlation effects (such as ionic contributions observed by Magnko et al.⁸)

In the next step the sr-DFT orbitals are used in the MP2, CCSD, and CCSD(T) calculations, resulting in lr-MP2, lr-CCSD, and lr-CCSD(T), which are added to the sr-DFT part. These are compared with the "full-range" WFT approaches HF+MP2, HF+CCSD, and HF+CCSD(T). The comparison is shown in Figure 6a for $(\text{AuH})_2$ and Figure 6b for $(\text{HAuPH}_3)_2$.

We observed before (see Table 2) that in full-range coupled cluster treatment the role of triples was overwhelming: 40% of the well depth in $(\text{AuH})_2$ and 42% in $(\text{HAuPH}_3)_2$. Furthermore,

Table 3. Energy Minimum Parameters of the (AuH)₂ and (HAuPH₃)₂ Dimers Obtained with Different Levels of Theory

theory	(AuH) ₂		(HAuPH ₃) ₂	
	E(min), kcal/mol	R(min), Å	E(min), kcal/mol	R(min), Å
HF+MP2	−6.01	2.89	−9.15	2.91
HF+CCSD	−2.04	3.16	−3.28	3.25
HF+CCSD(T)	−3.03	3.09	−4.95	3.11
HF+Disp	−6.04	2.91	−7.39	3.00
sr-PBE-EP	−0.52	3.18	−0.01	4.99
sr-PBE-EP + lr-MP2	−5.42	2.92	−6.68	3.06
sr-PBE-EP + lr-CCSD	−4.75	2.93	−5.37	3.10
sr-PBE-EP+ lr-CCSD(T)	−5.01	2.93	−5.72	3.09
sr-LDA	−1.08	3.10	−0.04	3.58
sr-LDA + lr-MP2	−8.03	2.84	−9.84	2.89
sr-LDA + lr-CCSD	−6.38	2.87	−7.14	2.97
sr-LDA + lrCCSD(T)	−6.84	2.86	−7.79	2.95
sr-PBE	−2.55	2.95	−1.05	3.20
sr-PBE + lr-MP2	−8.23	2.85	−9.86	2.90
sr-PBE + lr-CCSD	−7.51	2.86	−8.28	2.93
sr-PBE + lr-CCSD(T)	−7.80	2.86	−8.71	2.92

HF+MP2 overestimated the well depth of both complexes by nearly 2-fold. By contrast, our best candidate for dispersion-free rangehybrid, sr-PBE-EP, in combination with lr-WFT shows to be a quickly convergent treatment of electron correlation. Specifically, in both complexes, the sensitivity to triples has practically disappeared. The performance of lr-MP2 also appears to be much improved compared to the full-range approach, particularly for (HAuPH₃)₂. For (AuH)₂ the three range-separated sr-DFT + lr-WFT potentials which run very close to each other are considerably deeper than HF+CCSD(T). This again may be related to the fact that already the dispersion-free sr-PBE-EP is slightly attractive. Overall, the results show that the sr-PBE-EP rangehybrid combined with lr-WFT provides an excellent description of auophilic interactions. To see how the remaining sr-DFT approaches behave in combination with lr-WFT we summarize the equilibrium well-depth parameters in Table 3.

The sr-LDA approach combined with lr-WFT leads to deeper potential wells (by up to 2 kcal/mol for lr-CCSD/CCSD(T) and 3 kcal/mol for lr-MP2) with minima shifted toward shorter distances. For sr-PBE combined with lr-WFT the deepening is even more pronounced. Still the effect of triples in lr-CCSD(T) is quite small, on the order of 0.5 kcal/mol or less. Interestingly, it is not the lr-correlation which differentiates between the potentials. For (AuHPH₃)₂, the lr-CCSD contribution at R = 3.0 Å is almost the same if obtained from the sr-PBE-EP or sr-PBE orbitals, namely, −7.46 vs −7.57 kcal/mol, respectively. The differences reside in the sr-DFT, i.e., the putative dispersion-free part of the interaction. Let us also mention that the lr-CCSD contribution obtained from the sr-LDA orbitals amounts to −8.35 kcal/mol and that from the HF orbitals amounts to −8.45 kcal/mol. Finally, the lr-CCSD(T) calculations lead to the dramatically improved T1 and D1 diagnostics. For example, the sr-PBE-EP + lr-CCSD(T) calculations give at R = 3 Å the T1/D1 diagnostics of 0.0146/0.04689 and 0.0102/0.0339 for (AuH)₂ and (HAuPH₃)₂, respectively.

From the computational point of view, there is at present no particular time advantage of the sr-DFT + lr-WFT approach if one compares the calculations in the same basis set. However, since the basis set gets saturated much faster, a “converged” result is much cheaper to obtain.

IV. CONCLUSIONS

Auophilic interactions present a formidable challenge for computational treatments in that they include a large number of strongly correlated, relativistically contracted electrons. We explored a number of possible treatments for these interactions, which include symmetry-adapted perturbation theory, supermolecular full-range WFT methods, and the hybrid sr-DFT + lr-WFT approaches based on the concept of range separation.

SAPT(DFT) leads to a divergent description of these interactions due in large part to the spurious overpolarization of monomers which cannot be constrained by the exchange. A contributing factor is the use of the effective core potentials for the Au centers. The important SAPT(DFT) terms that appear to be unaffected are the dispersion and exchange–dispersion terms. This fact can be traced to the dispersion energy being related to the *two-electron* part of the intermolecular interaction operator as first noted in ref 72. When combined with the HF interaction energy, HF+Disp provides a reasonable approximation for ³Au₂ but becomes erratic for (AuH)₂ and (HAuPH₃)₂, presumably due to the larger role of the neglected intramonomer correlation. The dispersion attraction of two Au atoms is increased at long intersystem distances and reduced in the short range by inclusion of relativistic effects. These changes are consistent with the effects of relativity on the polarizability and ionization potential of Au. However, the strongest effect relativity exerts on the HF interaction potential, causing the shift of the repulsive wall toward shorter distances.

Of the full-range WFT approaches examined in this work, HF +MP2 leads to a significant overbinding because of a considerable overestimation of the dispersion energy by the UCHF approximation for auophilic interactions, as demonstrated in this work. The results of the HF+CCSD and HF+CCSD(T) calculations point to a significant role of triples. This indicates that to saturate correlation effects it may be necessary to include in the cluster expansion not only the iterative triple excitations but also the quadruples.

By contrast, the range-separated sr-DFT + lr-WFT approaches examined in this work show a much weaker dependence on the triple excitations as seen by the small difference between lr-CCSD and lr-CCSD(T) contributions. One plausible explanation is that the triple excitations are more important in the short range of the e–e interactions, but their contribution to the long range is smaller. In the combined sr-DFT + lr-WFT approach these short-range interactions are efficiently accounted for by a sr-DFT component.

The sr-DFT approach offers the added benefit of serving as a dispersion-free approximation—the DFT analog of the HF interaction energy. Of the three candidates examined here the sr-PBE-EP functional with a range-separation parameter of 0.4 appears to fit the characteristics of such an approximation for gold. We base this determination on the values of the sr-PBE-EP+lr-CCSD(T) interaction energies for (AuH)₂ and (HAuPH₃)₂, which appear to be the most sensible. The sr-PBE-EP+lr-CCSD(T) approach places the strength of auophilic interaction in the (HAuPH₃)₂ gauche dimer of the two unrelaxed monomers at

5.7 kcal/mol at $R = 3.09 \text{ \AA}$. This value is somewhat larger than the HF+CCSD(T) one of 4.95 kcal/mol ($R = 3.1 \text{ \AA}$) and considerably smaller than the HF+Disp value of 7.4 kcal/mol ($R = 2.9 \text{ \AA}$). The 5.7 kcal/mol estimate fits reasonably well within the prediction of the empirical relationship proposed by Schwerdtfeger et al.,⁸⁵ which gives ca. 6 kcal/mol at this distance.

The aurophilic interactions in $\text{Au}_2(^3\Sigma_g^+)$ and $(\text{AuH})_2$ models are very similar (-3.6 vs -2.9 kcal/mol). Addition of the PH_3 ligands to $(\text{AuH})_2$ results in a dramatic increase in the dispersion stabilization by 4 kcal/mol, which is counterbalanced by a 2.4 kcal/mol net increase in the first-order repulsion in the minimum region.

AUTHOR INFORMATION

Corresponding Author

*E-mail: Janos.Angyan@crm2.uhp-nancy.fr (J.G.A.), bryant@oakland.edu (M.M.S.).

ACKNOWLEDGMENT

This work was supported by the Agence Nationale de Recherche (France) in the framework of the project Metal–Metal (ANR-06-BLAN-0410), by the U.S. National Science Foundation (Grant No. CHE-0719260), and by the Polish Ministry of Science and Higher Education (Grant N N204 248440).

REFERENCES

- Schmidbaur, H.; Schier, A. *Chem. Soc. Rev.* **2008**, *37*, 1931.
- Pyykkö, P. *Angew. Chem., Int. Ed.* **2004**, *43*, 4412.
- Pyykkö, P. *Chem. Soc. Rev.* **2008**, *37*, 1967.
- Walter, M.; Akola, J.; Lopez-Acevedo, O.; Jadzinsky, P. D.; Calero, G.; Ackerson, C. J.; Whetten, R. L.; Gronbeck, H.; Hakkinen, H. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 9157.
- Muniz, J.; Wang, C.; Pyykkö, P. *Chem.—Eur. J.* **2011**, *17*, 368.
- Pyykkö, P.; Zhao, Y. F. *Angew. Chem., Int. Ed. Engl.* **1991**, *30*, 604.
- Pyykkö, P.; Runeberg, N.; Mendizabal, F. *Chem.—Eur. J.* **1997**, *3*, 1451.
- Magnko, L.; Schweizer, M.; Rauhut, G.; Schutz, M.; Stoll, H.; Werner, H. J. *Phys. Chem. Chem. Phys.* **2002**, *4*, 1006.
- Pyykkö, P.; Zaleski-Ejgierd, P. *J. Chem. Phys.* **2008**, *128*, 124309.
- O'Grady, E.; Kaltsoyannis, N. *Phys. Chem. Chem. Phys.* **2004**, *6*, 680.
- Assadollahzadeh, B.; Schwerdtfeger, P. *Chem. Phys. Lett.* **2008**, *462*, 222.
- Koch, W.; Holthausen, M. C. *A Chemists' Guide to Density Functional Theory*, 2nd ed.; Wiley–VCH: Weinheim, 2001.
- Dobson, J. F.; McLennan, K.; Rubio, A.; Wang, J.; Gould, T.; Le, H. M.; Dinte, B. P. *Aust. J. Chem.* **2002**, *54*, 513.
- Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.
- Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- Langreth, D. C.; Lundqvist, B. I.; Chakarova-Kack, S. D.; Cooper, V. R.; Dion, M.; Hyldgaard, P.; Kelkkanen, A.; Kleis, J.; Kong, L.; Li, S.; Moses, P. G.; Murray, E.; Puzder, A.; Rydberg, H.; Schroder, E.; Thonhauser, T. *J. Phys. Condens. Matter* **2009**, *21*, 084203.
- Vydrov, O. A.; Voorhis, T. V. *J. Chem. Phys.* **2010**, *132*, 164113.
- Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
- Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
- Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005.
- Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *122*, 154104.
- Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- Zhang, Y.; Xu, X.; Goddard, W. A., III *Proc. Natl. Acad. Sci.* **2009**, *106*, 4963.
- Ángyán, J. G.; Gerber, I. C.; Savin, A.; Toulouse, J. *Phys. Rev. A* **2005**, *72*, 012510.
- Goll, E.; Werner, H.-J.; Stoll, H. *Chem. Phys.* **2008**, *346*, 257.
- Toulouse, J.; Gerber, I. C.; Jansen, G.; Savin, A.; Ángyán, J. G. *Phys. Rev. Lett.* **2009**, *102*, 096404.
- Janesko, B. G.; Henderson, T. M.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *130*, 081105.
- Goll, E.; Stoll, H.; Thierfelder, C.; Schwerdtfeger, P. *Phys. Rev. A* **2007**, *76*, 032507.
- Burns, L. A.; Vázquez-Mayagoitia, Á.; Sumpter, B. G.; Sherrill, C. D. *J. Chem. Phys.* **2011**, *134*, 084107.
- Schwerdtfeger, P. *Inorg. Chem.* **1991**, *30*, 1660.
- Lieb, E. H.; Oxford, S. *Int. J. Quantum Chem.* **1981**, *19*, 427.
- Lacks, D. J.; Gordon, R. G. *Phys. Rev. A* **1993**, *47*, 4681.
- Pernal, K.; Podeszwa, R.; Patkowski, K.; Szalewicz, K. *Phys. Rev. Lett.* **2009**, *103*, 263201.
- Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2009**, *5*, 719.
- Rajchel, L.; Żuchowski, P. S.; Szczesniak, M. M.; Chałasiński, G. *Chem. Phys. Lett.* **2010**, *486*, 160.
- Rajchel, L.; Żuchowski, P. S.; Szczesniak, M. M.; Chałasiński, G. *Phys. Rev. Lett.* **2010**, *104*, 163001.
- Kamiya, M.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2002**, *117*, 6010.
- Figgen, D.; Rauhut, G.; Dolg, M.; Stoll, H. *Chem. Phys.* **2005**, *311*, 227.
- Peterson, K. A.; Puzzarini, C. *Theor. Chem. Acc.* **2005**, *114*, 283.
- Basis sets were obtained from the Basis Set Exchange Database, Version 1.2.2, as developed and distributed by the Environmental and Molecular Sciences Laboratory which is part of the Pacific Northwest Laboratory, P.O. Box 999, Richland, WA 99352 and funded by the U.S. Department of Energy; see: Feller, D. *J. Comput. Chem.* **1996**, *17*, 1571.
- Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorathi, V.; Chase, J.; Liand, J.; Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045.
- Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- Hampel, C.; Peterson, K.; Werner, H.-J. *Chem. Phys. Lett.* **1992**, *190*, 1.
- Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887.
- Żuchowski, P. S.; Bussery-Honvault, B.; Moszynski, R.; Jeziorski, B. *J. Chem. Phys.* **2003**, *119*, 10497.
- Jansen, G.; Hesselmann, A. *J. Phys. Chem. A* **2001**, *105*, 646.
- Misquitta, A. J.; Szalewicz, K. *Chem. Phys. Lett.* **2002**, *357*, 301.
- Werner, H. J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. et al. *MOLPRO, version 2010.1, a package of ab initio programs*; <http://www.molpro.net> (accessed June 30, 2011).
- Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, 033201.
- Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778.
- Grüning, M.; Gritsenko, O. V.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2001**, *114*, 652.
- Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029.
- Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.
- Gerber, I. C.; Ángyán, J. G. *Chem. Phys. Lett.* **2005**, *415*, 100.
- Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.
- Fromager, E.; Toulouse, J.; Jensen, H. J. A. *J. Chem. Phys.* **2007**, *126*, 074111.
- Savin, A. In *Recent developments and Applications of Modern Density Functional Theory*; Seminario, J. M., Ed.; Elsevier: Amsterdam, 1996; pp 327–357.
- Paziani, S.; Moroni, S.; Gori-Giorgi, P.; Bachelet, G. B. *Phys. Rev. B* **2006**, *73*, 155111.
- Toulouse, J.; Colonna, F.; Savin, A. *J. Chem. Phys.* **2005**, *122*, 14110.
- Goll, E.; Werner, H.-J.; Stoll, H. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3917.
- Goll, E.; Werner, H.-J.; Stoll, H.; Leininger, T.; Gori-Giorgi, P.; Savin, A. *Chem. Phys.* **2006**, *329*, 276.

- (58) Ernzerhof, M.; Perdew, J. P. *J. Chem. Phys.* **1998**, *109*, 3313.
- (59) Henderson, T. M.; Janesko, B. G.; Scuseria, G. E. *J. Chem. Phys.* **2008**, *128*, 194105.
- (60) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 8207.
- (61) Huber, K. P.; Herzberg, G. Constants of Diatomic Molecules (data prepared by Gallagher, J. W.; Johnson, R. D., III). In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; Linstrom, P. J.; Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD; <http://webbook.nist.gov> (accessed June 30, 2011).
- (62) Because of the complex's gauche orientation, the monomer geometry has only a minimal effect on the interaction energy of $(\text{HAuPH}_3)_2$. The dispersion energy differs only by 0.07 kcal/mol in the two geometries, and the HF+Disp interaction differs by 0.16 kcal/mol at $R = 3.1\text{\AA}$.
- (63) Danovich, D.; Shaik, S. *J. Chem. Theory Comput.* **2010**, *6*, 1479.
- (64) Misquitta, A. J.; Szalewicz, K. *J. Chem. Phys.* **2005**, *122*, 214109.
- (65) Schwerdtfeger, P.; Bowmaker, G. A. *J. Chem. Phys.* **1994**, *100*, 4487(nonrelativistic). Neogrady, P.; Kello, V.; Urban, M.; Sadlej, A. *J. Int. J. Quantum Chem.* **1997**, *63*, 557(relativistic).
- (66) Granatier, J.; Urban, M.; Sadlej, A. *J. Chem. Phys. Lett.* **2010**, *484*, 154.
- (67) Pluta, T.; Sadlej, A. *J. Chem. Phys. Lett.* **1998**, *297*, 391.
- (68) Chałasiński, G.; Szczesniak, M. M.; Kendall, R. A. *J. Chem. Phys.* **1994**, *101*, 8860.
- (69) Klos, J.; Żuchowski, P.; Rajchel, L.; Chałasiński, G.; Szczesniak, M. M. *J. Chem. Phys.* **2009**, *129*, 134302.
- (70) Lee, T. J.; Taylor, P. R. *Int. J. Quant. Chem. Symp.* **1989**, *23*, 199.
- (71) Janssen, C. L.; Nielsen, I. M. B. *Chem. Phys. Lett.* **1998**, *290*, 423. Lee, T. J. *Chem. Phys. Lett.* **2003**, *372*, 362.
- (72) Patkowski, K.; Szalewicz, K. *J. Chem. Phys.* **2007**, *127*, 164103.
- (73) <http://www.molpro.net/info/current/doc/manual/node425.html> (accessed June 30, 2011).
- (74) Gutowski, M.; Piela, L. *Mol. Phys.* **1988**, *64*, 337.
- (75) Szabo, A.; Ostlund, N. S. *J. Chem. Phys.* **1977**, *67*, 4351.
- (76) Chałasiński, G.; Szczesniak, M. M. *Mol. Phys.* **1988**, *63*, 205.
- (77) Gerber, I. C.; Angyan, J. G. *J. Chem. Phys.* **2007**, *126*, 044103.
- (78) Gerber, I. C.; Angyan, J. G. *Chem. Phys. Lett.* **2005**, *416*, 370.
- (79) Toulouse, J.; Zhu, W.; Angyan, J. G.; Savin, A. *Phys. Rev. A* **2010**, *82*, 032502.
- (80) Fromager, E.; Cimraglia, R.; Jensen, H. J. A. *Phys. Rev. A* **2010**, *81*, 024502.
- (81) Janesko, B. G.; Scuseria, G. E. *Phys. Chem. Chem. Phys.* **2009**, *11*, 9677.
- (82) Janesko, B. G.; Henderson, T. M.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 034110.
- (83) Zhu, W.; Toulouse, J.; Savin, A.; Angyan, J. G. *J. Chem. Phys.* **2010**, *132*, 244108.
- (84) Riedel, S.; Pyykko, P.; Mata, R. A.; Werner, H. J. *Chem. Phys. Lett.* **2005**, *405*, 148.
- (85) Schwerdtfeger, P.; Bruce, A. E.; Bruce, M. R. M. *J. Am. Chem. Soc.* **1998**, *120*, 6587.

NOTE ADDED IN PROOF

The notation for the sr-DFT term should be understood as containing also the long-range HF component.

Charge-Transfer-Like $\pi \rightarrow \pi^*$ Excitations in Time-Dependent Density Functional Theory: A Conundrum and Its Solution

Natalia Kuritz,^{†,§} Tamar Stein,^{‡,§} Roi Baer,^{*,‡} and Leeor Kronik^{*,†}[†]Department of Materials and Interfaces, Weizmann Institute of Science, Rehovoth 76100, Israel[‡]Fritz Haber Center for Molecular Dynamics, Institute of Chemistry, Hebrew University, Jerusalem 91904, Israel

ABSTRACT: We address the conundrum posed by the well-known failure of time-dependent DFT (TDDFT) with conventional functionals for “charge-transfer-like” excitations in oligoacenes. We show that this failure is due to a small spatial overlap in orbitals obtained from the underlying single-electron orbitals by means of a unitary transformation. We further show that, as in true charge-transfer excitations, this necessarily results in failure of linear-response TDDFT with standard functionals. Range-separated hybrid functionals have been previously shown to mitigate such errors but at the cost of an empirically adjusted range-separation parameter. Here, we explain why this approach should succeed where conventional functionals fail. Furthermore, we show that optimal tuning of a range-separated hybrid functional, so as to enforce the DFT version of Koopmans’ theorem, restores the predictive power of TDDFT even for such difficult cases, without any external reference data and without any adjustable parameters. We demonstrate the success of this approach on the oligoacene series and on related hydrocarbons. This resolves a long-standing question in TDDFT and extends the scope of molecules and systems to which TDDFT can be applied in a predictive manner.

I. INTRODUCTION

Time-dependent density functional theory (TDDFT) is an approach for the calculations of excited-state properties that is based on mapping the time-dependent Schrödinger equation into an equivalent set of Schrödinger-like equations for fictitious, noninteracting electrons.^{1–5} In principle, TDDFT is an exact theory. In practice, the above-mentioned mapping relies on an exchange–correlation potential, which is a functional of the electron density, but whose exact dependence on the density is unknown. The success or failure of TDDFT therefore hinges entirely on the availability of practical and reliable approximate forms for the exchange–correlation potential.

Almost all practical TDDFT calculations are performed using the adiabatic approximation, i.e., assuming that at each moment in time the exchange–correlation potential depends only on the contemporaneous density. In local or semilocal approximations, such as the local density approximation (LDA)⁶ or the generalized gradient approximation (GGA),⁷ it is further assumed that at each point in space the exchange–correlation potential depends only on the density at this point (in LDA) or also on its gradient (in GGA). In hybrid functionals (e.g., B3LYP,^{8,9} a functional of great popularity in organic chemistry), a fraction of a nonlocal Fock-exchange operator is also used. Linear-response TDDFT calculations with these standard approximations have proven to be a remarkably accurate tool for first principles calculations of valence excitations in broad classes of molecular and nanoscale systems.^{2,4,10–13} Despite this impressive success, lingering doubts about the true predictive power of TDDFT using these approximations remain, because failures are sometimes encountered in simple and seemingly straightforward scenarios.

Perhaps the best-known example of a failure without an obvious root in the underlying formalism is the prediction of $\pi \rightarrow \pi^*$ singlet excitation energies in the oligoacene series,

$C_{2+4n}H_{4+2n}$ ($n = 2–6$).¹⁴ The two lowest such excitations, usually labeled 1L_a and 1L_b , a notation due to Platt,¹⁵ differ in character. 1L_a is dominated by the highest occupied molecular orbital (HOMO)–lowest unoccupied molecular orbital (LUMO) transition and is short-axis polarized. 1L_b is dominated by a mixture of two transitions, usually the HOMO–1–LUMO and HOMO–LUMO+1 ones, and is long-axis polarized. Grimme and Parac noticed that whereas the 1L_b excitation energy is reasonably well-predicted by TDDFT using a GGA functional (BP86)^{16,17} and very-well predicted by using a hybrid functional (B3LYP)^{8,9}, the 1L_a excitation energy is consistently and substantially underestimated by either functional. These conclusions have been confirmed since by numerous additional studies and also extended to related systems exhibiting $\pi \rightarrow \pi^*$ excitations, notably nonlinear polyaromatic hydrocarbons (see, e.g., refs 5 and 18–23). They are underscored in Figure 1, where TDDFT with both BP86 and B3LYP are compared to the approximate coupled-cluster singles and doubles method (CC2) results of Grimme and Parac,¹⁴ which we take as a reference.²⁴ For the 1L_b excitation energy, the mean error between the TDDFT-B3LYP and the CC2 results across the naphthalene to hexacene series is a satisfactory 0.04 eV. But for the 1L_a excitation energy, TDDFT-B3LYP consistently underestimates the CC2 results, with an unsatisfactory mean error of 0.47 eV.

A different arena where linear-response TDDFT calculations with the same approximate functionals are known to fail is the prediction of charge-transfer excitation energies.^{25–27} These excitations are characterized by a small spatial overlap between the initial and the final orbital of the excited electron. Here and throughout, we define the spatial overlap between two orbitals, O_{12} , as the inner product of the moduli of two orbitals

Received: April 22, 2011

Published: June 22, 2011

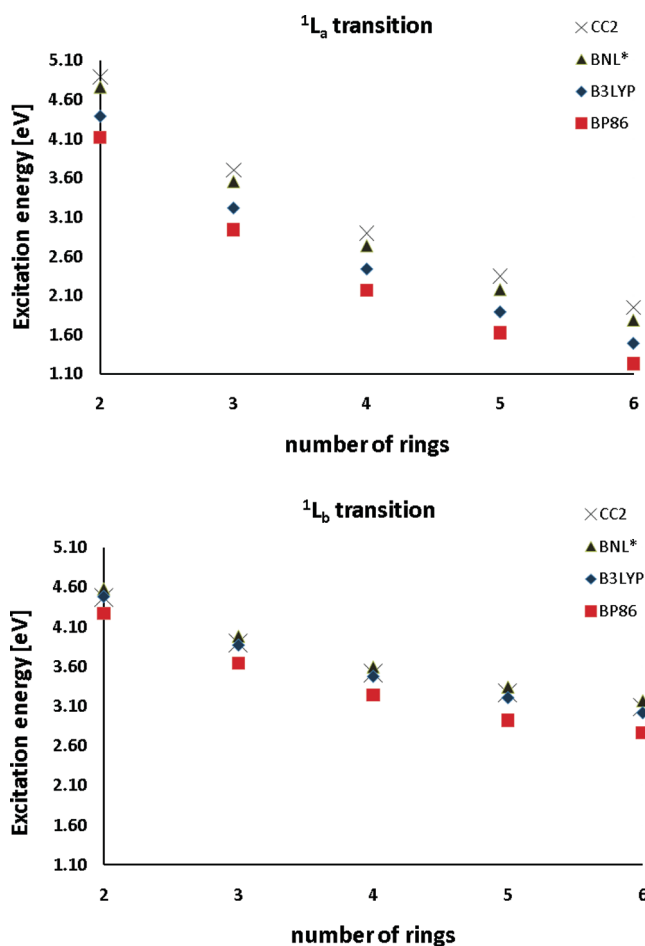


Figure 1. Excitations energies of the 1L_a (a) and 1L_b (b) transitions in the oligoacene series, $C_{2+4n}H_{4+2n}$ ($n = 2$ to 6). TDDFT data obtained with the BP86 GGA functional (red squares), the B3LYP standard hybrid functional (blue diamonds), and the optimally tuned BNL range-separated hybrid functional (green triangles) are compared to reference CC2 values, taken from ref 14 (black 'x' signs).

ψ_1 and ψ_2 ,²⁸ namely

$$O_{12} \equiv \langle \psi_1 | | \psi_2 \rangle \quad (1)$$

One perspective for this failure, given by Dreuw and Head-Gordon,²⁵ is that it is a direct consequence of the above-defined small spatial overlap. For such a case, exact (Fock) exchange would yield the correct electron–hole attraction term, but Fock exchange is completely absent in semilocal functionals and only a fraction of it is present in standard hybrid functionals. Therefore, at least within adiabatic linear response theory,²⁹ they cannot be expected to yield the correct result.

In the past few years, many studies^{30–42} have shown that the charge-transfer excitation problem can be remedied with the aid of a novel class of hybrid functionals known as range-separated hybrid (RSH) functionals.^{43–49} In this class of functionals, the repulsive Coulomb potential is split into a long-range (LR) and short-range (SR) term, e.g., via $r^{-1} = r^{-1} \text{erf}(\gamma r) + r^{-1} \text{erfc}(\gamma r)$. The components are treated differently in the generation of the exchange term. The SR exchange is represented by a local potential, typically derived from a GGA expression, whereas the LR part is treated via an “explicit” or “exact” exchange term. This affords a natural way for providing the missing Fock term for

long-ranged interactions between nonspatially overlapping orbitals while maintaining the compatibility between exchange and correlation for short-ranged interactions.

Very recently, Wong and Hsieh⁵⁰ have shown that use of a RSH functional also cures the above-discussed underestimate of 1L_a excitation energies in oligoacenes and still performs well for 1L_b excitations energies. Subsequently, Richard and Herbert⁵¹ have confirmed these findings and extended them to a wide range of nonlinear polyaromatic hydrocarbons. This success is not at all trivial because 1L_a excitations are certainly not charge-transfer excitations in the sense of eq 1, and even more refined quantitative measures of the nature of the excitation²⁸ clearly identify them as regular valence excitations.^{19,50–52} This has led Richard and Herbert⁵¹ to postulate that such excitations possess a “charge-transfer character in disguise” and to pose two important and related questions, which we paraphrase as follows: (1) How can this charge-transfer-like character be detected a priori? and (2) In the absence of accurate (ab initio or experimental) reference data, can we really trust TDDFT to have predictive power in such cases?

In this article, we propose a solution to this conundrum. First, we show that the elusive charge-transfer-like characteristics are due to a small spatial overlap in orbitals obtained from the underlying single-electron orbitals by means of a unitary transformation. Second, we show that the optimally tuned RSH functional, which we have previously established for charge-transfer excitations,^{53,54} restores the predictive power of TDDFT even for such difficult cases without any external reference data and without any adjustable parameters. The success of this approach is demonstrated on the oligoacene series as well as on related molecules.

II. THEORY OF CHARGE-TRANSFER-LIKE EXCITATIONS

To understand what is a “charge-transfer-like” excitation and how it may arise, consider the general form of the linear-response TDDFT equations based on a RSH functional of the type described above. By straightforward extension of the formalism of ref 55 from a conventional hybrid functional to an RSH one, we obtain

$$\begin{pmatrix} C & D \\ -D & -C \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \hbar\omega \begin{pmatrix} X \\ Y \end{pmatrix} \quad (2)$$

Where X , Y are the electron–hole and hole–electron components, respectively, of the eigenvector in the molecular orbital representation with

$$C_{k\sigma, j\sigma'} = (k_\sigma s_\sigma | r_{12}^{-1} | j_{\sigma'} t_{\sigma'}) + (k_\sigma s_\sigma | f'_{\sigma\sigma'} | j_{\sigma'} t_{\sigma'}) - \delta_{\sigma\sigma'} (k_\sigma j_{\sigma'} | u_\gamma(r_{12}) | s_\sigma t_{\sigma'}) + (\epsilon_{s\sigma} - \epsilon_{k\sigma}) \delta_{st} \delta_{kj} \delta_{\sigma\sigma'} \quad (3)$$

and

$$D_{k\sigma, j\sigma'} = (k_\sigma s_\sigma | r_{12}^{-1} | j_{\sigma'} t_{\sigma'}) + (k_\sigma s_\sigma | f'_{\sigma\sigma'} | j_{\sigma'} t_{\sigma'}) - \delta_{\sigma\sigma'} (k_\sigma t_{\sigma'} | u_\gamma(r_{12}) | j_{\sigma'} s_\sigma) \quad (4)$$

where σ, σ' are spin indices, k, j and s, t are, respectively, indices for occupied and unoccupied molecular orbitals, ψ , and eigenvalues, ϵ , and

$$(k_\sigma s_\sigma | r_{12}^{-1} | j_{\sigma'} t_{\sigma'}) = \iint \frac{\psi_k^\sigma(r) \psi_s^\sigma(r) \psi_j^{\sigma'}(r') \psi_t^{\sigma'}(r')}{|r - r'|} d^3r d^3r' \quad (5)$$

$$(k_{\sigma\sigma'}|f_{\sigma\sigma'}^{\gamma}|j_{\sigma'}t_{\sigma'}) = \int \psi_k^{\sigma}(r)\psi_s^{\sigma}(r)f_{\text{XC}}^{\gamma}(r;\sigma,\sigma')\psi_j^{\sigma'}(r)\psi_t^{\sigma'}(r)d^3r \quad (6)$$

$$(k_{\sigma j_{\sigma'}|u_{\gamma}(r_{12})|s_{\sigma'}t_{\sigma'}) = \iint \psi_k^{\sigma}(r)\psi_j^{\sigma'}(r)u_{\gamma}(|r-r'|)\psi_s^{\sigma}(r')\psi_t^{\sigma'}(r')d^3rd^3r' \quad (7)$$

where for simplicity all orbitals are assumed to be real and finally, where:

$$u_{\gamma}(r_{12}) = \frac{\text{erf}(\gamma r_{12})}{r} \quad \bar{u}_{\gamma}(r) = \frac{\text{erfc}(\gamma r_{12})}{r} \quad (8)$$

and $f_{\text{XC}}^{\gamma}(r;\sigma,\sigma')$ is the (semi-)local exchange–correlation kernel arising from the combination of the (semi-)local exchange corresponding to the short-range potential, $\bar{u}_{\gamma}(r)$, and the (semi-)local correlation. For the simple case of an excitation dominated by a singlet HOMO–LUMO transition, such that the contribution of all other transitions can be neglected, eq 2 reduces to a 2×2 matrix equation involving only the HOMO (H) and LUMO (L) orbitals:

$$\begin{pmatrix} c & d \\ -d & -c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \hbar\omega \begin{pmatrix} x \\ y \end{pmatrix} \quad (9)$$

where, using H(L) for the HOMO (LUMO) index, both of same spin, and $\epsilon_{\text{LH}} = \epsilon_{\text{L}} - \epsilon_{\text{H}}$, we have

$$\begin{aligned} c &= \epsilon_{\text{LH}} + d + (\text{HL}|u_{\gamma}(r_{12})|\text{HL}) - (\text{HH}|u_{\gamma}(r_{12})|\text{LL}) \\ d &= (\text{HL}|\bar{u}_{\gamma}(r_{12}) + f_{\sigma\sigma'}^{\gamma}|\text{HL}) \end{aligned} \quad (10)$$

which, after straightforward algebra yields

$$(\hbar\omega)^2 = c^2 - d^2 = \epsilon_{\text{cg}}(\epsilon_{\text{cg}} + 2d) \quad (11)$$

where we defined a “corrected gap”, ϵ_{cg} as

$$\begin{aligned} \epsilon_{\text{cg}} &= c - d \\ &= \epsilon_{\text{LH}} - (\text{HH}|u_{\gamma}(r_{12})|\text{LL}) + (\text{HL}|u_{\gamma}(r_{12})|\text{HL}) \end{aligned} \quad (12)$$

In the “extreme CT excitation case” the HOMO and LUMO orbitals are separate, i.e., their spatial overlap is vanishingly small and $|\psi_{\text{H}}(r)||\psi_{\text{L}}(r)| \approx 0$ for all r . In this case, $d = 0$, and the excitation energy equals the corrected gap: $\hbar\omega = \epsilon_{\text{cg}} = \epsilon_{\text{LH}} - (\text{HH}|u_{\gamma}(r_{12})|\text{LL})$, i.e., the HOMO–LUMO gap corrected by subtracting the long-range Coulomb energy between the electron density and the hole density. When the underlying functional is GGA the correction term is zero, and the excitation energy is equal to the KS HOMO–LUMO gap. This should be compared to the exact optical excitation energy, which in the extreme charge-transfer case is given by the Mulliken limit,⁵⁶ $\text{IP} - \text{EA} - 1/R$, where IP is the ionization potential of the donor, EA is the electron affinity of the acceptor, and R is the (large) distance between the electron and the hole. The quantity $\text{IP} - \text{EA}$ is often referred to as the fundamental gap.⁴³ Because ϵ_{LH} obtained from a local potential is much smaller than the fundamental gap⁴³ and because it is independent of R for large R , the charge-transfer excitation energy predicted from TDDFT based on GGA is usually much too low when compared to the experimental gap. We stress that this is a fundamental limitation of the formalism (as opposed to, e.g., an insufficiently accurate choice of parameters). The GGA functional possesses no mechanism that would allow for either increasing the fundamental gap value or including the $1/R$ dependence. With

B3LYP the problem is somewhat mitigated, but ϵ_{LH} is still significantly smaller than the fundamental gap and, owing to the fraction of exact exchange, only a fraction of the $1/R$ term is captured. With RSH, ϵ_{LH} can be quantitatively close to the fundamental gap,⁵⁷ and the electron–hole binding energy is close to $1/R$ for large R , which immediately explains why charge-transfer excitations are described realistically.

The above explanation of the failure of conventional functionals in describing charge-transfer excitations within linear-response TDDFT cannot be carried over, as is, to a similar phenomenon concerning L_a excitations in oligoacenes, because the HOMO and LUMO spatially overlap strongly, and so there is no charge-transfer to begin with. For example, using B3LYP for anthracene, the spatial overlap between the HOMO and the LUMO orbitals, as defined in eq 1, is 0.88. Thus a different explanation is needed.

There are many cases in the analysis of molecular orbitals where improved intuitive understanding as well as further quantitative analysis is possible with the aid of auxiliary orbitals obtained from the original ones via a unitary transformation. An important and well-known example is the effect of unitary transformations on the degree of orbital localization.^{58–61} It is therefore interesting to examine the effect of such transformations on the degree of spatial overlap. As a first step, consider two auxiliary orbitals, ψ_1 and ψ_2 , which are obtained from the HOMO and LUMO orbitals via the following simple unitary transformation:

$$\psi_1 = (\psi_{\text{H}} + \psi_{\text{L}})/\sqrt{2}, \quad \psi_2 = (\psi_{\text{H}} - \psi_{\text{L}})/\sqrt{2} \quad (13)$$

In terms of these auxiliary orbitals, eqs 10 and 12 yield

$$\begin{aligned} \epsilon_{\text{cg}} &= \epsilon_{\text{LH}} - (11|u_{\gamma}(r_{12})|22) + (12|u_{\gamma}(r_{12})|12) \\ d &= \frac{1}{4} [(11|\bar{u}_{\gamma}(r_{12}) + f_{\sigma\sigma'}^{\gamma}|11) + (22|\bar{u}_{\gamma}(r_{12}) + f_{\sigma\sigma'}^{\gamma}|22) \\ &\quad - 2(11|\bar{u}_{\gamma}(r_{12}) + f_{\sigma\sigma'}^{\gamma}|22)] \end{aligned} \quad (14)$$

and ϵ_{LH} can be expressed as

$$\begin{aligned} \epsilon_{\text{LH}} &= \langle \psi_{\text{L}} | \hat{H}_{\text{DFT}} | \psi_{\text{L}} \rangle - \langle \psi_{\text{H}} | \hat{H}_{\text{DFT}} | \psi_{\text{H}} \rangle \\ &= -2\langle \psi_1 | \hat{H}_{\text{DFT}} | \psi_2 \rangle \end{aligned} \quad (15)$$

where \hat{H}_{DFT} is the ground-state single-electron Hamiltonian corresponding to the approximate exchange–correlation functional chosen.

Importantly, the GGA-based linear-response TDDFT equations are obtained from eq 9 by considering the limit $\gamma \rightarrow 0$, in which $u_{\gamma} \rightarrow 0$ and $\bar{u}_{\gamma} \rightarrow 1/r$. In this limit, eq 14 shows that the corrected gap is equal to the HOMO–LUMO gap, and so the excitation energy is given by

$$(\hbar\omega)^2 = \epsilon_{\text{LH}}(\epsilon_{\text{LH}} + 2d) \quad (16)$$

Just as we defined the “extreme charge-transfer excitation case” to correspond to completely nonspatially overlapping HOMO and LUMO orbitals, let us now define the “extreme charge-transfer-like excitation case” to correspond to completely nonspatially overlapping auxiliary orbitals, i.e., $|\psi_1(r)||\psi_2(r)| = 0$ for every r . In practice, in oligoacenes one does not have an extreme charge-transfer-like case, but as shown in Figure 2 for the representative case of anthracene, this scenario is approximately obeyed—the spatial overlap, O_{12} , is 0.30 with B3LYP (0.29 with BNL), compared to 0.88 between the original HOMO and

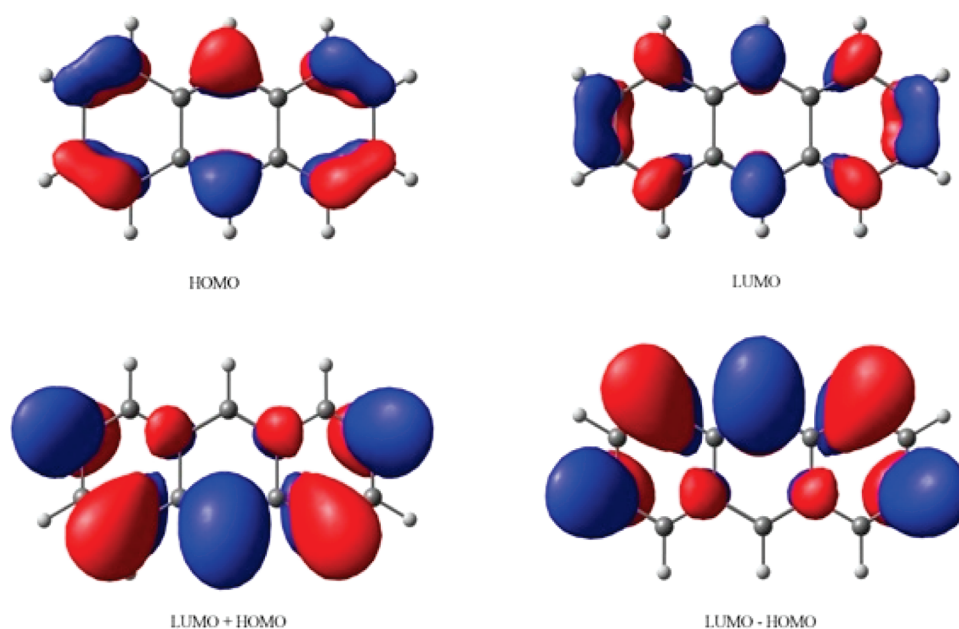


Figure 2. Orbital maps of the HOMO and LUMO orbitals (top) and their normalized sum and difference (bottom) for anthracene, as obtained from the optimally tuned BNL functional.

LUMO. In fact within a simple Hückel picture the zero overlap scenario is fully obeyed. Now, because the GGA Hamiltonian contains no long-range components, we find from eqs 15 and 16 that in this limit, perhaps counterintuitively, both the HOMO–LUMO gap, ϵ_{LH} , and the optical excitation energy, $\hbar\omega$, become vanishingly small! Therefore, the GGA-based calculation necessarily yields very small excitation energies. Just like in the true charge-transfer case, GGA is flawed here. The small spatial overlap of the auxiliary orbitals is a necessity of symmetry and thus inevitably leads to very small gaps, an error for which GGA fundamentally offers no “mechanism” of correction and therefore fails to produce realistic optical gaps. At the same time, the RSH-based calculation, where $\gamma \neq 0$, is saved from such failure because the long-range exchange terms in \hat{H}_{DFT} of eq 15 prevent the HOMO–LUMO gap from vanishing. Furthermore, similar to the CT case, the corrected gap in eq 14 is: $\epsilon_{\text{cg}} = \epsilon_{\text{LH}} - (11|u_{\gamma}(r_{12})|22)$, which differs from the HOMO–LUMO gap precisely by an “exciton binding energy”, but between the auxiliary orbital charge distributions, $|\psi_1(r)|^2$ and $|\psi_2(r)|^2$.

The similarities and differences between a charge-transfer-like and a true charge-transfer excitation are now clear: In both types of excitations, the serious errors that GGA-based calculations may arise from the presence of weakly spatially overlapping orbitals and the absence of nonlocality in the exchange–correlation kernel. And in both types of excitations, use of a Fock exchange term results in excitonic terms that correctly describe the physics of the transition. But in charge-transfer-like excitations, unlike in true charge-transfer ones, all this does not involve the orbitals obtained directly from the ground-state DFT calculation but rather a unitary transformation thereof. Consequently, the charge-transfer-like character cannot be exposed by considering only the untransformed orbitals or the density difference induced by the excitation.

We note that the above-scenario is clearly just one out of an entire family of charge-transfer-like scenarios, in which weakly spatially overlapping orbitals are obtained via a unitary transformation of the molecular ones. For example, the pertinent unitary transformation does not have to be one of a 45° rotation in the

orbital space. In this sense, the true charge-transfer excitation is simply the one obtained with the trivial (identity) unitary transformation. Furthermore, if the transition is dominated by more than one pair of orbitals, the requisite unitary matrix will be larger. An interesting special case of the latter scenario, analyzed in detail by Hieringer and Görling,^{62–64} is that of excitations in a spatially separated homodimer. There, the transition is dominated by four orbitals, two corresponding to a linear combination of the HOMO of each monomer and two corresponding to a linear combination of the LUMO of each monomer. Also in this case the excitation does not involve charge-transfer that can be deduced from density differences, and yet linear-response TDDFT based on GGA fails. But a 4×4 unitary transformation exposes the absence of spatial overlap between the HOMO of one monomer and the LUMO of the other as the true source of this failure.⁶⁵

Determining whether, and which, unitary transformation minimizes the spatial overlap of pertinent orbitals and whether weakly spatially overlapping orbitals can be obtained, then emerges as a path to deciding a priori whether a TDDFT failure associated with charge-transfer-like excitations may occur, thus answering the first challenge posted by Richard and Herbert.⁵¹

We further note that Peach et al.²⁸ introduced a “spatial overlap measure”, Λ , to be used as a diagnostic tool in a general scenario involving multiple molecular orbitals:

$$\Lambda = \frac{\sum_{ia} \kappa_{ia}^2 O_{ia}}{\sum_{ia} \kappa_{ia}^2} \quad (17)$$

where $\kappa_{ia} = X_{ia} + Y_{ia}$ (X and Y are the vector solutions of the linear-response TDDFT equations defined in eq 2 and O_{ia} is the spatial overlaps defined in eq 1 with ψ_1 the i^{th} (occupied) molecular orbital and ψ_2 the a^{th} (unoccupied) molecular orbital. A value of Λ that is too small indicates a charge-transfer situation, warning the user not to rely on standard functionals. The above discussion immediately explains the observation^{50,51} that Λ does not “sound the alarm” when a charge-transfer-like situation

arises, because the spatial overlap between the original orbitals is not small. However, the present work suggests that a relatively straightforward remedy is to seek the unitary transformation that minimizes a spatial overlap criterion such as that of Peach et al., then use as a diagnostic warning tool the value computed using the orbitals obtained with this unitary transformation.

III. PREDICTING CHARGE-TRANSFER-LIKE EXCITATIONS FROM AN OPTIMALLY TUNED RANGE-SEPARATED HYBRID FUNCTIONAL

Having laid out the general principles of charge-transfer-like excitations in the previous section, we make no attempt here at creating a comprehensive catalogue of charge-transfer-like scenarios. Instead, we move to the second and practically more pertinent challenge raised by Richard and Herbert:⁵¹ How can the predictive power of TDDFT, even in the presence of such “difficult cases”, be restored? eq 14 and its associated analysis, together with the excellent numerical results of Wong and Hsieh⁵⁰ and of Richard and Herbert,⁵¹ clearly suggest the use of an RSH functional. However, a serious difficulty remains, which is the choice of the range separation parameter, γ . The above-mentioned previous TDDFT studies of oligoacenes with RSH functionals have deduced appropriate values for γ from either coupled cluster or experimental data. While this is a perfectly valid approach, it limits the application of the method to new and unknown systems, especially given that the best choice of γ is known to generally vary with system.^{30,54,57,66,67}

In previous work on inter- and intramolecular charge-transfer excitations,^{53,54} we have shown that full predictive power can be obtained by *optimally tuning* the range-separation parameter, γ . The suggested tuning procedure has been discussed in detail elsewhere.^{53,54,57} Briefly, as mentioned above, in the limit of an infinite donor–acceptor separation, the lowest excitation energy of true charge-transfer systems reduces to the difference between the ionization potential of the donor and the electron affinity of the acceptor. Therefore, these quantities must be well predicted by ground-state DFT eigenvalues if the computation is to reduce to the correct limit. If the exact exchange–correlation functional is used, then the DFT version of Koopmans’ theorem establishes that the highest-occupied eigenvalue is equal and opposite to the ionization potential.^{68–70} This implies that an optimal choice for obtaining the correct ionization potential of an N electron system from the highest occupied RSH eigenvalue is to enforce Koopmans’ theorem, i.e., to find γ such that

$$-\epsilon_{\text{H}}^{\gamma}(N) = I^{\gamma}(N) \equiv E_{\text{gs}}(N-1; \gamma) - E_{\text{gs}}(N; \gamma) \quad (18)$$

where $I^{\gamma}(M)$ is the ionization potential of an M electron system, calculated as a ground-state energy difference, $\epsilon_{\text{H}}^{\gamma}(M)$ is the HOMO energy, and $E_{\text{gs}}(M; \gamma)$ is the ground-state energy of an M electron system. For determining the electron affinity, we employ Koopmans’ theorem also for the ionization potential of the negatively charged system, which, barring relaxation effects, is the same as electron affinity of the original system. Because there is one parameter but two conditions, we seek the γ that minimizes the overall deviation expressed in the target function:

$$J^2(\gamma) = (\epsilon_{\text{H}}^{\gamma}(N) + I^{\gamma}(N))^2 + (\epsilon_{\text{H}}^{\gamma}(N+1) + I^{\gamma}(N+1))^2 \quad (19)$$

The down-side of using eq 19 is that the optimal value of γ needs to be determined for each system of interest separately. Among

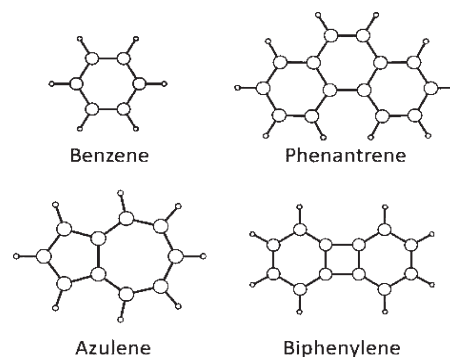


Figure 3. Schematic representation of benzene, azulene, phenanthrene, and biphenylene.

other things, this does not allow for size consistency of the functional. But a crucial observation as far as predictive power is concerned is that using eq 19 to choose the optimal γ does not require any empirical input and that the procedure contains no adjustable parameters. In other words, eq 19 is a *tuning* procedure but not a *fitting* procedure. Furthermore, it is based on upholding a physical criterion (enforcing a known and pertinent limit the exact functional must obey) rather than on semiempirical considerations. In the following, we examine whether the same approach is also useful for charge-transfer-like scenarios.

All RSH calculations presented in this work were performed using the Baer, Neuhauser, and Livshits (BNL) functional,^{30,47} as implemented in version 3.2 of Q-CHEM.⁷¹ In this functional, the long-range exchange term is a Fock-like term based on the $r^{-1} \text{erf}(\gamma r)$ potential, whereas the short-range exchange is a local expression, due to Toulouse et al.,⁷² which is based on the $r^{-1} \text{erfc}(\gamma r)$ potential. The correlation term is the standard Lee, Yang, and Parr (LYP) expression.⁷³ All BNL calculations were performed using the correlation-consistent triple- ζ basis set, cc-PVTZ,⁷⁴ which was carefully tested for convergence. All ground-state structures were optimized within a B3LYP calculation. For the oligoacenes, we used the coordinates provided by Wong and Hsieh.⁵⁰ For other molecules, we performed our own optimization.

For oligoacenes, we have previously shown⁵⁷ that optimal tuning using eq 19 (or using eq 18 for molecules without a positive electron affinity) does yield a quantitatively accurate prediction for the fundamental gaps (i.e., differences between the ionization potential and the electron affinity) throughout the series. Importantly, we found that the optimally tuned γ values decrease monotonically with system size; for naphthalene, the optimal value is 0.28, whereas for hexacene it is a significantly smaller 0.19. We have interpreted this physically as being due to the increase of electron delocalization with increasing size of this conjugated system, which renders the necessary weight of exact exchange smaller.

The TDDFT results obtained from the optimally tuned BNL calculations for the ${}^1\text{L}_a$ and ${}^1\text{L}_b$ excitation energies are shown in Figure 1, along with the previously discussed GGA and B3LYP results and the reference CC2 values. Unlike the GGA or B3LYP results, the optimally tuned BNL calculations are on par with those of the CC2 reference results for the ${}^1\text{L}_a$ excitation energy. Specifically, the GGA and B3LYP results significantly underestimate the CC2 ones by a mean value of 0.74 and 0.47 eV, respectively. But the optimally tuned BNL results only

Table 1. Lowest Singlet Excitations for Benzene, Azulene, Phenanthrene, and Biphenylene^a

molecule	BP86	B3LYP	BNL*	CC2 ⁷⁶	dominant transition	optimal γ	transition dipole moment
“Well-Predicted” Excitation							
benzene	5.32	5.47	5.50	5.27	H→L H-1→L+1	0.310	0.00
azulene	2.34	2.41	2.35	2.31	H→L	0.262	0.37 (short axis)
phenanthrene	3.66	3.98	4.19	4.04	H-1→L H→L+1	0.247	0.10 (short axis)
biphenylene	3.67	3.93	4.11	3.88	H-1→L H→L+1	0.253	1.08 (long axis)
“Poorly Predicted” Excitation							
benzene	6.07	6.16	6.40	6.56	H→L+1 H-1→L	0.310	0.00
azulene	3.49	3.64	3.79	3.95	H-1→L H→L+1	0.262	0.19 (long axis)
phenanthrene	3.92	4.22	4.57	4.70	H→L H-1→L+1	0.247	0.85 (long axis)
biphenylene	3.14	3.32	3.57	3.69	H→L	0.253	0.00

^a Excitation energies, in eV, obtained from TDDFT calculations with the BP86 GGA functional, the B3LYP standard hybrid functional, and the optimally tuned BNL range-separated hybrid functional are compared to reference CC2 values, taken from ref 71. Transitions where TDDFT with B3LYP provides satisfactory agreement with CC2 results (“well-predicted” excitations) are grouped separately from transitions where agreement was not satisfactory (“poorly predicted” excitations). Also given are the dominating transitions (where “H” stands for HOMO and “L” for LUMO) as well the optimal range-separation parameter, γ , and the resulting dipole moment (in atomic units) for the optimally tuned BNL calculations.

underestimate the CC2 ones by 0.15 eV, which is within the accepted error margin of either approach.^{54,75} Exactly as in true charge-transfer excitations, the fact that B3LYP contains a fraction of exact exchange has mitigated, but not solved, the quantitative failure resulting from the charge-transfer-like character of the excitation. We note that in refs 50 and 51 a smaller mean average error between RSH-based and CC2 results, of ~ 0.05 eV, was obtained with some of the RSH-based functionals used. Some of this difference is likely due to details of the local exchange and correlation used. However, given the limits of accuracy of the CC2 reference data themselves, our results are on par with the previous ones, without recourse to empirical parameters.

Equally importantly, the transition from B3LYP to optimally tuned BNL functional does not compromise accuracy for the ¹L_b excitation energies, where the mean error with B3LYP and BNL is 0.04 and 0.09 eV, respectively (GGA produces a less satisfactory, but perhaps tolerable, mean error of 0.28 eV). In other words, one obtains quantitative agreement with experiment (~ 0.1 – 0.2 eV), irrespective of the presence or absence of charge-transfer-like characteristics. Again, this is on par with the previous RSH-based results of refs 50 and 51 without recourse to empiricism. Thus, the second challenge raised by Richard and Herbert—achieving true predictive power—can be met even without going through the mathematical tedium of identifying the unitary transformation which minimizes the spatial overlap of pertinent orbitals, from which the nature of the failure of standard functionals becomes apparent.

For further confirmation of our computational approach, we performed similar calculations for four additional molecules: benzene, azulene, phenanthrene, and biphenylene, shown in Figure 3. These four molecules were chosen for several reasons. First, these molecules represent scenarios that are more general than that afforded by the oligacene series. Benzene is of higher

symmetry, azulene is a nonalternant hydrocarbon, phenanthrene is a nonlinear hydrocarbon, and biphenylene exhibits antiaromaticity. Second, Falden et al. recently provided wave function-based reference values for these molecules (of which we use the CC2 results for consistency).⁷⁶ Third, as summarized in Table 1, when using B3LYP each molecule exhibits both “well-predicted” low-lying excitation energies (i.e., showing differences of ~ 0.1 – 0.2 eV from the CC2 values) and “poorly-predicted” low-lying excitation energies (i.e., showing differences of ~ 0.3 – 0.5 eV from the CC2 values).

Table 1 shows that the optimized BNL results provide for a balanced and satisfactory level of accuracy (a mean error of ~ 0.15 eV for both types of excitations), despite the different nature of both the problematic and the nonproblematic transitions. Furthermore, as also shown in Table 1, each molecule possesses its own different optimally tuned range-separation parameter, underscoring the importance of optimal, molecule-specific tuning. They also lend further support our above presented theory of charge-transfer-like excitation: For azulene, unlike its alternant analogue, naphthalene, the HOMO–LUMO dominated transition is well-described by B3LYP. And indeed, for azulene no unitary transformation of the HOMO and LUMO orbitals was found to result in weakly spatially overlapping orbitals. The minimal spatial overlap was found to be 0.55 with B3LYP (0.57 with BNL), for the original HOMO and LUMO, and any 2×2 unitary transformation of the HOMO and LUMO orbitals was found to merely increase this number. This again underscores the diagnostic role unitary transformations play in uncovering whether, and which, excitations may be prone to charge-transfer-like errors. At the same time, it shows that with an optimally tuned RSH functional, such diagnostics are not essential to obtaining quantitatively predictive results.

IV. CONCLUSIONS

In conclusion, we have addressed the conundrum posed by “charge-transfer excitations in disguise”. We have shown that such excitations are due to a small spatial overlap in orbitals obtained from the underlying single-electron orbitals by means of a unitary transformation. Furthermore, we have shown that, as in true charge-transfer excitations, this necessarily results in failure of linear-response TDDFT with standard functionals. Second, we show that with optimal tuning of a range-separated hybrid functional, so as to enforce the DFT version of Koopmans’ theorem, the predictive power of TDDFT is restored even for such difficult cases without any external reference data and without any adjustable parameters. We demonstrated the success of this approach on the oligoacene series and on related hydrocarbons. This resolves a long-standing question in TDDFT and extends the scope of molecules and systems to which TDDFT can be applied in a predictive manner.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: leeor.kronik@weizmann.ac.il; roi.baer@huji.ac.il.

Author Contributions

[§]These authors contributed equally

ACKNOWLEDGMENT

Work in Rehovoth was supported by the Israel Science Foundation and the Lise Meitner Minerva Center for Computational Chemistry. Work in Jerusalem was supported by the Israel Science Foundation.

REFERENCES

- Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997–1000.
- Marques, M.; Rubio, A.; Ullrich, C. A.; Burke, K.; Nogueira, F.; Gross, A. *Time-dependent density functional theory*; Springer: Berlin, Germany, 2006.
- Onida, G.; Reining, L.; Rubio, A. *Rev. Mod. Phys.* **2002**, *74*, 601–659.
- Burke, K.; Werschnik, J.; Gross, E. K. U. *J. Chem. Phys.* **2005**, *123*, 062206.
- Elliott, P.; Furche, F.; Burke, K. In *Excited states from time-dependent density functional theory*; Wiley: Hoboken, NJ, 2009; Vol. 26.
- Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.
- Perdew, J. P.; Kurth, S. In *A Primer in Density Functional Theory*; Springer: Berlin, Germany, 2003.
- Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.
- Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- Chelikowsky, J. R.; Kronik, L.; Vasiliev, I. J. *Phys.: Condens. Mater.* **2003**, *15*, R1517–R1547.
- Silva, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*.
- Adamo, C.; Scuseria, G. E.; Barone, V. *J. Chem. Phys.* **1999**, *111*, 2889–2899.
- Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2002**, *117*, 7433–7447.
- Grimme, S.; Parac, M. *ChemPhysChem* **2003**, *4*, 292–295.
- Platt, J. R. *J. Chem. Phys.* **1949**, *17*, 484.
- Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- Parac, M.; Grimme, S. *Chem. Phys.* **2003**, *292*, 11–21.
- Kuhlman, T. S.; Mikkelsen, K. V.; Møller, K. B.; Sølling, T. I. *Chem. Phys. Lett.* **2009**, *478*, 127–131.
- Mallico, G.; Mulas, G.; Cappellini, G.; Joblin, C. *Chem. Phys.* **2007**, *340*, 43–58.
- Heinze, H. H.; Goerling, A.; Roesch, N. *J. Chem. Phys.* **2000**, *113*, 2088–2099.
- Marian, C. M.; Gilka, N. *J. Chem. Theory Comput.* **2008**, *4*, 1501–1515.
- Wang, Y. L.; Wu, G. S. *Int. J. Quantum Chem.* **2008**, *108*, 430–439.
- Although comparison to experimental results obtained in solution results in similar trends,¹⁴ here we prefer the comparison to single-molecule CC2 results in order to avoid issues related to modeling of solvent effects.
- Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007–4016.
- Tozer, D. J. *J. Chem. Phys.* **2003**, *119*, 12697–12699.
- Autschbach, J. *ChemPhysChem* **2009**, *10*, 1757–1760.
- Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.
- Ziegler, T.; Seth, M.; Krykunov, M.; Autschbach, J.; Wang, F. (*Theochem*) *J. Mol. Struct.* **2009**, *914*, 106–109.
- Livshits, E.; Baer, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932–2941.
- Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943–2946.
- Wong, B. M.; Cordaro, J. G. *J. Chem. Phys.* **2008**, *129*, 214703.
- Chai, J. D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.
- Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425–8433.
- Lange, A. W.; Rohrdanz, M. A.; Herbert, J. M. *J. Phys. Chem. B* **2008**, *112*, 6304–6308.
- Epifanovsky, E.; Polyakov, I.; Grigorenko, B.; Nemukhin, A.; Krylov, A. I. *J. Chem. Theory Comput.* **2009**, *5*, 1895–1906.
- Govind, N.; Valiev, M.; Jensen, L.; Kowalski, K. *J. Phys. Chem. A* **2009**, *113*, 6041–6043.
- Akinaga, Y.; Ten-No, S. *Int. J. Quantum Chem.* **2009**, *109*, 1905–1914.
- Andzelm, J.; Rinderspacher, B. C.; Rawlett, A.; Dougherty, J.; Baer, R.; Govind, N. *J. Chem. Theory Comput.* **2009**, *5*, 2835.
- Mohammed, A.; Agren, H.; Norman, P. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4539–4548.
- Pastore, M.; Mosconi, E.; De Angelis, F.; Gratzel, M. *J. Phys. Chem. C* **2010**, *114*, 7205–7212.
- Kerkines, I. S. K.; Petsalakis, I. D.; Theodorakopoulos, G.; Rebek, J. *J. Phys. Chem. A* **2011**, *115*, 834–840.
- Kümmel, S.; Kronik, L. *Rev. Mod. Phys.* **2008**, *80*, 3–60.
- Leininger, T.; Stoll, H.; Werner, H.-J.; Savin, A. *Chem. Phys. Lett.* **1997**, *275*, 151–160.
- Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.
- Baer, R.; Neuhauser, D. *Phys. Rev. Lett.* **2005**, *94*, 043002.
- Henderson, T. M.; Janesko, B. G.; Scuseria, G. E. *J. Chem. Phys.* **2008**, *128*, 194105.
- Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.
- Wong, B. M.; Hsieh, T. H. *J. Chem. Theory Comput.* **2010**, *6*, 3704–3712.
- Richard, R. M.; Herbert, J. M. *J. Chem. Theory Comput.* **2011**, *7*, 1296–1306.
- As explained above, here we take charge-transfer to imply a small spatial overlap between the initial and final state of the excited electron. In valence-bond language, ¹L_a transitions are “ionic” whereas ¹L_b transitions are “covalent”, hinting at charge transfer.^{14,50} However the spatial overlap of the initial and final state is still significant.
- Stein, T.; Kronik, L.; Baer, R. *J. Am. Chem. Soc.* **2009**, *131*, 2818–2820.
- Stein, T.; Kronik, L.; Baer, R. *J. Chem. Phys.* **2009**, *131*, 244119–5.
- Tretiak, S.; Chernyak, V. *J. Chem. Phys.* **2003**, *119*, 8809–8823.
- Mulliken, R. S. *J. Am. Chem. Soc.* **1950**, *72*, 600–608.

- (57) Stein, T.; Eisenberg, H.; Kronik, L.; Baer, R. *Phys. Rev. Lett.* **2010**, *105*, 266802.
- (58) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, *56*, 12847–12865.
- (59) Foster, J. M.; Boys, S. F. *Rev. Mod. Phys.* **1960**, *32*, 300.
- (60) Edmiston, C.; Ruedenberg, K. *Rev. Mod. Phys.* **1963**, *35*, 457.
- (61) Körzdörfer, T.; Kümmel, S.; Mundt, M. *J. Chem. Phys.* **2008**, *129*, 014110.
- (62) Hieringer, W.; Goerling, A. *Chem. Phys. Lett.* **2006**, *419*, 557–562.
- (63) Dreuw, A.; Head-Gordon, M. *Chem. Phys. Lett.* **2006**, *426*, 231–233.
- (64) Hieringer, W.; Goerling, A. *Chem. Phys. Lett.* **2006**, *426*, 234–236.
- (65) Kuhlman et al. (ref 19) have previously speculated on a possible connection between the failure observed in oligoacenes and that observed in the symmetric dimer but have not provided a mathematical analysis of the similarities and differences between the two cases.
- (66) Baer, R.; Livshits, E.; Neuhauser, D. *Chem. Phys.* **2006**, *329*, 266–275.
- (67) Livshits, E.; Baer, R. *J. Phys. Chem. A* **2008**, *112*, 12789–12791.
- (68) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. *Phys. Rev. Lett.* **1982**, *49*, 1691–1694.
- (69) Almbladh, C.-O.; von-Barth, U. *Phys. Rev. B* **1985**, *31*, 3231–3244.
- (70) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. T. *Phys. Rev. B* **2008**, *77*, 115123.
- (71) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C. P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L.; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.
- (72) Toulouse, J.; Savin, A.; Flad, H. J. *Int. J. Quantum Chem.* **2004**, *100*, 1047–1056.
- (73) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (74) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (75) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110–25.
- (76) Falden, H. H.; Falster-Hansen, K. R.; Bak, K. L.; Rettrup, S.; Sauer, S. P. A. *J. Phys. Chem. A* **2009**, *113*, 11995–12012.

Spin-Restriction in Explicitly Correlated Coupled Cluster Theory: The Z-Averaged CCSD(2)_{R12} Approach

Jeremiah J. Wilke and Henry F. Schaefer III*

Center for Computational Chemistry, University of Georgia, Athens, Georgia, United States

S Supporting Information

ABSTRACT: R12 methods have now been established to improve both the efficiency and accuracy of wave function-based theories. While closed-shell and spin-orbital methodologies for coupled cluster theory are well-studied, R12 corrections based on an open-shell, spin-restricted formalism have not been well developed. We present an efficient spin-restricted R12 method based on the symmetric exchange or Z-averaged approach that reduces the number of variational parameters. The current formalism reduces spin contamination relative to unrestricted methods but remains rigorously size consistent in contrast to other spin-adapted formulations. The theory is derived entirely in spin-orbital quantities, but Z-averaged symmetries are exploited to minimize the computational work in the residual equations. R12 corrections are formulated in a perturbative manner and are therefore obtained with little extra cost relative to the standard coupled cluster problem. R12 results with only a triple- ζ basis are competitive with conventional aug-cc-pVSZ and aug-cc-pV6Z results, demonstrating the utility of the method in thermochemical problems for high-spin open-shell systems.

1. INTRODUCTION

With only modestly sized basis sets, explicitly correlated methods can achieve remarkable accuracy, providing superior results for molecular energies,^{1–6} geometries,^{7,8} vibrational frequencies,^{9,10} and other properties.^{11–13} Approaches based on one-particle basis sets are unable to accurately treat the coalescence region between two electrons, which requires a cusp and a surrounding depletion of electron density (Coulomb hole) near $r_{12} = 0$.^{14–17} In many extrapolation schemes^{18,19} the error decreases as n^{-3} in the cardinal number of a Dunning correlation consistent (cc-pVnZ) basis set.²⁰ For methods with an N^4 basis set dependence, the correlation energy error only converges as $t^{-1.74}$ in the computational time!^{21,22} Supplementing the orbital basis with pair functions depending explicitly on the interelectronic distance, r_{12} , directly includes the correct Coulomb hole shape, rapidly accelerating basis set convergence. Inclusion of r_{12} -dependent terms leads to a large number of difficult three- and four-electron integrals,^{23–25} which originally restricted application to small molecules. R12 methods avoid these difficult integrals through a resolution of the identity (RI) approximation,^{26–28} factoring the numerous many-electron integrals into products of simple two-electron integrals. Closed-shell coupled-cluster singles and doubles CCSD-R12 and unrestricted CCSD-R12 based methods are now well established, and the R12 corrections add little extra cost relative to the conventional computation.

While spin restriction is trivially imposed in closed-shell coupled cluster methods, extending R12 corrections to spin-restricted and multireference methods presents several new challenges. While the concepts of normal ordering and similarity transformation are trivially formulated in a spin-free manner in closed-shell problems, open-shell problems are much more difficult since the singly occupied orbitals can correspond to both hole (occupied) and particle (unoccupied) indices. While much

focus has been given recently to multireference coupled cluster theory,²⁹ even the single-reference, high-spin problem for open shells is nontrivial. This is clearly illustrated by the numerous high-spin open-shell approaches introduced for coupled cluster, including spin-orbital restricted open-shell Hartree-Fock (ROHF-CCSD),³⁰ partially spin-adapted (PSA-CCSD),^{31–33} spin-restricted (SR-CCSD),³⁴ and spin-adapted (SA-CCSD)³⁵ coupled cluster methods. Although ROHF references spin-project correlation energy,^{36–38} they do not necessarily eliminate spin contamination in the actual wave function. Important progress was made on state specific, rigorously spin-adapted methods for high-spin open-shell problems by Li and Paldus^{39,40} and Bartlett and Nooijen.⁴¹ The normal-ordered exponential and Hausdorff expansion that forms the basis of closed-shell coupled cluster was generalized in an elegant manner but lost the simplicity of the conventional Hausdorff expansion. The majority of high-spin open-shell applications therefore still employ a spin-orbital framework even if the cluster operator is expanded in CSFs.^{32,34,35} In R12 methods, the Hausdorff expansion is central to simplifying approximations which ignore particular commutators,⁴² making some rigorously spin-adapted approaches incompatible with current R12 methods. A spin-orbital approach which uses spin restriction to both reduce computational cost and spin contamination therefore presents the most tractable framework for developing efficient coupled cluster R12 methods, especially in extending the treatment beyond CCSD.

Beyond the theoretical simplifications, spin-orbital methods can even improve accuracy in some cases. Specifically in symmetry-breaking problems, unrestricted Hartree-Fock (UHF)-like solutions that break orbital symmetries often provide

Received: April 27, 2011

Published: June 13, 2011

superior results, most notably for vibrational frequencies.⁴³ In particular, there has recently been renewed interest in unrestricted methods combined with spin constraints,⁴⁴ exploiting the physical picture of spin polarization while still avoiding the problematic spin contamination inherent in UHF.⁴⁵ For symmetry breaking problems, Brueckner methods can also improve results, but even for ROHF, references will break α, β symmetry when reoptimizing the orbitals.^{46,47} Due to the unrestricted nature of the orbitals and the need to recompute integrals on each iteration, however, tractable open-shell R12 extensions for Brueckner methods have not yet been proposed.

An “ideal” R12 coupled cluster formulation would therefore be spin restricted in nature, even in orbital-optimized approaches, like Brueckner theory, reducing computational cost and spin contamination while still maintaining a spin–orbital framework to make the R12 corrections tractable. This compromise is achieved in the Z-averaged approach of Lee and Jayatilaka^{48–51} which employs symmetric spin functions:

$$\sigma^+ = \frac{1}{\sqrt{2}}(\alpha + \beta) \quad (1)$$

$$\sigma^- = \frac{1}{\sqrt{2}}(\alpha - \beta) \quad (2)$$

in the singly occupied space, ensuring that the open-shell exchange is symmetric for α and β orbitals. Proceeding in an unrestricted formalism with no a priori assumptions, amplitude symmetry relations are automatically obtained similar to the closed-shell case which drastically reduce the computational cost. By spin-averaging exchange, ZA-CCSD can also be extended to a Brueckner theory which does not break orbital symmetries, drastically reducing the number of integrals that need to be recomputed. As an initial investigation of ZA-CCSD in explicitly correlated coupled cluster theory, we therefore present a partially spin-adapted formulation of ZA-CCSD for high-spin open-shell systems. R12 corrections are obtained based on the CCSD(2)_{RI2} ansatz of Valeev,^{52–54} which provides a noniterative, perturbative R12 correction to the conventional CCSD energy. In previous formulations⁵¹ based on symmetric spin functions, the theory was greatly complicated by the presence of α, β -exchange excitations that do not conserve M_S . We make important simplifications regarding the α, β -exchange excitations which both increase computational efficiency and actually decrease spin contamination. Excellent basis set convergence is observed for a set of atomization energies, rivaling conventional aug-cc-pVSZ and aug-cc-pV6Z computations with only an aug-cc-pVTZ orbital basis set. The new ZA-CCSD(2)_{RI2} theory therefore maximizes efficiency through accelerated basis set convergence, reduction in the number of amplitudes and residuals, and use of a determinantal formalism.

2. THEORETICAL BACKGROUND

2.1. Notation. In the current discussion, the following notation is used to distinguish orbital subspaces:

i, j, k, l, \dots	DOCC	doubly occupied orbitals
a, b, c, d, \dots	VIR	virtual orbitals
s, t, \dots	SOCC	singly occupied orbitals
p, q, \dots	ORB	computational orbital basis

Subspaces specific to R12 methods are also needed:

x, y, \dots		geminal space
p', q', \dots	CABS	external orbitals
α, β, \dots	CBS	complete basis set
p', q', \dots	RI	resolution of the identity

The RI space approximates the complete basis obtained within a finite computational basis. The complementary auxiliary basis set (CABS) space represents those orbitals in the RI that are orthogonal to the orbital basis set. The indices α, β denote orbitals in a formally complete basis but which never enter the programmable equations.

2.2. Z-averaged Coupled Cluster. For closed-shell coupled cluster, the amplitudes are related by

$$T_{i\alpha j\alpha}^{a\alpha b\alpha} = T_{i\alpha j\beta}^{a\alpha b\beta} - T_{j\alpha i\beta}^{a\alpha b\beta} \quad (3)$$

which guarantees that even if the individual determinants in the N -particle basis are not spin eigenfunctions, the total wave function will be. For open-shell cases using a truncated model space, such as CCSD or CISD, the determinantal basis does not automatically produce a spin eigenfunction even if a restricted open-shell Hartree–Fock (ROHF) reference is chosen. Due to asymmetric exchange of α and β orbital subspaces, the amplitude symmetry in eq 3 is broken, resulting in unrestricted $T_{i\alpha j\alpha}^{a\alpha b\alpha}$, $T_{i\alpha j\beta}^{a\alpha b\beta}$, and $T_{i\beta j\beta}^{a\beta b\beta}$ amplitudes. Furthermore, the projection manifold usually includes only configurations in the Møller–Plesset (MP2) first-order interacting space so that connected pseudo-triple excitations:

$$\Phi_{i\alpha j\beta s\alpha}^{a\alpha b\alpha s\beta} \quad (4)$$

are neglected despite being necessary for an exact spin eigenfunction. This presents both an efficiency concern due to the extra amplitudes and an accuracy concern due to spin contamination in the wave function. Practically speaking, for high-spin open-shell CCSD with ROHF or UHF references, spin contamination is very small^{34,55} except for molecules, such as NO₂, which have severely spin-contaminated UHF reference functions. The model space for CCSD therefore seems to be complete enough that Schrödinger projection:

$$\langle \Phi_{ij}^{ab} | \bar{H} | 0 \rangle = 0 \quad (5)$$

is enough to eliminate much of the spin contamination.

Generally speaking, the most common approach is the unrestricted coupled cluster with restricted orbitals scheme (ROHF-CCSD) described above.^{30,56,57} The coupled cluster equations are derived in a purely spin–orbital form, and separate amplitudes are allowed for each spin case. No efficiency is gained relative to UHF-based methods except that only a single set of integrals is required. Many methods were therefore developed to overcome the “unrestricted” problem of ROHF-based coupled cluster. Janssen and Schaefer³³ proposed a spin-adapted formulation employing the unitary group generators. For singly occupied orbitals, the unitary group operators contain both excitation and de-excitation operators with respect to the ROHF reference. The commutativity of the individual terms in \hat{T} is then no longer guaranteed, and the Hausdorff expansion does not truncate at quadruply nested commutators. Despite having fewer independent amplitudes than the unrestricted approach, the theory does not seem to provide a major efficiency advantage. Similar problems are encountered in the approach of Li and

Table 1. Doublet CSFs Used in SR-CCSD

$$\begin{aligned}
{}^D\tilde{\Phi}_i^a(1) &= \frac{1}{\sqrt{2}}(\Phi_{i\alpha}^{a\alpha} + \Phi_{i\beta}^{a\beta}) \\
{}^D\tilde{\Phi}_i^a(2) &= \frac{1}{\sqrt{6}}(\Phi_{i\alpha}^{a\alpha} - \Phi_{i\beta}^{a\beta} - 2\Phi_{s\alpha i\beta}^{a\alpha\beta}) \\
{}^D\tilde{\Phi}_s^a(1) &= \Phi_{s\alpha}^{a\alpha} \\
{}^D\tilde{\Phi}_i^s(1) &= \Phi_{i\beta}^{s\beta} \\
{}^D\tilde{\Phi}_{ij}^{ab}(1) &= \frac{1}{2\sqrt{3}}(2\Phi_{i\alpha j\alpha}^{a\alpha b\alpha} + 2\Phi_{i\beta j\beta}^{a\beta b\beta} + \Phi_{i\alpha j\beta}^{a\alpha b\beta} - \Phi_{j\alpha i\beta}^{a\alpha b\beta} + \Phi_{i\beta j\alpha}^{a\beta b\alpha} - \Phi_{j\beta i\alpha}^{a\beta b\alpha}) \\
{}^D\tilde{\Phi}_{ij}^{ab}(2) &= \frac{1}{2}(\Phi_{i\alpha j\beta}^{a\alpha b\beta} + \Phi_{j\alpha i\beta}^{a\alpha b\beta} + \Phi_{i\beta j\alpha}^{a\beta b\alpha} + \Phi_{j\beta i\alpha}^{a\beta b\alpha}) \\
{}^D\tilde{\Phi}_{si}^{ab}(1) &= \frac{1}{\sqrt{6}}(2\Phi_{s\alpha i\alpha}^{a\alpha b\alpha} + \Phi_{s\alpha i\beta}^{a\alpha b\beta} - \Phi_{s\alpha i\beta}^{b\alpha a\beta}) \\
{}^D\tilde{\Phi}_{si}^{ab}(2) &= \frac{1}{\sqrt{2}}(\Phi_{s\alpha i\beta}^{a\alpha b\beta} + \Phi_{s\alpha i\beta}^{b\alpha a\beta}) \\
{}^D\tilde{\Phi}_{ij}^{as}(1) &= \frac{1}{\sqrt{6}}(2\Phi_{i\beta j\beta}^{a\beta s\beta} - \Phi_{i\alpha j\beta}^{a\alpha s\beta} + \Phi_{j\alpha i\beta}^{a\alpha s\beta}) \\
{}^D\tilde{\Phi}_{ij}^{as}(2) &= \frac{1}{\sqrt{2}}(\Phi_{i\alpha j\beta}^{a\alpha s\beta} + \Phi_{j\alpha i\beta}^{a\alpha s\beta})
\end{aligned}$$

Paldus.^{39,40} A simpler partially spin-adapted (PSA-CCSD) approach was therefore proposed by Janssen,³³ Werner,^{32,58} and Neogrady,^{31,59,60} who employed spin-adapted \hat{T} operators for connected excitations resulting in spin adaptation of the linear terms. A reduction in the number of independent amplitudes is achieved, but no de-excitations are included in the cluster operator, restoring the conventional Hausdorff expansion. The independent amplitudes are determined from configuration state function (CSF) projections, producing a minimal number of residual equations. For the doublet case, CSFs are listed in Tables 1 and 2 with doublet configurations denoted by D and the external configurations denoted by Q , following the notation of Gauss and Szalay.³⁴ While the approach is almost entirely free of spin contamination, the spin expectation value is not rigorously, for example, 0.75 for doublet states, since the wave function is only spin adapted in the linear excitations.

This partial spin adaptation was extended by Gauss and Szalay^{34,35,61,62} in SR-CCSD. The Schrödinger projections (eq 5) for CSFs of the desired spin are supplemented by spin constraints on the external configurations:

$$\langle {}^Q\tilde{\Phi}_{ij}^{ab} | \exp(-\hat{T}) \hat{S}_N^2 \exp(\hat{T}) | 0 \rangle = 0 \quad (6)$$

Here subscript N denotes normal ordering with respect to the Hartree–Fock reference. The spin projections add almost no extra computational cost relative to the Schrödinger projections, and the SR-CCSD method is therefore similar in computational cost to the PSA-CCSD method. For SR-CCSD, the expectation value $\langle \hat{S}^2 \rangle$ computed from differentiation of the CCSD Lagrangian is rigorously $S(S+1)$ for a given multiplicity. We must emphasize, however, that the coupled cluster wave function is not a spin eigenfunction unless the spin constraints are extended to all configurations in the full configuration interaction (FCI) wave function. Efforts along these lines were pursued by Heckert et al.,³⁵ leading to the SA-CCSD approach. Gauss and Szalay suggested only including spin constraints for those configurations in the first-order interacting space, limiting the cluster operator to connected double excitations.³⁴ For accurate treatment of certain

Table 2. External Quartet and Sextet CSFs Used in SR-CCSD

$$\begin{aligned}
{}^Q\tilde{\Phi}_i^a(1) &= \frac{1}{\sqrt{3}}(\Phi_{i\alpha}^{a\alpha} - \Phi_{i\beta}^{a\beta} + \Phi_{s\alpha i\beta}^{a\alpha\beta}) \\
{}^Q\tilde{\Phi}_{ij}^{ab}(1) &= \frac{1}{2\sqrt{3}}(2\Phi_{i\alpha j\alpha}^{a\alpha b\alpha} - 2\Phi_{i\beta j\beta}^{a\beta b\beta} + \Phi_{i\alpha j\beta}^{a\alpha b\beta} + \Phi_{j\beta i\alpha}^{a\beta b\alpha} - \Phi_{j\alpha i\beta}^{a\alpha b\beta} - \Phi_{i\beta j\alpha}^{a\beta b\alpha}) \\
{}^Q\tilde{\Phi}_{ij}^{ab}(2) &= \frac{1}{2}(\Phi_{i\beta j\alpha}^{a\beta b\alpha} + \Phi_{j\beta i\alpha}^{a\beta b\alpha} - \Phi_{i\alpha j\beta}^{a\alpha b\beta} - \Phi_{j\alpha i\beta}^{a\alpha b\beta}) \\
{}^Q\tilde{\Phi}_{ij}^{ab}(3) &= \frac{1}{\sqrt{6}}(\Phi_{i\alpha j\alpha}^{a\alpha b\alpha} + \Phi_{i\beta j\beta}^{a\beta b\beta} - \Phi_{i\alpha j\beta}^{a\alpha b\beta} - \Phi_{i\beta j\alpha}^{a\beta b\alpha} + \Phi_{j\beta i\alpha}^{a\beta b\alpha} + \Phi_{j\alpha i\beta}^{a\alpha b\beta}) \\
{}^Q\tilde{\Phi}_{ij}^{ab}(4) &= \frac{1}{\sqrt{6}}(\Phi_{i\alpha j\alpha}^{a\alpha b\alpha} - \Phi_{i\beta j\beta}^{a\beta b\beta} - \Phi_{i\alpha j\beta}^{a\alpha b\beta} + \Phi_{i\beta j\alpha}^{a\beta b\alpha} - \Phi_{j\beta i\alpha}^{a\beta b\alpha} + \Phi_{j\alpha i\beta}^{a\alpha b\beta}) \\
{}^Q\tilde{\Phi}_{si}^{ab}(1) &= \frac{1}{\sqrt{3}}(\Phi_{s\alpha i\alpha}^{a\alpha b\alpha} - \Phi_{s\alpha i\beta}^{a\alpha b\beta} + \Phi_{s\alpha i\beta}^{b\alpha a\beta}) \\
{}^Q\tilde{\Phi}_{ij}^{as}(1) &= \frac{1}{\sqrt{3}}(\Phi_{i\beta j\beta}^{a\beta s\beta} + \Phi_{i\alpha j\beta}^{a\alpha s\beta} - \Phi_{j\alpha i\beta}^{a\alpha s\beta})
\end{aligned}$$

excited states in the equation-of-motion (EOM) method, pseudotriple excitations were also important,⁶¹ and the method was further extended through SR-CCSDT.⁶²

A fundamentally different approach to reducing the number of independent parameters was given by Lee and Jayatilaka^{50,51} by introducing a symmetric spin basis:

$$\sigma^+ = \frac{1}{\sqrt{2}}(\alpha + \beta) \quad (7)$$

$$\sigma^- = \frac{1}{\sqrt{2}}(\alpha - \beta) \quad (8)$$

for the singly occupied orbitals. This technique was denoted either open-shell CCSD (OCCSD) or Z-averaged perturbation theory (ZAPT).^{48,49} We here choose the name Z-averaged CCSD (ZA-CCSD) to emphasize that symmetry relations among the amplitudes are introduced by Z-averaging the SOCC space against the DOCC space. The exchange operator, however, no longer conserves M_S in the DOCC space, producing nonzero spin–exchange excitations:

$$K_{i\alpha}^{a\beta} = \frac{1}{2} \sum_s \langle si | as \rangle \quad (9)$$

which couple α and β subspaces via the symmetric spin orbitals. The ZA-CCSD generalization of eq 3 is therefore

$$T_{i\alpha j\alpha}^{a\alpha b\alpha} + T_{i\alpha j\alpha}^{a\beta b\beta} = T_{i\alpha j\beta}^{a\alpha b\beta} - T_{j\alpha i\beta}^{a\alpha b\beta} \quad (10)$$

where an additional set of independent amplitudes $T_{i\alpha j\alpha}^{a\beta b\beta}$ is nonzero. The ZA-CCSD approach can be greatly simplified by noting from second-order perturbation theory that α , β -exchange amplitudes are not included in the first-order interacting space and might be neglected in the cluster operator as higher-order excitations. We therefore proceed by ignoring these α , β -exchange terms.

In practice, CSF residuals are usually computed as linear combinations of single determinant projections rather than directly computed from spin-adapted matrix elements. The single CSF residual:

$$\begin{aligned}
\langle {}^D\tilde{\Phi}_{ij}^{ab}(1) | \bar{H} | 0 \rangle &= \frac{1}{2\sqrt{3}}(2R_{i\alpha j\alpha}^{a\alpha b\alpha} + 2R_{i\beta j\beta}^{a\beta b\beta} + R_{i\alpha j\beta}^{a\alpha b\beta} \\
&\quad - R_{j\alpha i\beta}^{a\alpha b\beta} + R_{i\beta j\alpha}^{a\beta b\alpha} - R_{j\beta i\alpha}^{a\beta b\alpha}) \quad (11)
\end{aligned}$$

$$R_{i\alpha j\alpha}^{a\alpha b\alpha} = \langle \Phi_{i\alpha j\alpha}^{a\alpha b\alpha} | \bar{H} | 0 \rangle \quad (12)$$

would therefore require three separate spin cases to be computed. Because of the extra ZA-CCSD symmetry in eq 10, the CSF projection in eq 11 only requires the spin-orbital residuals $R_{i\alpha j\beta}^{a\alpha b\beta}$ and $R_{i\alpha j\alpha}^{a\beta b\beta}$ (although the amplitudes $T_{i\alpha j\alpha}^{a\beta b\beta}$ are chosen zero, the corresponding residuals are not). The computational cost of $R_{i\alpha j\alpha}^{a\beta b\beta}$ is negligible, only requiring terms of the form:

$$R_{i\alpha j\alpha}^{a\beta b\beta} \leftarrow g_{c\beta d\beta}^{a\beta b\beta} T_{i\alpha}^{c\beta} T_{j\alpha}^{d\beta} \quad (13)$$

$$R_{i\alpha j\alpha}^{a\beta b\beta} \leftarrow g_{s-t}^{a\beta b\beta} T_{i\alpha j\alpha}^{s-t} \quad (14)$$

which involve either a T_1 contraction or a SOCC index, and therefore only N^5 work. Only a single N^6 spin-orbital residual $R_{i\alpha j\beta}^{a\alpha b\beta}$ must therefore be computed, drastically reducing the computational cost. The ZA-CCSD approach therefore imposes useful amplitude symmetries via the symmetric spin basis, and residuals are computed entirely in terms of determinantal rather than CSF projections.

2.3. Brueckner and CCSD(T) Extensions. The spin-restricted formalism in ZA-CCSD provides the foundation for two important, cost-saving extensions. First, Brueckner theory can be reformulated such that the orbitals are not spin polarized, drastically reducing the computational cost by maintaining spin restriction in the integral transformation on each iteration. In general, Brueckner theory reformulates the cluster operator:

$$\exp(\hat{T}) = \exp(\hat{T}_2 + \hat{T}_3 + \dots) \exp(\hat{T}_1 - \hat{T}_1^\dagger) \quad (15)$$

such that \hat{T}_1 now introduces a unitary transformation. In ZA-CCSD we have two separate components: a conventional \hat{T}_1 which preserves M_S and a $\hat{T}_1^{\alpha,\beta}$ which introduces the spin-exchange excitations. We can redefine the Brueckner cluster operator in ZA-CCSD as

$$\exp(\hat{T}) = \exp(\hat{T}_1^{\alpha,\beta} + \hat{T}_2 + \hat{T}_3 + \dots) \exp(\hat{T}_1 - \hat{T}_1^\dagger) \quad (16)$$

with only the conventional \hat{T}_1 generating a unitary transformation. Because ZA-CCSD has the amplitude symmetry:

$$T_{i\alpha}^{a\alpha} = T_{i\beta}^{a\beta} \quad (17)$$

the orbital transformation maintains equivalent α and β orbitals. This partitioning of \hat{T}_1 is consistent with the perturbative analysis of Lee and Jayatilaka who argue that the excitation rank of the operator should include not only the number of orbital substitutions but also the number of spin flips. In this regard, even though $\hat{T}_1^{\alpha,\beta}$ corresponds to a single *substitution*, it should be considered a double *excitation*. We therefore consider $\hat{T}_1^{\alpha,\beta}$ to be of rank two. Initial attempts along these lines were attempted by Crawford et al.⁴⁷ and yielded promising results, particularly for vibrational frequencies of difficult radicals like NO₃. Rigorous benchmarking of these Brueckner approximations, however, is not included in this work, as we are here only concerned with validating the R12 approximations described above within the basic Z-averaged formalism.

The ZA-CCSD formalism can also be further extended to a spin-restricted CCSD(T) for open-shells.⁶³ The CCSD(T) formalism requires the full Hamiltonian to be partitioned into a diagonal \hat{H}_0 and a perturbation \hat{V} . For ROHF references, \hat{H}_0 is usually constructed by semicanonicalization of the orbitals, diagonalizing the Fock operator separately within the occupied and the virtual subspaces. The mixed occupied-virtual terms, f_i^a ,

are neglected in \hat{H}_0 instead being included in the perturbation. Because of asymmetric exchange, however, the orbital canonicalization breaks α , β -symmetry requiring several different spin cases to be treated. In contrast, diagonalization of the symmetric-exchange Fock operator in ZA-CCSD maintains equivalent orbitals and eigenvalues for α and β subspaces, requiring only a single spin case to be treated. Additionally, in line with the Brueckner arguments given above for T_1 , for ZA-CCSD canonical orbitals, the matrix elements $f_{i\alpha}^{a\alpha}$ are rigorously zero. Unlike most ROHF-based theories, ZA-CCSD therefore has a pseudo-Brillouin condition similar to the closed-shell case, and only the off-diagonal Fock matrix elements $f_{i\alpha}^{a\beta}$ must be included in the perturbation. Again, we delay a complete discussion of the CCSD(T) extension and focus here on the R12 treatment.

2.4. Explicitly Correlated CCSD. To illustrate the essential concepts in R12 theory, we first briefly review the MP2 equations.^{16,26,64–66} In MP2-R12, the conventional residual equations are generalized to

$$R_{ab}^{ij} = f_a^c T_{cb}^{ij} + f_b^c T_{ac}^{ij} - f_k^i T_{ab}^{kj} - f_k^j T_{ab}^{ik} + C_{ab}^{xy} T_{xy}^{ij} + g_{ab}^{ij} \quad (18)$$

where f_p^a are Fock matrix elements and the T_{xy}^{ij} are explicit amplitudes associated with the pair correlations:

$$|ij\rangle \rightarrow \hat{Q}_{12} F(r_{12}) |xy\rangle \quad (19)$$

with x, y the geminal generating space, usually taken to be occupied orbitals. The correlation factor, $F(r_{12})$, is usually of Slater-type^{11,67–70} and approximated as a sum a Gaussian geminals:

$$F(r_{12}) = \exp(-\gamma r_{12}) \approx \sum_n c_n \exp(-\alpha_n r_{12}^2) \quad (20)$$

The projection operator, \hat{Q}_{12} , is included to ensure the explicit geminals are orthogonal to the reference function, \hat{o} , and virtual pairs, \hat{v}

$$\hat{Q}_{12} = (1 - \hat{o}_1)(1 - \hat{o}_2)(1 - \hat{v}_1 \hat{v}_2) \quad (21)$$

We enforce in the current work that the R12 terms contribute to the manifold of doubly excited configurations only and do not directly contribute to the singles correlation energy. Schemes have been developed for including R12 terms to correct the one-particle contribution, and we refer the reader to refs 71–73. The standard MP2 expression is augmented by the coupling matrix between conventional and R12 terms:

$$C_{ab}^{xy} = \langle ab | (\hat{f}_1 + \hat{f}_2) \hat{Q}_{12} F(r_{12}) | xy \rangle \quad (22)$$

For the explicit amplitudes, we must solve

$$R_{xy}^{ij} = V_{xy}^{ij} + B_{xy}^{wz} T_{wz}^{ij} - X_{xy}^{wz} (f_k^i T_{wz}^{kj} + f_k^j T_{wz}^{ik}) + C_{xy}^{ab} T_{ab}^{ij} \quad (23)$$

Here V_{xy}^{ij} is the R12 generalization of the electron repulsion integral:

$$V_{xy}^{ij} = \left\langle xy \left| F(r_{12}) \hat{Q}_{12} \frac{1}{r_{12}} \right| ij \right\rangle \quad (24)$$

where the conventional pair correlation function is replaced by the R12 geminal. B_{xy}^{wz} is the generalization of the energy denominator:

$$B_{xy}^{wz} = \langle xy | F(r_{12}) \hat{Q}_{12} (\hat{f}_1 + \hat{f}_2) \hat{Q}_{12} F(r_{12}) | wz \rangle \quad (25)$$

and X_{xy}^{wz} is the geminal overlap matrix:

$$X_{xy}^{wz} = \langle xy|F(r_{12})\hat{Q}_{12}F(r_{12})|wz\rangle \quad (26)$$

For further details, we refer the reader to refs 26, 64, or 65.

Rigorous development of the CCSD-R12^{74–77} model with RI approximations leads to new, difficult R12 intermediates with a severe auxiliary basis set dependence⁷⁸ which are absent in MP2. The CCSD(2)_{RTZ}^{52–54} approach formulates the R12 terms as a perturbative correction to the conventional CCSD wave function. Performing a Löwdin partitioning of the similarity transformed Hamiltonian:

$$\bar{H} = \begin{pmatrix} \bar{H}_{PP} & \bar{H}_{PQ} \\ \bar{H}_{QP} & \bar{H}_{QQ} \end{pmatrix} \quad (27)$$

The R12 geminal excitations are given as an external space, Q , while the conventional CCSD determinants comprise the reference space, P . The Hamiltonian is partitioned into zeroth- and first-order parts as

$$\bar{H}^{(0)} = \begin{pmatrix} \bar{H}_{PP} & 0 \\ 0 & \bar{H}_{QQ}^{(0)} \end{pmatrix} \quad (28)$$

$$\bar{H}^{(1)} = \begin{pmatrix} 0 & \bar{H}_{PQ} \\ \bar{H}_{QP} & \bar{H}_{QQ}^{(1)} \end{pmatrix} \quad (29)$$

Similar approximations are followed in CCSD(F12)^{79–82} and CCSD-F12a,b approaches which^{4,83} directly include R12 geminal excitations:

$$\hat{T}'_2 = T_{xy}^{ij} F_{\alpha\beta}^{xy} a_{ij}^{\alpha\beta} \quad (30)$$

$$F_{\alpha\beta}^{xy} = \langle \alpha\beta|\hat{Q}_{12}F(r_{12})|xy\rangle \quad (31)$$

in the cluster operator but approximate R12 commutators as

$$[\hat{H}, \hat{T}'_2] \approx [\hat{J}, \hat{T}'_2] \quad (32)$$

neglecting contractions between \hat{T}'_2 and the fluctuation potential. By partitioning the Hamiltonian for R12 terms, the basis set dependence of R12 matrix elements is drastically reduced, and MP2-R12 codes can be straightforwardly generalized for CCSD. Although all CCSD-R12 approaches are essentially equivalent in cost to a conventional CCSD computation, the approach of Valeev has the extra theoretical appeal of being noniterative and generated automatically by the underlying CCSD wave function. In this regard, the CCSD(2)_{RTZ} correction is more easily generalized to spin-restricted open-shell techniques or multireference coupled cluster ansätze and has therefore been called an “universally explicitly correlated coupled cluster.”⁸⁴

While the zeroth-order $H_{QQ}^{(0)}$ is trivially selected as the Fock operator for closed-shell molecules, the exact definition of \hat{H}_0 is greatly complicated for ROHF references by the nonzero occupied virtual block f_{ia} of the Fock operator. The ROHF reference function is not an eigenfunction of the Fock operator, requiring a new partition of the Hamiltonian. Numerous ROHF-based perturbation theories have therefore been developed,^{49,85,86} including open-shell perturbation theory (OPT1 and OPT2),^{86–88} invariant open-shell perturbation theory (IOPT),⁸⁹ restricted open-shell Møller-Plesset (ROMP),⁹⁰ restricted Møller-Plesset (RMP),^{37,91} and Z-averaged perturbation

theory (ZAPT).^{48–51} The most straightforward perturbation theory, RMP, currently forms the basis for the CCSD(F12), CCSD-F12a,b, and CCSD(2)_{RTZ} methods. Projection operators eliminate the undesired f_{ia} terms:

$$\hat{H}_0 = \hat{\sigma}\hat{f}\hat{\sigma} + \hat{\nu}\hat{f}\hat{\nu} \quad (33)$$

Asymmetric exchange in \hat{f} , however, breaks the spin restriction for α and β orbitals and is therefore not appropriate for the spin-restricted CCSD methods developed here. In ZA-CCSD we choose for DOCC and VIR orbitals:

$$\hat{H}_0 = \hat{h} + \hat{J} - \hat{K}_c - \frac{1}{2}\hat{K}_o \quad (34)$$

which Z-averages the open-shell exchange

$$K_o^{pq} = \sum_{s \in \text{SOCC}} \langle sp|qs\rangle \quad (35)$$

Here \hat{h} is the core Hamiltonian, \hat{J} is the Coulomb operator, and \hat{K}_c is the closed-shell exchange:

$$K_c^{pq} = \sum_{i \in \text{DOCC}} \langle ip|qi\rangle \quad (36)$$

This is conceptually equivalent to using σ^+ and σ^- orbitals within the SOCC space. For occupied SOCC orbitals:

$$\hat{H}_0 = \hat{h} + \hat{J} - \hat{K}_c - \hat{K}_o \quad (37)$$

and for unoccupied SOCC orbitals:

$$\hat{H}_0 = \hat{h} + \hat{J} - \hat{K}_c \quad (38)$$

The partition enforces α , β symmetry in the DOCC space. However, in contrast to other ROHF-based partitions,^{86–89} an eigenvalue splitting is introduced between σ^+ and σ^- orbitals. Compared to other choices of partition,^{37,88,91} the Z-averaged Fock operator here is therefore unique in maintaining α , β spin-restriction without unphysically treating occupied and virtual orbitals within the SOCC space equivalently. We emphasize again that the difference choices of \hat{H}_0 are only relevant for open-shell molecules. For closed shell, the ZA-CCSD(2)_{RTZ} method becomes equivalent to the original CCSD(2)_{RTZ} method of Valeev.⁵²

Applying Rayleigh–Schrödinger perturbation theory through second order, the energy correction is given as

$$E_2 = \bar{H}_{PQ}(\bar{H}_{QQ}^{(0)} - S_{QQ}E_0)^{-1}\bar{H}_{QP} \quad (39)$$

For the R12 correction in CCSD, the P space can be partitioned into the reference determinant Φ_0 and the set of excited determinants $\{\Phi_i^a\}$ and $\{\Phi_{ij}^{ab}\}$. The external space Q is the set of R12 configurations:

$$|\Phi_{ij}^{xy}\rangle = \Phi_{ij}^{\alpha\beta} F_{\alpha\beta}^{xy} \quad (40)$$

This leads to the matrix elements:

$$\langle \Phi_0|\bar{H}|\Phi_{ij}^{xy}\rangle = V_{ij}^{xy} \quad (41)$$

$$T_k^a \langle \Phi_a^k|\bar{H}|\Phi_{ij}^{xy}\rangle = T_i^a V_{aj}^{xy} \quad (42)$$

$$T_{kl}^{ab} \langle \Phi_{ab}^{kl}|\bar{H}|\Phi_{ij}^{xy}\rangle = V_{ij}^{xy} + P(ij)V_{aj}^{xy}T_i^a + \frac{1}{2}V_{ab}^{xy}\tau_{ij}^{ab} + P(ij)F_{ap'}^{xy}\mathcal{S}_{jb}^{p'k}T_{ik}^{ab} + \dots \quad (43)$$

where RI indices have been substituted for CBS indices and

$$\tau_{ij}^{ab} = T_{ij}^{ab} + T_i^a T_j^b - T_i^b T_j^a \quad (44)$$

Following Valeev,⁵⁴ two important approximations are made. First, to be consistent with our perturbative approach, cluster products, such as \hat{T}_1^2 , are neglected as higher order, resulting in the replacement of T_{ij}^{ab} for τ_{ij}^{ab} . For closed shell, \hat{T}_1 terms could also be neglected. However, for consistency with ROHF references, in contrast to the original formulation of Valeev,⁵² we include these terms even for closed-shell molecules following the spirit of the CCSD(T) correction.^{92,93} The more important approximation is the neglect of the term:

$$F_{ap'}^{xy} \delta_{jb}^{p'k} T_{ik}^{ab} \quad (45)$$

Under the original R12 standard approximation,¹⁶ the projection operator annihilates a, p' virtual-CABS pairs. Although the favored projection operator for MP2 applications would not annihilate this term, the current work and the work of Valeev^{53,54} demonstrate that neglecting this term still yields excellent results. Following similar arguments, we find

$$\langle \Phi_{xy}^{ij} | \bar{H} | \Phi_0 \rangle \approx V_{xy}^{ij} + T_a^i V_{xy}^{aj} + \frac{1}{2} V_{xy}^{ab} T_{ab}^{ij} \quad (46)$$

The final amplitudes are solved from the residuals equation:

$$R_{xy}^{ij} = B_{xy}^{wz} T_{wz}^{ij} - T_{wz}^{ij} X_{xy}^{wz} (f_k^i T_{xy}^{kj} + f_k^j T_{xy}^{ik}) + \tilde{V}_{xy}^{ij} \quad (47)$$

where the MP2 matrix element V_{xy}^{ij} is replaced by the intermediate:

$$\tilde{V}_{ij}^{xy} = V_{ij}^{xy} + P(ij) V_{aj}^{xy} T_i^a + \frac{1}{2} V_{xy}^{ab} T_{ab}^{ij} \quad (48)$$

which incorporates the first interference terms between conventional CCSD terms and the R12 geminals. Details of computing ZAPT matrix elements have been given previously.⁶⁵

Only the diagonal R12 amplitudes T_{ij}^{ij} are necessary to satisfy the cusp condition and may be fixed at their asymptotic cusp values of 1/2 and 1/4 for singlet and triplet pairs, respectively.^{94–98} Using the rational generator of Ten-no,^{99–101} a pair correlation can be asymptotically represented in spin-orbital form:

$$|ij\rangle \rightarrow \frac{3}{8} |ij\rangle + \frac{1}{8} p(12) |ij\rangle \quad (49)$$

where $p(12)$ is a permutation operator that permutes the spatial orbitals but not the spins. For closed shells, this leads, for example, to the pair correlation:

$$|i_\alpha j_\beta\rangle \rightarrow \frac{3}{8} |i_\alpha j_\beta\rangle + \frac{1}{8} |j_\alpha i_\beta\rangle \quad (50)$$

while for the case of ZAPT in the symmetric spin basis we have⁶⁵

$$|i_\alpha s_+\rangle \rightarrow \frac{3}{8} |i_\alpha s_+\rangle + \frac{1}{8} |s_\alpha i_+\rangle \quad (51)$$

The fixed amplitude ansatz leads to a severe dependence of the energy on the geminal exponent, γ .^{82,83} It is therefore necessary to define a second-order Lagrangian (Hylleraas functional in the case of MP2), to soften the dependence of the correlation energy on the “error” associated with fixed amplitudes. Following approximations similar to those adopted by Tew in CCSD-F12,⁸² a Lagrangian energy functional can be constructed for

the CCSD(2)_{RTZ} approach:⁵⁴

$$E_2^{\text{R12}} = 2\tilde{V}_{ij}^{xy} T_{xy}^{ij} + T_{ij}^{wz} B_{wz}^{xy} T_{xy}^{ij} - T_{ij}^{wz} X_{wz}^{xy} (f_k^i T_{xy}^{kj} + f_k^j T_{xy}^{ik}) \quad (52)$$

A similar approach is followed by Werner and Adler,^{4,83} although the explicit residual is approximated by the MP2 residual (CCSD-F12a) or additional terms from CEPA-0 (CCSD-F12b). Superior results are usually obtained within the diagonal approximation due to the absence of geminal basis set superposition errors.^{17,66}

Some comment is necessary on the nature of spin-exchange terms, such as $V_{i_\alpha j_\beta}^{\alpha\gamma}$.⁶⁵ Such terms will be nonzero by virtue of contraction with spin-exchange amplitudes $t_{i_\alpha}^{\alpha\beta}$. Clearly, the R12 correction should not be generating correlations of the form

$$|i_\alpha j_\beta\rangle \rightarrow |a_\alpha b_\alpha\rangle \quad (53)$$

as our chosen ZA-CCSD model enforces these to be zero in the basis set limit. As has been discussed previously,⁶⁵ the correlation factor is spinless and cannot produce spin-exchange configurations. We can consider the projection of the pair correlation function onto a formally complete virtual space:

$$(1 - \hat{o}_1)(1 - \hat{o}_2)F(r_{12})|xy\rangle = \sum_{\kappa\nu} |\kappa\nu\rangle \langle \kappa\nu | F(r_{12}) | xy \rangle \quad (54)$$

where the sum runs over all virtual orbitals κ, ν . Clearly, projections of the form:

$$\langle \kappa_\alpha \nu_\alpha | F(r_{12}) | x_\alpha y_\beta \rangle \quad (55)$$

are zero so that α, β spin exchanges are not generated by the correlation factor. However, virtual pairs of the form:

$$\langle \kappa_\alpha s_- | F(r_{12}) | x_\alpha y_\beta \rangle \quad (56)$$

$$\langle s_- t_- | F(r_{12}) | x_\alpha y_\beta \rangle \quad (57)$$

are nonzero due to the symmetric spin basis, allowing nonzero spin-exchange terms $V_{i_\alpha j_\beta}^{\alpha\gamma}$ from the contraction:

$$V_{i_\alpha j_\beta}^{\alpha\gamma} \leftarrow \delta_{i_\alpha j_\beta}^{\kappa_\alpha s_-} F_{\kappa_\alpha s_-}^{\alpha\gamma} \quad (58)$$

No true α, β spin exchanges are included, and the correlation factor is actually generating SOCC pairs. These terms are actually not included in the context of the diagonal ansatz anyway but will be important for other ansätze, such as those employed in response theory.¹² For the diagonal terms, by virtue of the amplitude symmetries in the conventional CCSD amplitudes, the spin-restriction symmetry, viz.:

$$\tilde{V}_{i_\alpha j_\alpha}^{i_\alpha j_\alpha} = \tilde{V}_{i_\alpha j_\beta}^{i_\alpha j_\beta} - \tilde{V}_{j_\alpha i_\beta}^{i_\alpha j_\beta} \quad (59)$$

therefore automatically extends to R12 matrix elements.

2.5. Spin Contamination. To assess spin contamination,⁵⁵ we compute expectation values as energy derivatives:

$$\hat{H}(\lambda) = \hat{H} + \lambda \hat{S}^2 \quad (60)$$

$$\langle \hat{S}^2 \rangle = \left. \frac{dE(\lambda)}{d\lambda} \right|_{\lambda=0} \quad (61)$$

The CCSD energy functional must therefore be replaced by the Lagrangian:

$$L = \langle 0 | (1 + \hat{\Lambda}) \bar{H} | 0 \rangle \quad (62)$$

Table 3. Doublet CSFs Used in ZA-CCSD

$$\begin{aligned}
D\tilde{\Phi}_i^a(1) &= \Phi_{i\alpha}^{a\alpha}, \Phi_{i\beta}^{a\beta} \\
D\tilde{\Phi}_i^a(2) &= \Phi_{i\alpha}^{a\beta}, \Phi_{i\beta}^{a\alpha} \\
D\tilde{\Phi}_s^a(1) &= \Phi_{s+\alpha}^{a\alpha}, \Phi_{s+\beta}^{a\beta} \\
D\tilde{\Phi}_s^s(1) &= \Phi_{i\alpha}^{s-}, \Phi_{i\beta}^{s-} \\
D\tilde{\Phi}_{ij}^{ab}(1) &= \frac{1}{2\sqrt{3}}(2\Phi_{i\alpha j\alpha}^{a\alpha b\alpha} + 2\Phi_{i\beta j\beta}^{a\beta b\beta} + \Phi_{i\alpha j\beta}^{a\alpha b\beta} - \Phi_{j\alpha i\beta}^{a\alpha b\beta} + \Phi_{i\beta j\alpha}^{a\beta b\alpha} - \Phi_{j\beta i\alpha}^{a\beta b\alpha}) \\
D\tilde{\Phi}_{ij}^{ab}(2) &= \frac{1}{2}(\Phi_{i\alpha j\beta}^{a\alpha b\beta} + \Phi_{j\alpha i\beta}^{a\alpha b\beta} + \Phi_{i\beta j\alpha}^{a\beta b\alpha} + \Phi_{j\beta i\alpha}^{a\beta b\alpha}) \\
D\tilde{\Phi}_{si}^{ab}(1) &= \frac{1}{2\sqrt{3}}(2\Phi_{s+\alpha i\alpha}^{a\alpha b\alpha} + 2\Phi_{s+\beta i\beta}^{a\beta b\beta} + \Phi_{s+\alpha i\beta}^{a\alpha b\beta} - \Phi_{s+\beta i\alpha}^{a\beta b\alpha} + \Phi_{s+\alpha i\alpha}^{a\beta b\alpha} - \Phi_{s+\beta i\beta}^{a\alpha b\beta}) \\
D\tilde{\Phi}_{si}^{ab}(2) &= \frac{1}{2}(\Phi_{s+\alpha i\beta}^{a\alpha b\beta} + \Phi_{s+\beta i\alpha}^{a\beta b\alpha} + \Phi_{s+\alpha i\alpha}^{a\beta b\alpha} + \Phi_{s+\beta i\beta}^{a\alpha b\beta}) \\
D\tilde{\Phi}_{ij}^{as}(1) &= \frac{1}{2\sqrt{3}}(2\Phi_{i\alpha j\alpha}^{a\alpha s-} - 2\Phi_{i\beta j\beta}^{a\beta s-} + \Phi_{i\beta j\alpha}^{a\beta s-} - \Phi_{j\beta i\alpha}^{a\beta s-} - \Phi_{i\alpha j\beta}^{a\alpha s-} + \Phi_{j\alpha i\beta}^{a\alpha s-}) \\
D\tilde{\Phi}_{ij}^{as}(2) &= \frac{1}{2}(\Phi_{i\beta j\alpha}^{a\beta s-} + \Phi_{j\beta i\alpha}^{a\beta s-} - \Phi_{i\alpha j\beta}^{a\alpha s-} - \Phi_{j\alpha i\beta}^{a\alpha s-}) \\
D\tilde{\Phi}_{si}^{at}(1) &= \Phi_{s+\alpha i\alpha}^{a\alpha t-}, \Phi_{s+\beta i\beta}^{a\beta t-}, \Phi_{s+\alpha i\beta}^{a\alpha t-}, \Phi_{s+\beta i\alpha}^{a\beta t-}
\end{aligned}$$

$$\langle 0|\hat{\Lambda} = \lambda_i^a \langle \Phi_i^a | + \lambda_{ij}^{ab} \langle \Phi_{ij}^{ab} | \quad (63)$$

where the λ_i^a and λ_{ij}^{ab} are the Lagrange multipliers. In ZA-CCSD, individual Lagrange multipliers are associated with projection by the CSFs given in Table 3. The λ_{ij}^{ab} amplitudes therefore have the same Z-averaged symmetries as the T_{ij}^{ab} amplitudes. Defining density matrix elements as

$$\gamma_r^p = \langle 0|(1 + \hat{\Lambda}) \exp(-\hat{T}) a_r^p \exp(\hat{T})|0\rangle \quad (64)$$

$$\gamma_{rs}^{pq} = \langle 0|(1 + \hat{\Lambda}) \exp(-\hat{T}) a_{rs}^{pq} \exp(\hat{T})|0\rangle \quad (65)$$

the spin expectation value can be computed. The density matrices are not Hermitian and are therefore symmetrized in the actual implementation. Further details can be found in refs 55 and 102.

3. THEORETICAL METHODS

The ZA-CCSD-R12 method was implemented within the MPQC^{103,104} and the PSI3¹⁰⁵ packages for a Slater-type correlation factor with exponent 1.4 approximated as a sum of 6 Gaussian geminals.⁶⁷ Only valence electrons were correlated. R12 intermediates were computed within approximation C so that kinetic energy commutator integrals were not necessary.¹⁰⁶ All R12 results are computed with fixed amplitudes, extended Brillouin condition (EBC), and hybrid RI approximation as described in greater detail in ref 65. Integrals were implemented in the CINTS/LIBINT¹⁰⁷ package within MPQC through a modification of the Obara–Saika scheme.^{69,108–110} Conventional ZA-CCSD computations were performed with the Dunning aug-cc-pVnZ^{20,111} family of basis sets (denoted aVnZ). The basis set limit is estimated from an aVSZ-aV6Z extrapolation using an n^{-3} extrapolation.¹⁹ All R12 computations use the Peterson F12¹¹² and F12-RI basis sets¹¹³ (hereafter denoted R12-XZ). All molecules in the current work were optimized at the CCSD(T)/cc-pVQZ level using the ACESII package.^{114–118}

Table 4. Comparison of Total Energy (Hartree)^a and \hat{S}^2 Values for Open-Shell CCSD Methods Computed with cc-pVTZ Basis

Method	Energy	$\langle \hat{S}^2 \rangle$
CH (² Π)		
ROHF-CCSD	−38.41802	0.750111
PSA-CCSD ^b	−38.41788	0.750000
SR-CCSD ^b	−38.41788	0.750000
ZA-CCSD	−38.41789	0.750006
CN (² Σ ⁺)		
ROHF-CCSD	−92.57190	0.754327
PSA-CCSD	−92.57115	0.750004
SR-CCSD	−92.57147	0.750000
ZA-CCSD	−92.57116	0.750038
NO (² Π)		
ROHF-CCSD	−129.72282	0.750371
PSA-CCSD	−129.72232	0.750001
SR-CCSD	−129.72234	0.750000
ZA-CCSD	−129.72232	0.750010
NO ₂ (² A ₂)		
ROHF-CCSD	−204.72711	0.768465
PSA-CCSD	−204.72447	0.750000
SR-CCSD	−204.72440	0.750000
ZA-CCSD	−204.72467	0.751310
HCCO (² Π)		
ROHF-CCSD	−151.69353	0.751820
PSA-CCSD	−151.69257	0.750020
SR-CCSD	−151.69274	0.750000
ZA-CCSD	−151.69262	0.750145

^a Geometries from ref 34. ^b PSA-CCSD and SR-CCSD values from ref 34.

All ROHF-CCSD and PSA-CCSD results were computed using the Molpro2006.1 package.¹¹⁹ Spin contamination values were computed by modification of the λ equation and the CCSD two-particle density codes within PSI3. See Supporting Information for a complete list of geometries, total energies, and ZA-CCSD correlation energies for all computed molecules.

4. RESULTS AND DISCUSSION

4.1. Comparison of CCSD Methodologies. Of primary concern is the performance of the newly proposed ZA-CCSD method with respect to previously established high-spin open-shell CCSD methodologies. For a series of atomic and molecular energies, the difference between the ZA-CCSD correlation energy from spin-orbital ROHF-CCSD is shown in Figure 1. Because the energy differences between the various CCSD methodologies are so small (less than 1 kcal mol^{−1}), it is not possible to identify one as intrinsically more accurate. We therefore only assert that the ZA-CCSD method should perform similar to ROHF-CCSD, and the primary benefit for ZA-CCSD will therefore be a minimum number of N^6 contractions. For the hydrocarbon radicals CH and CH₃, the difference is only 0.1 kcal mol^{−1}. However, for CN, NO₂, and NO with much greater spin contamination, deviations are as large as 0.7 kcal mol^{−1}, which is

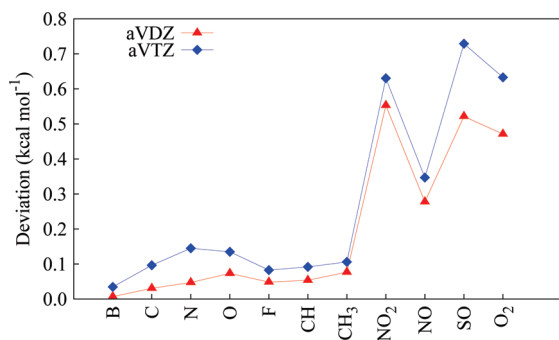


Figure 1. Correlation energy difference of Z-averaged CCSD from spin-orbital ROHF-CCSD for aug-cc-pVXZ (aVXZ) basis sets.

consistent with the partial spin-restrictions imposed. The difference is similar to that observed previously for PSA-CCSD and SR-CCSD.³⁴ A more subtle effect is observed for higher spin multiplicities, which generally show greater deviations than doublets as seen for nitrogen, oxygen, SO, and O₂. The α , β spin-exchange amplitudes occur through contractions involving symmetric spin orbitals. Since these spin-exchange terms are neglected in the current treatment, larger deviations from ROHF-CCSD will occur as the number of singly occupied orbitals is increased.

For a series of model systems considered in previous work,³⁴ we have also computed $\langle \hat{S}^2 \rangle$. For CH, CN, NO, and HCCO, the ZA-CCSD results are essentially equivalent to PSA-CCSD, both of which demonstrate significant reductions in spin contamination relative to ROHF-CCSD. Energies match PSA-CCSD within 0.1 mH and even SR-CCSD within 0.3 mH. The ZA-CCSD spin contamination is slightly greater than PSA-CCSD, consistent with ZA-CCSD imposing fewer spin constraints. For the severe case of NO₂, ZA-CCSD deviates from SR-CCSD and PSA-CCSD but still exhibits a drastic reduction in spin contamination relative to ROHF-CCSD.

There is some ambiguity in separate trade-offs between the spin-restricted and unrestricted approaches. It has long been known that perturbation theories based on restricted reference functions converge much more quickly with respect to perturbation order than those based on unrestricted references.^{120–124} One therefore anticipates that reduced spin contamination will accelerate convergence with respect to the coupled cluster CCSD, CCSDT, and CCSDTQ hierarchy. Conversely, for certain symmetry breaking problems, UHF or Brueckner solutions which break orbital symmetries may provide more accurate results at the truncated CCSD level.^{43,47,125,126} Given the small energy differences observed here between the different methodologies, a discussion of which method is intrinsically more accurate by comparison to experiment would not be meaningful at the CCSD level. We therefore emphasize instead that the approximations made in the Z-averaged ansatz will not degrade the accuracy relative to other CCSD methods (indeed, spin-contamination is reduced!), but a great deal of efficiency is gained by minimizing the number of independent amplitudes and residuals. The theoretical three-fold speedup is intrinsic to the ZA-CCSD method regardless of the inclusion of R12 corrections.

4.2. Size Consistency of Spin-Adapted Approaches. The introduction of amplitude constraints in PSA-CCSD can actually introduce size-consistency errors when the cluster operator \hat{T} is truncated at double excitations. For example, the computed

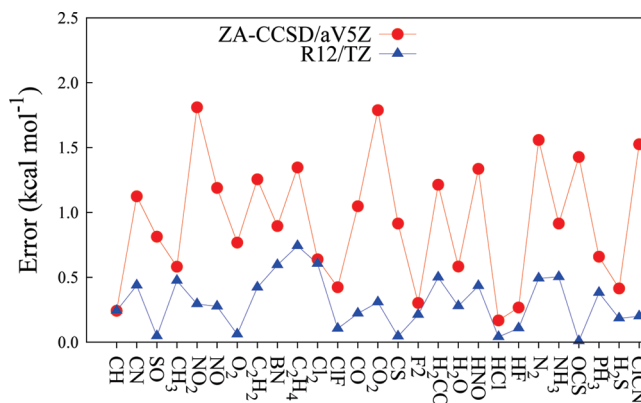


Figure 2. Error of correlation energy increments ($\delta[\text{CCSD}]$, see text) to atomization energies for Z-averaged CCSD with aug-cc-pVSZ (ZA-CCSD/aVSZ) and ZA-CCSD(2)_{R12}/TZ with cc-pVTZ-F12 (R12/TZ) basis set with respect to aVSZ/aV6Z extrapolation benchmark.

energy of two fluorine atoms at infinite separation coupled to a triplet will not be exactly twice the energy of an isolated fluorine atom in its doublet ground state. The imposed spin constraints effectively introduce an infinite-range interaction between the monomers. For multiplicities of triplet or higher, PSA-CCSD employs the “semi-internal” triplet CSF:

$$T \tilde{\Phi}_{is}^{as} = \Phi_{i\alpha}^{a\alpha} - \Phi_{i\beta}^{a\beta} - \frac{m+1}{m+2} \Phi_{s\alpha i\beta}^{a\alpha s\beta} + \frac{1}{m+2} \sum_{t \neq s} \Phi_{\alpha i\beta}^{a\alpha t\beta} \quad (66)$$

along with the spin-symmetric single excitation

$$T \tilde{\Phi}_i^a = \Phi_{i\alpha}^{a\alpha} + \Phi_{i\beta}^{a\beta} \quad (67)$$

In this regard, semi-internal doubly excited determinants are coupled by spin constraints to singly excited determinants. Assuming localized orbitals for centers A and B, suppose we have orbitals i, a, s on center A and singly occupied orbital t on center B. Upon convergence, we will have nonzero spin-orbital amplitudes $T_{i\alpha}^{a\alpha}$ and $T_{i\beta}^{a\beta}$ which by virtue of asymmetric exchange will be spin-polarized such that $T_{i\alpha}^{a\alpha} \neq T_{i\beta}^{a\beta}$. To introduce spin polarization of the single amplitudes, we must have a nonzero T_{si}^{as} for the CSF in eq 66. By virtue of the spin constraints, we must have a nonzero amplitude $T_{td\beta}^{a\alpha t\beta}$ where the *connected* T_2 amplitude now contains mixed indices for centers A and B, leading to the size consistency error. The Z-averaged formalism only ensures $T_{i\alpha}^{a\alpha} = T_{i\beta}^{a\beta}$ and does not constrain the T_1 amplitudes in any way relative to semi-internal doubly excited configurations. The numerical importance of this size-consistency error is probably negligible in many systems. For triplet fluorine dissociation the size consistency error is only 0.043 mH. Still, ZA-CCSD has the additional theoretical appeal of being rigorously size consistent at all levels of truncation in the cluster operator.

4.3. R12 Corrections. For simplicity, we here focus only on the correlation increment for the energy difference $A \rightarrow B$

$$\delta[\text{CCSD}] = E_{\text{CCSD}}^{\text{corr}}(B) - E_{\text{CCSD}}^{\text{corr}}(A) \quad (68)$$

and neglect the basis set error associated with the Hartree-Fock reference. The accelerated basis set convergence of the CCSD-R12 corrections is demonstrated for a series of atomization energies with the CBS limit taken as an aVSZ-aV6Z extrapolation. Figure 2 clearly demonstrates that even with the cc-pVTZ-F12 basis, the method even outperforms the conventional aVSZ

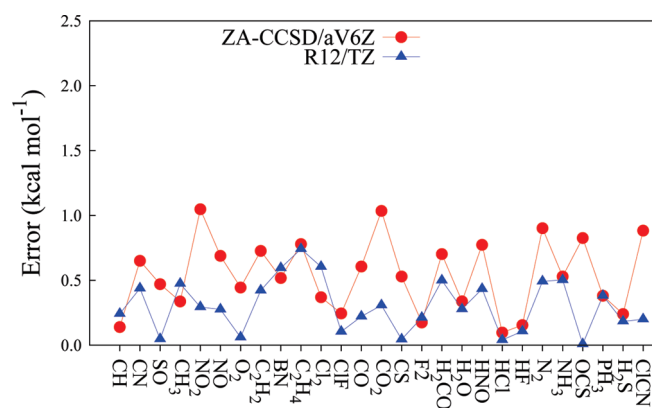


Figure 3. Error of correlation energy increments ($\delta[\text{CCSD}]$, see text) to atomization energies for Z-averaged CCSD with aug-cc-pV6Z (ZA-CCSD/aV6Z) and ZA-CCSD(2) $_{\text{RTZ}}$ with cc-pVTZ-F12 (R12/TZ) basis set with respect to aV5Z/aV6Z extrapolation benchmark.

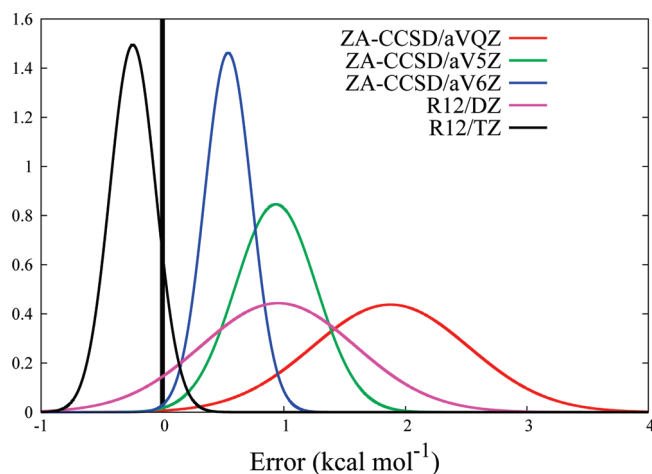


Figure 4. Error distribution of correlation energy increments ($\delta[\text{CCSD}]$, see text) to atomization energies for Z-averaged CCSD with aug-cc-pVXZ (ZA-CCSD/aVXZ) and ZA-CCSD(2) $_{\text{RTZ}}$ with cc-pVXZ-F12 (R12/XZ) basis sets with respect to aV5Z/aV6Z extrapolation benchmark.

basis. In Figure 3, the conventional aV6Z results are nearly equivalent to the TZ-R12 method but have a dramatically higher computational cost. The aV6Z basis is four times larger! The R12 results could be further improved by extending to the cc-pVQZ-F12 basis and performing a n^{-4} $T-Q$ extrapolation,^{112,127} which has been demonstrated previously to yield excellent results for atomization energies.⁶⁵ The total correlation energies of individual molecules for the cc-pVTZ-F12 basis themselves should be converged to be better than 0.1 mH per valence electron. For example, for oxygen atom, the correlation energy difference between TZ-R12 and QZ-R12 computations is only 0.5 mH (see Supporting Information). Although cancellation of errors likely contributes in some fashion for the R12 corrections (as it does for the conventional methods), the observed basis set convergence results primarily from the intrinsic accuracy of the individual correlation energies.

The accuracy and precision of the results are summarized in Figure 4. The observed basis set convergence of the Z-averaged scheme closely parallels atomization energy results observed for

the RMP formulation of R12 matrix elements.⁶⁶ Although not explicitly used for atomization energies, formulations based on UHF have demonstrated excellent basis set convergence as well.¹⁰⁰ In this regard, the basis set convergence properties do not seem sensitive to the particular choice of \hat{H}_0 used in computing R12 matrix elements. We therefore again emphasize that the primary advantage of the ZA-CCSD(2) $_{\text{RTZ}}$ approach is greatly increased computational efficiency and reduced spin contamination from spin restriction in the wave function.

The success of the cc-pVTZ-F12 results is quite dramatic when all of the approximations involved are considered. For the R12 matrix elements, the EBC assumes the virtual space is closed under the action of the Fock operator. The virtual space is relatively more complete in describing the mean field correlation of the Fock operator than the cusp region, and the EBC is therefore nearly exact for basis sets of at least triple- ζ quality.²⁴ The R12 amplitudes themselves are also not optimized, being fixed at their asymptotic value. However, this fixed amplitude approach yields precise results provided a residuals correction is included via a Lagrangian functional. A Hermiticity simplification is further made on the left- and right-hand eigenvectors. However, to a first approximation (linearized coupled cluster), the left and right amplitudes are equal so that the symmetric amplitude assumption appears justified. All of the above approximations therefore seem appropriate for the desired level of accuracy.

5. CONCLUSION

The broad range of thermochemical and spectroscopic problems associated with high-spin, open-shell radicals clearly motivates maximizing the efficiency of current CCSD methods. The ZA-CCSD(2) $_{\text{RTZ}}$ approach provides results competitive with CCSD/aV6Z with only a triple- ζ quality basis. For rigorous thermochemistry, a $T-Q$ R12 extrapolation should therefore provide unrivaled accuracy in computing energies at the basis set limit. We have employed the Z-averaged ansatz to make two important simplifications. Z-averaged symmetries are exploited in the conventional CCSD equations to give spin-restricted amplitudes, yielding a theoretical 3-fold speedup over current open-shell R12 methods. The Z-averaged partitioning of the Hamiltonian is also employed in formulating perturbative R12 corrections, leading to greatly simplified, spin-restricted R12 intermediates. Beyond conventional singles and doubles, the Z-averaged approach also provides the basis for an improved Brueckner theory and a perturbative triples correction based on spin-restricted amplitudes. While some complications arise associated with α , β -exchange excitations, ZA-CCSD remains inherently simple in its spin-orbital form. Results are similar to existing CCSD methods, and ZA-CCSD therefore improves computational efficiency without any reduction in accuracy.

■ ASSOCIATED CONTENT

S Supporting Information. Supporting Information is available containing a complete list of geometries, total energies, and ZA-CCSD correlation energies for all computed molecules. This information is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: fri@uga.edu.

ACKNOWLEDGMENT

This work was supported by U.S. Department of Energy Office of Basic Energy Sciences, Chemical Sciences Division, Combustion Program, grant no. DEF602-97-ER14748. We thank Dr. Andrew Simmonett for insightful discussions. Special thanks goes out to Dr. David Tew, Dr. Florian Bischoff, and Professor Willem Klopper whose instruction in explicitly correlated methods made this work possible. We also thank Professor Edward Valeev and Dr. Curtis Janssen for their help with the MPQC program.

REFERENCES

- Höfener, S.; Tew, D. P.; Klopper, W.; Helgaker, T. *Chem. Phys.* **2009**, *356*, 25–30.
- Aguilera-Ipparraguirre, J.; Boese, A. D.; Klopper, W.; Ruscic, B. *Chem. Phys.* **2008**, *346*, 56–68.
- Klopper, W.; Ruscic, B.; Tew, D. P.; Bischoff, F. A.; Wolfsegger, S. *Chem. Phys.* **2009**, *356*, 14–24.
- Knizia, G.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054104.
- Marchetti, O.; Werner, H.-J. *J. Phys. Chem. A* **2009**, *113*, 11580–11585.
- Flores, J. R.; Gdanitz, R. J. *J. Chem. Phys.* **2005**, *123*, 144316.
- Heckert, M.; Kállay, M.; Tew, D. P.; Klopper, W.; Gauss, J. J. *Chem. Phys.* **2006**, *125*, 044108.
- Cardoen, W.; Gdanitz, R. J. *J. Chem. Phys.* **2005**, *123*, 024304.
- Rauhut, G.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054105.
- Tew, D. P.; Klopper, W.; Heckert, M.; Gauss, J. J. *J. Phys. Chem. A* **2007**, *111*, 11242–11248.
- Hanauer, M.; Köhn, A. J. *Chem. Phys.* **2009**, *131*, 124118.
- Köhn, A. J. *Chem. Phys.* **2009**, *130*, 201103.
- Kordel, E.; Villani, C.; Klopper, W. *J. Chem. Phys.* **2005**, *122*, 214306.
- Hylleraas, E. A. *Z. Phys.* **1929**, *54*, 347–366.
- Kutzelnigg, W. *Theor. Chim. Acta* **1985**, *68*, 445–469.
- Kutzelnigg, W.; Klopper, W. *J. Chem. Phys.* **1991**, *94*, 1985–2001.
- Tew, D. P.; Klopper, W. *J. Chem. Phys.* **2006**, *125*, 094302.
- Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.
- Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.
- Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- Samson, C. C. M.; Klopper, W.; Helgaker, T. *Comput. Phys. Commun.* **2002**, *149*, 1–10.
- Klopper, W.; Röhse, R. *Theor. Chim. Acta* **1992**, *83*, 441–453.
- Boys, S. F.; Handy, N. C. *Proc. R. Soc. London, Ser. A* **1969**, *311*, 309–329.
- May, A. J.; Valeev, E. F.; Polly, R.; Manby, F. R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2710–2713.
- Ten-No, S. *Chem. Phys. Lett.* **2000**, *330*, 175–179.
- Klopper, W.; Samson, C. C. M. *J. Chem. Phys.* **2002**, *116*, 6397–6410.
- Valeev, E. F. *Chem. Phys. Lett.* **2004**, *395*, 190–195.
- Klopper, W. *J. Chem. Phys.* **2004**, *120*, 10890–10895.
- Evangelista, F. A.; Simmonett, A. C.; Schaefer, H. F.; Mukherjee, D.; Allen, W. D. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4728–4741.
- Gauss, J.; Lauderdale, W. J.; Stanton, J. F.; Watts, J. D.; Bartlett, R. J. *Chem. Phys. Lett.* **1991**, *182*, 207–215.
- Neogrady, P.; Urban, M.; Hubac, I. *J. Chem. Phys.* **1992**, *97*, 5074–5080.
- Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1993**, *99*, 5219–5227.
- Janssen, C. L.; Schaefer, H. F. *Theor. Chim. Acta* **1991**, *79*, 1–42.
- Szalay, P. G.; Gauss, J. *J. Chem. Phys.* **1997**, *107*, 9028–9038.
- Heckert, M.; Heun, O.; Gauss, J.; Szalay, P. G. *J. Chem. Phys.* **2006**, *124*, 124105.
- Rittby, M.; Bartlett, R. J. *J. Phys. Chem.* **1988**, *92*, 3033–3036.
- Lauderdale, W. J.; Stanton, J. F.; Gauss, J.; Watts, J. D.; Bartlett, R. J. *Chem. Phys. Lett.* **1991**, *187*, 21–28.
- Watts, J. D.; Bartlett, R. J. *J. Chem. Phys.* **1990**, *93*, 6104–6105.
- Li, X.; Paldus, J. *J. Chem. Phys.* **1994**, *101*, 8812–8826.
- Li, X.; Paldus, J. *J. Chem. Phys.* **1995**, *102*, 8897–8905.
- Nooijen, M.; Bartlett, R. J. *J. Chem. Phys.* **1996**, *104*, 2652–2668.
- Köhn, A.; Tew, D. P. *J. Chem. Phys.* **2010**, *133*, 174117.
- Sherrill, C. D.; Krylov, A. I.; Byrd, E. F. C.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 4171–4181.
- Tsuchimochi, T.; Scuseria, G. E. *J. Chem. Phys.* **2011**, *134*, 064101.
- Tsuchimochi, T.; Scuseria, G. E. *J. Chem. Phys.* **2010**, *133*, 141102.
- Hampel, C.; Peterson, K. A.; Werner, H.-J. *Chem. Phys. Lett.* **1992**, *190*, 1–12.
- Crawford, T. D.; Lee, T. J.; Handy, N. C.; Schaefer, H. F. *J. Chem. Phys.* **1997**, *107*, 9980–9984.
- Lee, T. J.; Jayatilaka, D. *Chem. Phys. Lett.* **1993**, *201*, 1–10.
- Lee, T. J.; Rendell, A. P. L.; Dyall, K. G.; Jayatilaka, D. *J. Chem. Phys.* **1994**, *100*, 7400–7409.
- Jayatilaka, D.; Lee, T. J. *Chem. Phys. Lett.* **1992**, *199*, 211–219.
- Jayatilaka, D.; Lee, T. J. *J. Chem. Phys.* **1993**, *98*, 9734–9747.
- Valeev, E. F. *Phys. Chem. Chem. Phys.* **2008**, *10*, 106–113.
- Valeev, E. F.; Crawford, T. D. *J. Chem. Phys.* **2008**, *128*, 244113.
- Torheyden, M.; Valeev, E. F. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3410–3420.
- Stanton, J. F. *J. Chem. Phys.* **1994**, *101*, 371–374.
- Watts, J. D.; Gauss, J.; Bartlett, R. J. *Chem. Phys. Lett.* **1992**, *200*, 1–7.
- Scuseria, G. E. *Chem. Phys. Lett.* **1991**, *176*, 27–35.
- Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **2000**, *112*, 3106–3107.
- Neogrady, P.; Urban, M.; Hubac, I. *J. Chem. Phys.* **1994**, *100*, 3706–3716.
- Neogrady, P.; Urban, M. *Int. J. Quantum Chem.* **1995**, *55*, 187–203.
- Szalay, P. G.; Gauss, J. *J. Chem. Phys.* **2000**, *112*, 4027–4036.
- Berente, I.; Szalay, P. G.; Gauss, J. *J. Chem. Phys.* **2002**, *117*, 7872–7881.
- Crawford, T. D.; Lee, T. J.; Schaefer, H. F. *J. Chem. Phys.* **1997**, *107*, 7943–7950.
- Werner, H.-J.; Adler, T. B.; Manby, F. R. *J. Chem. Phys.* **2007**, *126*, 164102.
- Wilke, J. J.; Schaefer, H. F. *J. Chem. Phys.* **2009**, *131*, 244116.
- Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 154103.
- Tew, D. P.; Klopper, W. *J. Chem. Phys.* **2005**, *123*, 074101.
- Ten-No, S. *Chem. Phys. Lett.* **2004**, *398*, 56–61.
- Höfener, S.; Bischoff, F. A.; Gloess, A.; Klopper, W. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3390–3399.
- Ten-No, S. *J. Chem. Phys.* **2007**, *126*, 014108.
- Köhn, A.; Tew, D. P. *J. Chem. Phys.* **2010**, *132*, 024101.
- Noga, J.; Šimunek, J. *Chem. Phys.* **2009**, *356*, 1–6.
- Noga, J.; Kedzuch, S.; Šimunek, J. *J. Chem. Phys.* **2007**, *127*, 034106.
- Noga, J.; Kutzelnigg, W. *J. Chem. Phys.* **1994**, *101*, 7738–7762.
- Noga, J.; Tunega, D.; Klopper, W.; Kutzelnigg, W. *J. Chem. Phys.* **1995**, *103*, 309–320.
- Noga, J.; Valiron, P. *Chem. Phys. Lett.* **2000**, *324*, 166–174.
- Müller, H.; Kutzelnigg, W.; Noga, J. *Mol. Phys.* **1997**, *92*, 535–546.
- Köhn, A.; Richings, G. W.; Tew, D. P. *J. Chem. Phys.* **2008**, *129*, 201103.
- Fliegl, H.; Hättig, C.; Klopper, W. *Int. J. Quantum Chem.* **2006**, *106*, 2306–2317.
- Fliegl, H.; Klopper, W.; Hättig, C. *J. Chem. Phys.* **2005**, *122*, 084107.

- (81) Tew, D. P.; Klopper, W.; Neiss, C.; Hättig, C. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1921–1930.
- (82) Tew, D. P.; Klopper, W.; Hättig, C. *Chem. Phys. Lett.* **2008**, *452*, 326–332.
- (83) Adler, T. B.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2007**, *127*, 221106.
- (84) Torheyden, M.; Valeev, E. F. *J. Chem. Phys.* **2009**, *131*, 171103.
- (85) Crawford, T. D.; Schaefer, H. F.; Lee, T. J. *J. Chem. Phys.* **1996**, *105*, 1060–1069.
- (86) Murray, C. W.; Handy, N. C. *J. Chem. Phys.* **1992**, *97*, 6509–6516.
- (87) Murray, C. W.; Davidson, E. R. *Int. J. Quantum Chem.* **1992**, *43*, 755–768.
- (88) Murray, C. W.; Davidson, E. R. *Chem. Phys. Lett.* **1991**, *187*, 451–454.
- (89) Kozłowski, P. M.; Davidson, E. R. *Chem. Phys. Lett.* **1994**, *226*, 440–446.
- (90) Amos, R. D.; Andrews, J. S.; Handy, N. C.; Knowles, P. J. *Chem. Phys. Lett.* **1991**, *185*, 256–264.
- (91) Knowles, P. J.; Andrews, J. S.; Amos, R. D.; Handy, N. C.; Pople, J. A. *Chem. Phys. Lett.* **1991**, *186*, 130–136.
- (92) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.
- (93) Raghavachari, K.; Pople, J. A.; Replogle, E. S.; Head-Gordon, M. *J. Phys. Chem. A* **1990**, *94*, 5579–5586.
- (94) Kato, T. *Comm. Pure Appl. Math.* **1957**, *10*, 151–177.
- (95) Pack, R. T.; Brown, W. B. *J. Chem. Phys.* **1966**, *45*, 556–559.
- (96) King, H. F. *Theor. Chim. Acta* **1996**, *94*, 345–381.
- (97) Rosas-Garcia, V. M.; Crawford, T. D. *J. Chem. Phys.* **2003**, *118*, 2491–2497.
- (98) Tew, D. P. *J. Chem. Phys.* **2008**, *129*, 014104.
- (99) Ten-No, S. *J. Chem. Phys.* **2004**, *121*, 117–129.
- (100) Bokhan, D.; Ten-No, S.; Noga, J. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3320–3326.
- (101) Bokhan, D.; Bernadotte, S.; Ten-No, S. *Chem. Phys. Lett.* **2009**, *469*, 214–218.
- (102) Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **1995**, *103*, 3561–3577.
- (103) Valeev, E. F.; Janssen, C. L. *J. Chem. Phys.* **2004**, *121*, 1214–1225.
- (104) Janssen, C. L.; Nielsen, I. M. B.; Leininger, M. L.; Valeev, E. F.; Kenny, J. P.; Seidl, E. T. *The Massively Parallel Quantum Chemistry Program (MPQC)*, version 3.0; Sandia National Laboratories: Livermore, CA, 2008; <http://www.mpqc.org>.
- (105) Crawford, T. D.; Sherrill, C. D.; Valeev, E. F.; Fermann, J. T.; King, R. A.; Leininger, M. L.; Brown, S. T.; Janssen, C. L.; Seidl, E. T.; Kenny, J. P.; Allen, W. D. *J. Comput. Chem.* **2007**, *28*, 1610–1616.
- (106) Kedzuch, S.; Milko, M.; Noga, J. *Int. J. Quantum Chem.* **2005**, *105*, 929–936.
- (107) Valeev, E. F.; Fermann, J. T. *Libint 1.0*; Virginia Tech Department of Chemistry: Blacksburg, VA, 2007; <http://www.files.chem.vt.edu/chem-dept/valeev/software/libint/libint.html>.
- (108) Obara, S.; Saika, A. *J. Chem. Phys.* **1986**, *84*, 3963–3974.
- (109) Obara, S.; Saika, A. *J. Chem. Phys.* **1988**, *89*, 1540–1559.
- (110) Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3072–3077.
- (111) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1994**, *100*, 2976–2989.
- (112) Peterson, K. A.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 084102.
- (113) Yousaf, K. E.; Peterson, K. A. *J. Chem. Phys.* **2008**, *129*, 184108.
- (114) Stanton, J. F.; Gauss, J.; Watts, J. D.; Szalay, G.; Bartlett, R. J.; Auer, A. A.; Bernholdt, D. B.; Christiansen, O.; Harding, M. E.; Heckert, M.; Heun, O. *ACES II*, Mainz–Austin–Budapest version; The University of Texas at Austin: Austin, TX, 2005; <http://www.aces2.de>.
- (115) J. Almlöf, Taylor, P. *MOLECULE*, an integral package.
- (116) Helgaker, T.; Jensen, H. J.; Jørgensen, P.; Olsen, J. *ABACUS*, an integral package.
- (117) Taylor, P. *PROPS*, an integral package.
- (118) Stanton, J. F.; Gauss, J.; Watts, J. D.; Lauderdale, W. J.; Bartlett, R. J. *Int. J. Quantum Chem. Symp* **1992**, *26*, 879–894.
- (119) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; M. Schütz, Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Palmieri, P.; Nicklass, A.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO*, version 2006.1, a package of ab initio programs; 2006; <http://www.molpro.net>.
- (120) Schlegel, H. B. *J. Chem. Phys.* **1986**, *84*, 4530–4534.
- (121) Gill, P. M. W.; Pople, J. A.; Radom, L.; Nobes, R. H. *J. Chem. Phys.* **1988**, *89*, 7307–7314.
- (122) Nobes, R. H.; Pople, J. A.; Radom, L.; Handy, N. C.; Knowles, P. J. *Chem. Phys. Lett.* **1987**, *138*, 481–485.
- (123) Knowles, P. J.; Handy, N. C. *J. Phys. Chem. A* **1988**, *92*, 3097–3100.
- (124) Wheeler, S. E.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2008**, *128*, 074107.
- (125) Crawford, T. D.; Kraka, E.; Stanton, J. F.; Cremer, D. *J. Chem. Phys.* **2001**, *114*, 10638–10650.
- (126) Allen, W. D.; Horner, D. A.; Dekock, R. L.; Remington, R. B.; Schaefer, H. F. *Chem. Phys.* **1989**, *133*, 11–45.
- (127) Hill, J. G.; Peterson, K. A.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2009**, *131*, 194105.

S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures

Jan Řezáč,^{*,†} Kevin E. Riley,[†] and Pavel Hobza^{†,‡}

[†]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems, 166 10 Prague, Czech Republic

[‡]Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Palacky University, 771 46 Olomouc, Czech Republic

S Supporting Information

ABSTRACT: With numerous new quantum chemistry methods being developed in recent years and the promise of even more new methods to be developed in the near future, it is clearly critical that highly accurate, well-balanced, reference data for many different atomic and molecular properties be available for the parametrization and validation of these methods. One area of research that is of particular importance in many areas of chemistry, biology, and material science is the study of noncovalent interactions. Because these interactions are often strongly influenced by correlation effects, it is necessary to use computationally expensive high-order wave function methods to describe them accurately. Here, we present a large new database of interaction energies calculated using an accurate CCSD(T)/CBS scheme. Data are presented for 66 molecular complexes, at their reference equilibrium geometries and at 8 points systematically exploring their dissociation curves; in total, the database contains 594 points: 66 at equilibrium geometries, and 528 in dissociation curves. The data set is designed to cover the most common types of noncovalent interactions in biomolecules, while keeping a balanced representation of dispersion and electrostatic contributions. The data set is therefore well suited for testing and development of methods applicable to bioorganic systems. In addition to the benchmark CCSD(T) results, we also provide decompositions of the interaction energies by means of DFT-SAPT calculations. The data set was used to test several correlated QM methods, including those parametrized specifically for noncovalent interactions. Among these, the SCS-MI-CCSD method outperforms all other tested methods, with a root-mean-square error of 0.08 kcal/mol for the S66 data set.

INTRODUCTION

In the past decade, we have seen a great acceleration in the development of new quantum chemical methods, resulting in dozens of new computational techniques that potentially improve the accuracy of results and/or computational efficiency.^{1–3} Among these methods, many contain at least one empirical parameter fitted to reference data. As the value of all scientific models must ultimately be determined by comparison with experimental observations, it would be ideal if the empirical models were based on experimental data. Unfortunately, in many cases, the computed quantity cannot be isolated in an experiment, and direct comparison is thus not possible.⁴ In such cases, it is necessary to establish a set of very accurate, well-characterized computational data that can be used to parametrize and validate empirical models. These benchmark data also serve as a valuable tool for the assessment of the accuracy of nonempirical, but more approximate, methods.

One area of intensive development of computational methods in the past decade is in the proper treatment of noncovalent interactions. As these types of interactions occur throughout nature, designing accurate, computationally efficient, quantum chemical techniques that give accurate interaction energies and geometries for intermolecular interactions is critical in the treatment of a vast number of systems relevant in areas of chemistry, biodisciplines, and material science.

As it has become increasingly clear in the past several years that there are many possibilities for developing new computational

methods that give improved accuracy at lower computational costs, it has also become clear that there should exist standard databases of high-quality data against which new methods can be parametrized and validated. Many such databases have recently been developed for several different molecular properties, including heats of formation, ionization potentials, electron affinities, and intermolecular interaction energies. The existence of reliable data for the latter of these properties has historically been very limited because of the large computational expense associated with the calculation of accurate interaction energies for all but the smallest molecular complexes. Only recently has it become possible to compute interaction energies for medium-sized complexes (up to approximately 40 atoms) with accuracy that is sufficient for benchmark data. There are two main reasons that it has now become possible to get these types of data. The first of these is the development of new computer hardware (including computer parallelization) that allows for much more efficient computations on molecules and molecular clusters. The second reason is the development of a hybrid estimated CCSD(T)/CBS method, which requires only computation of the extrapolated MP2/CBS result as well as the CCSD(T) interaction energy with a small/medium basis set.^{5–8} This method is the most computationally inexpensive technique that has been shown to

Received: April 28, 2011

Published: July 01, 2011

give accurate interaction energies for many different types of molecular complexes.

Our main goal in this work is to create a well-balanced database of benchmark interaction energies for noncovalent interactions relevant to biological chemistry. Noncovalent interactions are extremely important in biomolecules, as they play large roles in determining their structure and dynamics, and are also responsible for recognition processes in biological systems. Thus, the development of new computational tools that accurately describe noncovalent interactions within biomolecules in an efficient way is crucial if significant advances are to be made in computational biophysical chemistry. The development of such methods, which generally contain several empirical parameters, critically depends on the availability of high-quality reference data.

There are several interaction energy databases that have been developed in the past 5–10 years, each with distinct strengths and weaknesses.^{9–16} The most important of these databases are the ones that use the estimated CCSD(T)/CBS method, or other similar (CCSD(T)-based) methods for the reference values. Below we will describe some of the most notable of these data sets.

The S22 reference set,⁹ which was developed in this laboratory in 2006, has become the most popular interaction energy database and has been used extensively in the parametrization and validation of many different computational techniques (more in discussion below). At the time of its creation, this database represented the state-of-the-art in terms of the level of accuracy that could be attained for relatively large complexes (the largest complex contains 30 atoms). The interaction energies in this test set were computed using the estimated CCSD(T)/CBS procedure, which had recently been developed.

Despite the fact that S22 is extremely useful and has served as a model in the development of newer databases, there are some problems associated with this data set. Although the interaction energies published in the original S22 paper were very accurate relative to the standard methods of the time, it has been shown that, using more modern computers, it is possible to improve the accuracies of these values by using larger basis sets. Thus, S22 interaction energies were recalculated by Sherrill and co-workers¹⁷ and by Szalewicz and co-workers¹⁸ using larger basis sets consistently for all of the complexes. Another potential problem with the S22 set is that it is heavily weighted toward nucleic acid-like structures, containing many base pair-like (cyclic) hydrogen bonds and many examples of stacked aromatic (especially heterocyclic aromatic) species. There are several interaction motifs that are strongly under-represented, such as single hydrogen bonds and aromatic–aliphatic dispersion interactions, or practically missing, such as aliphatic–aliphatic dispersion interactions.

Several years after the development of S22, dissociation curves of the 22 complexes were calculated by Merz and co-workers¹⁹ and in this laboratory.¹⁰ In our approach, the resulting data set (named S22 × 5) contains five examples of each of the S22 complexes, with relative displacements of 0.9, 1.0 (ie the equilibrium geometry), 1.2, 1.5, and 2.0. In this study, DFT-SAPT analyses were employed, principally to determine the relative contributions of electrostatic and dispersion terms to the total interaction energy of each complex.

Recently, Grimme and co-workers published the GMTKN30 (General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions) data set, which might actually be classified as a superdatabase containing 30 distinct data sets collected from the literature.^{11,12} As indicated by the title, the GMTKN30 set

contains data sets for several different molecular properties, including barrier heights, reaction energies, and properties for noncovalent interactions. Among the 30 data sets, there are 10 that deal explicitly with noncovalent interactions, with five interaction energy databases (containing a total of 89 complexes) and five databases of relative energies for molecules containing intramolecular noncovalent interactions (58 molecules). It should be noted that S22 is one of the five interaction energy data sets contained within the GMTKN30 database.

Zhao and Truhlar have also developed a superdatabase of atomic and molecular properties that is divided into six categories: thermochemistry, barrier heights, electronic spectroscopy, transition metal reaction energies, structural data, and noncovalent interactions.³ The noncovalent interaction category (NCIE53) contains eight separate subsets with a total of 53 complexes.^{20–22} As in the case of the GMTKN30 database, the S22 set is contained within the NCIE53 set. The additional subsets contain hydrogen bonds, charge-transfer complexes, dipole interactions, weak (dispersion dominated) interactions, and π – π stacking complexes. The NCIE53 database can be said to be better balanced than the S22 set, mainly because of its inclusion of single hydrogen-bonding complexes, dipole interactions, and charge transfer complexes. In terms of dispersion-dominated interactions, this data set, like the S22, heavily favors aromatic complexes. This is true because the weak interactions subset contains mainly noble gas dimers, which have extremely low interaction energies (often less than 0.1 kcal/mol); thus, the S22 dispersion complexes dominate this category.

In an ambitious project, Friesner and co-workers constructed an extremely large database of interaction energies containing 2027 complexes.¹⁵ This set was constructed by collecting almost all of the interaction energy data that had been produced at (at least) the estimated CCSD(T)/CBS level at the time (December 2010). Also included within this test set are several potential energy curves for various complexes. This collection is, of course, very valuable, as it represents the largest single repository of interaction energy data. There are, however, several reasons that this database is not well suited to certain applications. Because of the data set's enormous size, it is not practical to routinely use it for the parametrization of new methods; this is especially true when the method is computationally demanding. Another issue is that the database is not very well balanced in terms of inclusion of different interaction motifs. For example, among the 2027 complexes contained in the set, 1892 of them include at least one aromatic molecule (93.3%, 59.8% contain benzene), while there are only 66 examples of aliphatic–aliphatic interactions (3.3%). Finally, the data collected from various sources were calculated using different setups; this may have a non-negligible impact on the quality of the data found in the set. Most importantly, deficiencies in the size of the basis set might lead to inconsistencies of the order of magnitude of the accuracy of the methods parametrized on these data.

Here, we present a database of accurate interaction energies for 66 molecular complexes, which we refer to as the S66 database, computed at the estimated CCSD(T)/CBS level of theory. The complexes contained within the database represent a wide distribution of interaction motifs, including electrostatic dominated (hydrogen bonding), dispersion dominated, as well as mixed (electrostatic/dispersion) interactions. Several variations of each interaction type are also taken into account; for example, both single and double (cyclic) hydrogen bonds are included. Among the dispersion dominated interactions, examples of

Table 1. Monomers Used To Construct the Complexes in the S66 Data Set

molecule	model for
acetic acid	cyclic hydrogen bonds with OH donor, electrostatic interactions
acetamide	cyclic hydrogen bonds with NH donor, electrostatic interactions
benzene	π - π and X - π interactions – aromatic
cyclopentane	aliphatic dispersion – cyclic hydrocarbons
ethene	π - π and X - π interactions – nonaromatic
ethyne	π - π and X - π interactions of triple bond
neopentane	aliphatic dispersion – branched hydrocarbons
<i>n</i> -pentane	aliphatic dispersion – linear hydrocarbons
methylamine	hydrogen bonding – NH group
methanol	hydrogen bonding – OH group
<i>N</i> -methylacetamide	peptide bond model, carbonyl hydrogen bonds
pyridine	π - π and X - π interactions in heterocycles
uracil	π - π and X - π interactions, base pairing
water	hydrogen bonds and other interactions with water

aromatic–aromatic (stacking), aromatic–aliphatic, and aliphatic–aliphatic interactions are incorporated into the set. We include molecules containing only carbon, oxygen, nitrogen, and hydrogen, as these are the most commonly encountered elements in biochemistry. One valuable property of this database is that it is easily expandable and the addition of complexes containing additional elements should be straightforward.

We include not only accurate interaction energies at the potential energy minima, but also a set of 8 points along the dissociation curve, referred to as the $S66 \times 8$ data set. The accurate description of the entire potential energy surface is of great importance for any method that is applied to calculations on nonequilibrium geometries, or that is used for geometry minimizations, vibration analyses, or molecular dynamics simulations. The former is especially important in the case of large systems where a given moiety may interact with a great number of other chemical groups, with the number of interactions quickly increasing as a function of distance.

DESCRIPTION OF THE S66 DATA SET

S66 Data Set Construction. Our goal is to design a new, larger, data set that covers noncovalent interactions in bioorganic molecules in a balanced way. The data set consists of 66 complexes formed by combining 14 monomers in various configurations. The monomers were chosen so that they represent the motifs and functional groups most commonly found in biomolecules (see Table 1). The smaller molecules considered are generally carriers of the functional group of interest (i.e., methanol, methylamine, etc.), while the larger ones are actual biomolecular building blocks (uracil, *N*-methylacetamide as a peptide bond model). A more detailed list of the interactions obtained by combining these is given below; the complexes are depicted in Figure S1 in the Supporting Information. The size of the data set was chosen so that various types of interactions are well represented, yet it is small enough to make more demanding calculations on it practical. Only complexes with interactions stronger than approximately 1.5 kcal/mol were included in the set to minimize the number of systems that contribute negligibly

to the statistical analysis of the errors (when an absolute error measure is used). Also, duplicate entries (for the same interaction (i.e., hydrogen-bond donor/acceptor group combination) were removed from the set, usually keeping the smaller complex.

In this work, we do not consider charged species, for which interaction energies calculated in small gas-phase models are not directly applicable to real system in the condensed phase. Also, such interactions are an order of magnitude stronger than the ones in neutral complexes, which would distort the even distribution of interaction energies desired in the data set. Moreover, most computational methods can describe ionic interactions well. This topic may be addressed by a separate data set.

One of the main goals of this work is to produce a interaction energy data set that is very well balanced. Toward that end, it is our goal to include roughly equal amounts of electrostatic-rich, dispersion-rich, and mixed (electrostatic/dispersion) interactions in the set. The S66 set is divided into three categories: hydrogen bonding (23 complexes), dispersion-dominated (23), and “other” (20). This classification is somewhat arbitrary but is consistent with previous works in the field. In addition, we provide interaction energy decompositions from DFT-SAPT calculations that allow quantifying the ratio between electrostatics and dispersion when more accurate characterization of an interaction is desired. Each interaction is assigned a category, SAPT-electrostatic (23 complexes), SAPT-dispersion (27), or SAPT-mixed (16), based on the relative contributions of electrostatic and dispersion forces. The heuristic categorization and the actual calculations are in good agreement.

In the first group, there are 23 hydrogen-bonded complexes. The single hydrogen bonds cover all possible combinations of donors and acceptors in the water molecule, hydroxyl group, amine group, and carbonyl group, plus some other hydrogen bonds possible within our set of monomers. Our selection of complexes therefore allows detailed examination of how a given method performs for different types of hydrogen bonds. Five cyclic hydrogen bonds, represented by both small models (acetic acid and acetamide) and the uracil dimer, are included to cover hydrogen bonding in nucleic acid base pairs. This category contains only strong $X-H \cdots Y$ ($X = O, N$; $Y = O, N$) hydrogen bonds; there are several electrostatic interactions that can also be classified as hydrogen bonds, such as in ethyne \cdots water, included in the “others” category.

The group of dispersion-dominated complexes (23 systems) is built from two types of monomers with different properties: planar, often aromatic molecules and aliphatic hydrocarbons, which results in three possible interaction classes: π - π stacking (10 systems), aliphatic–aliphatic (5 systems), and π -aliphatic (8 systems) interactions. These interactions are often described differently by approximate computational methods; it is therefore very important to include all of them in the data set. The aliphatic hydrocarbons are represented by three different isomers of pentane to cover linear, branched, and cyclic hydrocarbon chains.

The last group, named “others”, contains 20 complexes that do not fit to the two categories above. Generally, an interaction in this category contains a combination of dispersion and electrostatics. This group includes $X-H \cdots \pi$ ($X = C, O, N$) interactions, T-shaped aromatic ring complexes, nonspecific interactions of polar molecules, and others.

Data Set Properties. In a well-designed data set, the interaction energy should be equally distributed among all of the systems. The histogram of interaction energies in the S66 data

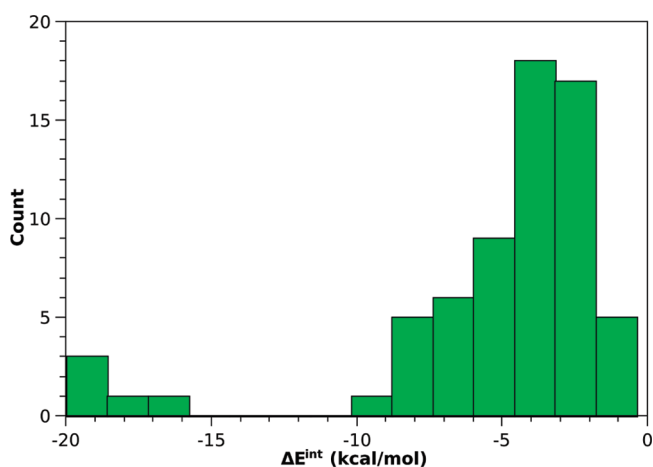


Figure 1. Distribution of interaction energies in the 66 complexes of the S66 data set.

set is shown in Figure 1. The majority of complexes have interaction energies that are around -4 kcal/mol, approximately following a normal distribution. The only outlying points are the double hydrogen bonds with interaction energies between -15 and -20 kcal/mol; these cannot be eliminated because we want to include this type of interaction in the set.

All of the groups of complexes should ideally have equal interaction energy sums if the data set is used to parametrize or test methods that should describe all types of interaction equally. This is hard to achieve in a set of this size when we also would like to include all of the most important interactions. In S66, the sum of interaction energies in hydrogen-bonded complexes (-205 kcal/mol) is more than twice as large as that in dispersion-dominated (-80 kcal/mol) or “other” (-70 kcal/mol) complexes. Although the hydrogen bonds dominate in such a summation, there are other arguments to consider: In the group “other”, dispersion still makes a very important contribution, and it cannot be neglected even in the H-bonded complexes. Overall, the DFT-SAPT decomposition shows that the dispersion to electrostatics ratio for the entire set is 0.86:1, which is not far from being an even representation of both interaction types. Therefore, we consider the S66 to be well balanced for general use. In cases where more control over the separate components is needed, the desired weighting can be applied.

S66 \times 8 Data Set. In addition to equilibrium geometries, we also provide data for eight points along the dissociation curve of each complex. The resulting set, named S66 \times 8, contains 528 points. The displaced complexes are created by scaling the intermolecular distance in the optimized structure; details are given in the Methods. Using eight points for each complex enables an accurate reconstruction of the dissociation curve by interpolation. The sampling of the region around equilibrium was improved to allow accurate determination of the minimum of the dissociation curve. This information is very important for parametrization of new methods where fitting to the extended set should lead to better reproduction of the geometries of non-covalent complexes. The possibility to interpolate the optimal distance accurately allows for assessment of the performance of a method based on comparison of equilibrium intermolecular distances with benchmark data, the minimum at the CCSD(T)/CBS level obtained from the S66 \times 8 set. The interaction energies in this set are all calculated at the CCSD(T)/CBS level.

The S66 and S66 \times 8 data sets do not overlap exactly, but the S66 geometries are always close to one of the S66 \times 8 points. Therefore, they should be used separately, S66 \times 8 to explore the entire dissociation curve and S66 when a more accurate description of the minima is needed.

Comparison to the S22 Data Set. In the past 6 years, the S22 database⁹ developed in our laboratory has been widely adopted as a standard data set used to test and develop methods focused on noncovalent interactions. Other databases of benchmark data covering interaction energies often include the S22 set, as described above. Therefore, a detailed comparison of the S22 and S66 data sets clearly highlights all of the issues S66 attempts to correct, and this comparison also partially applies to other data sets based on S22.

- (1) The S66 set contains 3 times more complexes than S22. This becomes important when one focuses on some particular type of interaction, for example, hydrogen bonds. In such a case, the S22 set does not contain enough complexes of a given type for reliable statistical processing of the data.
- (2) The S22 set is focused mainly on interactions of nucleic acid bases and does not include other types of interactions with comparable weights. Regarding hydrogen bonds, most of the complexes, and an even larger fraction of the total H-bonding energy, feature double hydrogen bonds that are stronger than the single ones. This has been improved by extending the set by four more complexes, forming the S26 set.²³ In dispersion-dominated complexes, the S22 set contains only stacked aromatic molecules, with the only exception being the methane dimer, for which the magnitude of the interaction energy is very small, making its contribution to the entire set negligible. This is, in our opinion, the most important drawback of the S22 set, because many methods parametrized on S22, or performing well on it, fail to describe dispersion interactions between aliphatic hydrocarbon groups (see the discussion of the results below). This has been noted previously but has not been solved systematically.²⁴
- (3) In contrast to S22, the same basis sets are used for benchmark calculations on all of the complexes in the S66 data set, regardless of the size of the system. Although more accurate calculations are possible for the smaller complexes, our approach eliminates possible method-dependent errors.
- (4) The same applies to the geometries; all complexes in the S66 data set were optimized using the same protocol. With the intermolecular distance interpolated from CCSD(T)/CBS calculations, the geometries of the larger complexes should be more accurate than the ones used previously.
- (5) The DFT-SAPT interaction energy decomposition provided for the S66 data set allows unbiased categorization of the nature of the interactions and more detailed analysis of the results, that is, correlating the errors in a tested method with a numerical descriptor of the nature of the interaction. This decomposition is available for the S22 set, but it has been published only very recently.²⁵
- (6) The S66 \times 8 data set provides a better description of the dissociation curve than our extension of the S22 data set, the S22 \times 5 set.¹⁰ Better sampling around the minimum allows accurate interpolation of the potential energy surface.

Planned Extensions. Despite the size of the data set, we do not consider it complete. Now, it covers the most common interactions in biomolecules containing only H, C, N, and O elements. We are working on extension of the data set to other elements and functional groups often present in bioorganic systems. The same methodology is being applied so that the new data will be perfectly compatible with the S66 set. These extensions are another important reason for our choice of a reference method that can routinely be applied to a large number of systems. Another extension we are working on is a better coverage of the potential energy surface of the S66 complexes. In addition to the dissociation curves presented in the S66 data set, we will also sample other intermolecular degrees of freedom.

METHODS

Interaction Energy Calculations. Interaction energies for all of the complexes considered here were computed using fixed monomer geometries, meaning that deformation energies of monomers are not included. The structures of the monomers were taken directly from the dimer optimizations (described below), and no further geometry optimizations were performed on them. The counterpoise correction was employed in all interaction energy calculations to minimize the effects of the basis set superposition error (BSSE).

Basis Sets. Dunning's correlation-consistent series of basis sets²⁶ with diffuse functions²⁷ are used throughout this study. The use of diffuse functions is crucial for the accurate description of noncovalent interactions. In the following text, we use an abbreviated form aXZ (X = D,T,Q) instead of the full names (aug-cc-pVXZ). The cc-pVTZ basis set (abbreviated TZ) is used in specific cases discussed below.

CCSD(T)/CBS. To calculate interaction energies in the S66 and S66 × 8 data sets (66 and 528 entries, respectively), we had to choose a method that balances accuracy and available computational resources. We use the method of estimating the CCSD(T)/CBS limit described in refs 5–8. The Hartree–Fock, MP2, and $\Delta\text{CCSD(T)} = \text{CCSD(T)} - \text{MP2}$ terms are calculated separately in suitable basis sets, and the total energy is composed as follows:

$$E(\text{CCSD(T)/CBS}) = E(\text{HF}) + E^{\text{corr}}(\text{MP2/CBS}) + \Delta\text{CCSD(T)} \quad (1)$$

The HF energy converges with the basis set faster than the correlation energy, and one calculation using a large basis set is adequate.²⁸ Here, we use the aQZ basis set. The MP2 correlation energy is extrapolated to the CBS limit using Helgaker's formula²⁹ from the aTZ and aQZ basis sets. We have tested other extrapolation schemes,^{30,31} but none produced smaller errors when interaction energies were compared to accurate CBS limit estimates in a set of small complexes. Therefore, we conservatively choose the Helgaker scheme, which is robust and free of empirical parameters. The most important part of the CCSD(T)/CBS scheme is the choice of basis set for the CCSD(T) calculation, for which we are much more limited by the steep scaling of the calculation with system size. In contrast to some other works that use customized basis sets, we wanted to use a standard basis so that our protocol can be easily reproduced. In the series of correlation-consistent basis sets, the largest basis sets suitable for these calculations are TZ and aDZ; the latter has a lower maximum quantum number, but includes diffuse functions.

We tested both basis sets against more accurate calculations on a set of small complexes, and aDZ performed noticeably better. It has been previously noted that aDZ is the smallest basis set that can be used for the $\Delta\text{CCSD(T)}$ correction that gives errors of less than 0.1 kcal/mol.^{32,33} For these reasons, we have chosen to use this basis throughout this work.

Geometries. We have followed a multistep protocol to prepare high-quality geometries for complexes of this size.

- (1) Complex preparation: For complexes for which the minimum geometry was not known beforehand, we performed a search along the most important degrees of freedom at the SCC-DFTB-D³⁴ level to identify possible conformations. If multiple minima close in energy were found, we applied the following steps to all of them until we were able to select the one with the lowest energy using a method with accuracy better than the energy difference between them.
- (2) Preliminary optimization of the geometry has been performed using density functional theory with an empirical dispersion correction³⁵ (DFT-D), using the TPSS functional³⁶ and the TZVP basis set³⁷ along with a dispersion correction optimized for this combination of functional and basis set. No symmetry was assumed in any calculation, and the starting structures were perturbed randomly to remove any possible symmetry. This procedure is used to avoid possible optimization to a saddle point instead of a minimum.
- (3) Final optimization of the complexes was carried out at the counterpoise-corrected MP2/TZ level. The resolution of the identity (RI-MP2 method) was used to accelerate the calculations. This setup has been shown to yield geometries close to those obtained at the coupled clusters level.^{24,38} Tight optimization limits (energy change 3×10^{-4} kcal/mol (5×10^{-7} au), max. gradient component 0.06 kcal/mol/Å (5×10^{-5} au), root-mean-square (RMS) gradient 0.03 kcal/mol/Å (2.5×10^{-5} au)) were used to ensure good convergence, even in the intermolecular degrees of freedom. Such an optimization requires well converged energy calculations, so an SCF convergence threshold of 10^{-9} au was used.
- (4) From these geometries, the S66 × 8 set was prepared by scaling the closest intermolecular distance in the complex along an intermolecular axis. The definition of the axis is different for different types of complexes. For hydrogen bonds, it is defined by the hydrogen and the acceptor atom. For cyclic hydrogen bonds, the average of both hydrogen bonds is used. For most other complexes, the centers of mass of the monomers are used, with some exceptions in, for example, T-shaped complexes, where only some atoms have been arbitrarily chosen as the centers to conserve the original arrangement in the displaced geometries. Details on the displacement coordinate are provided in the Supporting Information, Table S3. One of the monomers is moved along the axis so that the minimum distance between them is 0.9, 0.95, 1.05, 1.1, 1.25, 1.5, and 2.0 times the equilibrium value. These seven extensions, along with the MP2 equilibrium geometry, form the S66 × 8 data set.
- (5) Once the CCSD(T)/CBS interaction energies for the S66 × 8 set were calculated, we used the first five points of each dissociation curve (factors of 0.9–1.1 multiplying the equilibrium distance) to obtain a new minimum in the

distance coordinate at the coupled clusters level. The selected points were interpolated with a fourth-order polynomial, and the position of the minimum of this function was used to construct a new geometry. These geometries form the S66 data set.

DFT-SAPT. The DFT-SAPT^{39–42} interaction energy decomposition has been performed with the aDZ basis set to make calculations on larger complexes practical. It is known that the dispersion is slightly underestimated in this setup. We have shown that this can be addressed by scaling of the dispersion component;⁴³ in this work, we apply the scaling factor of 1.193 recommended for this basis set (all DFT-SAPT dispersion energies listed in this Article are already scaled). Such a scaling has been shown to improve the results consistently; the accuracy that can be reached is sufficient for our analysis, which accompanies more accurate CCSD(T)/CBS calculations. The PBE0AC functional recommended by the authors of the method was used for the DFT-SAPT calculations;⁴⁴ the calculations of the monomers have been performed in the basis set of the dimer. The shift needed to correct the asymptotic behavior of the functional was calculated as the difference between the HOMO energy of each monomer and the true ionization potential obtained from the calculation of its neutral and ionized form, using the same functional and basis set. Density fitting was used to speed up these calculations.⁴⁵

The interactions of each of the complexes in the S66 set were characterized as being electrostatics dominated, dispersion dominated, or mixed (electrostatic/dispersion). The complexes with dispersion/electrostatics ratios lower than 0.59 are categorized as electrostatic, those with D/E ratios higher than 1.7 (1/0.59) are categorized as dispersion bound, and complexes with D/E ratios between 0.59 and 1.7 are categorized as mixed. The threshold of 0.59 generates groups that agree well with empirical categorization.

Other Methods Tested in This Study. We have used the newly obtained benchmark data to analyze the performance of several advanced wave function methods that are supposed to closely reproduce CCSD(T)/CBS data. The complete basis set limit for these methods has been calculated analogously as in the CCSD(T)/CBS scheme by combining the MP2/CBS term extrapolated from aTZ and aQZ basis set and the higher order correction calculated in aDZ basis set. In addition to plain MP2, MP3, and CCSD calculations, we have focused on correlated WFT methods that use empirical parameters, MP2.⁵⁴⁶ SCS-MP2,⁴⁷ SCS-MI-MP2,⁴⁸ SCS-CCSD,⁴⁹ SCS-MI-CCSD,⁵⁰ dispersion-weighted MP2⁵¹ (DW-MP2), and nonempirically corrected MP2C.^{52,53} All of the spin-component-scaled results have been derived from the calculations already performed to obtain the benchmark data. MP2 with a time-dependent DFT based dispersion correction (MP2C) represents another approach to improve the performance of the relatively efficient MP2 method.

In the spin component scaled methods, we used the following scaling coefficients for the other-spin (singlet) and same-spin (triplet) terms: The SCS-MP2 method uses $c_{os} = 6/5$ and $c_{ss} = 1/3$ regardless of the basis set used. The SCS-MI-MP2 was optimized to reproduce interaction energies in the S22 set. We use coefficients published for extrapolation from the TZ to QZ basis sets⁴⁸ ($c_{os} = 0.4$, $c_{ss} = 1.29$). We extrapolate from the equivalent, but augmented, basis sets; the difference in the CBS value is negligible. The SCS-CCSD method ($c_{os} = 1.27$, $c_{ss} = 1.13$) was developed using the QZ basis set, while we use extrapolation to the CBS limit; we expect the parameters to be transferable

because the quadruple- ζ basis should not be far from the CBS limit. The SCS-MI-CCSD coefficients ($c_{os} = 1.11$, $c_{ss} = 1.28$) were optimized on the S22 data set, which uses different basis sets for complexes of varying size; here, the transferability of the parameters is an open question.

The DW-MP2 method was originally based on explicitly correlated MP2 (MP2-F12) calculations; this approach improves the convergence of the MP2 energy with basis set size and allows the use of a smaller basis set (aDZ). In this work, we have replaced it with the extrapolated MP2 results as another means to approach the complete basis set limit. Therefore, minor differences in our DW-MP2 results can be expected as compared to ref 51.

Computational Details. All DFT and RI-MP2 optimizations have been carried out in Turbomole 6.2.⁵⁴ Interaction energies at the MP2, MP3, MP2C, CCSD, and CCSD(T) levels and DFT-SAPT interaction energy decompositions have been calculated using the MOLPRO program⁵⁵ in versions 2009 and 2010. A threshold for SCF convergence of at least 10^{-8} au was used for all of the calculations. Density fitting was used for the MP2 calculations used to obtain the MP2/CBS correlation energy term. For the SCC-DFTB-D calculations, the DFTB+ program⁵⁶ was used.

Error Analysis. The performance of the studied methods, with respect to the benchmark calculations, can be described by multiple statistical tools. For the S66 data set, we provide multiple error measures that often carry different information. We consider the root-mean-square error (RMSE) as the most important one, because it reflects the overall quality of the tested method well and is widely used in the field. It is also the variable optimized in parametrization of a method using the least-squares algorithm. Additionally, we list the mean unsigned error, the average (signed) error, which indicates the systematic component of the error, and the largest (maximum unsigned) error, expressed as a percentage, representing the worst case scenario.

We also provide the error separately for different groups of complexes of the data set. Because the average interaction energy in these groups differs, an error in the units of energy does not allow comparison between these groups. Therefore, we use relative errors, calculated as an RMSE divided by the average interaction energy in the group, expressed as percentages.

RESULTS AND DISCUSSION

Geometries. Construction of the $S66 \times 8$ set allowed us to interpolate the CCSD(T)/CBS energies around the MP2 minimum to obtain an accurate estimate of the equilibrium distance at the CCSD(T)/CBS level (see Methods for description of the procedure).

Comparing the energy predicted by the interpolation with the actual calculation on the new geometry can be used to assess the quality of the fit. Even the largest difference in the set is only 0.002 kcal/mol (0.05% of the interaction energy), which indicates that our polynomial fit accurately represents the potential energy surface and that the obtained geometry can be safely considered a minimum in this coordinate.

Comparison of the MP2 and CCSD(T) minima, in terms of distance and CCSD(T)/CBS interaction energies, allows us to measure the quality of the (counterpoise corrected) MP2/TZ geometries. The relative difference in geometries, measured as the change of the closest distance, is on average 0.03 Å. In the worst case, for the stacked benzene dimer, it is 0.11 Å or 3% of the distance.

Table 2. List of the Benchmark CCSD(T)/CBS Interaction Energies (in kcal/mol), the Dispersion/Electrostatics Ratio from the DFT-SAPT Decomposition, and the Interaction Type (E, Electrostatics-Dominated; D, Dispersion-Dominated; and M, Mixed) Based on It for the S66 Data Set

	hydrogen bonds	ΔE	disp/elec	category
1	water...water	-4.92	0.29	E
2	water...MeOH	-5.59	0.35	E
3	water...MeNH ₂	-6.91	0.30	E
4	water...peptide	-8.10	0.37	E
5	MeOH...MeOH	-5.76	0.40	E
6	MeOH...MeNH ₂	-7.55	0.38	E
7	MeOH...peptide	-8.23	0.42	E
8	MeOH...water	-5.01	0.34	E
9	MeNH ₂ ...MeOH	-3.06	0.71	M
10	MeNH ₂ ...MeNH ₂	-4.16	0.71	M
11	MeNH ₂ ...peptide	-5.42	0.79	M
12	MeNH ₂ ...water	-7.27	0.33	E
13	peptide...MeOH	-6.19	0.56	E
14	peptide...MeNH ₂	-7.45	0.50	E
15	peptide...peptide	-8.63	0.56	E
16	peptide...water	-5.12	0.42	E
17	uracil...uracil (BP)	-17.18	0.35	E
18	water...pyridine	-6.86	0.34	E
19	MeOH...pyridine	-7.41	0.40	E
20	AcOH...AcOH	-19.09	0.30	E
21	AcNH ₂ ...AcNH ₂	-16.26	0.32	E
22	AcOH...uracil	-19.49	0.31	E
23	AcNH ₂ ...uracil	-19.19	0.31	E
	dispersion	ΔE	disp/elec	category
24	benzene...benzene ($\pi-\pi$)	-2.82	3.83	D
25	pyridine...pyridine ($\pi-\pi$)	-3.90	2.41	D
26	uracil...uracil ($\pi-\pi$)	-9.83	1.35	M
27	benzene...pyridine ($\pi-\pi$)	-3.44	2.86	D
28	benzene...uracil ($\pi-\pi$)	-5.71	2.16	D
29	pyridine...uracil ($\pi-\pi$)	-6.82	1.75	M
30	benzene...ethene	-1.43	4.57	D
31	uracil...ethene	-3.38	1.86	D
32	uracil...ethyne	-3.74	1.33	M
33	pyridine...ethene	-1.87	3.00	D
34	pentane...pentane	-3.78	4.46	D
35	neopentane...pentane	-2.61	5.42	D
36	neopentane...neopentane	-1.78	4.28	D
37	cyclopentane...neopentane	-2.40	4.33	D
38	cyclopentane...cyclopentane	-3.00	3.98	D
39	benzene...cyclopentane	-3.58	3.10	D
40	benzene...neopentane	-2.90	3.17	D
41	uracil...pentane	-4.85	3.44	D
42	uracil...cyclopentane	-4.14	3.72	D
43	uracil...neopentane	-3.71	2.80	D
44	ethene...pentane	-2.01	4.46	D
45	ethyne...pentane	-1.75	3.11	D
46	peptide...pentane	-4.26	3.32	D
	others	ΔE	disp/elec	category
47	benzene...benzene (TS)	-2.88	2.60	D
48	pyridine...pyridine (TS)	-3.54	1.83	D

Table 2. Continued

	others	ΔE	disp/elec	category
49	benzene...pyridine (TS)	-3.33	2.13	D
50	benzene...ethyne (CH... π)	-2.87	1.60	M
51	ethyne...ethyne (TS)	-1.52	0.79	M
52	benzene...AcOH (OH... π)	-4.71	1.25	M
53	benzene...AcNH ₂ (NH... π)	-4.36	0.98	M
54	benzene...water (OH... π)	-3.28	1.08	M
55	benzene...MeOH (OH... π)	-4.19	1.54	M
56	benzene...MeNH ₂ (NH... π)	-3.23	2.07	D
57	benzene...peptide (NH... π)	-5.28	1.74	M
58	pyridine...pyridine (CH...N)	-4.15	0.88	M
59	ethyne...water (CH...O)	-2.85	0.40	E
60	ethyne...AcOH (OH... π)	-4.87	0.54	E
61	pentane...AcOH	-2.91	3.60	D
62	pentane...AcNH ₂	-3.53	3.09	D
63	benzene...AcOH	-3.80	1.97	D
64	peptide...ethene	-3.00	1.70	M
65	pyridine...ethyne	-3.99	0.46	E
66	MeNH ₂ ...pyridine	-3.97	1.21	M

Of course, the changes in geometry affect the interaction energies. Because the intermolecular potentials are rather flat, such small changes in geometry translate into comparably small improvements in the interaction energies. Over the entire set, we measured an RMSE of 0.020, and a maximum of 0.086 kcal/mol (3% of E_{int}). This is comparable to the estimated accuracy of the method used for the interaction energy calculation; therefore, the MP2 optimization can be safely used in cases when we are interested only in the interaction energy at higher levels. These results confirm previous studies that recommend the TZ basis set for MP2 optimizations.³⁸

The interaction remains attractive even in the geometries with shortest distance (displacement factor 0.9). In all but the $\pi-\pi$ stacked systems, the interaction remains very strong; on average, it amounts to 83% of the interaction energy in equilibrium. In the stacked complexes, the repulsion is steeper due to the large contact area and short equilibrium distance, and the interaction energy becomes as low as 10% of the equilibrium value.

Benchmark Calculations. The final CCSD(T)/CBS interaction energies are provided in Table 2. To estimate the accuracy of our benchmark data, we compare the scheme used in the S66 data set with more accurate estimates of CCSD(T)/CBS. This has been done on a set of ten small complexes introduced in our previous work.⁴³ For these, we extrapolated the CCSD(T) correlation energy to the complete basis set limit from calculations using the aTZ and aQZ basis sets. Table S4 in the Supporting Information gives the results of the comparison of the current benchmark technique with the extrapolated CCSD(T)/CBS(aTZ→aQZ) method. For this small test set, the S66 benchmark method gives an average error value of 1.2% with the largest error being 2.5%. These results indicate that the errors associated with the current benchmark method should be reasonably small; for the S66 test set, we expect that the errors should generally be below 3%.

We are aware of the fact that the errors present in our benchmark data are close to the errors of some of the studied methods to which they are compared. However, in the comparison of similar methods, using the same extrapolation scheme and basis sets, the major part of the error coming from the

Table 3. Errors of the Studied Methods with Respect to the Benchmark CCSD(T)/CBS Calculations on the S66 Data Set^a

method	RMSE, kcal/mol	MUE, kcal/mol	AVG, kcal/mol	MAX %
MP2/TZ	0.70	0.56	0.43	29
MP2/aDZ	0.79	0.58	0.31	32
MP2/CBS	0.69	0.45	-0.44	40
MP2C/CBS	0.71	0.47	-0.01	174
SCS-MP2/CBS	0.87	0.74	0.73	79
SCS-MI-MP2/CBS	0.38	0.28	0.21	54
DW-MP2/CBS	0.40	0.27	0.09	58
MP3/CBS	0.62	0.45	0.44	64
MP2.5/CBS	0.16	0.12	0.00	16
CCSD/CBS	0.70	0.62	0.62	73
SCS-CCSD/CBS	0.25	0.15	0.12	6
SCS-MI-CCSD/CBS	0.08	0.06	-0.04	6

^aThe errors are reported as RMSE, mean unsigned error (MUE), average signed error (AVG), and largest error in the set relative to the interaction energy (MAX).

contributions shared by both methods is canceled. For example, the CCSD(T)/CBS approach is consistent with all of the CCSD-based methods where the MP2 and Δ CCSD terms are the same and the protocol differs only in omitting the triples.

DFT-SAPT. Table 2 gives the DFT-SAPT dispersion/electrostatic ratios and the SAPT-electrostatic/dispersion/mixed (SAPT-E/D/M) category for all of the complexes in the S66 set.

It can be seen in this table that among the interactions in the hydrogen-bonding category, 20 are classified as SAPT-electrostatic, while three are classified as SAPT-mixed. The complexes containing cyclic hydrogen bonds or water exhibit the smallest D/E ratios. The range of D/E ratios for single H-bonds involving NH₂ or NH donors is 0.33–0.79, while the corresponding ratios for single H-bonds involving OH donors are 0.29–0.42. The three SAPT-mixed complexes in this group, MeNH₂···MeOH, MeNH₂···MeNH₂, and MeNH₂···peptide, represent three of the four complexes that have MeNH₂ as the hydrogen-bond donor. Here, the hydrogen bond itself is weaker, while the secondary (dispersion) interactions in the systems are not negligible.

The D/E values for the interactions in the dispersion category are in the range from 1.33 to 5.42. Within this group, there are 20 SAPT-dispersion interactions and three SAPT-mixed interactions, which are for the uracil···ethyne, uracil···uracil, and pyridine···uracil complexes. Generally the interactions that are most electrostatic in nature are π - π interactions involving the two heterocyclic aromatic molecules (uracil and pyridine). This trend stands to reason as the uracil and pyridine heteroatoms give these molecules large charge separations. As would be expected, the aliphatic-aliphatic interactions exhibit the least amount of electrostatic character.

There are three SAPT-electrostatic, 10 SAPT-mixed, and seven SAPT-dispersion complexes within the mixed category. Here, it can be seen that interactions involving ethyne are generally electrostatic in nature, with three of these, ethyne···water, ethyne···AcOH, and pyridine···ethyne, being categorized as SAPT-electrostatic.

Methods Tested. Selected error measures of the tested methods for the S66 test set are listed in Table 3 and plotted in Figure 2. Relative errors for different interaction categories are plotted in Figure 3; the dispersion category is further divided to

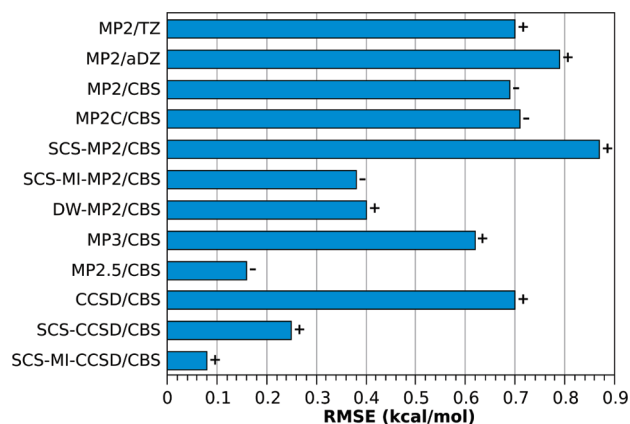


Figure 2. The RMSE (kcal/mol) with respect to the CCSD(T)/CBS benchmark. The symbol next to the bar is the sign of the average error. Plus indicates that the method underestimates the strength of the binding over the whole data set; minus indicates systematic overbinding.

π - π , aliphatic-aliphatic, and π -aliphatic interactions, for which the relative errors are plotted in Figure 4. The full listing of the errors for these groups is provided in Table S1 in the Supporting Information. The performance of the individual methods is discussed in the following text.

1. MP2. The MP2 method has long served as the workhorse in calculations on molecular complexes. This method is generally regarded as yielding qualitatively, or semiquantitatively, accurate results (when used with the counterpoise correction), with the quality of its interaction energy values depending strongly on the basis set with which it is used. In a previous study in this laboratory, we have shown that MP2 generally yields its best results when it is paired with either the aDZ or the TZ basis.⁵⁷ Here, we investigate the performance of this method with the aDZ and TZ bases, as well as at the CBS limit.

It can be seen in Figure 2 that similar errors are produced by the MP2/aDZ (0.79 kcal/mol), MP2/CBS (0.69 kcal/mol), and MP2/TZ (0.70 kcal/mol) methods. These errors are rather high, as compared to those of many of the other methods considered here. However, it should be kept in mind that MP2/aDZ and MP2/TZ are the least computationally expensive methods included in the study and have much better scaling properties than any method that includes any higher order terms (especially when density fitting is used).

Inspection of Figure 3 reveals that, despite the fact that MP2 produces similar overall errors with TZ, aDZ, and at the CBS limit, MP2/TZ gives the most balanced description of the various interaction categories, giving its largest relative error for the dispersion category (14%). MP2/CBS gives extremely accurate results for hydrogen bonding complexes (2%), while producing very large errors for interactions in the dispersion category (29%).

2. SCS-MP2 and SCS-MI-MP2. The SCS-MP2 method makes use of a separate scaling of the singlet and triplet MP2 correlation and was originally developed for reaction energies. The SCS-MI-MP2 method uses the same scaling scheme, but was parametrized for improved performance in the description of noncovalent interactions (using the S22 data set).

Results for SCS-MP2/CBS and SCS-MI-MP2/CBS are depicted in Figure 2. Not surprisingly, SCS-MP2/CBS gives errors that are relatively high (RMSE 0.87 kcal/mol), as the method is not parametrized for intermolecular interactions.

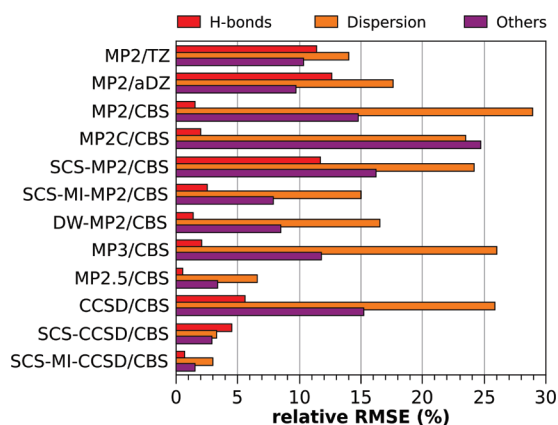


Figure 3. Relative errors (%) for the three groups of complexes: hydrogen bonds, dispersion-dominated, and others. The error is calculated as a RMSE relative to average interaction energy in the group so that the errors can be compared between the groups.

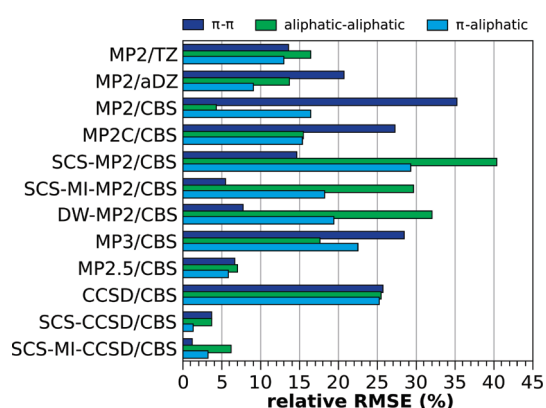


Figure 4. Relative errors (%) for the three types of dispersion-dominated complexes: π - π , aliphatic-aliphatic, and π -aliphatic interactions. The error is calculated as a RMSE relative to average interaction energy in the group so that the errors can be compared between the groups.

SCS-MI-MP2/CBS, on the other hand, yields an RMSE value of 0.38 kcal/mol, which represents a significant improvement over the unscaled MP2 method (for any of the basis sets tested here). Comparing SCS-MI-MP2/CBS to MP2/TZ (the best MP2 performer), it can be seen in Figure 3 that the scaled MP2 method gives improved results for hydrogen bonds and interactions in the “other” category, with especially large improvements for the hydrogen bonds (11% vs 3%). Dispersion interactions are slightly better described by MP2/TZ (14% vs 15%).

3. MP3 and MP2.5. The most computationally expensive part of the MP2.5 method (as utilized here) is the MP3/aDZ calculation, meaning that this technique is only more computationally intensive than the MP2-based methods. MP3 is known to strongly underestimate dispersion interactions, as opposed to MP2/CBS, which is known to overbind dispersion bound complexes. The basis for the MP2.5 method, which is constructed as an average of MP2 and MP3, is a mutual cancellation of these errors.

As might be expected, MP3/CBS yields relatively high RMSEs (0.62 kcal/mol). On the other hand, MP2.5/CBS yields surprisingly accurate results for this data set, considering its

relatively low cost, and the fact that it does not (technically) contain any empirical parameters. Figure 2 reveals that the only method producing RMSEs lower than MP2.5 (0.16 kcal/mol) is SCS-MI-CCSD/CBS (0.08 kcal/mol). Considering the data presented in Figure 3, it can be seen that MP2.5 gives a well-balanced description of the three interaction categories, producing relative errors of no more than 7% for any particular interaction type. The lowest errors occur for hydrogen bonds (1%), while the largest errors occur for the dispersion complexes (7%). MP2.5 gives a very well-balanced description of the dispersion-bound complexes, with errors between 6% and 7% for all three dispersion subcategories.

4. CCSD, SCS-CCSD, and SCS-MI-CCSD. Like SCS-MP2, SCS-CCSD was parametrized to improve CCSD’s description of reaction energies. It has also been noted that this method gives improved results for noncovalent interactions. As in the case of SCS-MP2, SCS-CCSD has also been parametrized (against the S22 set) to give improved results for molecular complexes; with the new parameters the method is designated SCS-MI-CCSD. This reparameterization leads to significant improvement of the accuracy from RMSE of 0.25 kcal/mol in SCS-CCSD/CBS to 0.08 kcal/mol in SCS-MI-CCSD/CBS.

It can be seen in Figure 2 that the RMSEs produced by both SCS-CCSD/CBS and SCS-MI-CCSD/CBS are much lower than that of CCSD/CBS (0.70 kcal/mol). To highlight the accuracy that can be obtained with each of these techniques, it will be noted that the maximum relative error produced by each of these methods is 6%. Figure 3 shows that both SCS-CCSD/CBS and SCS-MI-CCSD/CBS produce very small errors for all interaction categories. The SCS-MI-CCSD technique gives particularly low errors for all interaction categories, producing its largest relative RMSE for the dispersion category (3%). Both SCS-CCSD and SCS-MI-CCSD produce small relative errors for the dispersion subcategories, with SCS-CCSD giving its largest error for the π - π and aliphatic-aliphatic categories (4%) and SCS-MI-CCSD giving its largest error for the aliphatic-aliphatic category (6%).

The SCS-MI-CCSD/CBS is the most accurate method from the studied set. The error with which it reproduces the CCSD(T)/CBS benchmark is smaller than the estimated accuracy of the benchmark calculations. The method is also very robust, as indicated by the narrow range between maximum and minimum error. Therefore, it can be recommended as an alternative to CCSD(T) calculations for larger systems or for any purpose where the ultimate accuracy is not required.

5. Dispersion Weighted MP2. The DW-MP2 method, which utilizes a (system dependent) weighted average of MP2/CBS and SCS-MP2/CBS results to compute interaction energies, yields results that are improved with respect to its parent methods, both in terms of overall errors (RMSE 0.40 kcal/mol) and in terms of errors for the three interaction categories. This is somewhat surprising; given the relatively poor performance of SCS-MP2 for dispersion bound complexes, it would be expected that DW-MP2 should give low errors for hydrogen-bonding complexes (as does MP2/CBS) and higher errors for dispersion bound complexes (as does SCS-MP2/CBS). However, it should be noted that the signs of the errors given by MP2/CBS and SCS-MP2 for dispersion bound complexes are generally opposite in sign; that is, MP2/CBS tends to overbind, while SCS-MP2/CBS tends to underbind. Thus, any interaction energy constructed as a linear combination of results from these two methods will exhibit some inherent error cancellation, which is likely responsible for this method’s relatively low errors.

6. *MP2C*. The *MP2C* method, which incorporates a TD-DFT description of dispersion, yields an RMSE (0.71 kcal/mol) that is comparable to those of *MP2/CBS* (0.69 kcal/mol) and *MP2/TZ* (0.70 kcal/mol). This technique gives very low errors for hydrogen-bonding complexes (2%) but rather high errors for both the dispersion (24%) and the “other” (25%) categories. This is somewhat surprising because, from the theoretical point of view, the *MP2C* offers a well justified improvement over *MP2* itself. On the other hand, it is still only a second-order perturbation treatment, which cannot describe higher order contributions.

DFT-SAPT Decomposition as a Tool for Method Assessment. The availability of DFT-SAPT electrostatic and dispersion data for all of the *S66* complexes allows for the assessment of various methods in terms of the errors they produce for varying dispersion/electrostatic ratios. As an example, we have prepared a plot depicting the (percentagewise) errors as a function of the dispersion/electrostatic ratio for the *MP2/CBS* and *MP2/TZ* methods (Figure 5). The results shown here are in good agreement with the known properties of these two methods. *MP2/CBS* gives a very good description of electrostatically driven interactions while overbinding substantially for dispersion-bound complexes. *MP2/TZ*, on the other hand, underestimates electrostatically driven interactions and generally gives a better overall description of dispersion dominated and mixed interactions (although the errors are still significant), yielding errors that indicate no specific tendency to overbind or underbind. The type of plot shown for *MP2/CBS* and *MP2/TZ* can be used for the assessment and development of new methods and as a tool for better determining their strengths and weaknesses.

Tests on $S66 \times 8$ Set. Figure 6 shows the RMSEs produced by the tested methods for the $S66 \times 8$ data set. The *MP2C* method was not applied to these sets because the improvement as compared to *MP2* itself is expected to be small, while the computational demands are much larger. It will be noted here that errors for the $S66 \times 8$ data set are generally smaller than those for the *S66* set, which is attributable to the fact that interaction energies for structures far from their equilibrium geometries are generally very small, and the corresponding errors for these structures will also tend to be small.

It can be seen in Figures 2 and 6 that, in terms of the relative performance of each of the tested methods, the same general trends are followed for the *S66* and $S66 \times 8$ data sets. Table S2 (Supporting Information) gives the $S66 \times 8$ RMSEs for each of the tested methods along with a description of the errors that occur at long intermolecular separations (displacement factors 1.1–2.0) and short intermolecular separations (factor of 0.9). Here, it can be seen that the relative errors at the shorter intermolecular distances are always about 2–3 times larger than those at the longer separations. This is not a surprising result, as it would be expected that larger errors occur in regions where the potential energy curves are the steepest.

CONCLUSIONS

Here, we have presented a database of interaction energies for 66 intermolecular complexes, each in 9 distinct geometric configurations. The data set was constructed to include a balanced set of commonly encountered interaction motifs involved in biomolecular structures containing C, O, N, and H. This data set was designed to be expandable, meaning that complexes containing additional binding motifs, or additional elements, can easily be added. Reference data were obtained at a

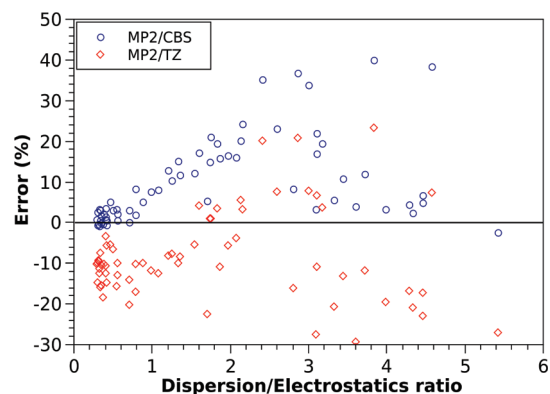


Figure 5. Errors of selected methods plotted against the ratio of dispersion to electrostatic term from the DFT-SAPT decomposition.

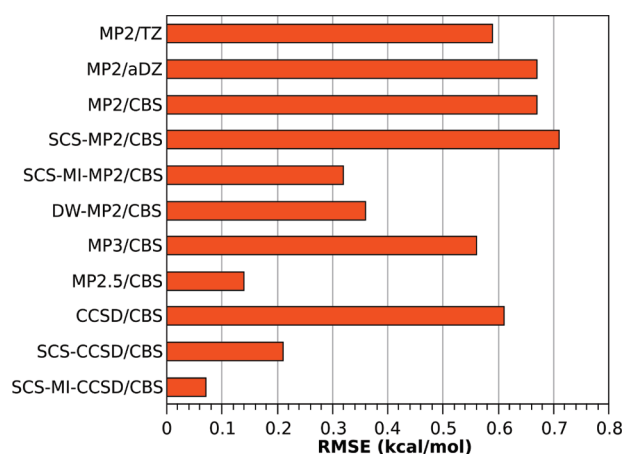


Figure 6. The RMSE (kcal/mol) in the $S66 \times 8$ (dissociation curves of the 66 complexes).

high level of theory using the *CCSD(T)/CBS* scheme consistently for all 594 points. Importantly, the reference method and basis sets used for each of the complexes are identical to avoid the introduction of additional random error.

Data Availability. Geometries of the complexes in *S66* and $S66 \times 8$ data sets, the benchmark *CCSD(T)/CBS* interaction energies, and results of all of the methods tested here are available through the BEGDB Web site⁵⁸ (www.begdb.com) for download and interactive browsing.

Geometries. The geometries of the complexes in the *S66* data set have been carefully optimized, and the intermolecular distance is the minimum at the *CCSD(T)/CBS* level. This is an important advantage over previous data sets where geometries of all but very small complexes had been optimized at only the *MP2* level. Accurate equilibrium geometries are a valuable tool for parametrization of new methods that should yield not only accurate energies, but also equilibrium geometries, and potential energy surfaces in general.

$S66 \times 8$ Geometries. The $S66 \times 8$ data set contains eight points on the dissociation curves of each of the *S66* systems (528 in total). For a given complex, the region around the equilibrium distance is sampled preferentially, which allows for accurate interpolation of the true minimum of the curve.

***S66* Benchmark Results.** The *S66* data set contains 66 *CCSD(T)/CBS* interaction energies obtained using aTZ and

aQZ basis sets for the extrapolation of MP2 correlation energy and aDZ for the Δ CCSD(T) term. These values have been calculated on refined geometries where the intermolecular distance is a true minimum at this computational level.

S66 \times 8 Benchmark Results. For the 528 nonequilibrium geometries contained in these sets, we provide CCSD(T)/CBS interaction energies obtained using the same setup.

Tested Methods. The enormous computational expense associated with the CCSD(T) method, even with a relatively small basis set such as aDZ, makes it necessary to seek less costly methods that can produce comparable results. Of the 12 wave function-based methods tested in this work, there are three that stand out, in terms of their performance on the S66 and S66 \times 8 sets, as producing particularly low errors at a given computational cost point; these are SCS-MI-CCSD/CBS, MP2.5/CBS, and SCS-MI-MP2/CBS. Both of the SCS-MI- (MP2 and CCSD) methods yield errors that are substantially lower than their parent methods, with SCS-MI-CCSD/CBS giving errors that are extremely low for both data sets (RMSE 0.08 kcal/mol in the S66 set). MP2.5/CBS produces errors that are only slightly higher than those of SCS-MI-CCSD/CBS (0.16 kcal/mol), at a much lower computational cost.

■ ASSOCIATED CONTENT

S Supporting Information. A detailed error analysis of the tested methods (Tables S1 and S2), pictures of all of the complexes in the S66 set (Figure S1), definition of the displacement coordinates used to construct the S66 \times 8 set (Table S3), and results of benchmark CCSD(T) calculations on model complexes used to assess the accuracy of the methodology applied to S66 and S66 \times 8 data sets (Table S4). This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Fax: +420 220 410 320. E-mail: rezac@uochb.cas.cz.

■ ACKNOWLEDGMENT

This work was a part of Research Project No. Z40550506 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, and was supported by Grant nos. LC512 and MSM6198959216 from the Ministry of Education, Youth, and Sports of the Czech Republic. It was also supported by the operational program Research and Development for Innovations of European Social Fund (CZ.1.05/2.1.00/03.0058). The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged. We thank anonymous reviewers who read the manuscript with great care and suggested important corrections and improvements.

■ REFERENCES

- (1) Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P. *Chem. Rev.* **2010**, *110*, 5023–5063.
- (2) Riley, K. E.; Hobza, P. *WIREs Comput. Mol. Sci.* **2011**, *1*, 3–17.
- (3) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.
- (4) Šponer, J.; Riley, K. E.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595.

- (5) Koch, H.; Fernández, B.; Christiansen, O. *J. Chem. Phys.* **1998**, *108*, 2784.
- (6) Hobza, P.; Šponer, J. *J. Am. Chem. Soc.* **2002**, *124*, 11802–11808.
- (7) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887–10893.
- (8) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104–112.
- (9) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
- (10) Gráfová, L.; Pitoňák, M.; Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 2365–2376.
- (11) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2010**, *6*, 107–126.
- (12) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2011**, *7*, 291–309.
- (13) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.
- (14) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz, K. M. *J. Chem. Theory Comput.* **2011**, *7*, 790–797.
- (15) Schneebeli, S. T.; Bochevarov, A. D.; Friesner, R. A. *J. Chem. Theory Comput.* **2011**, *7*, 658–668.
- (16) Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrášek, J. *J. Chem. Theory Comput.* **2009**, *5*, 982–992.
- (17) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 144104.
- (18) Podeszwa, R.; Patkowski, K.; Szalewicz, K. *Phys. Chem. Chem. Phys.* **2010**, *12*, 5974.
- (19) Molnar, L.; He, X.; Wang, B.; Merz, K. J. *Chem. Phys.* **2009**, *131*.
- (20) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415–432.
- (21) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656–5667.
- (22) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 289–300.
- (23) Riley, K. E.; Hobza, P. *J. Phys. Chem. A* **2007**, *111*, 8257–8263.
- (24) Riley, K. E.; Pitoňák, M.; Černý, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 66–80.
- (25) Hohenstein, E. G.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *133*, 014101.
- (26) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.
- (27) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1994**, *100*, 2975.
- (28) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Olsen, J. *Chem. Phys. Lett.* **1999**, *302*, 437–446.
- (29) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.
- (30) Martin, J. M. L. *Chem. Phys. Lett.* **1996**, *259*, 669–678.
- (31) Schwenke, D. W. *J. Chem. Phys.* **2005**, *122*, 014107.
- (32) Sherrill, C. D.; Takatani, T.; Hohenstein, E. G. *J. Phys. Chem. A* **2009**, *113*, 10146–10159.
- (33) Pitoňák, M.; Riley, K. E.; Neogrády, P.; Hobza, P. *ChemPhysChem* **2008**, *9*, 1636–1644.
- (34) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149.
- (35) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. *J. Comput. Chem.* **2007**, *28*, 555–569.
- (36) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (37) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.
- (38) Dabkowska, I.; Jurečka, P.; Hobza, P. *J. Chem. Phys.* **2005**, *122*, 204322.
- (39) Williams, H. L.; Chabalowski, C. F. *J. Phys. Chem. A* **2001**, *105*, 646–659.
- (40) Jansen, G.; Hesselmann, A. *J. Phys. Chem. A* **2001**, *105*, 11156–11157.
- (41) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *357*, 464–470.
- (42) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *362*, 319–325.
- (43) Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 685–689.

- (44) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778–784.
- (45) Hesselmann, A.; Jansen, G.; Schütz, M. *J. Chem. Phys.* **2005**, *122*, 014103.
- (46) Pitoňák, M.; Neogrády, P.; Černý, J.; Grimme, S.; Hobza, P. *ChemPhysChem* **2009**, *10*, 282–289.
- (47) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.
- (48) Distasio, R.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 1073–1083.
- (49) Takatani, T.; Hohenstein, E. G.; Sherrill, C. D. *J. Chem. Phys.* **2008**, *128*, 124111.
- (50) Pitoňák, M.; Řezáč, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 9611.
- (51) Marchetti, O.; Werner, H.-J. *J. Phys. Chem. A* **2009**, *113*, 11580–11585.
- (52) Hesselmann, A. *J. Chem. Phys.* **2008**, *128*, 144112.
- (53) Pitoňák, M.; Hesselmann, A. *J. Chem. Theory Comput.* **2010**, *6*, 168–178.
- (54) *TURBOMOLE v6.2*; University of Karlsruhe and Forschungszentrum Karlsruhe GmbH: Karlsruhe, Germany, 2010.
- (55) Werner, H.-J.; Knowles, P. J.; Manby, F. R.; Schütz, M.; Celani, P.; Knizia, G.; Korona, T.; Lindh, R.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hesselmann, A.; Hetzer, G.; Hrenar, T.; Jansen, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO, version 2010.1, a package of ab initio programs*; MOLPRO, 2010.
- (56) Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.
- (57) Riley, K. E.; Hobza, P. *J. Phys. Chem. A* **2007**, *111*, 8257–8263.
- (58) Řezáč, J.; Jurečka, P.; Riley, K. E.; Černý, J.; Valdes, H.; Pluháčková, K.; Berka, K.; Řezáč, T.; Pitoňák, M.; Vondrášek, J.; Hobza, P. *Collect. Czech. Chem. Commun.* **2008**, *73*, 1261–1270.

NOTE ADDED AFTER ASAP PUBLICATION

This article was published ASAP on July 19, 2011. Changes have been made to entries 35, 36, 37, and 38 in Table 2. The correct version was published on July 26, 2011.

Generalized Gradient Approximations of the Noninteracting Kinetic Energy from the Semiclassical Atom Theory: Rationalization of the Accuracy of the Frozen Density Embedding Theory for Nonbonded Interactions

S. Laricchia,[†] E. Fabiano,[‡] L. A. Constantin,[†] and F. Della Sala^{*,†,‡}

[†]Center for Biomolecular Nanotechnologies @UNILE, Istituto Italiano di Tecnologia (IIT), Via Barsanti, 73010 Arnesano (LE), Italy

[‡]National Nanotechnology Laboratory (NNL), Istituto Nanoscienze-CNR, Via per Arnesano 16, 73100 Lecce, Italy

 Supporting Information

ABSTRACT: We present a new class of noninteracting kinetic energy (KE) functionals, derived from the semiclassical-atom theory. These functionals are constructed using the link between exchange and kinetic energies and employ a generalized gradient approximation (GGA) for the enhancement factor, namely, the Perdew–Burke–Ernzerhof (PBE) one. Two of them, named APBEK and revAPBEK, recover in the slowly varying density limit the modified second-order gradient (MGE2) expansion of the KE, which is valid for a neutral atom with a large number of electrons. APBEK contains no empirical parameters, while revAPBEK has one empirical parameter derived from exchange energies, which leads to a higher degree of nonlocality. The other two functionals, APBEKint and revAPBEKint, modify the APBEK and revAPBEK enhancement factors, respectively, to recover the second-order gradient expansion (GE2) of the homogeneous electron gas. We first benchmarked the total KE of atoms/ions and jellium spheres/surfaces: we found that functionals based on the MGE2 are as accurate as the current state-of-the-art KE functionals, containing several empirical parameters. Then, we verified the accuracy of these new functionals in the context of the frozen density embedding (FDE) theory. We benchmarked 20 systems with nonbonded interactions, and we considered embedding errors in the energy and density. We found that all of the PBE-like functionals give accurate and similar embedded densities, but the revAPBEK and revAPBEKint functionals have a significant superior accuracy for the embedded energy, outperforming the current state-of-the-art GGA approaches. While the revAPBEK functional is more accurate than revAPBEKint, APBEKint is better than APBEK. To rationalize this performance, we introduce the reduced-gradient decomposition of the nonadditive kinetic energy, and we discuss how systems with different interactions can be described with the same functional form.

1. INTRODUCTION

Density functional theory (DFT)^{1–3} is one of the most widely used approaches for theoretical calculations in solid-state physics⁴ and quantum-chemistry.⁵ In its original orbital-free (OF) formulation,^{6,7} DFT allows one to describe the ground state of a many-electron system as a function of the electron density (ρ) alone, through the solution of the Euler equation:²

$$\frac{\delta T_s[\rho]}{\delta \rho(\mathbf{r})} + v_{\text{ext}}(\mathbf{r}) + v_J(\mathbf{r}; [\rho]) + \frac{\delta E_{\text{xc}}[\rho]}{\delta \rho(\mathbf{r})} = \mu \quad (1)$$

where $T_s[\rho]$ is the noninteracting kinetic energy (KE) density functional; $v_{\text{ext}}(\mathbf{r})$ and $v_J(\mathbf{r}; [\rho])$ are the external (i.e., nuclear) and Coulomb potentials, respectively; $E_{\text{xc}}[\rho]$ is the exchange–correlation (XC) energy functional; and the Lagrange multiplier μ is the chemical potential, which takes into account that the number of electrons is fixed.

Equation 1 is of limited practical utility, since only $v_J(\mathbf{r}; [\rho])$ is known as an explicit functional of the electron density, while the exact functional forms of $T_s[\rho]$ and $E_{\text{xc}}[\rho]$ are not known. For the XC term, many different successful approximations have been developed for different kinds of systems (molecules,⁸ solids,^{9,10} surfaces,¹⁰ interfaces,^{11,12} etc.), mainly taking advantage of the fact that E_{xc} is by far the smallest term in eq 1. For the noninteracting kinetic energy functional, on the contrary, reliable approximations

are still lacking. For this reason, OF-DFT calculations have been so far only performed for selected solid-state systems,^{7,13–20} and DFT calculations are instead routinely performed within the Kohn–Sham (KS) scheme.¹ This requires, however, the introduction of orthonormalized orbitals and implies a formal $\mathcal{O}(N^3)$ scaling.

In recent years, further interest in the KE functionals and related approximations has been motivated by the development of density-based embedding methods^{21–33} and, in particular, the frozen density embedding (FDE).²⁵ In the FDE approach, a many-electron system with electron density ρ is partitioned into two subsystems A and B, such that the total electron density is $\rho = \rho_A + \rho_B$. The KS equations of subsystem A under the influence of subsystem B (or equivalently those of subsystem B under the influence of A) can be solved exactly within the FDE formalism if the embedding (local) potential

$$v_{\text{emb}}(\mathbf{r}, [\rho_A; \rho_B]) = v_{\text{ext}}^B(\mathbf{r}) + v_J(\mathbf{r}; [\rho_B]) + \frac{\delta E_{\text{xc}}^{\text{nadd}}[\rho_A; \rho_B]}{\delta \rho_A(\mathbf{r})} + \frac{\delta T_s^{\text{nadd}}[\rho_A; \rho_B]}{\delta \rho_A(\mathbf{r})} \quad (2)$$

is added to the standard KS equations. In eq 2, $v_{\text{ext}}^B(\mathbf{r})$ and $v_J(\mathbf{r}; [\rho_B])$ are respectively the external (i.e., nuclear) and the Coulomb potentials due to subsystem B, and the nonadditive XC and kinetic energy

Received: June 7, 2011

Published: July 08, 2011

terms are

$$\begin{aligned} E_{xc}^{\text{add}}[\rho_A, \rho_B] &= E_{xc}[\rho_A + \rho_B] - E_{xc}[\rho_A] - E_{xc}[\rho_B] \\ T_s^{\text{add}}[\rho_A, \rho_B] &= T_s[\rho_A + \rho_B] - T_s[\rho_A] - T_s[\rho_B] \end{aligned} \quad (3)$$

For the nonadditive XC energy (and potential), generalized gradient approximations (GGA) are often successfully employed. For the nonadditive kinetic energy (and potential), accurate approximations expressing T_s as a functional of the electron density are needed. Actually, for molecular systems, GGA approximations of T_s only provide good accuracy for FDE calculations for subsystems with nonbonded interactions,^{34–38} while strongly interacting systems or charge-transfer interactions cannot be properly described.^{37–39} Despite these shortcomings, GGA functionals are the method of choice for, e.g., biological systems: in this context, weak interactions dominate, and FDE can outperform current hybrid quantum mechanics/molecular mechanics (QM/MM) empirical models.⁴⁰

Exact KE functionals are well-known for the homogeneous electron gas, i.e., the Thomas–Fermi (TF) local functional,^{41,42} and for one- and two-electron systems, i.e., the von Weizsäcker⁴³ functional. In between these two extreme cases, different semilocal^{44–53} or nonlocal^{14,17,18,54–58} approximations were derived in recent years. Nonlocal KE functionals are rather accurate, but they are derived mainly from the linear response of the homogeneous electron gas (the Lindhard function³) and are thus only suitable for solid-state simulations. Semilocal KE functionals are instead usually developed on the basis of the generalized gradient expansion or its resummation (for recent reviews, see refs 38 and 59).

The development of GGA KE functionals was also guided by the so-called *conjointness conjecture*, which expresses a hypothetical link between exchange and kinetic energy functionals.^{60–62} Therefore, semilocal KE approximations were constructed with the general form

$$T_s[\rho] = \int d\mathbf{r} \tau_s^{\text{HEG}}(\rho) F_s(s) \quad (4)$$

with $\tau_s^{\text{HEG}} = 3/10(3\pi^2)^{2/3}\rho^{5/3}$ being the kinetic energy density of the homogeneous electron gas (HEG),^{41,42} $s = |\nabla\rho|/\{2(3\pi^2)^{1/3}\rho^{4/3}\}$ being the dimensionless gradient, and F_s being a suitable kinetic enhancement factor resembling a corresponding exchange enhancement factor. We note that, because of the spin scaling relations,⁶³ the exchange and kinetic energies for a spin-polarized system can be evaluated from the spin-unpolarized versions, so in this paper, we only use equations for spin-unpolarized systems. In addition, unless otherwise stated, atomic units are used throughout, i.e., $e^2 = \hbar = m_e = 1$.

Following the approach outlined above, Lembarki and Chermette⁶⁴ constructed a GGA kinetic energy functional (named LC94 or PW91K) using for the enhancement factor the same form as the Perdew–Wang (PW91)⁶⁵ exchange functional:

$$F_s^{\text{LC94}}(s) = \frac{1 + a_1 s \operatorname{arcsinh}(a_2 s) + (a_3 - a_4 e^{a_5 s^2})s^2}{1 + a_1 s \operatorname{arcsinh}(a_2 s) + a_6 s^4} \quad (5)$$

where the parameters $a_1 = 0.093907$, $a_2 = 76.32$, $a_3 = 0.26608$, $a_4 = 0.0809615$, $a_5 = -100$, and $a_6 = 0.57767 \times 10^{-4}$ were refitted to kinetic energies of atoms. For a slowly varying density, LC94 formally recovers the correct second-order gradient expansion (GE2)⁶⁶ and is considered one of the most accurate

GGA functionals for FDE simulations of weakly bounded molecular systems.^{27,34,38,67} Later, Tran and Wesolowski⁶⁸ constructed a PBE-like functional (named TW02) for the kinetic energy, with an enhancement factor of the form

$$F_s^{\text{PBE}}(s) = 1 + \kappa_s - \frac{\kappa_s}{1 + \frac{\mu_s s^2}{\kappa_s}} \quad (6)$$

where $\kappa_s = 0.8438$ and $\mu_s = 0.2319$ were fitted to the kinetic energies of various noble gases. The TW02 functional also shows very good performance for FDE calculations of weakly interacting molecules.³⁸ For a complete list of KE functionals based on the conjointness conjecture, see refs 38, 59, and 69.

Recently,⁷⁰ we proposed two different PBE-like approximations for the KE (APBEK and revAPBEK) based on the conjointness conjecture and the semiclassical atom reference system.⁷¹ In these approximations, the value of $\mu_s = 0.23899$ was not obtained empirically from a fit but fixed by imposing exact constraints on the KE. In APBEK, the value of $\kappa_s = 0.804$ was also fixed nonempirically, by numerical analysis of the KE of heavy atoms, and it coincides with the one used in the PBE exchange (fixed by the Lieb–Oxford bound^{8,70}). In revAPBEK, the value of κ_s was obtained in analogy with the revised PBE exchange⁷² ($\kappa_s = 1.245$), and it was found that this value leads to very accurate FDE energies for eight weakly interacting systems, outperforming LC94.⁷⁰

In this paper, we develop further the work of ref 70 and discuss in detail the performance of the APBEK and revAPBEK functionals for a larger set of systems within the FDE theory. Moreover, we introduce new PBE-like KE approximations (named revAPBEKint and APBEKint) which recover the correct second-order gradient expansion coefficient μ_s^{GE2} in the slowly varying density regime. The APBEKint/revAPBEKint functionals are developed in the spirit of the conjointness conjecture following the approach used in the development of the PBEint exchange functional¹¹ (which solves the analogous problem for the exchange and provides a link between the rapidly and slowly varying density regimes). The APBEKint/revAPBEKint functionals show, in general, performances close to the ones of APBEK/revAPBEK. However, for various tests, differences are significant. A direct comparison of these four PBE-like functionals allows us to rationalize the importance of the enhancement factor for small and large reduced gradients. To this end, we introduce the reduced-gradient decomposition of the nonadditive kinetic energy (see section 3.3).

The paper is organized as follows: In section 2, we explain the APBEK and revAPBEK constructions, and we present the revAPBEKint and APBEKint KE functionals. Section 3 reports computational details and the definition of the reduced-gradient decomposition. In section 4.1, we test the performance of our functionals to compute the total kinetic energy of 51 atoms and ions, jellium spheres, and surfaces. In section 4.2, we apply them to compute the nonadditive kinetic energy and potential in FDE calculations of 20 molecular systems with different kinds of nonbonded interactions: dispersion, dipole–dipole, and hydrogen bonds. Finally, in section 6, conclusions are drawn.

2. THEORY

2.1. The Large-Z Asymptotic Expansion of the Kinetic Energy.

For many-electron nonrelativistic neutral atoms, the noninteracting kinetic energy has the following asymptotic

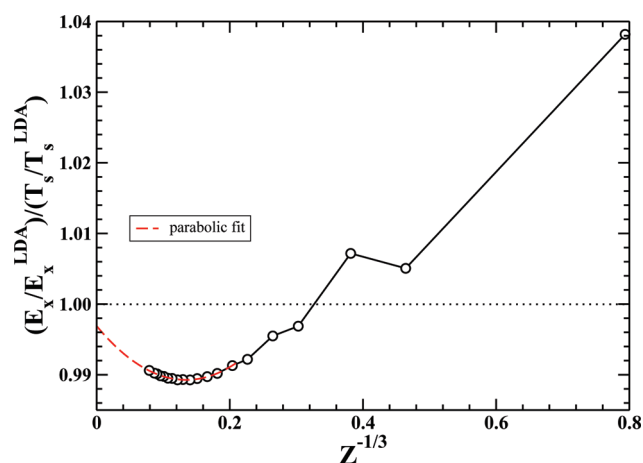


Figure 1. Comparison between exchange and kinetic energies for noble gases.

expansion

$$T_s = c_0 Z^{7/3} + c_1 Z^2 + c_2 Z^{5/3} + \dots \quad (7)$$

where Z is the number of electrons, $c_0 = 0.768745$ is given by the Thomas–Fermi theory,^{41,42} $c_1 = -0.5$ is the Scott correction,⁷³ and $c_2 = 0.2699$ was found in the semiclassical theory.^{74,75} Equation 7 is very accurate, even for a small Z , with a typical error on the order of 0.2–0.5%⁷⁶ for atoms of the periodic table with $Z > 6$. The coefficients of the asymptotic expansion (eq 7) cannot be recovered from GE2 with $\mu_s^{\text{GE2}} = 5/27 = 0.185^{\text{66}}$ ($c_1^{\text{GE2}} = -0.5362$ and $c_2^{\text{GE2}} = 0.3360$ are obtained⁷⁶), but the c_1 coefficient is exactly recovered by the modified second-order gradient expansion (MGE2)⁷⁶

$$T_s^{\text{MGE2}}[\rho] = \int d\mathbf{r} \tau_s^{\text{HEG}}[\rho] (1 + \mu_s^{\text{MGE2}} s^2) \quad (8)$$

with $\mu_s^{\text{MGE2}} = 0.23899$.

For the c_2 coefficient, gradient expansions are less useful,⁷⁶ and $c_2 = 0.3217$ is found from MGE2, slightly better than GE2. Since MGE2 yields exactly the c_1 coefficient of eq 7 and for most systems the important energetic region is $s \leq 2$, we can expect any reasonable functional reducing to MGE2 in the small- s limit to reproduce fairly well eq 7 and thus to be rather accurate for the KE of atoms.

2.2. The Conjointness Conjecture. In ref 70, we discussed the relation between the kinetic and exchange enhancement factors. In order to understand this link better, we report in Figure 1 the ratio $(E_x^{\text{exact}}/E_x^{\text{LDA}})/(T_s^{\text{exact}}/T_s^{\text{LDA}})$ for nonrelativistic noble gas atoms with $Z = 2$ to $Z = 2022$. (Here, LDA is the popular acronym for the local density approximation.¹) We also extrapolate the curve to $Z = \infty$, using a parabolic fit ($a_1 + a_2 Z^{-1/3} + a_3 Z^{-2/3}$); see the red dashed line in Figure 1. In this calculation, we used accurate Kohn–Sham exact-exchange orbitals and densities.

The quantities $E_x^{\text{exact}}/E_x^{\text{LDA}}$ and $T_s^{\text{exact}}/T_s^{\text{LDA}}$ approximate the integrated average values of exact enhancement factors for exchange and kinetic energies, respectively. Thus, when their ratio approaches 1, we have a clear indication that the exchange and kinetic enhancement factors are very similar. Figure 1 thus shows how well the conjointness conjecture works for all of the noble gas atoms; for the He atom ($Z^{-1/3} = 0.79$), the ratio is in fact below 1.04, and for larger Z values, it is very close to unity, because quantum oscillations reduce for heavier atoms. At $Z = \infty$, the extrapolated ratio is about 0.997, whereas the *exact* ratio is 1,

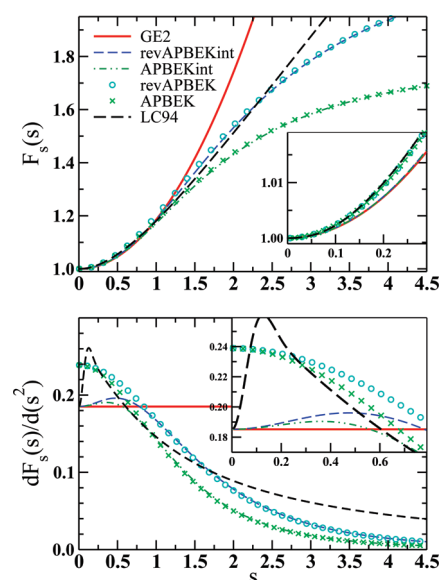


Figure 2. (Top panel) Kinetic energy enhancement factors and (lower panel) their derivative with respect to s^2 . Note that $dF_s(s=0)/ds^2 = \mu_s$.

because LDA becomes exact in this limit for both exchange and kinetic energies.⁷⁷ These results indicate that the kinetic and exchange enhancement factors must be very similar on average, especially at small values of the reduced gradient s that dominate the core of heavy atoms.

2.3. PBE-Like Kinetic Functionals. We recently proposed⁷⁰ the APBEK (asymptotic PBE KE) GGA functional with the enhancement factor of eq 6 and $\mu_s = \mu_s^{\text{MGE2}} = 0.23889$. The value of κ_s^{APBEK} was fixed from a numerical analysis⁷⁰ of atoms with large Z to 0.804, i.e., like the one for the PBE exchange, despite no Lieb–Oxford bound holds for the KE. Thus, in APBEK, both μ_s and κ_s are nonempirical parameters, and it turned out that they are *very close to that of the TW02 functional, which was fitted to noble gas atoms*.

We also introduced the revAPBEK functional,⁷⁰ which is similar to the APBEK but with $\kappa_s^{\text{revAPBEK}} = \kappa_s^{\text{revPBE}} = 1.245$. In analogy with the revPBE exchange GGA functional, the revAPBEK functional takes advantage of the larger nonlocality granted by a higher value of κ_s and yields very accurate FDE energies.⁷⁰

The APBEK and revAPBEK functionals satisfy several exact constraints of the KE but violate others.^{50,55,78–82} In particular, they recover in the slowly varying density limit the MGE2 but not the GE2 expansion: within a simple GGA scheme, it is not possible to satisfy exactly both conditions. This situation resembles that of the exchange PBE functional, which yields μ_x^{PBE} and not μ_x^{GE2} for $s \rightarrow 0$. In this case, one possible solution was recently proposed by introducing the PBEint functional,¹¹ which correctly yields μ_x^{GE2} for small s while reproducing μ_x^{PBE} for slightly larger s (more important for atoms and molecules). Here, we follow a similar path and propose new KE GGA approximations (revAPBEKint and APBEKint) with the enhancement factor given by eq 6 and μ_s defined by

$$\begin{aligned} \mu_s^{\text{int}}(s) &= \mu_s^{\text{GE2}} + (\mu_s^{\text{MGE2}} - \mu_s^{\text{GE2}}) \frac{\tau_s^{\text{W}}}{\tau_s^{\text{HEG}} + \tau_s^{\text{W}}} \\ &= \frac{3\mu_s^{\text{GE2}} + 5s^2\mu_s^{\text{MGE2}}}{3 + 5s^2} \end{aligned} \quad (9)$$

where we used the von Weizsäcker kinetic energy density $\tau_s^W = \tau_s^{\text{HEG}} S_s^2/3$, and $\kappa_s^{\text{APBEKint}} = 0.804$ ($\kappa_s^{\text{revAPBEKint}} = 1.245$). We note that $\tau_s^{\text{HEG}} + \tau_s^W$ is the correct limit for a uniform density perturbed by a small-amplitude, short-wavelength density wave.⁸³

In the upper panel of Figure 2, we report the enhancement factors of different KE functionals versus the reduced gradient s . In the lower panel of Figure 2, the derivatives of the enhancement factors with respect to s^2 versus the reduced gradient s are reported. Note that $dF_s(s=0)/ds^2 = \mu_s$.

For small s ($s < 0.5$), the APBEK and revAPBEK functionals have the behavior dictated by eq 7 and $F_s \rightarrow 1 + \mu_s^{\text{MGE2}} s^2$, thus recovering the MGE2 (in analogy with the APBE exchange functional⁷⁰). On the other hand, by construction, both revAPBEKint and APBEKint KE functionals recover the correct GE2 for a slowly varying density. The LC94 functional recovers the correct GE2, but only at very small values of s ($s < 0.1$), as we can see in the lower panel of Figure 2. In the region $s < 0.5$, the enhancement factor of APBEK, revAPBEK, and LC94 cannot be distinguished; in a similar way, we have revAPBEKint \approx APBEKint \approx GE2. Significant differences are instead present in the lower panel of Figure 2. The LC94 functional displays a pronounced unphysical peak at small values of s , whereas the PBE-like functionals (APBEK and revAPBEK) have a smooth monotonic behavior. Two small peaks around $s = 0.5$ and $s = 0.4$ are observed for revAPBEKint and APBEKint, respectively.

In the range $0.5 < s < 2$, GE2 increases more rapidly than the other functionals. GE2 crosses the APBEK, revAPBEK, and LC94 curves at $s \approx 0.99, 1.23$, and 0.94 , respectively.

For large s ($s > 2$), revAPBEK has a much larger enhancement factor than APBEK, as all the PBE-like functionals must approach the asymptotic value $1 + \kappa_s$. In this region, no differences can be observed between revAPBEK and revAPBEKint and between APBEK and APBEKint.

A totally different behavior is observed for LC94, which increases very rapidly for $s \geq 1$ but then decays asymptotically to zero as $s \rightarrow \infty$ (not shown in the figure). We note that this limit, where the density varies rapidly over a Fermi wavelength, is usually present only in evanescent regions such as the exponential decay of the density far in the tail. Several studies^{2,3} pointed out that for atoms and molecules it is not energetically important, but it can be important to limit numerical errors.⁸⁴

Finally, we calculate the c_1 and c_2 coefficients of eq 7 for different GGA approximations, using the same approach as in ref 76. All functionals that recover μ^{MGE2} give very accurate c_1 coefficients, whereas revAPBEKint, APBEKint, and GE2 yield larger deviations from the exact c_1 , as expected (see Table 1). For the c_2 coefficient, a similar trend is found, but the agreement is not as good since the c_2 coefficient accounts for quantum oscillations that are not well described with gradient expansions. Notably, LC94 shows a good value for both c_1 and c_2 , despite it formally recovering GE2 for $s \rightarrow 0$. This is due to the fact that LC94 recovers GE2 only at very small values of s ($s \leq 0.1$), but in fact it is very close to the MGE2 behavior over a wide range $0.1 \leq s \leq 1.2$.

3. COMPUTATIONAL DETAILS

3.1. Atoms and Ions, Jellium Spheres, and Clusters. Densities and orbitals of atoms and ions (section 4.1) were obtained by analytic Hartree–Fock calculations.⁸⁵ For jellium clusters, we used accurate numerical exact exchange Kohn–Sham

Table 1. Deviations (Multiplied by 10^3) of c_1 and c_2 Coefficients, Computed for Different GGA Approximations, from Exact Ones (eq 7)^a

	GE2	revAKi	AKi	revAK	AK	TW02	LC94
c_1	−36.2	−30.8	31.5	−2.2	−3.8	−8.0	−2.4
c_2	65.8	62.6	53.7	27.0	16.8	21.6	7.2

^aThe smallest values are in bold style. The following shorthands have been used for the functional names: AK for APBEK; revAK for revAPBEK; AKi for APBEKint; revAKi for revAPBEKint.

orbitals and densities,⁸⁶ for jellium surfaces, accurate numerical LDA Kohn–Sham orbitals and densities.⁵¹

3.2. FDE Calculations. The frozen density embedding calculations were performed within the KSCED formalism^{25,32} with freeze-and-thaw cycles⁸⁷ to guarantee the full relaxation of the embedded ground-state electron densities: this approach is equivalent to Cortona’s approach.²³ We used a development version of the TURBOMOLE program package.⁸⁸ Details of our FDE implementation in TURBOMOLE are discussed in ref 32. In the FDE calculations, the PBE⁸ XC functional and def2-TZVPPD supermolecular basis set⁸⁹ were used. The def2-TZVPPD basis set adds diffuse basis functions to the def2-TZVPP⁹⁰ basis set and grants a very accurate description of weakly interacting systems. We used a very accurate integration grid (*gridsize* = 7, *radsize* = 14) to minimize numerical errors.

Several sets of molecular complexes with different kinds of interactions were considered in FDE calculations. The geometries of dispersion, dipole–dipole, and hydrogen bonding interacting systems were taken from the literature.^{27,91,92}

The embedding error in the total energy was computed as

$$\Delta E = E^{\text{FDE}}[\tilde{\rho}_A^e, \tilde{\rho}_B^e] - E^{\text{KS}}[\rho^{\text{KS}}] \quad (10)$$

where $\tilde{\rho}_A^e(\mathbf{r})$ and $\tilde{\rho}_B^e(\mathbf{r})$ are the (approximated) embedded densities, E^{KS} is the total KS energy of total supermolecular system with density ρ^{KS} , and

$$\begin{aligned} E^{\text{FDE}}[\tilde{\rho}_A^e, \tilde{\rho}_B^e] = & T_s^{\text{KS}}[\tilde{\rho}_A^e] + T_s^{\text{KS}}[\tilde{\rho}_B^e] \\ & + \tilde{T}_s^{\text{madd}}[\tilde{\rho}_A^e, \tilde{\rho}_B^e] + V_{\text{ext}}[\tilde{\rho}_A^e + \tilde{\rho}_B^e] \\ & + J[\tilde{\rho}_A^e + \tilde{\rho}_B^e] + E_{\text{xc}}[\tilde{\rho}_A^e + \tilde{\rho}_B^e] \end{aligned} \quad (11)$$

In eq 11, E_{xc} is the GGA exchange–correlation energy functional, J is the Coulomb energy, V_{ext} is the energy associated with the external potential, $T_s^{\text{KS}}[\tilde{\rho}_j^e]$ is the exact KS kinetic energy of the embedded subsystem j (actually computed from KS embedded orbitals), and finally the (approximated) nonadditive kinetic energy functional is

$$\tilde{T}_s^{\text{madd}}[\tilde{\rho}_A^e, \tilde{\rho}_B^e] = \tilde{T}_s[\tilde{\rho}_A^e + \tilde{\rho}_B^e] - \tilde{T}_s[\tilde{\rho}_A^e] - \tilde{T}_s[\tilde{\rho}_B^e] \quad (12)$$

In this work, approximated functionals or densities are indicated with a tilde ($\tilde{}$). Equation 12 has two degrees of approximation: the functional form (\tilde{T}_s) and the embedded densities ($\tilde{\rho}_A^e, \tilde{\rho}_B^e$), which in turn depend on the approximated nonadditive kinetic energy potential.

A quantitative measurement of the absolute error associated with a given embedding density was obtained by computing the error

$$\xi = \frac{1000}{N} \int |(\tilde{\rho}_A^e(\mathbf{r}) + \tilde{\rho}_B^e(\mathbf{r})) - \rho^{\text{KS}}(\mathbf{r})| \, \text{d}\mathbf{r} \quad (13)$$

with N being the number of electrons. In evaluating ξ , we considered only the valence electron density (ξ_v): the core density

Table 2. Percent Mean Absolute Relative Error of Kinetic Energies of 51 Atoms and Ions (See Text), Jellium Surfaces (see ref 51) and Neutral Spherical Jellium Clusters (see refs 51 and 52)^a

	GE2	revAKi	AKi	revAK	AK	TW02	LC94	ref.
atoms/ions	1.10	0.79	0.84	1.24	0.79	0.72	0.82	0.83 ^b
surfaces	3.30	4.23	4.40	3.55	3.93	4.00	3.80	1.17, ^c 2.47, ^d 1.77 ^e
clusters	0.99	1.02	1.14	0.83	0.96	0.98	0.92	1.76 ^c
error	1.63	1.70	1.80	1.71	1.62	1.60	1.59	1.94 ^d

^a Also shown is the total error given by eq 20. Best (worst) results are indicated in boldface (underlined). The following shorthands have been used for the functional names: AK for APBEK; revAK for revAPBEK; AKi for APBEKint; revAKi for revAPBEKint. ^b The A0.185 functional from the Airy gas, see ref 52. ^c GE4 see ref 51. ^d MGGA kinetic functional, see ref 51. ^e The A(1/6) functional from the Airy gas, see ref 52.

is not important for the determination of chemical and physical properties of interaction between the subsystems.

3.3. Reduced Gradient Kinetic Energy Decomposition. To analyze the performance of different GGA kinetic functionals, we follow the idea proposed in ref 93, where the XC functional is decomposed in terms of its reduced gradient contributions, and apply it to kinetic energy functionals. Thus, we define the local (Thomas–Fermi) KE density in the s space

$$t[\rho](s) = \int d\mathbf{r} \tau^{HEG}[\rho](\mathbf{r}) \delta(s - s(\mathbf{r})) \quad (14)$$

so that

$$T_s^{HEG}[\rho] = \int ds t[\rho](s) \quad (15)$$

We call $t[\rho](s)$ the s -decomposed HEG kinetic energy distribution. (Note that for a constant density, $t[\rho](s)$ becomes a Dirac function centered at $s = 0$.) For a GGA KE functional, we have

$$T_s^{GGA}[\rho] = \int d\mathbf{r} \tau_s^{HEG}[\rho](\mathbf{r}) F_s(s[\rho](\mathbf{r})) \quad (16)$$

and thus

$$T_s^{GGA}[\rho] = \int ds t[\rho](s) F_s(s) \quad (17)$$

Equation 17 is very insightful: it states that the total kinetic energy is the scalar product (in the s space) of $t(s)$ and the kinetic enhancement factor $F_s(s)$.

A similar decomposition can also be obtained for the non-additive kinetic energy (see eq 3). Hence,

$$T_s^{nadd}[\rho_A, \rho_B] = \int ds t^{nadd}[\rho_A, \rho_B](s) F_s(s) \quad (18)$$

with

$$t^{nadd}[\rho_A, \rho_B](s) = t[\rho_A + \rho_B](s) - t[\rho_A](s) - t[\rho_B](s) \quad (19)$$

Equation 18 states that the nonadditive kinetic energy is the scalar product (in the s space) of the s -decomposed nonadditive HEG kinetic energy distribution $t^{nadd}(s)$ and the kinetic enhancement factor $F_s(s)$.

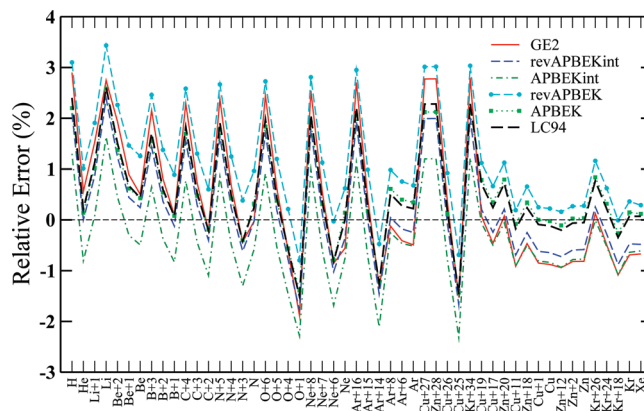


Figure 3. Percent mean absolute relative error for all atoms and ions considered, sorted by the exact kinetic energy.

In this paper, we present several plots of $t(s)$ and $t^{nadd}(s)$. These plots were obtained considering the values of the density and of the reduced gradient on a very accurate DFT quadrature grid.⁹⁴ To obtain a smooth curve, a Gaussian broadening in s ($\sigma = 0.07$) was used.

4. RESULTS

In this section, we present the results of the application of APBEK, revAPBEK, revAPBEKint, and APBEKint functionals to different problems, ranging from the calculation of kinetic energies of atoms/ions and jellium clusters/surfaces to FDE calculations on many different complexes. For comparison, results obtained with the GE2, LC94, and TW02 functionals are also reported.

Molecular atomization kinetic energies have already been reported in ref 70. This test however leads to a very large absolute error for GGA KE functionals^{51,52,95} and small differences among the PBE-like functionals.⁷⁰

4.1. Atoms, Jellium Spheres, and Jellium Surfaces. In Table 2, we show the accuracy of different GGA functionals for the calculation of noninteracting kinetic energies of atoms/ions, jellium clusters, and jellium surfaces (similar to Table I of ref 51 and Table II of ref 52). The total error displayed in the last row of the table is defined as in refs 51 and 52.

Error = 100

$$\times \left(\frac{1}{2} \text{MARE}(\text{atoms}) + \frac{1}{4} \text{MARE}(\text{clusters}) + \frac{1}{4} \text{MARE}(\text{LDM}) \right) \quad (20)$$

where MARE(atoms) is the mean absolute relative error (MARE) of the integrated kinetic energy of 51 atoms and ions (listed in ref 51; in addition, we included Zn^{2+} , see Figure 3), MARE(clusters) considers neutral spherical jellium clusters with bulk parameter $r_s = 3.93$ (listed in ref 51), and MARE(LDM) is related to the jellium surface KE for $r_s = 2, 4,$ and 6 , calculated in the liquid drop model (LDM).⁵¹

Table 2 shows that the total error of eq 20 is practically the same for most of the functional approximations considered, because errors relative to different classes of systems compensate each other.

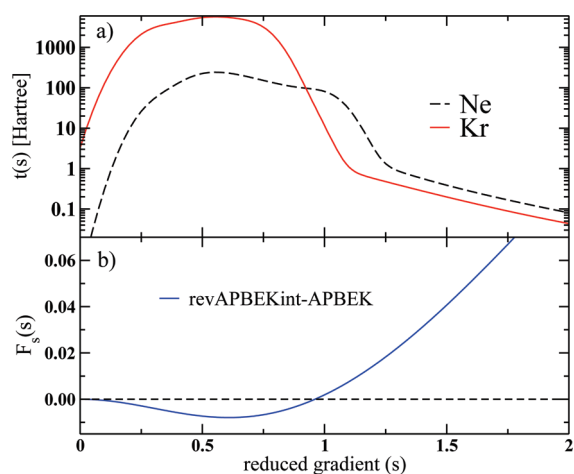


Figure 4. (Panel a) The s -decomposed HEG kinetic energy distribution $t(s)$ for Ne and Kr. (panel b) APBEK-revAPBEKint difference in the enhancement factor $F_s(s)$.

For atoms and ions, TW02 is the most accurate (MARE = 0.72), followed by APBEK, revAPBEKint, LC94, and APBEKint, all with similar errors (MARE = 0.79–0.82). We recall that TW02 and LC94 were fitted to atoms, in contrast to APBEKint and APBEK, which do not contain any empirical parameter. Instead, revAPBEK and GE2 show a lower accuracy. However, it is worth it to note that these functionals still outperform the fourth-order gradient expansion and meta-GGA functionals.⁵¹

To further investigate the errors in the kinetic energy for atoms and ions, in Figure 3, we report the relative error for all of the atoms and ions considered, sorted according to increasing (exact) kinetic energies.

Figure 3 shows that despite the MARE being below 1%, errors up to 3% can be present, in particular for small atoms and heavily charged ions. The relative error instead decreases for systems with a large number of electrons. In this case, the functionals recovering exactly or approximately $\mu_s^{\text{MGE}2}$ (i.e., APBEK, revAPBEK, TW02, LC94) are rather accurate, whereas those recovering $\mu_s^{\text{GE}2}$ (i.e., revAPBEKint, APBEKint) perform similar to the second-order gradient expansion and underestimate the kinetic energy. This can be easily seen by looking at the last two atoms (Kr and Xe) in Figure 3. This emphasizes the importance of the nonempirical parameter $\mu_s^{\text{MGE}2}$ for heavy atoms.

Note also that the total kinetic energy is always in the order revAPBEK > APBEK > revAPBEKint > APBEKint. While it can be easily expected that functionals with a larger κ_s or a large μ_s give larger kinetic energies (due to the larger enhancement factor), it is not obvious why APBEK > revAPBEKint, as the former has a larger μ_s while the latter has a larger κ_s . In addition, Figure 3 shows that the difference between these two functionals increases with the number of electrons: for Ne(Kr), the APBEK–revAPBEKint difference is 0.49% (0.61%).

To shine light on this issue, we report in Figure 4 (panel a) the s -decomposed HEG kinetic energy distribution $t(s)$ (see eq 14) and the difference of the two enhancement factors in panel b.

Figure 4a shows that the energy-relevant region for Ne is $0.2 < s < 1.3$: for larger s , $t(s)$ is vanishing exponentially (note that the energy axis is in a log scale). As the total kinetic energy can be obtained as the scalar product in the s space between $t(s)$ and $F_s(s)$, differences in $F_s(s)$ for $s < 1$ (where the revAPBEKint enhancement factor is lower than APBEK, due to a smaller μ_s parameter)

are more important than differences due to different κ_s values (high s region). Upon increasing the atom size, the $t(s)$ distribution moves to smaller s , making the role of κ_s less important.

Coming back to Table 2, we see that for jellium surfaces, on the contrary, GE2 shows the best performance, although the error is quite large, because in this case a strong nonlocality is required and the fourth order-gradient expansion of the kinetic energy⁹⁶ needs to be recovered. In fact, laplacian-dependent functionals can be expected to perform accurately for these systems.⁵¹ In addition, any GGA functional performing well for jellium surfaces will be rather inaccurate for atoms and molecules (see Figure 3 of ref 48). Because of the need for high nonlocality (i.e., large enhancement factor), better results are obtained with LC94, APBEK, and revAPBEK. Note also that the LC94 GGA has five empirical parameters fitted to the kinetic energies of He, He⁺, Ne, and Ne⁺ atoms.⁶⁴ Such a fitting set takes into account the ionization kinetic energy that contains a small contribution from atomic surface kinetic energy. Thus, for jellium surfaces, LC94 is favored with respect to the PBE-like GGAs.

Finally, for jellium clusters, which are in between surfaces and atoms, revAPBEK is the most accurate functional, while APBEK, TW02, and LC94 yield errors similar to GE2. Also, in this case, functionals with a higher κ_s are favored.

4.2. FDE Calculations. In this subsection, we present the results of FDE calculations on several sets of molecular complexes characterized by different types on nonbonded interactions:

WI weakly interacting systems (i.e., dispersion dominated). C₆H₆–Ne, (CH₄)₂, CH₄–Ne, He–Ne, Ar–Ne, and (Ne)₂, from the benchmark set WI7–05⁹¹

DI dipole interacting systems. CH₃Cl–HCl, CH₃SH–HCl, CH₃SH–NCH, (H₂S)₂, (HCl)₂, and H₂S–HCl, from the benchmark set DI6–04⁹²

HB hydrogen-bonded systems. HF–NCH, a strong hydrogen-bond system deeply investigated in the context of the FDE theory;²⁷ (H₂O)₂, (HCONH₂)₂, (HCOOH)₂, (HF)₂, (NH₃)₂, and NH₃–H₂O from the benchmark set HB6-04⁹²

Systems with a chemical bond and/or significant charge-transfer cannot be treated at the GGA level^{38,39} and thus are not considered in this work. For these systems, corrections to GGA are required, as discussed in refs 82, 97, and 98.

4.2.1. Embedding Error on Energy. In Table 3, we report the embedding error on the total energy (ΔE), see eq 10. The KS binding energy (E_b^{KS}) of the total system is reported in the third column of Table 3. This was corrected for the basis-set superposition error (BSSE) using a counterpoise correction procedure,⁹⁹ and all geometries are kept fixed. Note that the embedding error on the total energy coincides with the embedding error of the binding energy, as the contributions from isolated subsystems cancel and, with a supermolecular basis set, the BSSE corrections also cancel. The results reported in Table 3 for the LC94 and TW02 functionals reproduce those reported in ref 38 and obtained with different basis sets. This shows the correct convergence of our calculations with respect to the basis set. The last column of Table 3 reports also the mean average error (MAE) for the LC94 and the LLP91⁶¹ functionals from ref 38.

In each subgroup, the systems are sorted according to the reference E_b reported in the second column on the table. The E_b computed at the PBE level is in agreement with the reference values, since the PBE XC-functional correctly describes dipole–dipole and hydrogen-bond interactions.⁹² On the other hand, PBE fails to describe correctly

Table 3. Energy Deviations between Supermolecular FDE and Kohn–Sham Calculations ($\Delta E = E^{\text{FDE}} - E^{\text{KS}} = E_b^{\text{FDE}} - E_b^{\text{KS}}$) Corresponding to Different KE GGA Functionals for Different Interactions (Weak, Dipole, Hydrogen-Bonded Systems)^a

system	E_b		$\Delta E = E^{\text{FDE}} - E^{\text{KS}} = E_b^{\text{FDE}} - E_b^{\text{KS}}$							ref
	ref	E_b^{KS}	GE2	revAKi	AKi	revAK	AK	TW02	LC94	
weak interaction										
He–Ne	0.06	0.14	–1.12	0.08	0.12	0.08	0.12	0.12	–0.10	
He–Ar	0.10	0.14	–1.17	0.05	0.10	0.05	0.10	0.09	–0.13	
(Ne) ₂	0.13	0.18	–1.71	0.12	0.22	0.14	0.23	0.22	–0.15	
Ne–Ar	0.21	0.18	–1.90	0.10	0.22	0.11	0.24	0.22	–0.20	
CH ₄ –Ne	0.35	0.22	–2.13	0.10	0.24	0.12	0.26	0.24	–0.23	
C ₆ H ₆ –Ne	0.75	0.14 ^b	–4.71	–0.08	0.40	– 0.03	0.44	0.39	–0.73	
(CH ₄) ₂	0.81	0.02 ^b	–4.49	–0.45	0.07	–0.38	0.13	0.07	–0.98	
MAE			2.46	0.14	0.20	0.13	0.22	0.19	0.36	0.35, ^c 0.14 ^d
MARE (%)			831	46	83	49	88	81	98	88, ^{c,f} 29 ^{d,f}
dipole–dipole interaction										
(H ₂ S) ₂	2.65	2.60	–4.99	–0.64	0.28	–0.48	0.44	0.34	–0.86	
(HCl) ₂	3.20	3.06	–5.16	–0.12	0.97	0.07	1.15	1.03	–0.37	
H ₂ S–HCl	5.34	6.25	–5.62	0.12	1.57	0.40	1.85	1.70	0.11	
CH ₃ Cl–HCl	5.66	5.06	–7.77	–0.28	1.54	0.02	1.85	1.65	–0.45	
CH ₃ SH–HCN	5.72	5.39	–7.83	–1.41	–0.06	–1.18	0.16	0.02	–1.75	
CH ₃ SH–HCl	6.63	8.58	–7.54	0.34	2.47	0.73	2.87	2.66	0.48	
MAE			6.49	0.49	1.15	0.48	1.39	1.23	0.67	0.66, ^c 0.85 ^d
MARE (%)			139	11.1	21.3	9.6	26.2	22.3	15.7	16, ^{c,f} 15 ^{d,f}
hydrogen bond										
(NH ₃) ₂	5.02	4.83	–6.68	–1.15	0.04	–0.95	0.24	0.11	–1.43	
(HF) ₂	7.28	7.25	–6.43	0.54	2.05	0.79	2.29	2.12	0.18	
(H ₂ O) ₂	7.92	7.83	–7.35	–0.47	1.13	– 0.20	1.40	1.23	–0.69	
NH ₃ –H ₂ O	10.21	10.81	–7.91	–0.76	1.08	– 0.44	1.41	1.22	–0.77	
HF–NCH	11.33	12.32	–8.05	0.07	2.18	0.43	2.55	2.33	0.09	
(HCONH ₂) ₂	23.81	23.34	–19.36	–4.94	–0.75	–4.20	0.02	–0.41	–4.52	
(HCOOH) ₂	25.74	28.08	–18.95	–2.74	2.34	–1.87	3.29	2.80	–1.71	
MAE			10.67	1.53	1.37	1.27	1.60	1.46	1.34	1.44, ^{c,e} 1.07 ^{d,e}
MARE (%)			87.1	10.9	11.8	9.3	14.3	12.7	10.6	11, ^{c,e,f} 9 ^{d,e,f}
all systems										
MAE			6.54	0.73	0.89	0.63	1.05	0.95	0.80	
rwMAE			9.05	0.78	1.08	0.70	1.26	1.13	1.17	
rwMARE			8.75	0.72	1.10	0.66	1.29	1.14	1.03	

^a The first two columns report the binding energy (E_b) from reference data⁹¹ and from Kohn–Sham calculations with the PBE XC functional (E_b^{KS}). The mean absolute error (MAE) and the mean absolute relative error (MARE) are also shown for each set of molecules. In the last row, the relative weighted MAE (rwMAE) and the global relative weighted MARE (rwMARE) are reported. Bold style indicates the smallest error in each row. All energies are in mHa (1mHa = 0.62751 kcal/mol). The following shorthands have been used for the functional names: AK for APBEK; revAK for revAPBEK; AKi for APBEKint; revAKi for revAPBEKint. ^b E_b^{KS} differs significantly from the reference value, and it is very small. Thus, to compute the MARE, we use (in the denominator) the reference value of E_b . ^c LC94 functional, see ref 38. ^d LLP91 functional, see ref 38. ^e Without HF-NCH^f Recomputed from data in ref 38.

systems with significant dispersion interaction (see e.g. C₆H₆–Ne and (CH₄)₂), although for other systems the PBE functional can capture the short-range part of the dispersion interaction.

To compare results from different classes of systems, and have a global indicator for the performance of the different KE functionals, we consider the relative weighted MAE (rwMAE) among all of the subgroups, namely

$$rwMAE = \frac{1}{3} \sum_{i=WI,DI,HB} \left(\frac{MAE_i}{\langle MAE_i \rangle} \right) \quad (21)$$

where $\langle MAE_i \rangle$ is the average MAE among all of the different functionals (but GE2) for subgroup i (we used two-digit accuracy, thus $\langle MAE_{WI} \rangle = 0.2$, $\langle MAE_{DI} \rangle = 0.9$, $\langle MAE_{HB} \rangle = 1.4$). In this way, all of the different subgroups have the same mean influence on the rwMAE, and functionals with rwMAE larger (smaller) than 1 have a worse (better) performance than the average.

Alternatively, we also consider the mean absolute relative error (MARE) of ΔE relative to the E_b computed at the supermolecular PBE level. In this way, large absolute differences between systems with different sizes are eliminated, and a better

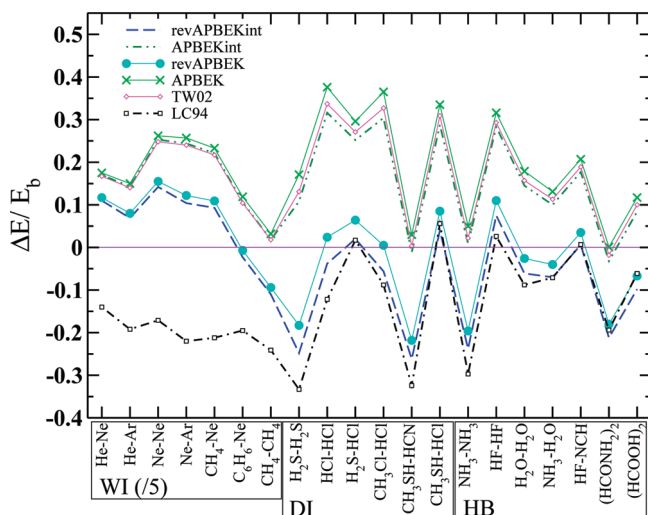


Figure 5. Relative embedding energy error ($\Delta E/E_b$) for all of the systems investigated. $\Delta E = E^{\text{FDE}} - E^{\text{KS}}$. E^{KS} is the total energy from supermolecular Kohn–Sham PBE calculations. E_b is the PBE Kohn–Sham binding energy, but for $\text{C}_6\text{H}_6\text{-Ne}$ and $(\text{CH}_4)_2$ where the reference binding energy⁹¹ was considered. For graphical reasons, the errors of the first seven systems (WI) have been divided by a factor of 5.

comparison is possible. However, for the systems $\text{C}_6\text{H}_6\text{-Ne}$ and $(\text{CH}_4)_2$, the E_b at the PBE level is very small, so that the MARE will almost diverge. For these two systems, we thus considered the error relative to the reference E_b (second column of Table 3). In any case, the accuracy of the MARE for the WI system is not very significant, and for this reason, we report in Table 3 only two digits. The relative error ($\Delta E/E_b$) for all systems is shown in Figure 5, where the error for WI systems was divided by a factor of 5.

Finally, we consider a global relative weighted error (rwMARE) defined as

$$rwMARE = \frac{1}{3} \sum_{i=WI,DI,HB} \left(\frac{MARE_i}{\langle MARE_i \rangle} \right) \quad (22)$$

with $\langle MARE_{WI} \rangle = 74$, $\langle MARE_{DI} \rangle = 18$, and $\langle MARE_{HB} \rangle = 12$.

Considering the MAE for all systems, we have, first of all, to emphasize that these errors are as small (0.6–1.0 mHa) as the errors due to the approximated XC functionals.⁹² This means that the FDE method can be successfully used to describe these systems. The revAPBEK functional yields the best overall results with a MAE of 0.63 mHa, outperforming the LC94 (and TW02) results. More importantly, the rwMAE (rwMARE) of revAPBEK is by 40% (36%) smaller than that of LC94.

On the other hand, the APBEK functional shows the worst overall performance. However, we must recall that the APBEK has *no empirical parameters* and was designed to recover the correct kinetic energy of atoms with an infinite number of electrons, whereas here we are considering small systems in the first two rows of the periodic table. Indeed, the APBEK functional has a global performance similar to that of the TW02 functional, which contains two empirical parameters. TW02 is slightly better due to the somehow larger κ_s value. The comparison between APBEK, TW02, and revAPBEK in fact clearly shows that a large value of κ_s is needed to obtain improved embedding energies.

For DI systems, functionals with a high enhancement factor in the large s region (i.e., revAPBEKint, revAPBEK and LC94) show a much better performance (with a MAE/MARE 2–3 times smaller) than functionals with a small κ_s (i.e., APBEK and TW02). The revAPBEK functional has a MAE/MARE which is 40% better than the LC94 one. The very good performance of revAPBEK can also be inspected in Figure 5: APBEK and TW02 largely overestimate ΔE , and LC94 underestimates it, while revAPBEK is in between them with a MARE of only 9%.

Larger κ_s values also improve ΔE for WI systems, but with a smaller ratio (less than a factor of 2). For these systems, LC94 seems to have a too large enhancement factor, as it yields largely underestimated energies (see Figure 5) with a MAE = 0.36 mHa. The MAE of the revAPBEK is almost 3 times smaller than that of the LC94 functional and revAPBEK shows a similar performance to that of the LLP91 functional, which is the most accurate one for this benchmark.³⁸

For HB systems, the role of κ_s is further reduced, and differences among functionals are smaller. Figure 5 shows that for HB systems all of the functionals give almost exactly the same profile, except for a constant shift. The APBEK (revAPBEK) is the best (the worst) for $(\text{HCONH}_2)_2$ and $(\text{NH}_3)_2$. Considering all HB systems, the revAPBEK functional gives the best MAE and the same MARE as the LLP91 functional.

It is worth it to note that revAPBEK also gives the same MARE (9%) for both DI and HB, meaning that it can provide a balanced description of all kinds of interactions.

Concerning the other important limit, i.e., the one for small s , the revAPBEKint functional (which recovers GE2 at small s) is, considering all systems, also much better than LC94/TW02, but it is slightly worse than revAPBEK (rwMAE/rwMARE 0.08/0.06 higher). This is more evident for the HB systems. On the other hand, APBEKint is better than APBEK (rwMAE/rwMARE 0.16/0.19 smaller). This shows that there is an inter-relation between the large- and small- s regimes and that *both the MGE2 limit for μ_s and the large κ_s value are required to obtain a very good performance*. Finally, we note that ΔE values are always in the order APBEK > APBEKint > revAPBEK > revAPBEKint, see Table 3 and Figure 5. This order is different from that observed for atoms. We will come back to this important issue in section 5.

4.2.2. Embedding Error on Density. To further analyze the ability of different KE functionals to yield accurate nonadditive kinetic energy potentials, we inspected the errors in embedding densities (see eq 13) and reported them in Table 4. The analysis of the deformation densities is frequently employed to assess the approximations made within the FDE scheme.^{32,100–103} In fact, while the analysis of interaction energies can be affected by error-compensation between approximated kinetic energy functionals and approximated embedding densities, the errors on the embedding densities give a direct benchmark of the quality of the kinetic energy potential.

The last two rows of Table 4 report the MAE for all systems and the rwMAE, computed as in eq 21. These data show that *all* functionals yield very accurate densities. GE2 fails for the WI systems (as it excessively favors charge-transfer and covalent bonding), but it is instead quite accurate for HB and DI ones.

Functionals with a high enhancement factor in the large- s region (i.e., revAPBEKint, revAPBEK and LC94) show a slightly worse performance than the ones with small κ_s (i.e., APBEK, APBEKint, and TW02), but the differences in the global MAE are very small (less than 5%). On the other hand, differences between different μ_s values can be hardly distinguished.

Table 4. Global Absolute Errors on Valence Embedding Density ξ_v (See eq 13) Resulting from Supermolecular FDE Calculations with Different KE Functionals on Different Classes of Systems (Weak, Dipole, Hydrogen-Bonded Systems)^a

system	GE2	revAKi	AKi	revAK	AK	TW02	LC94
weak interaction							
He–Ne	0.60	0.04	0.08	0.05	0.09	0.08	0.10
He–Ar	0.79	0.06	0.07	0.06	0.07	0.07	0.17
(Ne) ₂	0.50	0.03	0.09	0.04	0.09	0.09	0.08
Ne–Ar	0.64	0.06	0.10	0.06	0.11	0.10	0.12
CH ₄ –Ne	0.77	0.07	0.10	0.07	0.11	0.10	0.14
C ₆ H ₆ –Ne	0.70	0.15	0.08	0.13	0.09	0.08	0.17
(CH ₄) ₂	1.45	0.65	0.35	0.61	0.31	0.34	0.59
MAE	0.78	0.15	0.12	0.15	0.12	0.12	0.20
dipole–dipole interaction							
(H ₂ S) ₂	2.18	1.88	1.75	1.86	1.76	1.76	1.80
(HCl) ₂	2.00	1.88	1.85	1.89	1.87	1.87	1.92
H ₂ S–HCl	3.50	3.70	3.73	3.74	3.78	3.77	3.75
CH ₃ Cl–HCl	2.53	2.40	2.33	2.40	2.35	2.35	2.40
CH ₃ SH–HCN	2.03	1.75	1.57	1.72	1.54	1.55	1.61
CH ₃ SH–HCl	3.95	4.11	4.11	4.15	4.15	4.15	4.13
MAE	2.70	2.62	2.56	2.63	2.58	2.58	2.60
hydrogen bond							
(NH ₃) ₂	2.20	1.83	1.60	1.79	1.58	1.59	1.69
(HF) ₂	1.76	1.54	1.54	1.55	1.57	1.57	1.64
(H ₂ O) ₂	2.25	2.04	1.95	2.03	1.96	1.97	2.04
NH ₃ –H ₂ O	3.12	3.14	3.07	3.14	3.08	3.08	3.07
HF–HCN	2.66	2.84	2.81	2.84	2.82	2.82	2.79
(HCONH ₂) ₂	2.78	2.79	2.60	2.76	2.57	2.58	2.59
(HCOOH) ₂	3.39	3.47	3.38	3.47	3.38	3.39	3.39
MAE	2.59	2.52	2.42	2.51	2.42	2.43	2.46
all							
MAE	1.99	1.72	1.66	1.72	1.66	1.67	1.71
rwMAE	2.55	1.04	0.95	1.03	0.96	0.95	1.13

^aThe mean absolute error (MAE) for each set of molecules and, in the last row, the rwMAE for these weakly-bonded systems are also reported. Bold style indicates the smallest error in each row. The following shorthands have been used for the functional names: AK for APBEK; revAK for revAPBEK; AKi for APBEKint; revAKi for revAPBEKint.

These small modifications in the description of the density indicate that all functionals yield a reasonably good kinetic energy potential, and we can suppose that the corresponding density differences will therefore play a minor role in the determination of errors on embedding energies (see section 5).

In addition, we found that the embedding errors in the density have similar spatial distribution. This can be inspected from Figures S1–S4 of the Supporting Information, where we reported the plane averaged absolute error in the embedded valence density, i.e. $\iint dx dy |\rho_{\text{val}}^{\text{FDE}}(\mathbf{r}) - \rho_{\text{val}}^{\text{KS}}(\mathbf{r})|$, where z is the interaction axis, for different systems.

5. DISCUSSION

For the exact T_s^{nadd} , the embedded density will be exact (i.e., the sum $\rho_A^e(\mathbf{r}) + \rho_B^e(\mathbf{r})$ will yield the total KS density of the total

system $\rho^{\text{KS}}(\mathbf{r})$ and $\Delta E = 0$. For an approximate $\tilde{T}_s^{\text{nadd}}$, we have instead

$$\Delta E[\tilde{\rho}_A^e, \tilde{\rho}_B^e] = \tilde{T}_s^{\text{nadd}}[\tilde{\rho}_A^e, \tilde{\rho}_B^e] + \Delta W[\tilde{\rho}_A^e, \tilde{\rho}_B^e] \quad (23)$$

where

$$\begin{aligned} \Delta W[\tilde{\rho}_A^e, \tilde{\rho}_B^e] &= T_s^{\text{KS}}[\tilde{\rho}_A^e] + T_s^{\text{KS}}[\tilde{\rho}_B^e] - T_s^{\text{KS}}[\rho^{\text{KS}}] \\ &+ V_{\text{ext}}[\tilde{\rho}_A^e + \tilde{\rho}_B^e] - V_{\text{ext}}[\rho^{\text{KS}}] \\ &+ J[\tilde{\rho}_A^e + \tilde{\rho}_B^e] - J[\rho^{\text{KS}}] \\ &+ E_{\text{xc}}[\tilde{\rho}_A^e + \tilde{\rho}_B^e] - E_{\text{xc}}[\rho^{\text{KS}}] \end{aligned} \quad (24)$$

The subdivision in eq 23 is interesting because ΔW depends *only* on the approximation for the nonadditive kinetic potential, as it depends only on the embedded densities. In other words, ΔW is a *known* bifunctional of the embedded densities.

If we consider a set of KE functionals which yield very similar embedding densities (as is the case for the PBE-like ones considered in this work), then for any given enhancement factor $F_s(s)$ in this set, ΔW can be approximately considered independent of $F_s(s)$ (i.e., it is a constant) and

$$\begin{aligned} \Delta E[F_s] &\approx \tilde{T}_s^{\text{nadd}}[F_s] + \Delta W \\ &\approx \int ds t^{\text{nadd}}(s) F_s(s) + \Delta W \end{aligned} \quad (25)$$

where we used eq 18. Equation 25 means that differences in ΔE among different KE functionals are *directly related to differences in the enhancement factor $F_s(s)$* for the kinetic energy, if embedding densities are not changing.

In order to verify the validity of eq 25 for the systems and the functionals under investigation in this work, we consider the relation between the following two quantities:

$$\begin{aligned} e[F_s] &= \frac{\Delta E[F_s] - \Delta E[F_s^{\text{APBEK}}]}{\Delta T} \\ t[F_s] &= \frac{\tilde{T}_s^{\text{nadd}}[F_s] - \tilde{T}_s^{\text{nadd}}[F_s^{\text{APBEK}}]}{\Delta T} \end{aligned} \quad (26)$$

where

$$\Delta T = \max_{F_s} \{\tilde{T}_s^{\text{nadd}}\} - \min_{F_s} \{\tilde{T}_s^{\text{nadd}}\} \quad (27)$$

is used to normalize the range of variation of the nonadditive kinetic energies. With the above definitions, $e[F_s^{\text{APBEK}}] = t[F_s^{\text{APBEK}}] = 0$ (to set a reference value), $e[F_s] < e[F_s^{\text{APBEK}}]$ (as we found that APBEK gives the highest ΔE , see Figure 5), and $t[F_s] > -1$. If eq 25 holds exactly, then a plot of $e[F_s]$ against $t[F_s]$ gives a straight line with a slope of -1 . Tables with all values of $\tilde{T}_s^{\text{nadd}}$ and ΔW are reported in the Supporting Information.

In Figure 6, we consider four functionals in the order APBEK, APBEKint, revAPBEK, and revAPBEKint and all of the systems considered in this work. Figure 6 clearly shows that, (i) for all of the WI and HB systems, eq 25 holds almost exactly and, (ii) for DI systems, a different behavior is found: in particular (rev)APBEKint has a larger $\tilde{T}_s^{\text{nadd}}$ than (rev)APBEK.

In order to shine light on these two findings, we report in Figure 7 the s -decomposed nonadditive HEG kinetic energy distribution $t^{\text{nadd}}(s)$ (considering LC94 embedded densities) for four representative systems.

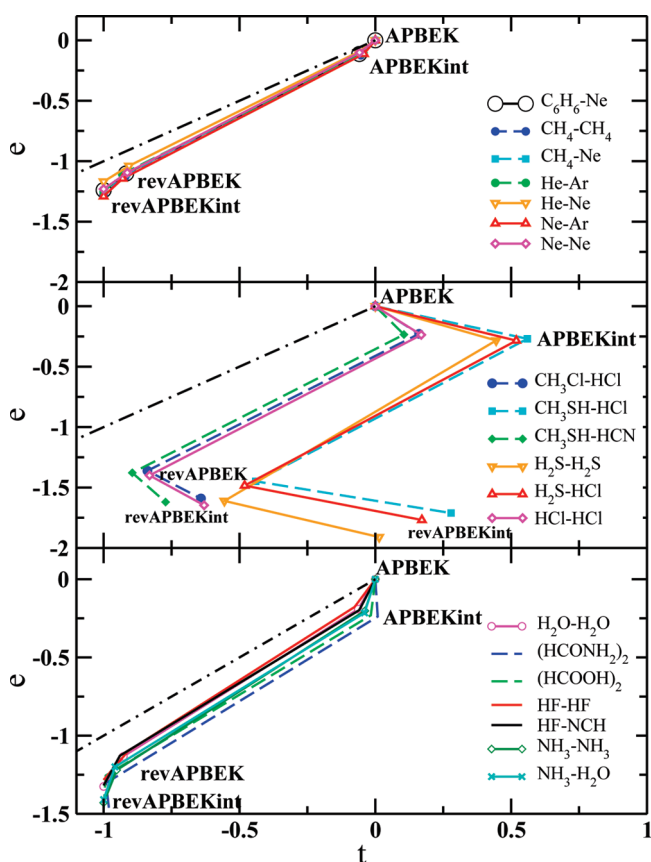


Figure 6. Relation between $e[F_s]$ and $t[F_s]$ (see text for definitions) for weakly interacting systems (panel a), dipole-interacting systems (panel b), and hydrogen-bonded systems (panel c). The values at $e = t = 0$ represent the (reference) APBEK results. Lines starting from it show the values for the APBEKint, revAPBEK, and revAPBEKint functionals (in this order). In each panel, the black dot-dashed line with slope -1 represents the ideal behavior according to eq 25.

Figure 7 shows that upon formation of a bond, there is a decrease of kinetic-energy density at large s (i.e., in the tail of the isolated molecules) and an increase at small s (i.e., in the bond region). The sign change of $t^{\text{nadd}}(s)$ occurs at about $s = 1.2\text{--}1.3$ for both DI and HB systems, and only the region with $s < 3$ is energetically important. For the WI system, the sign change happens at a larger s , due to the large intermolecular equilibrium distances. The alternating behavior of $t^{\text{nadd}}(s)$ for the systems illustrated in Figure 7 directly explains the differences in ΔE and $\tilde{T}_s^{\text{nadd}}$ among different functionals (see eq 18). In fact:

- KE functionals with a large enhancement factor at large s (revAPBEK and revAPBEKint) have lower ΔE and $\tilde{T}_s^{\text{nadd}}$ than the corresponding ones with smaller k_s (APBEK and APBEKint), because $t^{\text{nadd}}(s)$ is negative at large s .
- In a similar way, KE functionals with a higher enhancement factor at small s (APBEK and revAPBEK) have higher ΔE and $\tilde{T}_s^{\text{nadd}}$ than the corresponding ones with smaller μ_s (APBEKint and revAPBEKint), because $t^{\text{nadd}}(s)$ is positive at small s .

These findings rationalize the performance of different functionals. We discuss hereafter in more detail one representative example for the case where eq 25 holds, the water dimer. For this system, revAPBEK has a ΔE of only -0.2 mHa, whereas APBEK is much worse ($\Delta E = 1.4$ mHa). The

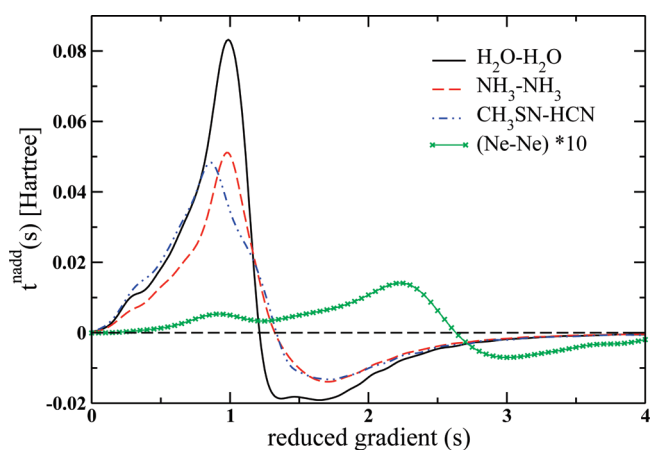


Figure 7. s -Decomposed nonadditive HEG kinetic energy density distribution t^{nadd} for $\text{H}_2\text{O}-\text{H}_2\text{O}$, NH_3-NH_3 , $\text{CH}_3\text{SH}-\text{HCN}$, and $(\text{Ne})_2$ (the latter is multiplied by a factor of 10 for graphical reasons).

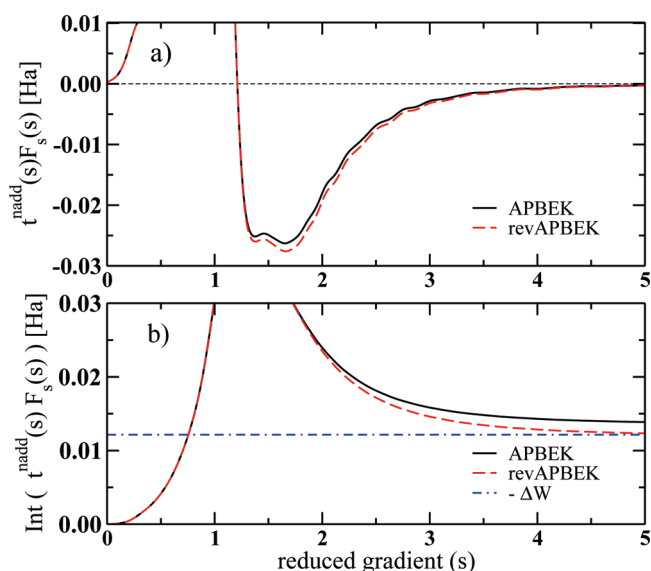


Figure 8. Contributions to the nonadditive kinetic energy for $\text{H}_2\text{O}-\text{H}_2\text{O}$, using the APBEK and the revAPBEK functionals, fixing the embedded density to the APBEK ones. (Panel a) The s -decomposed nonadditive HEG KE distribution t^{nadd} multiplied by the enhancement factor ($F_s(s)$) as a function of the reduced gradient (s). (Panel b) Cumulative integral of the data in panel a. Also shown, the $-\Delta W$ value (from APBEK).

corresponding values of ΔW are -12.44 and -12.15 mHa, respectively (see Supporting Information): thus differences in the embedded densities are negligible with respect to differences in $\tilde{T}_s^{\text{nadd}}$. We thus fix the density to the APBEK one, and in Figure 8, we plot $t^{\text{nadd}}(s) F_s(s)$ (panel a) and $\int_0^s t^{\text{nadd}}(s') F_s(s') ds'$ (panel b).

The plots clearly show that differences are significant only for $s > 1.5$. For this reason, if we define the correct value of the approximated $\tilde{T}_s^{\text{nadd}}$ as the one that gives $\Delta E = 0$, i.e., $\tilde{T}_s^{\text{nadd}} = -\Delta W$, we note that the larger $F_s(s)$ at high s of the revAPBEK reduces the $\tilde{T}_s^{\text{nadd}}$ energy toward the correct value. Clearly, this condition holds due to an error cancellation. In fact, for a given system, different approximate KE functionals (i.e., different enhancement factors) might exist which satisfy $\tilde{T}_s^{\text{nadd}} = -\Delta W$, although both

$\tilde{T}_s^{\text{nadd}}$ and ΔW can be significantly different from the ones obtained with the exact KE functional. Actually, accurate KE functionals should reproduce the *exact* T_s^{nadd} and ΔW *separately*, and not only their sum. However, these *exact* values cannot be easily obtained because they will require in practice the solution of an inverse-KS problem,^{104,105} in order to obtain the exact nonadditive kinetic potential and the exact embedded density.^{33,103,106–108} Nevertheless, within a simple (and efficient) GGA approach, we can consider a KE functional to be accurate for FDE applications if it can obtain a good performance over a very large set of systems, i.e., achieving similar error cancellation for different systems. Indeed, as shown in Table 3, the larger κ_s value of the revAPBEK functional improves ΔE for the largest number of systems and yields the best averaged performance (rwMARE = 0.66).

Clearly, a different empirically optimized value κ_s may also be found that yields even a better performance for the FDE energies for a given set of systems. To demonstrate our latest statement, we note that (see Figure 5) for HB systems all of the functionals give the same profile, with only a constant shift. This means that none of the considered PBE functional forms can achieve in this case a vanishing error for all of the HB systems, but further tuning of the κ_s value can then be used to reduce the MAE. In fact, we recomputed all seven HB systems for $\kappa_s = 1$ (and $\mu_s = \mu_s^{\text{MGE2}}$), and we found a MAE of 1.050 mHa, i.e., almost the same as that of the LLP91 functional, which yields the lowest MAE among different KE functionals considered in ref 38. We note however that with $\kappa_s = 1$ the description of $(\text{NH}_3)_2$, $(\text{HCOOH})_2$, and $(\text{HCONH}_2)_2$ is improved (with respect to revAPBEK), but for other systems, the energy is overestimated (see Supporting Information). Furthermore, we found that the MARE is not reduced (still 9%) with respect to revAPBEK. It is worth it to recall that in the revAPBEK functional the κ_s value was chosen without taking into account weakly bounded systems but just following the conjointness relation with the improved exchange energy of the revPBE functional.⁷² In this sense, the revAPBEK functional can also be considered *nonempirical*.

On the other hand, Table 3 and Figure 5 show that a modification of the enhancement factor for small s (i.e., a reduction of the μ_s parameter) does not lead to large differences in the embedding energies, as $t^{\text{nadd}}(s)$ is very small in this region. However, revAPBEKint (rwMARE = 0.72) is even less accurate than revAPBEK (rwMARE = 0.66), mainly due to the worse description of HB systems. Thus, functionals which satisfy the MGE2 limit provide *both* accurate embedding energies and accurate total KE, where, see Figure 4, the important region is $s < 1.3$. The improvement of APBEKint (rwMARE = 1.10) with respect to APBEK (rwMARE = 1.29) for embedding energies can be instead traced back to an error-compensation for the too small κ_s . In fact, $\tilde{T}_s^{\text{nadd}}$ can be reduced by decreasing μ_s or increasing κ_s . However, APBEKint is the worst functional for the benchmark in Table 2.

Coming back to Figure 6, we now discuss the DI systems, where eq 25 does not hold. First of all, we note that the APBEKint and revAPBEKint functionals lead to larger T_s^{nadd} values but lower ΔE values than APBEK and revAPBEK, respectively. This traces back to a significant reduction of ΔW from (rev)APBEK to (rev)APBEKint KE functionals (see Supporting Information), originating from relatively large variations of the embedded density. To better understand this point, we plot in Figure 9 the difference of t^{nadd} calculated at the APBEKint and APBEK levels (both different enhancement factors and densities were used here).

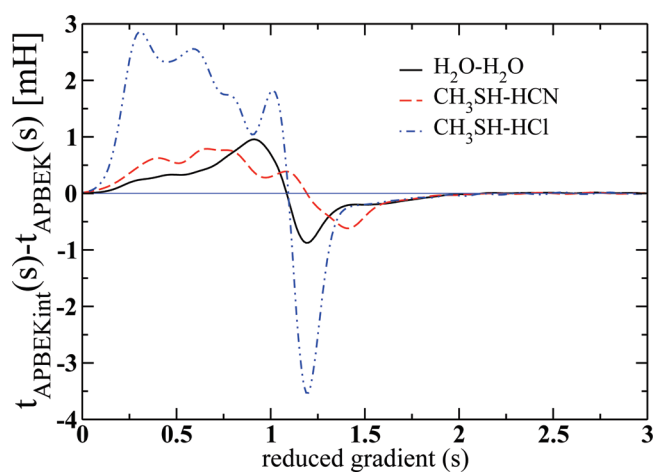


Figure 9. Difference in the s -decomposed nonadditive HEG kinetic energy density distribution, between the APBEKint and APBEK functionals, for $\text{H}_2\text{O}-\text{H}_2\text{O}$, $\text{CH}_3\text{SH}-\text{HCN}$, and $\text{CH}_3\text{SH}-\text{HCl}$.

The figure clearly shows that for $\text{CH}_3\text{SH}-\text{HCl}$ the kinetic energy density at small s values is much larger for APBEKint than for APBEK, which accounts for the large T_s^{nadd} variation (see Figure 6b). A similar effect is also present for $\text{CH}_3\text{SH}-\text{HCN}$, with lower intensity, while for the H_2O dimer only a small oscillation, with opposite peaks of equal intensity, is found, so that upon integration the difference between APBEKint and APBEK is averaged out (T_s^{nadd} changes by 0.1 mHa from APBEKint and APBEK, see Supporting Information). The higher kinetic energy density at small s observed for APBEKint indicates that a higher electron density is present in the bonding region when this functional is used. The reason for this is probably relayed in the behavior of the APBEKint enhancement factor at small s (recovering $1 + \mu^{\text{GE2}}s^2$), which favors delocalization of the density in the bond. This effect is larger for the DI systems because they contain highly polarizable atoms of the second row (S, Cl). We note finally that an increased density in the bond region is a small effect and does not necessarily correlate with improvements in the embedded density (see Table 4).

6. SUMMARY AND CONCLUSIONS

In this paper, we have reviewed the recently introduced APBEK and revAPBEK GGA kinetic energy functionals and have presented an extended study of their ability to compute kinetic energies of atoms and jellium clusters and surfaces and their performance in supermolecular FDE calculations.

These nonempirical functionals recover the MGE2 expansion and have been constructed mainly from the semiclassical theory of the many-electron neutral atoms, which incorporates a strong conjointness conjecture, as we showed in Figure 1. However, small perturbations of the homogeneous electron gas are correctly described by the second-order gradient expansion (GE2), and due to the simplicity of the GGA level, any semilocal functional cannot recover both expansions in the slowly varying density limit. Thus, we introduced two new functionals, APBEKint and revAPBEKint, which recover the GE2 limit at small s and which have enhancement factors similar to the PBEint exchange functional of ref 11.

The APBEK and revAPBEK functionals show performances comparable and in many cases superior to the current state-of-the-art for GGA KE functionals, which in this context were represented by functionals such as LC94 and TW02. The

accuracy of these latter two functionals relied on empirical parametrization to the kinetic energy of small atoms: we showed instead that this empirical parametrization led to enhancement factors which closely resemble the MGE2. This means that the MGE2 (exact for an atom with a large number of electrons) is very accurate also for small atoms so that the MGE2 represents a new paradigm for the construction of accurate and nonempirical kinetic energy functionals.

However, the MGE2 mainly defines the behavior of the enhancement factor for small values of the reduced density gradient, and furthermore GGA KE functionals are approximations of integrated KE T_s and not of KE density τ . Thus, GGA KE functionals rely on error cancellation. To shine light on the role of gradient corrections in KE functionals, we have introduced the s decomposition of KE, equivalent with the KE distribution in the s space (see eqs 14–19). Our approach can well show the behavior of a GGA for a given system (see Figures 4, 6, 7–9) and can be a useful tool in the FDE and kinetic energy development fields.

Our analysis showed that the performance of PBE-like functionals depends on a balance between the low and high s region: thus, for accurate results, μ_s and κ_s must be interrelated. Results presented in this work show that functionals which satisfy the exact MGE2 limit have to be preferred. In fact:

- For the total kinetic energy in atoms/ions and in jellium clusters/surfaces, APBEK performs best and its accuracy increases for large atoms; in these cases, revAPBEK is also accurate, while functionals that satisfy the GE2 limit strongly underestimate the kinetic energy.
- For a benchmark of 20 small molecules with dispersion, dipole–dipole, and hydrogen bond interactions (thus incorporating the behavior of a very large set of molecular systems), we have found that revAPBEK is the most accurate functional for FDE interaction energies (see Table 3), while all of the PBE-like functionals give similar (small) integrated embedding density errors (see Table 4). Recovering the GE2 does not lead to improved accuracy. The similarity of the embedding errors on the density and their similar spatial distribution (see Figures S1–S4 in the Supporting Information) suggests that all of the nonadditive kinetic energy potentials are rather similar.

Therefore, the revAPBEK GGA functional can be considered as the current best choice for FDE calculations of nonbonded systems, because it is nonempirical and it has a well balanced accuracy for any kind of weak interaction.

Our embedding energy s decomposition (eq 25) shows that the performance of the GGA KE functional might be still be improved. In fact, it should be possible to develop new enhancement factors¹⁰⁹ eventually with the inclusion of the laplacian of the density,^{51,53,80,109} which can improve the accuracy of the energy without changing the embedded density. In this context, the MGE2 is an important limit to be satisfied to reduce empiricism.

Future works should also verify the accuracy of the revAPBEK functional for the embedding energy of large systems containing heavy atoms, where the MGE2 limit is expected to work better. Furthermore, error cancellation can also occur with the approximate XC functional. In this work, we used PBE, but the new APBE XC functional⁷⁰ might be more appropriate.

■ ASSOCIATED CONTENT

Supporting Information. Tables containing $\tilde{T}_s^{\text{nadd}}[\tilde{\rho}_A^e, \tilde{\rho}_A^e]$ and $\Delta W[\tilde{\rho}_A^e, \tilde{\rho}_A^e]$ for all systems and functionals; a table of

embedding errors for hydrogen bond systems, using a revAPBEK kinetic energy functional with $\kappa_s = 1.0$; figures on the plane averaged absolute error on the embedded valence density for various systems. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: fabio.dellasala@unisalento.it

■ ACKNOWLEDGMENT

We thank TURBOMOLE GmbH for providing us the TURBOMOLE program package and M. Margarito for technical support. This work was funded by the ERC Starting Grant FP7 Project DEDOM, Grant Agreement No. 207441.

■ REFERENCES

- (1) Kohn, W.; Sham, L. *Phys. Rev.* **1965**, *140*, A1133.
- (2) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: Oxford, 1989; pp 1–333.
- (3) Dreizler, R. M.; Gross, E. K. U. *Density Functional Theory*; Springer: Heidelberg, 1990; pp 1–302.
- (4) Martin, R. *Electronic Structure: Basic Theory and Practical Methods*; Cambridge University Press: Cambridge, U. K., 2004; pp 1–624.
- (5) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: New York, 2001; pp 1–294.
- (6) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.
- (7) Wang, Y.; Carter, E. A. In *Progress in Theoretical Chemistry and Physics*; Schwartz, S., Ed.; Kluwer: Dordrecht, The Netherlands, 2000; p 117.
- (8) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (9) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K. *Phys. Rev. Lett.* **2008**, *100*, 136406.
- (10) Armiento, R.; Mattsson, A. E. *Phys. Rev. B* **2005**, *72*, 085108.
- (11) Fabiano, E.; Constantin, L. A.; Della Sala, F. *Phys. Rev. B* **2010**, *82*, 113104.
- (12) Fabiano, E.; Constantin, L. A.; Della Sala, F. *J. Chem. Phys.* **2011**, *134*, 194112.
- (13) Pearson, M.; Smargiassi, E.; Madden, P. A. *J. Phys. Condens. Matter* **1993**, *5*, 3221.
- (14) Smargiassi, E.; Madden, P. A. *Phys. Rev. B* **1994**, *49*, 5220.
- (15) Govind, N.; Wang, J.; Guo, H. *Phys. Rev. B* **1994**, *50*, 11175.
- (16) Foley, M.; Madden, P. A. *Phys. Rev. B* **1996**, *53*, 10589.
- (17) Wang, Y. A.; Govind, N.; Carter, E. A. *Phys. Rev. B* **1998**, *58*, 13465.
- (18) Wang, Y. A.; Govind, N.; Carter, E. A. *Phys. Rev. B* **2001**, *64*, 129901.
- (19) Watson, S. C.; Carter, E. A. *Comput. Phys. Commun.* **2000**, *128*, 67.
- (20) Zhou, B.; Ligneres, V. L.; Carter, E. A. *J. Chem. Phys.* **2005**, *122*, 044103.
- (21) Senatore, G.; Subbaswamy, K. R. *Phys. Rev. B* **1986**, *34*, 5754.
- (22) Johnson, M. D.; Subbaswamy, K. R.; Senatore, G. *Phys. Rev. B* **1987**, *36*, 9202.
- (23) Cortona, P. *Phys. Rev. B* **1991**, *44*, 8454.
- (24) Cortona, P. *Phys. Rev. B* **1992**, *46*, 2008.
- (25) Wesolowski, T. A.; Warshel, A. *J. Phys. Chem.* **1993**, *97*, 8050.
- (26) Wesolowski, T. A. In *Chemistry: Reviews of Current Trends*; Leszczynski, J., Ed.; World Scientific: Singapore, 2006; Vol. 10; pp 1–82.
- (27) Wesolowski, T. A.; Chermette, H.; Weber, J. *J. Chem. Phys.* **1996**, *105*, 9182.

- (28) Huang, P.; Carter, E. A. *J. Chem. Phys.* **2006**, *125*, 084102.
- (29) Hodak, M.; Lu, W.; Bernholc, J. *J. Chem. Phys.* **2008**, *128*, 014101.
- (30) Elliot, P.; Cohen, M. H.; Wasserman, A.; Burke, K. *J. Chem. Theory Comput.* **2009**, *5*, 827.
- (31) Neugebauer, J. *Phys. Rep.* **2010**, *489*, 1.
- (32) Laricchia, S.; Fabiano, E.; Della Sala, F. *J. Chem. Phys.* **2010**, *133*, 164111.
- (33) Goodpaster, J. D.; Ananth, N.; Manby, F. R.; Miller, T. F., III. *J. Chem. Phys.* **2010**, *133*, 084103.
- (34) Wesolowski, T. A. *J. Chem. Phys.* **1997**, *106*, 8516.
- (35) Wesolowski, T. A.; Tran, F. *J. Chem. Phys.* **2003**, *118*, 2072.
- (36) Kevorkyants, R.; Dulak, M.; Wesolowski, T. A. *J. Chem. Phys.* **2006**, *124*, 024104.
- (37) Dulak, M.; Wesolowski, T. A. *J. Mol. Model* **2007**, *13*, 631.
- (38) Götz, A. W.; Beyhan, S. M.; Visscher, L. *J. Chem. Theory Comput.* **2009**, *5*, 3161.
- (39) Beyhan, S. M.; Götz, A. W.; Jacob, C. R.; Visscher, L. *J. Chem. Phys.* **2010**, *132*, 044114.
- (40) Fradelos, G.; Wesolowski, T. A. *J. Chem. Theory Comput.* **2011**, *7*, 213.
- (41) Thomas, L. H. *Proc. Cambridge Phil. Soc.* **1926**, *23*, 542.
- (42) Fermi, E. *Rend. Accad. Naz. Lincei* **1927**, *6*, 602.
- (43) von Weizsäcker, C. F. Z. *Phys. A* **1935**, *96*, 431.
- (44) Acharya, P. K.; Bartolotti, L. J.; Sears, S. B.; Parr, R. G. *Proc. Natl. Acad. Sci.* **1980**, *77*, 6978.
- (45) DePristo, A. E.; Kress, J. D. *Phys. Rev. A* **1987**, *35*, 438.
- (46) Ou-Yang, H.; Levy, M. *Int. J. Quantum Chem.* **1991**, *40*, 379.
- (47) Thakkar, A. J. *Phys. Rev. A* **1992**, *46*, 6920.
- (48) Vitos, L.; Skriver, H. L.; Kollár, J. *Phys. Rev. B* **1998**, *57*, 12611.
- (49) Karasiev, V. V.; Ludeña, E. V.; Artemyev, A. N. *Phys. Rev. A* **2000**, *62*, 062510.
- (50) Karasiev, V. V.; Trickey, S. B.; Harris, F. E. *J. Comput.-Aided Mater. Des.* **2006**, *13*, 111.
- (51) Perdew, J. P.; Constantin, L. A. *Phys. Rev. B* **2007**, *75*, 155109.
- (52) Constantin, L. A.; Ruzsinszky, A. *Phys. Rev. B* **2009**, *79*, 115117.
- (53) Karasiev, V. V.; Jones, R. S.; Trickey, S. B.; Harris, F. E. *Phys. Rev. B* **2009**, *80*, 245120.
- (54) Wang, L.-W.; Teter, M. P. *Phys. Rev. B* **1992**, *45*, 13196.
- (55) Chai, J.-D.; Weeks, J. D. *Phys. Rev. B* **2007**, *75*, 205122.
- (56) Garca-Aldea, D.; Alvarelllos, J. E. *J. Chem. Phys.* **2008**, *129*, 074103.
- (57) Garca-Aldea, D.; Alvarelllos, J. E. *Phys. Rev. A* **2008**, *77*, 022502.
- (58) Huang, C.; Carter, E. A. *Phys. Rev. B* **2010**, *81*, 045206.
- (59) Garca-Aldea, D.; Alvarelllos, J. E. *J. Chem. Phys.* **2007**, *127*, 144109.
- (60) March, N.; Santamaria, R. *Int. J. Quantum Chem.* **1990**, *39*, 585.
- (61) Lee, H.; Lee, C.; Parr, R. G. *Phys. Rev. A* **1991**, *44*, 768.
- (62) March, N. *Electron Density Theory of Atoms and Molecules*; Academic Press: London, 1992; p 1.
- (63) Oliver, G. L.; Perdew, J. P. *Phys. Rev. A* **1979**, *20*, 397.
- (64) Lembarki, A.; Chermette, H. *Phys. Rev. A* **1994**, *50*, 5328.
- (65) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.
- (66) Kirzhnits, D. *Sov. Phys. JETP* **1957**, *5*, 64.
- (67) Bernard, Y. A.; Dulak, M.; Kamiński, W.; Wesolowski, T. A. *J. Phys. A: Math. Theor.* **2008**, *41*, 055302.
- (68) Tran, F.; Wesolowski, T. A. *Int. J. Quantum Chem.* **2002**, *89*, 441.
- (69) Lacks, D. J.; Gordon, R. G. *J. Chem. Phys.* **1994**, *100*, 4446.
- (70) Constantin, L. A.; Fabiano, E.; Laricchia, S.; Della Sala, F. *Phys. Rev. Lett.* **2011**, *106*, 186406.
- (71) Elliott, P.; Lee, D.; Cangi, A.; Burke, K. *Phys. Rev. Lett.* **2008**, *100*, 256406.
- (72) Zhang, Y.; Yang, W. *Phys. Rev. Lett.* **1998**, *80*, 890.
- (73) Scott, J. *Philos. Mag.* **1952**, *43*, 859.
- (74) Schwinger, J. *Phys. Rev. A* **1980**, *22*, 1827.
- (75) Englert, B.-G.; Schwinger, J. *Phys. Rev. A* **1984**, *29*, 2339.
- (76) Lee, D.; Constantin, L. A.; Perdew, J. P.; Burke, K. *J. Chem. Phys.* **2009**, *130*, 034107.
- (77) Perdew, J. P.; Constantin, L. A.; Sagvolden, E.; Burke, K. *Phys. Rev. Lett.* **2006**, *97*, 223002.
- (78) Hui, O.-Y.; Levy, M. *Phys. Rev. A* **1990**, *42*, 155.
- (79) Levy, M.; Hui, O.-Y. *Phys. Rev. A* **1988**, *38*, 625–629.
- (80) Wang, B.; Stott, M. J.; von Barth, U. *Phys. Rev. A* **2001**, *63*, 052501.
- (81) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540.
- (82) Lastra, J. M. G.; Kaminski, J. W.; Wesolowski, T. A. *J. Chem. Phys.* **2008**, *129*, 074107.
- (83) Jones, W.; Young, W. J. *Phys. C* **1971**, *4*, 1322.
- (84) Vela, A.; Medel, V.; Trickey, S. B. *J. Chem. Phys.* **2009**, *130*, 244103.
- (85) Clementi, E.; Roetti, C. *Atomic Data Nucl. Data Tables* **1974**, *14*, 177.
- (86) Constantin, L. A.; Snyder, J. C.; Perdew, J. P.; Burke, K. *J. Chem. Phys.* **2010**, *133*, 241103.
- (87) Wesolowski, T. A.; Warshel, A. *Chem. Phys. Lett.* **1996**, *258*, 71.
- (88) TURBOMOLE, V6.2; University of Karlsruhe and Forschungszentrum Karlsruhe GmbH: Karlsruhe, Germany, 1989–2007; TURBOMOLE GmbH: Karlsruhe, Germany, 2009.
- (89) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2010**, *133*, 134105.
- (90) Weigned, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
- (91) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656.
- (92) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput* **2005**, *1*, 415.
- (93) Zupan, A.; Burke, K.; Ernzerhof, M.; Perdew, J. P. *J. Chem. Phys.* **1997**, *106*, 10184.
- (94) Treutler, O.; Ahlrichs, R. *J. Chem. Phys.* **1995**, *102*, 346.
- (95) Iyengar, S. S.; Ernzerhof, M.; Maximoff, S. N.; Scuseria, G. E. *Phys. Rev. A* **2001**, *63*, 052508.
- (96) Murphy, D. R. *Phys. Rev. A* **1981**, *24*, 1682.
- (97) Tran, F.; Wesolowski, T. A. *Chem. Phys. Lett.* **2002**, *360*, 209.
- (98) Jacob, C. R.; Beyhan, S. M.; Visscher, L. *J. Chem. Phys.* **2007**, *126*, 234116.
- (99) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
- (100) Jacob, C. R.; Visscher, L. *J. Chem. Phys.* **2008**, *128*, 155102.
- (101) Kiewisch, K.; Eickerling, G.; Reiher, M.; Neugebauer, J. *J. Chem. Phys.* **2008**, *128*, 044114.
- (102) Govind, N.; Sushko, P.; Hess, W.; Valiev, M.; Kowalski, K. *Chem. Phys. Lett.* **2009**, *470*, 353.
- (103) Fux, S.; Jacob, C. R.; Neugebauer, J.; Visscher, L.; Reiher, M. *J. Chem. Phys.* **2010**, *132*, 164101.
- (104) Zhao, Q.; Morrison, R. C.; Parr, R. G. *Phys. Rev. A* **1995**, *50*, 238.
- (105) Wu, Q.; Yang, W. *J. Chem. Phys.* **2003**, *118*, 2498.
- (106) Roncero, O.; de Lara-Castells, M. P.; Villarreal, P.; Flores, F.; Ortega, J.; Paniagua, M.; Aguado, A. *J. Chem. Phys.* **2008**, *129*, 184104.
- (107) Roncero, O.; Zanchet, A.; Villarreal, P.; Aguado, A. *J. Chem. Phys.* **2009**, *131*, 234110.
- (108) Goodpaster, J. D.; Barnes, T. A.; Miller, T. F., III. *J. Chem. Phys.* **2011**, *134*, 164108.
- (109) Chan, G. K.-L.; Handy, N. C. *J. Chem. Phys.* **2000**, *112*, 5639.

Comparison of Methods to Obtain Force-Field Parameters for Metal Sites

LiHong Hu^{†,‡} and Ulf Ryde^{*,‡}[†]School of Computer Science and Information Technology, North-east Normal University, Changchun, 130024, People's Republic of China[‡]Department of Theoretical Chemistry, Lund University, Chemical Centre, Post Office Box 124, SE-221 00 Lund, Sweden Supporting Information

ABSTRACT: We have critically examined and compared various ways to obtain standard harmonic molecular mechanics (MM) force-field parameters for metal sites in proteins, using the 12 most common Zn²⁺ sites as test cases. We show that the parametrization of metal sites is hard to treat with automatic methods. The choice of method is a compromise between speed and accuracy and therefore depends on the intended use of the parameters. If the metal site is not of central interest in the investigation, for example, a structural metal far from the active site, a simple and fast parametrization is normally enough, using either a nonbonded model with restraints or a bonded parametrization based on the method of Seminario. On the other hand, if the metal site is of central interest in the investigation, a more accurate method is needed to give quantitative results, for example, the method by Norrby and Liljefors. The former methods are semiautomatic and can be performed in seconds, once a quantum mechanical (QM) geometry optimization and frequency calculation has been performed, whereas the latter method typically takes several days and requires significant human intervention. All approaches require a careful selection of the atom types used. For a nonbonded model, standard atom types can be used, whereas for a bonded model, it is normally wise to use special atom types for each metal ligand. For accurate results, new atom types for all atoms in the metal site can be used. Atomic charges should also be considered. Typically, QM restrained electrostatic potential charges are accurate and easy to obtain once the QM calculation is performed, and they allow for charge transfer within the complex. For negatively charged complexes, it should be checked that hydrogen atoms of the ligands get proper charges. Finally, water ligands pose severe problems for bonded models in force fields that ignore nonbonded interactions for atoms separated by two bonds. Complexes with a single water ligand can normally be accurately treated with a bonded potential, once it is ensured that the water H atoms have nonzero Lennard-Jones parameters. However, for metal sites with several water molecules, a nonbonded model with restraints (taken from the QM calculations) is more stable.

INTRODUCTION

Molecular mechanics (MM) simulations have become an important complement to experiments for obtaining structural and mechanistic information on biological systems at an atomic level.¹ For example, in the great majority of X-ray and NMR structure determinations, the experimental data are supplemented by MM calculations to give chemically reasonable bond lengths and angles.² Moreover, MM calculations and molecular dynamics (MD) simulations have become an important ingredient in biochemical and medicinal chemistry studies.^{3,4}

The advantage with MM methods is their speed: With today's computer resources and software, you can study even big protein complexes with full atomic detail and simulate medium-sized proteins for hundreds of nanoseconds. On the other hand, the MM methods need to be parametrized, that is, you need to have MM parameters for all atoms in the system of interest. For biochemical macromolecules, this does not pose any problem, because standard MM parameters are available for all normal amino acids and nucleic acids.^{5,6} Moreover, more general force fields are available for other molecules, such as carbohydrates and small druglike molecules.^{7–16}

However, metal sites constitute a major problem for force fields.^{1,17–22} The reason for this is that the strength of metal–ligand bonds is intermediate between that of covalent bonds

and nonbonded interactions, such as hydrogen bonds. Moreover, metals can have many different types of ligands and the number of bonds around a metal is often more than the number of covalent bonds around an atom. Therefore, the metal-coordination sphere is often flexible with several different geometries possible. Finally, for transition metals, quantum mechanical ligand-field, spin-state, trans, and Jahn–Teller effects also become important, which are hard to model in a standard MM force field.²³ Therefore, MM parametrizations for metals have traditionally been restricted to specific metal sites (with a given set of ligands), for which accurate results can be obtained,^{24–27} or metal-specific force fields requiring specialized software.^{23,28,29}

There are several approaches to incorporate metal ions into MM force fields. The simplest one is to describe the interaction between the ion and its ligand entirely by nonbonded interactions, that is, by electrostatics and Lennard-Jones terms (the nonbonded or ionic method). Such a model has been suggested for Zn, based on formal charges on the ion (+2).³⁰ This model has also been modified to include polarization and charge transfer,³¹ or dummy atoms between the metal and the ligands.^{32–34} Sometimes, the nonbonded potential has been supplemented by

Received: December 17, 2010

Published: June 30, 2011

metal–ligand restraints to ensure that the ligands stay bound.³⁵ The restraints can be of many types, for example, harmonic or flat-bottomed, single- or double-sided.

Alternatively, explicit bonds are defined between the metal and its ligand, a bonded (valence) model. Then, the metal–ligand bonds are treated the same way as covalent bonds, that is, with bond, angle, and dihedral terms (although the latter are often ignored for metal sites).^{24,25,36,37} This bonded model is typically combined with a nonbonded potential (electrostatics and Lennard-Jones terms), using either formal charges on the metal or charges derived from quantum mechanical (QM) calculations.²⁴ The points-on-a-sphere approach is a mixed model in which a metal–ligand stretching function is used, but the ligand–metal–ligand bond angles are replaced by 1,3-nonbonded interactions for the coordinating atoms.^{17,38,39} Other more specialized methods to treat metal ions also exist, for example, ligand-field MM,²³ Valbond,²⁸ the universal force field (UFF),^{40,41} YETI,^{42,43} or SIBFA.^{44–46}

Likewise, there exist several methods to perform a force-field parametrization, which depend on the force field used. For example, for a nonbonded force field with formal charges, only two parameters (the Lennard-Jones parameters for the metal) are available, and they can easily be optimized to reproduce QM data, for example, to give proper metal–ligand distances. However, for a bonded force field, you have many more parameters that need to be optimized. In principle, these parameters can also be obtained from QM calculations, for example, by taking the optimum metal–ligand bond lengths and angles from the values observed in QM-optimized structures of model complexes.^{24,25} The corresponding force constants can be obtained from QM scans of the potential surface for the bond or angle, but this is tedious for more than a few parameters. Instead, it is more common to obtain the force constants from a projection of the Hessian matrix (obtained from a QM frequency calculation) into internal coordinates.^{16,24,25} Unfortunately, such a procedure is ambiguous because the total number of internal coordinates (bond, angles, and dihedrals) around a metal is more than the degrees of freedom (for example, around a four-coordinate metal, four bonds and six angles can be defined, but there are only $5 \times 3 - 6 = 9$ degrees of freedom). This means that the internal coordinates are not independent and different choices of internal coordinates will give different force constants.^{47,48} Moreover, it is not certain that the optimum bond lengths and angles obtained from a QM optimization represent unstrained equilibrium parameters (as is assumed for a MM force field). Instead, they represent an optimum compromise for all interactions (bonded as well as nonbonded) between all atoms in the complex.

In 1996, Seminario⁴⁷ suggested an approach to obtain force constants directly from the Hessian matrix, thereby avoiding any use of internal coordinates. This procedure has been employed in automatic parametrization programs by at least three groups, for example, Hess2FF and the metal-center parameter builder.^{48–50} It makes the extraction of force constants from the Hessian unambiguous (in fact, the force constants can be obtained in two ways, but they typically give similar results that can be averaged).⁴⁸ Test calculations have shown that the Seminario approach gives better force constants than the method involving internal coordinates.⁴⁹ Unfortunately, the interdependence of the various internal coordinates still exists. Moreover, this approach involves a double-counting of electrostatic and Lennard-Jones interactions: The QM Hessian matrix contains all

interactions, including electrostatic and van der Waals interactions. However, the Hessian is used only to extract the bonded interactions, whereas electrostatics and van der Waals interactions are calculated by separate terms by the MM program. For bonds and angles, this is no problem, because most force fields ignore nonbonded interactions between atoms one or two bonds apart. However, for dihedrals, it is a serious problem, because most force fields complement dihedral terms with nonbonded energies, typically scaled down by a constant factor (for example, in the AMBER force field, used in this paper, 1,4-electrostatics are scaled down by a factor of 1.2 and 1,4-Lennard-Jones interactions by a factor of 2.0). This means that the dihedral parameters will already contain some nonbonded interactions, which then are double-counted in the MM calculations. Thus, the Seminario approach is only approximate.

Norrby and Liljefors and co-workers^{51,52} have suggested an approach that solves these problems. It involves a complete optimization of all parameters of the force field in an iterative manner. For every set of parameters, the MM structure is optimized, so there is no risk of double-counting any interactions and it is not assumed that the geometric parameters in the QM structure represent equilibrium values in the MM force field. However, QM data are still used as the reference, e.g. bond lengths, angles, dihedrals from the QM structure, as well as the Hessian elements, and the fitting procedure ensures that the optimized MM structure is as close as possible to the QM structure. This approach was originally developed for the MM3 force field,⁵³ but it has recently been implemented also for the AMBER force field and software.⁵⁴

In this paper, we compare five different approaches to obtain MM parameters for metal sites in proteins using standard nonpolarizable harmonic force fields, viz., two variants of a nonbonded potential, a restrained nonbonded potential, the Seminario approach,⁴⁷ as implemented in the Hess2FF software,⁴⁸ the zinc AMBER force field (ZAFF; also based on the Seminario approach),⁴⁹ and the ideal procedure of Norrby and Liljefors (NL).^{51,54} As test cases, we use 12 simple models for the most common Zn²⁺ sites in proteins. Sites of this type were also used in the previous studies.^{49,50} As a reference, we use the corresponding structures optimized by QM methods, and we compare with the structures and Hessian elements of the structures optimized with the various force fields. This comparison also allows us to discuss various problems that are typically encountered during the parametrization of metal sites. The take-home message is that the method of choice depends on the intended use of the force field and that the parametrization of metal sites is seldom an automatic procedure.

METHODS

Structures. As test cases, we used 10 models of the most common Zn²⁺ sites in proteins, taken from the previous investigation by Merz and co-workers:⁵⁰ ZnCys₄, ZnCys₃His, ZnCys₂His₂, ZnCysHis₃, ZnHis₄, ZnHis₃H₂O, ZnHis₂(H₂O)₂, ZnHis(H₂O)₃, ZnHis₃Asp, and ZnHis₂Asp₂. For two of the models (ZnCys₃His and ZnHis₃H₂O), alternative coordinating atoms of the His groups were also tested.⁵⁰ In these models, Cys was modeled by CH₃S[−], His by methylimidazole, and Asp by CH₃COO[−]. The starting structures for the optimizations were taken from typical crystal structures, selected as in the previous investigation.⁵⁰ The structures were protonated by the tleap

routine in AMBER-10⁵⁵ and then truncated with hydrogen atoms (these H atoms will be called HT below).

Quantum Mechanical Calculations. The structures were optimized with the hybrid density-functional theory method B3LYP,^{56–58} using the 6-31G* basis set^{59–61} (again following the previous study).⁵⁰ After the geometry optimization, analytic frequencies were calculated at the same level of theory. This structure and the Hessian matrix obtained from the frequency calculation were used as a reference in both the NL and Hess2FF parametrizations and as the target for all the MM minimizations. Charges on all atoms were calculated with the RESP approach, as implemented in AMBER-10,⁵⁵ using electrostatic potentials calculated at the B3LYP/6-31G* level and sampled with the Merz–Kollman scheme,⁶² albeit with a higher-than-default density of points (10 concentric layers with 17 points/Å²). All QM calculations were performed with Gaussian-09.⁶³ Coordinates of the optimized QM structures are given in Table S10 in the Supporting Information.

Force Fields, Parametrizations, and Molecular Mechanical Calculations. All the MM calculations were run with the sander (minimization) and nmode (Hessian calculation) modules of the AMBER-10 software.⁵⁵ Two sets of nonbonded force fields were studied (NB1 and NB2). They differed in the charges and Lennard-Jones parameters: One set used a formal charge of +2 on Zn and standard AMBER charges on all the ligand atoms (corrected to a proper net charge for each ligand on the HT atom; NB2). All the other force fields (including the bonded potentials) used the QM RESP charges for all atoms (note that ZAFF normally uses charges calculated for larger models that include full capped amino acids,⁵⁰ but for the truncated models we used the RESP charges also for ZAFF). Likewise, one set of nonbonded force field (NB1) used the Stote–Karplus Lennard-Jones parameters for Zn,³⁰ $R = 1.95$ Å and $\epsilon = 1.046$ kJ/mol, whereas all the other force fields employed the Merz parameters, $R = 1.1$ Å and $\epsilon = 0.0523$ kJ/mol.^{24,50} On all the other atoms, standard AMBER Lennard-Jones parameters were used, except sometimes for water, as will be discussed below.

Atom types and bonded parameters for ZAFF were extracted from the Supporting Information of the original publication.⁵⁰ The restrained nonbonded potential used RESP charges, the Merz Lennard-Jones parameters, and a double-sided harmonic potential with the minimum distance taken from the QM structure and the force constant taken from the Hess2FF⁴⁸ calculation (distinct distances and force constants were used for all four ligands).

For the Hess2FF and NL force fields, two sets of atom types were tested. The first (AT_{min}) used standard AMBER atom types for all ligands (and the same atom type for the HT atom as the other hydrogen atoms bound to the same carbon atom). A distinct atom type was used for each Zn-ligating atom. This means that only bonded interactions involving Zn were parametrized; all the other bonded parameters were taken from the AMBER-99SB force field.^{64,65} The second set (AT_{max}) used individual atom types for all atoms in the complex. Thus, all bonded parameters of the complex were optimized.

The Hess2FF force field was obtained with the Hess2FF program.⁴⁸ This program reads the output file of the Gaussian frequency calculation and calculates all the bonded parameters via the Seminario approach.⁴⁷ For the AT_{max} set, the program is completely automatic and generates AMBER topology and parameter files. For the AT_{min} set, a file with the desired atom types is provided as input. The program then gives the proper averaging of all parameters. Further instructions for the program

can be found in <http://www.teokem.lu.se/~ulf/Methods/parm.html>.

The NL force field was constructed according to the method developed by Norrby and Liljefors,⁵¹ using the recent implementation for AMBER.⁵⁵ This method minimizes a penalty function consisting of the deviation of geometries and Hessian elements between the QM and MM calculations, giving different weights to different kinds of data. The geometries were described as lists of all bonds, angles, and dihedral angles, rather than by absolute coordinates. The weight factors of the various data types were 100 Å⁻¹ for bonds, 2 deg⁻¹ for angles, 1 deg⁻¹ for torsions, and 0.01–0.1 mol·Å²/kcal for Hessian elements (0.01 for elements involving interactions of an atom with itself, 0.02 for atoms bound to each other, 0.04 for atoms connected by two bonds, 0.1 for atoms connected by three bonds, and 0.01 for all other elements).^{51,66}

The iterative NL optimizations were started from the corresponding Hess2FF force field. After convergence, the force field was checked. Typically, some bonds and angles get zero force constants in the first runs of the parametrizations. These were reset to reasonable values and force constants of the other angles around the same central atom were reduced, and then the parametrization was run again. This was repeated until all bonds and angles had nonzero force constants and all other parameters were reasonable. Further instructions for the procedure are found in <http://www.teokem.lu.se/~ulf/Methods/ponparm.html>. Both the NL and the Hess2FF programs are available from the authors upon request.

Molecular Dynamics Simulations. Two Zn-containing proteins were studied with molecular dynamics (MD) simulations, viz. the δ' subunit of the clamp-loader complex of DNA polymerase III (PDB file 1AST),⁶⁷ which contains a single Cys₄ Zn finger, and the spore coat polysaccharide biosynthesis protein SpSE (PDB file 1VLI),⁶⁸ which contains a ZnHis₂(H₂O)₂ site. The two structures were protonated with the tleap module in AMBER,⁵⁵ with the assumption that all Asp and Glu residues are negatively charged and all Lys and Arg residues are positively charged. The protonation state of the His residues was decided from a detailed study of the hydrogen-bond network and the local surroundings of these sites. The two His residues in 1VLI that coordinate to Zn were protonated on the ND1 atom, whereas the other His residues were assumed to be protonated on the NE2 atom. For 1AST, His residues 24, 73, 103, and 238 were assumed to be protonated on the NE2 atom and the other residues were assumed to be doubly protonated. A few side-chain atoms that were not resolved in the 1VLI crystal structure were built with the tleap software. Residues 65–73 were also missing in the structure and they were ignored in the calculations.

Both proteins were solvated in a periodic octahedral box with water molecules extending at least 9 Å from the protein on all sides, keeping the crystallographic water molecules. Six simulations were performed. First, the systems were subjected to a 1000-step minimization, keeping all heavy atoms in the proteins restrained toward their positions in the crystal structure with a force constant of 418 kJ·mol⁻¹·Å⁻². Then, two 20 ps MD simulations were run with the same restraints. The first simulation was run with a constant volume and the second with a constant pressure. Finally, the box size was equilibrated by a 50-ps MD simulation with a constant pressure and without any restraints. Finally, an equilibration of 200 ps and a production simulation of 5 ns were run with a constant volume. During the latter run, coordinates were collected every 2 ps.

Table 1. Performance of the Various Force Fields for the ZnHis₄ Complex^a

	NB1	NB2	restrained	ZAFF	Hess2FF		NL	
					AT _{min}	AT _{max}	AT _{min}	AT _{max}
<i>r</i> ² Hessian	0.861	0.978	0.967	0.979	0.972	0.963	0.987	0.985
rmsd bonds	0.973	0.045	0.025	0.011	0.016	0.010	0.017	0.006
rmsd angles	7.3	3.1	5.0	2.3	2.2	0.4	2.2	0.2
rmsd dihedrals	14.4	2.3	3.3	2.6	3.7	3.6	3.0	1.0
rmsd coordinates	2.45	0.26	0.33	0.18	0.20	0.20	0.21	0.05
max coordinates	6.05	0.44	0.57	0.40	0.46	0.47	0.52	0.12

^aThe six quality measures are correlation coefficient (*r*²) between all QM and MM Hessian elements; root-mean-squared deviations (rmsd) for all bonds, angles, dihedral angles, and coordinates between MM and QM optimized structures; and maximum deviation (max) for coordinates.

The MD calculations were run with the sander module in the Amber software,⁵⁵ using the Amber 1999SB force field (FF99).^{64,65} For 1AST, we used the ZAFF charges for the Zn site, whereas for 1VLL, the ZAFF charges are erroneous in the deposited files (e.g., $7 \times 10^{247} e$ for one of the two H atoms in water).⁵⁰ Therefore, we recalculated these charges, following the instructions in the ZAFF paper (the charges are listed in Table S9 in the Supporting Information).⁵⁰ The NB2 simulations employed a +2 charge for the Zn ion and Amber 1999SB charges⁶⁵ for the other residues. Water was described explicitly with the TIP3P model.⁶⁹ Long-range electrostatics were treated with particle-mesh Ewald method^{70,71} with a grid size of 80³, a fourth-order B-spline interpolation, a tolerance of 10⁻⁵, and a real-space cutoff of 8 Å. The temperature was kept constant at 300 K and the pressure was kept at 1 atm by use of the Berendsen weak-coupling algorithm⁷² with time constants of 1 ps. The nonbonded pair list was updated every 50 fs. Bond lengths involving hydrogen atoms were constrained with the SHAKE algorithm,⁷³ allowing for a MD time step of 2 fs. However, for some force fields, several of the simulations crashed with SHAKE failure. This was solved by turning off SHAKE and reducing the time step to 0.5 fs.

RESULTS AND DISCUSSION

In this paper, we compare the performance of eight different Zn force fields on 12 model complexes of the most common Zn sites in proteins.⁵⁰ We will first discuss the results of a model complex with typical behavior, followed by the average results of all complexes. Then we will describe three complexes that illustrate two types of problems that are encountered when parametrizing metal complexes. The performance of the force fields is judged by how closely they reproduce the structure and the Hessian of the QM-optimized structure used for the parametrization. For the Hessian, we calculated the correlation coefficient (*r*²) between all QM and MM Hessian elements. For the structure, we studied the root-mean-squared deviations (rmsd) for all bonds, angles, and dihedral angles (in Ångströms or degrees). Moreover, we will list the root-mean-squared deviations (rmsd) and maximum deviation for the coordinates (after a rmsd fit of the MM and QM structures).

Zn(His)₄ Model. The [Zn(CH₃-imidazole)₄]²⁺ model gives results that are typical for most of the complexes. From Table 1, it can be seen that the nonbonded potential with RESP charges and Stote–Karplus Zn Lennard-Jones parameters (NB1) gave the worst results, with a rmsd for the coordinates of 2.45 Å, illustrating that the structure has completely changed. The

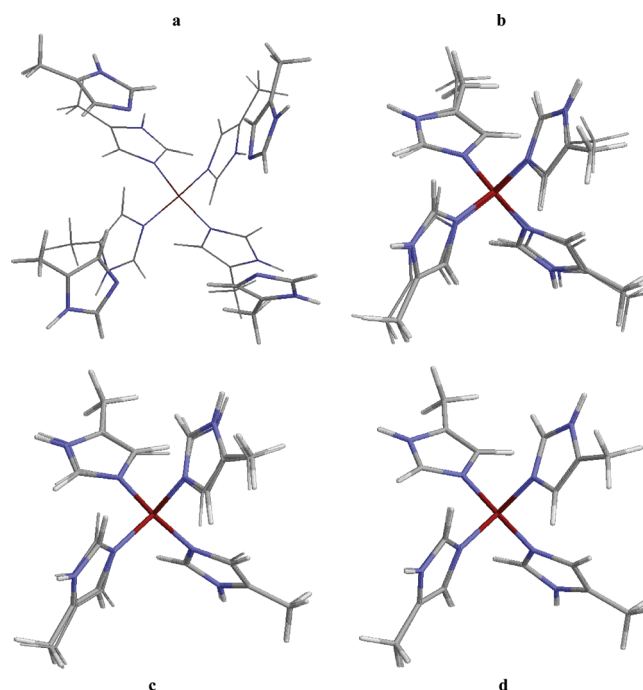


Figure 1. Comparison of QM structure (thin sticks) with MM structures obtained with (a) NB1, (b) NB2, (c) ZAFF, and (d) NL–AT_{max} force fields for the Zn(His)₄ model.

reason for this is that all Zn–N distances have increased (from 2.01 to 3.35, 3.56, 4.94, and 8.05 Å; cf. Figure 1a), illustrating that the parameters are not optimized for this complex. Much better results were obtained with a formal +2 charge for Zn (and standard AMBER charges for the other atoms), especially when combined with the Merz Zn Lennard-Jones parameters (NB2; cf. Figure 1b; the Stote–Karplus parameters gave appreciably worse results). Of course, even better results could be obtained by tuning the Lennard-Jones parameters for Zn, but because there are only two parameters available, only restricted improvements can be expected, especially for complexes with more than one type of ligands. This illustrates that nonbonded models are very sensitive to the nonbonded parameters and that a single set of parameters will not be optimal for all type of complexes.¹⁷

The restrained nonbonded potential gave much better result than NB1 but slightly worse than NB2 (because it is based on the RESP charges), although the rmsd for the bonds is better. This shows that the restraints ensure that the ligands do not dissociate.

Table 2. Average Performance of the Various Force Fields Tested for the 12 Model Complexes^a

	NB1	NB2	restrained	ZAFF ^b	Hess2FF		NL	
					AT _{min}	AT _{max}	AT _{min}	AT _{max}
<i>r</i> ² Hessian	0.742	0.882	0.833	0.907	0.896	0.928	0.985	0.992
rmsd bonds	0.213	0.048	0.020	0.022	0.032	0.012	0.012	0.003
rmsd angles	3.6	3.2	3.3	2.6	2.9	1.9	1.6	0.4
rmsd dihedrals	11.9	10.1	10.3	8.2	10.3	8.4	5.5	2.3
rmsd coordinates	1.25	0.56	0.50	0.33	0.34	0.26	0.16	0.07
max coordinates	2.31	1.10	1.04	0.79	0.71	0.57	0.39	0.15

^a Quality measures are the same as in Table 1. Raw data are given in Tables 1 and 3–5 and in Tables S1–S8 in Supporting Information. ^b ZAFF failed for two complexes [ZnHis₃(H₂O) and ZnHis₂(H₂O)₂], which are omitted from the average only for ZAFF.

Table 3. Performance of the Various Force Fields for the ZnCys₄ Complex^a

	NB1	NB2	restrained	ZAFF	Hess2FF		NL	
					AT _{min}	AT _{max}	AT _{min}	AT _{max}
<i>r</i> ² Hessian	0.869	0.836	0.384	0.853	0.901	0.904	0.995	0.993
rmsd bonds	0.330	0.245	0.062	0.076	0.016	0.014	0.006	0.004
rmsd angles	4.4	2.1	3.6	1.2	0.6	0.4	0.4	1.0
rmsd dihedrals	10.0	14.9	37.1	13.1	13.2	13.2	13.2	12.6
rmsd coordinates	0.49	0.46	0.30	0.32	0.32	0.32	0.31	0.29
max coordinates	0.74	0.76	1.29	0.44	0.54	0.54	0.54	0.53

^a Quality measures are the same as in Table 1.

All the bonded force fields gave better results than the nonbonded ones, with RMSDs for the coordinates of only 0.05–0.21 Å (cf. Figure 1c). As expected, the ZAFF and Hess2FF–AT_{min} approaches gave similar results; they employ essentially the same methodology and differ only in details of the implementation (for example, ZAFF have zeroed all force constants for dihedrals involving Zn). Interestingly, the NL–AT_{min} results are also similar (NL–AT_{min} gives better Hessian and angles, whereas ZAFF gives better bonds, dihedrals, and coordinates). Of course, NL approach can be tuned to give better results for the other quality measures by changing the weights in the NL penalty function.

Even better results can be obtained if all atom types are optimized (AT_{max}), with both Hess2FF and NL. This shows that the accuracy of ZAFF and the other AT_{min} force fields is mainly limited by standard AMBER force-field parameters, used for the imidazole ligands. However with AT_{max}, NL gives much better results than Hess2FF in all quality measures. In fact, NL–AT_{max} gives a nearly perfect fit, as can be seen in Figure 1d. This shows that NL is inherently a much better approach for parametrization and that the rather poor results with the AT_{min} atom types mainly reflected problems in the AMBER parameters.

Other Models. Most of the other models gave results similar to that of the Zn(His)₄ model. Therefore, the results are not discussed in detail (the performance of each model is shown in Tables S1–S8 in the Supporting Information). Instead we list the average quality measures of each MM force field for all 12 models in Table 2. From these data, it can be seen that NL–AT_{max} gives the best performance for all six quality measures. In fact, the rmsd for the bonds and angles are 4 times better than the second-best method, and the rmsd for the dihedrals and the rmsd and maximum deviation of the coordinates are half that of the

second-best method. This shows that this approach is considerably more accurate than the other methods.

The next method is NL–AT_{min}, which is second in all quality measures, even if the rmsd for the bonds is similar to that of Hess2FF–AT_{max}. The latter method is third in performance for all quality measures except the rmsd for dihedrals, for which ZAFF gives slightly better results.

The ZAFF force field gives slightly better results than the Hess2FF–AT_{min} force field in all quality measures except the maximum coordinate error. On the other hand, we will see below that ZAFF failed for two of the complexes, which are omitted from the ZAFF average but not from the averages of the other methods; if these two complexes are omitted also from the Hess2FF–AT_{min} average, the latter method is actually better than ZAFF in all quality measures (giving 0.951, 0.022 Å, 2.66°, 8.2°, 0.32 Å, and 0.44 Å for the six entries in Table 2). This shows that these two complexes are problematic also for the other methods and deteriorate the averages in Table 2.

The restrained and NB2 force fields give similar and somewhat varying results. NB2 gives a better Hessian and better angles and dihedrals, whereas the restrained model is better for the other quality measures. In fact, the restrained model gives better bonds than both ZAFF and Hess2FF–AT_{min}. The NB1 force field gives the worst results in all quality measures. Thus, we can conclude that the performance of the force fields is quite uniform over the 12 tested complexes and follows the order NL–AT_{max}, NL–AT_{min}, Hess2FF–AT_{max}, Hess2FF–AT_{min}, ZAFF, restrained, NB2, and NB1.

Zn(Cys)₄ Model. Three of the model complexes gave problems that often are encountered when parametrizing metal sites. Initial calculations on the [Zn(CH₃S)₄]²⁻ complex gave unexpectedly poor results with all methods. In particular, the dihedral angles of the methyl groups were poorly reproduced, giving large

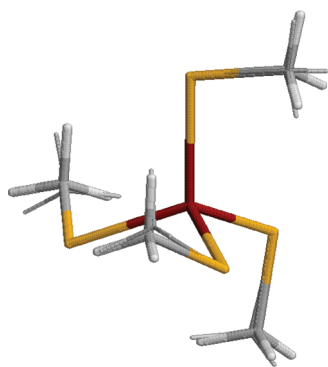


Figure 2. Comparison of QM (thin sticks) and Hess2FF-AT_{max} structures of the ZnCys₄ model.

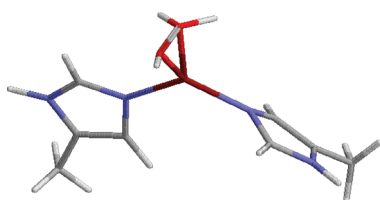


Figure 3. Structure of the Zn(His)₂(H₂O)₂ complex, optimized with the ZAFF force field, the step before the minimization crashes. Note that two of the H and O atoms overlap.

RMSDs of dihedrals and coordinates. The problem was traced to the charges of the hydrogen atoms, which turned out to be negative. In this small model, the general structure (the dihedrals) is completely determined by weak hydrogen bonds between the methyl groups and the sulfur atoms. If the hydrogen atoms have a negative charge, these interactions become repulsive, and the structure will completely change.

The improper charge of the hydrogen atoms is probably caused by the negative net charge (−2) of the complex; negatively charged complexes often give problems in QM calculations.⁷⁴ It can be avoided in many different ways, for example, by changing the method or basis set or by performing the calculations in a continuum solvent. In order to keep the calculations as similar as possible to the original ZAFF calculations, we decided to use a slightly different basis set, LanL2DZ.^{75,76} It gave a positive charge on the hydrogen atoms and a geometry that was similar to that obtained with the 6-31G* basis set.

Results of the parametrizations obtained from the structures optimized with this basis set are shown in Table 3. The results are quite similar to those obtained for the other models. The nonbonded models gave the worst results, but this time the two nonbonded models gave similar results.

The three AT_{min} force fields gave similar results. Moreover, the two AT_{max} force fields gave slightly better results, but the improvement is not so large, because the model is so small (21 atoms) and the number of AMBER parameters is few (two bonds and one angle). The differences in coordinates are quite high for all bonded potentials because they form a somewhat different pattern of the S–HC hydrogen bonds (Figure 2). However, we have checked with QM calculations that the QM pattern is more stable than that obtained with MM.

ZnHis₂(H₂O)₂ and ZnHis(H₂O)₃ Models. For the two [Zn(CH₃-imidazole)₂(H₂O)₂]²⁺ and [Zn(CH₃-imidazole)(H₂O)₃]²⁺ models,

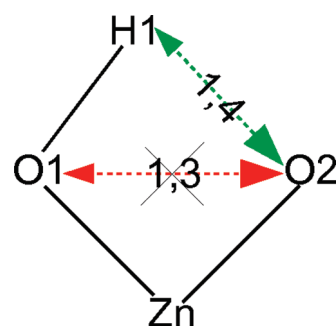


Figure 4. Schematic representation of nonbonded interactions between two Zn-bounded water molecules, illustrating that the O1–O2 1,3-interaction is ignored, whereas the H1–O2 1,4-interaction is included.

the ZAFF minimizations starting from the QM structures crashed because the O and H atoms from different water molecules overlap, as can be seen in Figure 3. This is caused by the Lennard-Jones parameters of the H atoms of water. The default water model in AMBER is TIP3P⁶⁹ and it has zeroed Lennard-Jones parameters on the hydrogen atoms (i.e., it interacts with other molecules only through the O atom, to increase computational speed). These parameters were also used in ZAFF. This is problematic for a metal complex, because there are typically other ligating atoms with a substantial negative charge (in our case, the coordinating O atom of the other water ligand; O2 in Figure 4). This atom is two bonds away from the O atom of the first water ligand (O1), implying that the Lennard-Jones and electrostatic interactions between these two atoms are ignored. On the other hand, the O2 atom and the H atom of the first water (H1) are three bonds away, implying that the electrostatics and Lennard-Jones interactions in AMBER are scaled down by factors of 1.2 and 2.0, respectively. However, if the Lennard-Jones parameters of H1 atoms are zero, it means that there is no repulsive interaction between the O2 and H1 atoms and therefore the electrostatic attraction may bring the two atoms together, until they reside on top on each other, giving an infinite electrostatic energy.

This problem can partly be avoided by adding nonzero Lennard-Jones parameters on the H atoms of water. For the Hess2FF and NL force fields, we use the same Lennard-Jones parameters as backbone HN groups (atom type H; this is the only polar H atom type in AMBER with nonzero Lennard-Jones parameters), $R = 0.6 \text{ \AA}$ and $\epsilon = 0.657 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$. From Table 4, it can be seen that this is enough to avoid problems in the minimizations.

However, it can be seen that the two Hess2FF force fields still give quite large rmsd values for the coordinates (0.9 and 1.6 Å). The reason for this is that the O–O distance becomes only 0.8–0.9 Å, giving strongly distorted structures (Figure 5a). This is caused by a related problem: Even if there are nonzero Lennard-Jones parameters on the H atoms, there is still no repulsion between the O atoms on different water molecules. Therefore, there will be a strong attraction between the H and O atoms on different water molecules, which only will be repelled at short distances, owing to the small radius of the H atom. In fact, there are four H–O interactions (hydrogen bonds) of 1.6 Å in the structure in Figure 5a, but with the wrong orientation (the O atoms are in the middle), owing to the missing O–O repulsion. This shows that the problem cannot be avoided by using a larger Lennard-Jones radius of the water H atom, because it will not

change the orientation (1.6 Å is a reasonable distance for a strong hydrogen bond).

On the other hand, the NL force fields gave excellent structures, especially with AT_{\max} as can be seen in Figure 5b. Apparently, a thorough parametrization of the bonded parameters can overcome the problems caused by water Lennard-Jones parameters, for example, by compensating the missing nonbonded O–O interaction by a large O–Zn–O force constant.

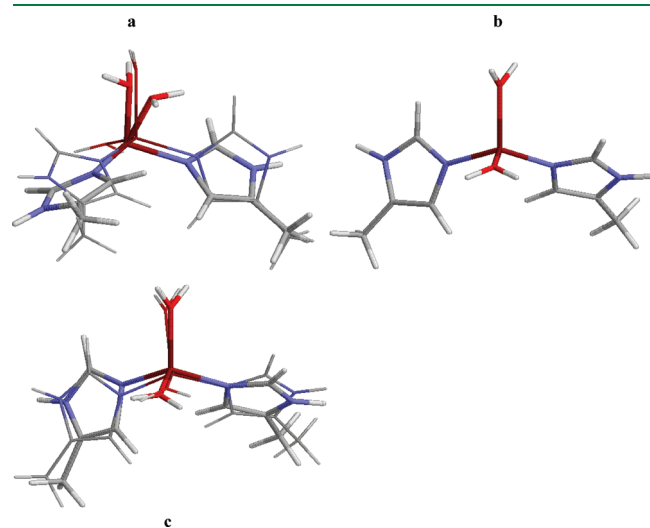


Figure 5. Comparison of QM (thin lines) with (a) Hess2FF- AT_{\min} , (b) NL- AT_{\max} or (c) restrained nonbonded structures of the $Zn(His)_2(H_2O)_2$ models. Panel a shows that the Hess2FF- AT_{\min} structure has a close O–O interaction and, hence, a strongly distorted structure. In panel b, the fit is so perfect (the rmsd of the coordinates is 0.003 Å) that the thin lines cannot be discerned.

It is also notable that the restrained nonbonded force field gave a better structure than Hess2FF, as can be seen in Table 4 and Figure 5c. The reason for this is that, in a nonbonded force field, there are no bonds to the Zn ligands, so the O–O interaction is fully active, ensuring that there is no great distortion of the structure. Moreover, the restraints on the Zn–ligand distances ensure that these distances are well reproduced, better than in the Hess2FF structures. On the other hand, there are no angles involving Zn, so the structure is far from perfect (the rmsd for the coordinates is 0.5 Å). This shows that the restrained nonbonded model is an alternative to a bonded model with more than one water molecule.

Similar results apply also to the $ZnHis(H_2O)_3$ complex (Table 5): The ZAFF minimization crashed with overlapping H and O atoms. Hess2FF- AT_{\min} (with nonzero H parameters for water) managed the minimization but gave a quite poor structure, this time caused by decreased Zn–O bond lengths (~ 1.6 Å, to improve the H–O hydrogen bonds). On the other hand, Hess2FF- AT_{\max} , NL- AT_{\min} , and especially NL- AT_{\max} gave accurate structures. The restrained nonbonded model is better than both nonbonded models.

A natural question is why the problems with the water molecule and the Lennard-Jones parameters are seen only with two or more water molecules and not for the $ZnHis_3(H_2O)$ complex (Table S1, Supporting Information). The reason is most likely that, for the latter complex, van der Waals interactions between the water O atom and the nonligating atoms in the imidazole ring prohibit the H and O atoms of water from coming too close to the N atom. However, our experience with other metal complexes involving a water ligand is that you often encounter problems with unstable MD trajectories if you use zeroed Lennard-Jones parameters for the water H atoms.

Table 4. Performance of the Various Force Fields for the $ZnHis_2(H_2O)_2$ Complex^a

	NB1	NB2	restrained	ZAFF	Hess2FF		NL	
					AT_{\min}	AT_{\max}	AT_{\min}	AT_{\max}
r^2 Hessian	0.924	0.914	0.920	crashed	0.413	0.669	0.971	0.997
rmsd bonds	0.375	0.048	0.014		0.060	0.011	0.011	0.002
rmsd angles	4.6	5.6	2.7		10.8	10.6	1.6	0.1
rmsd dihedrals	8.2	7.1	2.2		49.1	40.6	1.6	0.1
rmsd coordinates	0.95	0.52	0.50		1.58	0.94	0.10	0.00
max coordinates	1.51	0.87	0.81		3.48	2.44	0.21	0.01

^aQuality measures are the same as in Table 1.

Table 5. Performance of the Various Force Fields for the $ZnHis(H_2O)_3$ Complex^a

	NB1	NB2	restrained	ZAFF	Hess2FF		NL	
					AT_{\min}	AT_{\max}	AT_{\min}	AT_{\max}
r^2 Hessian	0.781	0.759	0.750	crashed	0.821	0.967	0.972	0.997
rmsd bonds	0.405	0.080	0.024		0.163	0.015	0.015	0.002
rmsd angles	4.4	5.5	4.5		2.6	2.0	2.5	0.1
rmsd dihedrals	21.2	23.2	24.4		14.8	2.9	3.6	0.1
rmsd coordinates	0.89	0.54	0.31		0.37	0.09	0.11	0.00
max coordinates	1.32	1.33	0.93		0.74	0.19	0.23	0.01

^aQuality measures are the same as in Table 1.

Molecular Dynamics Simulations. The tests in the previous sections were performed only for small model systems. Such tests have the advantage of having a well-defined reference, viz., the QM calculation. Therefore, small differences between the various force fields can be easily discerned. However, it is also of interest to see how the various Zn force fields behave in simulations of full proteins. Therefore, we have performed two sets of MD simulations with all eight force fields on two Zn-containing proteins: the δ' subunit of the clamp-loader complex of DNA polymerase III (1AST),⁶⁷ which contains a single Cys₄ Zn finger, and the spore coat polysaccharide biosynthesis protein SpsE (1VLI),⁶⁸ which contains a ZnHis₂(H₂O)₂ site. These two proteins were selected among those used for the model systems because they contain a single metal site and no other unusual ligands. One of them

contains a Zn site, for which we do not expect any problems in the simulations, whereas the other one is a hard test case. The proteins were simulated for 5 ns (after equilibration) and the structure of the Zn site was examined for 2500 snapshots sampled every 2 ps. The results are collected in Tables 6 and 7.

For the protein with the ZnCys₄ site, all force fields except NB1 gave stable Zn sites throughout the whole simulation. However, for the NB1 force field, the Zn–S bonds were always long (~ 3.2 Å); after 100 ps the ligands started to dissociate, and after 800 ps Zn had dissociated from all four Cys ligands and drifted around in the solvent. Such a dissociation of the metal ion is always a risk with a nonbonded model, especially if the Zn–ligand interactions are too weak, as in the NB1 force field.

On the other hand, the NB2 force field gave too strong Zn–S bonds, with average and maximum distances of 1.98 Å and 2.17 Å, respectively; these are appreciably shorter than the QM distances of 2.53 Å (Table 6). This shows that a nonbonded force field needs to be thoroughly calibrated for the complex of interest to give accurate results. A nonbonded force field with RESP charges and Merz Lennard-Jones parameters gave average Zn–S bond lengths of 2.43 Å (NB3 in Table 6), closer to the QM reference.

The other six force fields gave rather similar results with average Zn–S bonds of 2.40–2.52 Å (Table 6). It is notable that the NL force fields gave longer Zn–S bonds that are closer to the QM bond lengths than the other four force fields. The reason for this is that the other force fields simply use the QM bond length as the equilibrium value, which owing to interactions with the other atoms in the complex leads to too short bonds. On the other hand, the NL method optimizes the structure with MM in every step of the parametrization, so that if the other interactions in the complex tend to shorten the Zn–S bonds, this is compensated by longer equilibrium Zn–S bond lengths

Table 6. Variation of Zn–S Bond Lengths in Simulation of δ' Subunit of Clamp–Loader Complex of DNA Polymerase III^a

force field	avg	std dev	min	max
NB1 ^b				
NB2	1.98	0.04	1.86	2.17
NB3 ^c	2.32	0.13	1.99	3.40
restrained	2.43	0.09	2.11	2.74
ZAFF	2.40	0.09	2.08	2.74
Hess2FF–AT _{min}	2.47	0.09	2.13	2.79
Hess2FF–AT _{max}	2.48	0.09	2.11	2.85
NL–AT _{min}	2.50	0.09	2.15	2.85
NL–AT _{max}	2.52	0.08	2.20	2.85

^a Average, minimum, and maximum distances are given for the four Zn–S bonds, as well as their standard deviation (std dev). The QM distances were all 2.53 Å. ^b Zn dissociates. ^c A nonbonded force field with RESP charges, but Merz Zn Lennard-Jones parameters.

Table 7. Variation of Zn–Ligand Bond Lengths in Simulation of Spore Coat Polysaccharide Biosynthesis Protein SpsE^a

	His1			His2			Wat1			Wat2			O–O		
	avg	min	max	avg	min	max	avg	min	max	avg	min	max	avg	min	max
QM	1.95			1.95			2.05			2.05			3.08		
NB1 ^b															
NB2 ^c	2.11	1.87	2.52	2.20	1.89	2.84	1.90	1.73	2.18	1.92	1.73	2.34			
Simulations with Merz Zn Lennard-Jones Parameters															
restrained	1.98	1.84	2.13	1.97	1.84	2.13	2.03	1.84	2.28	2.03	1.85	2.25	3.85	2.54	4.36
ZAFF ^d															
Hess2FF–AT _{min}	1.92	1.75	2.18	1.94	1.73	2.13	1.73	1.47	1.92	1.72	1.49	1.96	0.83	0.69	1.05
Hess2FF–AT _{max}	1.95	1.74	2.16	1.98	1.76	2.16	2.02	1.9	2.16	2.03	1.89	2.18	0.82	0.66	1.09
NL–AT _{min}	1.92	1.75	2.12	1.94	1.76	2.12	2.02	1.86	2.14	2.02	1.89	2.14	0.97	0.77	1.43
NL–AT _{max}	1.99	1.79	2.22	1.96	1.78	2.17	2.02	1.92	2.13	2.02	1.88	2.14	1.74	1.41	2.12
Simulations with Stote–Karpus Zn Lennard-Jones Parameters															
restrained	2.49	2.40	2.58	2.49	2.42	2.59	2.57	2.43	2.71	2.56	2.45	2.70	4.67	2.93	5.25
ZAFF ^d															
Hess2FF–AT _{min}	1.87	1.84	2.54	1.90	1.84	2.09	1.64	1.84	1.91	1.79	1.85	1.95	0.85	0.71	1.18
Hess2FF–AT _{max}	1.90	1.70	2.08	1.90	1.71	2.10	2.01	1.87	2.13	2.02	1.88	2.14	0.85	0.68	1.10
NL–AT _{min}	1.89	1.76	0.66	1.91	1.90	2.11	2.01	1.89	2.13	2.02	1.77	2.15	1.08	0.79	1.85
NL–AT _{max}	1.94	2.16	0.21	1.91	2.16	2.11	2.02	2.18	2.14	2.00	2.36	2.13	1.74	1.32	2.35

^a Average, minimum, and maximum distances are given for the four Zn–ligand bonds, as well as the O–O distance between the two water ligands. ^b Zn dissociates. ^c The Wat2 ligand for this simulation dissociated and was replaced by three new water ligands. The results in the Wat2 column are the average, minimum, and maximum values for these three ligands. ^d Simulation crashed.

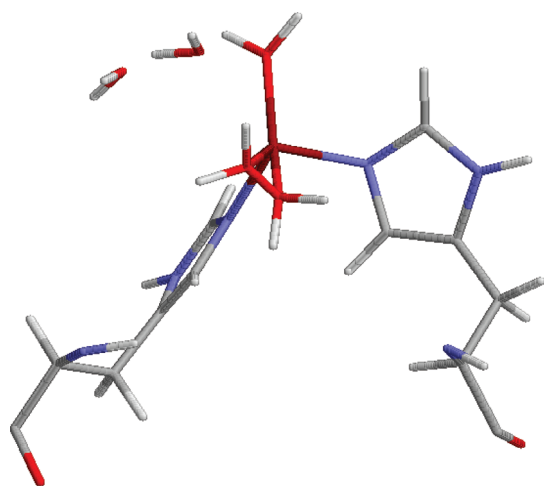


Figure 6. Snapshot of MD simulation of spore coat polysaccharide biosynthesis protein SpsE with NL-AT_{min} force field. The Zn ligands, as well as two second-sphere water molecules, are shown. Note the short O–O interaction between the two original water ligands (0.89 Å).

(e.g., ~ 2.56 Å with NL-AT_{min}). Consequently, the average Zn–S bond length is within 0.01 Å of the QM value for the NL-AT_{max} simulation and 0.03 Å too short with NL-AT_{min}, whereas it is 0.05–0.06 Å too short with Hess2FF, 0.10 Å too short for the restrained model, and 0.13 Å too short with ZAFF. This illustrates one of the advantages of the NL approach.

The Zn site in this protein is on the surface, exposed to the solvent. Therefore, it is not unexpected that water molecules occasionally come rather close to the Zn ion. This happens in a few snapshots for all force fields except NB2, but the water molecule never stays close to the Zn ion for more than one or two snapshots.

Finally, we can compare the Zn–S distances observed in the simulations with those reported in the crystal structure, 2.27–2.34 Å, with an average of 2.31 Å. Thus, the experimental distances are 0.22 Å shorter than the theoretical ones. This discrepancy between theory and experiment for a ZnCys₄ site has been discussed before;⁷⁴ it is caused by missing solvation effects in the vacuum QM calculations for this highly negatively charged complex. Of course, such problems need to be considered in an accurate parametrization of the metal site. In our case, it could be solved by performing the QM optimization and frequency calculations in a continuum solvent with a high dielectric constant. Alternatively, for the restrained nonbonded potential and the bonded potentials based on the Seminario approach, experimental data can easily be included in the force field, as the restraint target or the equilibrium bond lengths or angles. However, for the NL approach this is more problematic, because a target function needs to be defined and it is likely that if the QM bond lengths are changed, the Hessian elements and the angles will also change. Therefore, for NL it is better to use a QM method that gives accurate results.

The results of the simulation of the ZnHis₂(H₂O)₂ site in the spore coat polysaccharide biosynthesis protein spsE are described in Table 7. As expected, we encountered severe problems with the stability of the simulations with ZAFF, owing to overlapping H and O atoms from the water ligand (like the structure in Figure 3), and the resulting structures were distorted. However, the other bonded force fields also showed stability

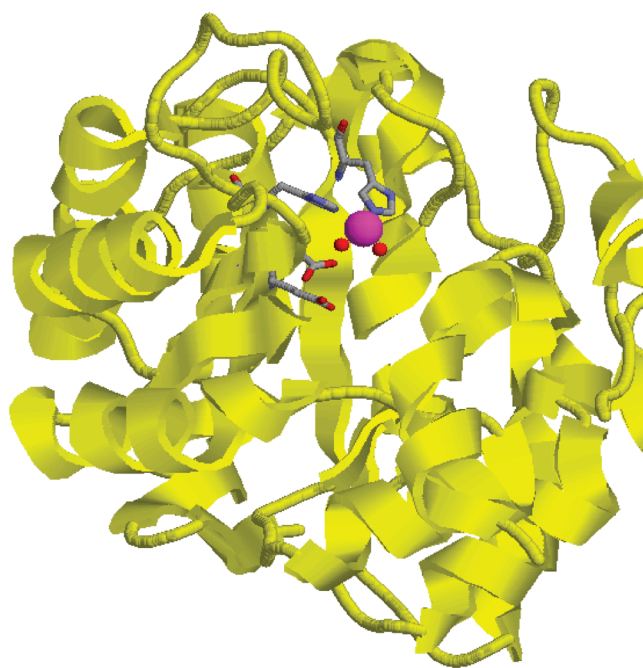


Figure 7. Crystal structure of spore coat polysaccharide biosynthesis protein SpsE with Zn ion (magenta), two water ligands (red balls), two His ligands (sticks), and Glu-22 and Glu-234 (sticks) emphasized.

problems, but these could be solved by running some (or in a few cases all) of the simulations without any bond-length constraints and a short time step (0.5 fs).

As for the other protein, the Zn ion dissociates from the original ligands with the NB1 force field, in this case already during the equilibration. It diffuses around in water solution and does not bind to the same ligand for more than 400 ps.

However, with the NB2 force field, the Zn ion remains in the original site but one of the water ligands dissociates during the equilibration. It is replaced by three additional water ligands, giving a six-coordinate Zn ion throughout the production simulation. The Zn–N bond lengths are 0.16–0.25 Å longer than in the QM calculations, whereas the Zn–O bonds are 0.13–0.15 Å too short. This shows that it is hard with a nonbonded model to get all metal–ligand bond lengths correct if there are several types of ligands, because only two Lennard-Jones parameters are available (unless you change the Lennard-Jones parameters also of the ligands, which will influence their interactions with the surrounding protein).

All five bonded potentials retained the bonds between Zn ion and the original ligands. However, in all simulations, an additional water molecule also bound to the Zn ion, although with the NL-AT_{min} force field it was frequently exchanged during the simulation. Sometimes, the coordination number increased even to six. All average Zn–ligand bond lengths are within 0.04 Å of the QM bond lengths, except the two Zn–O distances for Hess2FF-AT_{min} (0.33 Å too short). This may give the impression that the simulations are successful, but this is not the case, as the statistics of the O–O distance between the two water ligands show (also included in Table 7). In the QM structure it is 3.08 Å, but in the simulations this distance is never longer than 2.18 Å, and it is 0.8 Å on average in the Hess2FF simulations. This shows that the structures are strongly affected by the problem of the missing 1,3 O–O repulsion, discussed above, and the structures

are actually totally unrealistic, as is shown by the typical snapshot in Figure 6.

In fact, the only simulation that gives realistic results is the restrained nonbonded potential, for which the Zn–ligand distances are within 0.03 Å of the QM distances. However, also in this simulation, an additional water ligand tends to bind to the Zn site, exchanging several times during the 5 ns simulation (one water ligand stays for 2 ns; the others stay for less than 0.6 ns). On the other hand, such an increase in the coordination number of the Zn ion is not totally unrealistic considering the crystal structure. As can be seen in Figure 7, the four observed ligands are all nearly on the same hemisphere of the Zn ion, with a seemingly empty coordination site on the opposite side, with over 6 Å to the closest protein atom. Thus, it is possible that the simulation with the restrained nonbonded potential actually gives a good picture of the metal coordination.

The coordination number can be determined by the force field. The Zn ion is almost hidden by the ligands. Therefore, the Zn Lennard-Jones parameters for a bonded potential can be used to avoid undesired ligand. With the Zn charge used for this site (1.33 e), the Merz Zn Lennard-Jones parameters gives a minimum for the interaction with a water molecule of ~ 1.95 Å. On the other hand, the Stote–Karplus parameters give an ~ 0.9 Å longer interaction. Therefore, we rerun the bonded and restrained nonbonded simulations also with the latter parameters. For the bonded potentials, the results did not change significantly, except that the extra water ligand disappeared (i.e., the structures were still unrealistic), although there were still occasional water molecules approaching the Zn site, with distances down to 2.62 Å. However, for the restrained nonbonded potential, reasonable four-coordinate structures were obtained. Unfortunately, the Zn Lennard-Jones parameters also affect the distances of the four original ligands so that they became ~ 0.5 Å too long. The reason for this is, of course, that the restraints are harmonic (a r^2 term) whereas the repulsive Lennard-Jones term is a much steeper r^{-12} term. Of course, we can tune these interactions by changing the restraint or by using specific Lennard-Jones terms for the Zn–solvent interactions, but we did not pursue such calculations any further.

This shows that the MD simulations will give the results dictated by the parameters used. If we want to know whether the Zn site is four- or five-coordinate in this protein, we could optimize the structure of $\text{ZnHis}_2(\text{H}_2\text{O})_3$ with QM methods and then tune the Zn Lennard-Jones parameters so that we get the correct Zn–O bond lengths with the Zn charge used in the simulations (note that the minimum distance depends on the Zn charge, and therefore a new calibration has to be done for each new metal site when RESP charges are used). If the simulations then show an increased coordination number, the calculations indicate that that the site should be five-coordinate.

On the basis of the available results, the Merz parameters give Zn–O bond lengths closer to the QM results [for the $\text{ZnHis}_2(\text{H}_2\text{O})_2$ model] than the Stote–Karplus parameters. Therefore, it is likely that the site actually should be five-coordinate, although additional calculations are needed to settle this issue. The reason why only two of the three water molecules are observed in the crystal structure (at 2.38 Å resolution) is probably that the two observed water molecules are stabilized by two Glu residues (Glu-22 and Glu-224, cf. Figure 7), whereas the third water molecule does not form any strong interactions with the surrounding protein.

CONCLUSIONS

Recently, two softwares have been presented to obtain MM force-field parameters for metal sites.^{49,50} These may give the false impression that any metal site may be rapidly and accurately parametrized in an automatic manner. The intention of this paper is to give a more nuanced picture of the matter and to show that the parametrization of metal sites involves several pitfalls.

First, there are several levels of approximation for the parametrization of metal sites. The lowest is a nonbonded model with standard charges for the ligands and formal charges for the metal. It can be used without any parametrization for any metal site but is unlikely to give accurate results for a general metal site, as has been seen in Tables 1–7. In particular, there is a great risk for an unwanted exchange of ligands.^{49,77} Therefore, such a model is not recommended.

At the next level, the QM calculations can be used to obtain a full bonded model for the metal site. Such calculations can be made automatic with the approach of Seminario,⁴⁷ as has been suggested several times.^{48–50} Once the QM calculations are done, the parametrization takes only seconds. However, it must be remembered that this approach is only approximate: It assumes that the bond lengths, angles, and dihedrals observed in the QM structure are optimal, although they actually are a compromise of strain and nonbonded interactions caused by the other ligands, and it involves a double counting of nonbonded interactions, in particular for dihedrals.

These problems are avoided by the ideal iterative method by Norrby and Liljefors.⁵¹ In this approach, a penalty function is set up describing the goal of the force field and the importance of the various terms, for example, the reproduction of QM bond lengths, angles, dihedrals, and Hessian elements. Then the parameters are optimized by full minimization at each step, numerical derivatives, and nonlinear optimization algorithms. This is a much more involved method, taking hours to several days for complicated systems. Moreover, the optimization typically have to be run several times before all parameters are acceptable, and significant human intervention and judging are needed. On the other hand, essentially a perfect fit to the QM data can be obtained (cf. Figures 1d and 5b), which is also reflected in the Zn–ligand distances during the simulations. As a result, reliable energies can be extracted from such simulations of metal complexes.^{52,54}

Furthermore, the accuracy of a parametrization is affected by the choice of atom types. For a nonbonded model, it is natural to use standard atom types for all atoms. However, for a bonded model, the choice of atom types is more crucial. A reasonable choice for the Seminario approach is standard atom types for all ligand atoms, except those binding to the metal, for which new atom types are employed, different for all ligands (but using standard parameters within the ligand). For parametrizations with the Norrby–Liljefors approach, a more thorough consideration of atom types is necessary. This method is so accurate that the accuracy of the final model is typically limited by the force field of the ligands. Therefore, it is preferable to use new atom types also for all (not symmetry-equivalent) atoms of the ligands.

This study has also illustrated several possible pitfalls during the parametrization of metal sites. First, the ZnCys_4 complex showed that for negatively charged complexes, there is a risk that improper charges are obtained. In particular, it should be checked that ligand hydrogen atoms get a proper (positive) charge. For this model, QM calculations in vacuum also give poor results compared to experiments.

Second, we have showed that water molecules pose a serious problem for bonded models in force fields (like AMBER) that ignore nonbonded interactions for atoms separated by two bonds. The reason for this is that the repulsion between the ligating atoms is ignored. If these bind hydrogen atoms (as for water), these H atoms may form hydrogen bonds to the other ligating atoms with a wrong orientation, as is shown in Figures Sa and 6, leading to distorted structures.

With a single water molecule, it is possible to obtain reasonable structures, once it is ensured that the water H atoms have nonzero Lennard-Jones parameters. However, with several water molecules even the NL method gives strange structures in MD simulations. The only method that works in such a case is the restrained nonbonded model. It combines the best aspects of the nonbonded and bonded models and avoids the problem with the 1,3-interactions by not explicitly define any bonds. Clearly, it is only an approximate method (cf. Figure 5c), but it can be improved by adding more restraints, for example, for angles around the metal, at the end giving a model that is similar to the Seminario approach but avoiding the problem of ignored 1,3-interactions. The restrained nonbonded model is similar to the points-on-a-sphere (POS) model, frequently used for coordination compounds.^{17,38,39} Both methods include metal–ligand bonds, exclude ligand–metal–ligand angles, and include ligand–ligand nonbonded interactions. However, the POS model normally retains all the other bonded terms involving the metal, whereas the restrained model excludes them. Moreover, the POS model typically ignores many nonbonded interactions involving the metal, whereas the restrained model includes them all.

In conclusion, we have seen that the parametrization of metal sites is not an automatic approach and, as usual in computational methods, there is a trade-off between accuracy and speed. We would suggest the following general strategy: If the metal site is not of central interest (e.g., a structural metal far from a catalytic or ligand-binding site), a bonded parametrization based on the Seminario method is recommended. If the metal site involves more than one water ligand, a restrained nonbonded model must be used. However, if the metal site is of central interest, either in structural or energy terms, a more thorough parametrization is needed, based on the Norrby–Liljefors approach. Finally, we want to emphasize the need of testing the parametrization before use, that is, how well it reproduces the QM calculations used for the parametrization in terms of structures and Hessian elements.

■ ASSOCIATED CONTENT

Supporting Information. Ten tables showing performance of various force fields, charges used in MD simulations, and coordinates of QM-optimized structures. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Ulf.Ryde@teokem.lu.se. Tel: +46-46 2224502. Fax: +46-46 2228648.

■ ACKNOWLEDGMENT

This investigation has been supported by grants from the Swedish Research Council (Project 2010-5025) and the

Wenner-Gren Foundation. It has also been supported by computer resources of Lunarc at Lund University.

■ REFERENCES

- (1) Boeyens, J. C. A.; Comba, P. *Coord. Chem. Rev.* **2001**, *212*, 3–10.
- (2) Kleywegt, G. J.; Jones, T. A. *Methods Enzymol.* **1997**, *227*, 208–230.
- (3) Gohlke, H.; Klebe, G. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644–2676.
- (4) Gubbins, K. E.; Moore, J. D. *Ind. Eng. Chem. Res.* **2010**, *49*, 3026–3046.
- (5) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.
- (6) MacKerell, A. D. *J. Comput. Chem.* **2004**, *25*, 1584–1604.
- (7) Lifson, S.; Warshel, A. *J. Chem. Phys.* **1968**, *49*, 5116–5129.
- (8) Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127–8134.
- (9) Fabricius, J.; Engelsen, S. B.; Rasmussen, K. *J. Carbohydr. Chem.* **1997**, *16*, 751–772.
- (10) Pérez, S.; Imbert, A.; Engelsen, S. B.; Gruza, J.; Mazeau, K.; Jimenez-Barbero, J.; Poveda, A.; Espinosa, J.-F.; van Eyck, B. P.; Johnson, G.; French, A. D.; Kouwijzer, M. L. C. E.; Grootenuis, P. D. J.; Bernardi, A.; Raimondi, L.; Senderowitz, H.; Durier, V.; Vergoten, G.; Rasmussen, K. *Carbohydr. Res.* **1998**, *314*, 141–151.
- (11) Allinger, N. L.; Yuh, Y. H.; Lii, J.-H. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.
- (12) Maple, J. R.; Hwang, M.-J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. *J. Comput. Chem.* **1994**, *15*, 162–182.
- (13) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490–519.
- (14) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (15) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008**, *29*, 622–655.
- (16) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D. *J. Comput. Chem.* **2010**, *31*, 671–690.
- (17) Hay, B. P. *Coord. Chem. Rev.* **1993**, *126*, 177–236.
- (18) Comba, P.; Zimmer, M. *J. Chem. Educ.* **1996**, *73*, 108–110.
- (19) Norrby, P.-O.; Brandt, P. *Coord. Chem. Rev.* **2001**, *212*, 79–109.
- (20) Hambley, T. W.; Jones, A. R. *Coord. Chem. Rev.* **2001**, *212*, 35–59.
- (21) Marques, H. M.; Brown, K. L. *Coord. Chem. Rev.* **2002**, *225*, 123–158.
- (22) Zimmer, M. *Coord. Chem. Rev.* **2009**, *253*, 817–826.
- (23) Deeth, R. J.; Anastasi, A.; Diedrich, C.; Randell, K. *Coord. Chem. Rev.* **2009**, *253*, 795–816.
- (24) Hoops, S. C.; Anderson, K. W.; Merz, K. M. *J. Am. Chem. Soc.* **1991**, *113*, 8262.
- (25) Ryde, U. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 40–56.
- (26) Gresh, N.; Piquemal, J. P.; Krauss, M. *J. Comput. Chem.* **2005**, *26*, 1113.
- (27) Nilsson Lill, S. O.; Forbes, A.; Donoghue, P.; Verdolino, V.; Wiest, O.; Rydberg, P.; Norrby, P.-O. *Curr. Org. Chem.* **2010**, *14*, 1629–1645.
- (28) Root, D. M.; Landis, C. R.; Cleveland, T. *J. Am. Chem. Soc.* **1993**, *115*, 4201–4209.
- (29) Sabolovic, J.; Gomzi, V. *J. Chem. Theory Comput.* **2009**, *5*, 1940–1945.
- (30) Stote, R. H.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 12–31.
- (31) Sakharov, D. V.; Lim, C. *J. Am. Chem. Soc.* **2005**, *127*, 4921.
- (32) Åqvist, J.; Warshel, A. *J. Mol. Biol.* **1992**, *224*, 7.
- (33) Pang, Y. P. *J. Mol. Model.* **1999**, *5*, 196–202.
- (34) Pang, Y. P.; Xu, K.; Yazal, J. E.; Prendergas, F. G. *Protein Sci.* **2000**, *9*, 1857.
- (35) Singh, N.; Warshel, A. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1705–1723.
- (36) Hancock, R. D. *Prog. Inorg. Chem.* **1989**, *37*, 187.
- (37) Hancock, R. D. *Acc. Chem. Res.* **1990**, *23*, 253.

- (38) Kepert, D. L. *Inorg. Chem.* **1972**, *11*, 1561–1567.
- (39) Hambley, T. W.; Hawkins, C. J.; Palmer, J. A.; Snow, M. R. *Aust. J. Chem.* **1981**, *34*, 45–53.
- (40) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.
- (41) Sirovatka, J. M.; Rappe, A. K.; Finke, R. G. *Inorg. Chim. Acta* **2000**, *300*, 545.
- (42) Vedani, A.; Huhta, D. W. *J. Am. Chem. Soc.* **1990**, *112*, 4759.
- (43) Vedani, A.; Huhta, D. W.; Jacober, S. P. *J. Am. Chem. Soc.* **1989**, *111*, 4075.
- (44) Gresh, N. *Curr. Pharm. Des.* **2006**, *12*, 2121.
- (45) Gresh, N.; Piquemal, J.-P.; Krauss, M. *J. Comput. Chem.* **2005**, *26*, 1113–1130.
- (46) Piquemal, J.-P.; Chevreaux, H.; Gresh, N. *J. Chem. Theory Comput.* **2007**, *3*, 824–837.
- (47) Seminario, J. M. *Int. J. Quantum Chem.* **1996**, *60*, 1271.
- (48) Nilsson, K.; Lecerof, D.; Sigfridsson, E.; Ryde, U. *Acta Crystallogr. D* **2003**, *59*, 274–289.
- (49) Lin, F.; Wang, R. *J. Chem. Theory Comput.* **2010**, *6*, 1852–1870.
- (50) Peters, M. B.; Yang, Y.; Wang, B.; Fürstli-Molnár, L.; Weaver, M. N.; Merz, K. M. *J. Chem. Theory Comput.* **2010**, *6*, 2935–2947.
- (51) Norrby, P.-O.; Liljefors, T. *J. Comput. Chem.* **1998**, *19*, 1146–1166.
- (52) Norrby, P.-O.; Brandt, P. *Coord. Chem. Rev.* **2001**, *212*, 79.
- (53) Allinger, N. L.; Yuh, Y. H.; Li, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551.
- (54) Rydberg, P.; Olsen, L.; Norrby, P.-O.; Ryde, U. *J. Chem. Theory Comput.* **2007**, *3*, 1765–1773.
- (55) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. AMBER 10, University of California, San Francisco, 2008.
- (56) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (57) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (58) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (59) Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939.
- (60) Gordon, M. S.; Binkley, J. S.; Pople, J. A.; Pietro, W. J.; Hehre, W. J. *J. Am. Chem. Soc.* **1982**, *104*, 2797.
- (61) Dobbs, K. D.; Hehre, W. J. *J. Comput. Chem.* **1987**, *8*, 861.
- (62) Besler, B. H.; Merz, K. M.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431–439.
- (63) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision A.02, Gaussian, Inc., Wallingford CT, 2009.
- (64) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.
- (65) 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.
- (66) Norrby, P.-O. *J. Mol. Struct. (THEOCHEM)* **2000**, *506*, 9–16.
- (67) Guenther, B.; Onrust, R.; Sali, A.; O'Donnell, M.; Kuriyan, J. *Cell* **1997**, *91*, 335–345.
- (68) Joint Center for Structural Genomics; to be published; PDB structure 1VLI.
- (69) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (70) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (71) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8592.
- (72) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (73) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (74) Ryde, U. *Eur. Biophys. J.* **1996**, *24*, 213–221.
- (75) Dunning, T. H.; Hay, P. J. *Modern Theoretical Chemistry*; Schaefer, H. F., Ed.; Plenum Press: New York, 1976; Vol. 3., p 1.
- (76) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270.
- (77) Donini, O. A.; Kollman, P. A. *J. Med. Chem.* **2000**, *43*, 4180–4188.

Block Covariance Overlap Method and Convergence in Molecular Dynamics Simulation

Tod D. Romo and Alan Grossfield*

Department of Biochemistry and Biophysics, University of Rochester Medical Center, Rochester, New York 14642, United States

 Supporting Information

ABSTRACT: Molecular dynamics (MD) is a powerful tool for understanding the fluctuations of biomolecular systems. It is, however, subject to statistical errors in its sampling of the underlying distribution of states. One must understand these errors in order to draw meaningful conclusions from the simulation. This is becoming ever more critical as MD simulations of even larger systems are attempted. We present here a new method for determining the extent of convergence that relies on measures of the fluctuation space sampled by the simulation without any a priori knowledge of states or partitioning of the configuration space. This method reveals long correlation times, even for simple systems, and suggests caution when interpreting macromolecular simulations. We also compare this method with previous efforts to characterize the sampling of MD simulation.

1. INTRODUCTION

It has long been known that molecular dynamics (MD) is subject to statistical errors introduced by both the temporal sampling and the scale or length of the simulation. These statistical errors can lead to poor or erroneous estimations of the distribution of states and hence reduce the accuracy of values calculated from the simulation. Many biologically relevant questions that MD attempts to answer require understanding the statistical uncertainty in measurements derived from the simulation. Ideally, the best way to estimate the error is to run multiple independent simulations. However, this is not always practical, so we often wish to determine the error from a single simulation. The size of the statistical uncertainty (error bars) for these measurements is intimately related to the number of independent samples of the quantity in question that are present in the simulation. This number is implicitly dependent on the correlation time for the observable. Estimating the correlation time therefore leads to an estimate for the uncertainty in the measurement.

In the simplest case, determining the precision of measurements from the simulation focuses on a scalar quantity or an observable. In this case, block averaging¹ is the gold standard for determining the statistical error; it works by dividing the trajectory into blocks and computing the standard error of the observable for each block. As the block size is increased, the error estimated approaches the true error. However, methods that rely on a single observable can be misleading due to coupling between fast relaxation and other, slower processes.²

In order to safely estimate the standard error, one ought to focus on the slowest relevant relaxations known in the system. As such, a number of groups have attempted to develop measures of global sampling quality. A number of these are based on principal component analysis (PCA). Balsara et al.³ examined the overlap between fluctuation directions from different sampling windows and found that the dominant modes changed, suggesting a lack of convergence. Amadei et al.,⁴ in contrast, used the root-mean-squared inner product for the first 10 directions between

two halves of a 2 ns long simulation for protein L and cytochrome c551. This measure was applied to pairs of subtrajectories of increasing size as well as consecutive 50 ps windows and suggested convergence occurring within the subnanosecond regime, when compared to short (nanosecond-scale) simulations. Hess⁵ introduced a more detailed measure called the covariance overlap that considers both the directions of fluctuations and their relative magnitudes. He also presented the cosine content measure that compares the projection of the trajectory along a principal component to a cosine. The projection for a diffusive system will be more cosine-like than a system that has sampled multiple conformations. Using the covariance overlap between a trajectory of a protein and all subintervals as well as the cosine content, Hess found that while there was a suggestion of convergence on the order of 10 ns for the system used, a longer simulation was needed to accurately estimate the longest correlation times. Faraldo-Gómez et al.⁶ used the covariance overlap to compare consecutive nanosecond blocks as well as blocks of different sizes (1, 2, 4, 8, and 16 ns) as well as the cosine content and found that some regions of a protein (such as the membrane-embedded domains of smaller proteins) can be well-sampled on the order of tens of nanoseconds. They also found that overall undersampling leads to imprecise B-factor predictions. Grossfield et al.⁷ used the covariance overlap to characterize the similarity of fluctuations spaces between multiple 100 ns simulations of rhodopsin and found indications of a lack of convergence even with longer simulations. Sullivan et al.⁸ introduced a measure of the configuration space sampled by the simulation and examined both the dimensionality of this space over time as well as the dependence of the phase space volume with increasingly large windows sampled from the trajectory. Using this method, transitions between conformational substates were found to occur on the nanosecond time scale, implying that longer time scales are required for adequate statistical sampling.

Received: April 20, 2011

Published: June 22, 2011

There are alternative approaches that do not directly consider the fluctuation directions. Smith et al.⁹ compared a number of observables over time, such as intramolecular activation energy, root-mean-square distance, numbers of clusters, and numbers of hydrogen bonds. Lyman et al.^{10,11} used a random partitioning of the conformation space of the system along with an analysis of the variance associated with this binning to determine an “effective sample size” that is related to how well it has sampled the distributions of states. This method was also applied to multiple independent simulations of rhodopsin.⁷

The method we have developed is in essence an extension of one used previously to analyze the convergence of membrane proteins⁷ and is similar in spirit to the blocked covariance overlap used by Hess.⁵ The underlying concept is that if a system is well sampled by MD, then the fluctuation space for sufficiently large subsets of the trajectory should be very similar. The blocked covariance overlap method combines the best aspects of two existing methods: block averaging is the best-of-breed method for assessing statistical error in a single variable, and covariance overlap is a powerful tool for assessing similarity of global fluctuations as a single scalar value.

Here, we present the results of applying this method to long, state of the art simulations of three different classes of biomolecules, ranging in size from a dipeptide to an integral membrane protein. In aggregate, we analyze nearly 29 μ s of simulation time from all-atom MD. We show that this method indicates the quality of sampling in a simulation as well as gives an estimate of the rate of convergence, within certain limits. We also compare these results to previous efforts to characterize sampling quality, in particular, the decorrelation time¹⁰ and the effective sample size¹¹ methods developed by Zuckerman and co-workers. Finally, we compare these results to the cosine content measure.⁵

2. METHODS

2.1. Block Average Root-Mean-Square Distance. We used block averaging¹ to assess convergence of the average structure. Briefly, the entire trajectory is first aligned to an optimal average structure using an iterative scheme.⁷ We then divide the trajectory into contiguous blocks. We compute the average structure for each block, and the root-mean-square deviation (rmsd) between each average structure is calculated. The standard deviation for the rmsd at each block size is then plotted. The plateau, if present, indicates both the error in and the correlation time for the average structure.

2.2. Principal Component Analysis. The Cartesian coordinates for a structure at a given time point can be thought of as a $3N$ dimensional column vector, where N is the number of atoms, i.e., $[x_1, y_1, z_1, \dots, x_N, y_N, z_N]^T$. The trajectory (or ensemble of structures) can then be represented by concatenating these column vectors together forming a $3N \times L$ conformation matrix A , where L is the number of snapshots in the trajectory. To compute the principal component analysis (PCA) of A , the average structure must be removed (i.e., the row-average of A is subtracted from A). The principal components are then calculated by finding the eigendecomposition of the covariance matrix AA^T , i.e., $U \Lambda U^T = AA^T$. The eigenvectors (columns of U) give the direction of fluctuations, in a least-squares sense, and the eigenvalues (diagonal elements of Λ) give the magnitude of the corresponding fluctuation. In the case of a protein, or other complete molecule, it is necessary to first remove global rotations and translations. We do this by computing the average structure

from the trajectory (described above) and aligning each frame to it.

2.3. Covariance Overlap. For the proposed method to work, it is necessary to devise a mechanism to quantitatively compare two PCA results. This is typically done by comparing the subspaces (i.e., fluctuation directions for a set number of most significant modes) determined by the PCA. Here, we use the covariance overlap,^{5–7} which measures not only the similarity in the directions of motion (eigenvectors) but also their relative importance (eigenvalues). This measure ranges from 0, where the fluctuations are completely dissimilar, to 1, where the fluctuations are identical. The covariance overlap between two PCA results is defined as

$$\Omega_{A,B} = 1 - \left[\frac{\sum_i^{N_{\text{modes}}} (\lambda_i^A + \lambda_i^B) - 2 \sum_i^{N_{\text{modes}}} \sum_j^{N_{\text{modes}}} \sqrt{\lambda_i^A \lambda_j^B} (\vec{v}_i^A \cdot \vec{v}_j^B)^2}{\sum_i^{N_{\text{modes}}} (\lambda_i^A + \lambda_i^B)} \right]^{1/2} \quad (1)$$

where λ_i^A is the i^{th} eigenvalue from the PCA for ensemble A and \vec{v}_i^B is the i^{th} eigenvector from ensemble B .

2.4. Bootstrapping. Bootstrapping is a computational procedure used to determine the statistical error in a measurement.¹² The fundamental idea behind bootstrapping is that randomly drawing subpopulations from a distribution provides many different estimates for a statistical quantity. These estimates can then be used to estimate the error in the measure. It is particularly useful when the full distribution is unknown or complex but one has a statistical sampling of it.

In MD simulations, each structure is correlated to its neighbors in time. Bootstrapping can be used to remove this correlation from an observable. Structures are randomly drawn from the trajectory to create an ensemble and the observable calculated. This is repeated many times with the standard deviation in the distribution of averages computed becoming an estimate of the true uncertainty.

2.5. Block Covariance Overlap Method. The first step in the block covariance overlap method (BCOM) is to align the entire trajectory with an iterative alignment procedure⁷ using specific atoms as reference points (in this work, nonhydrogen atoms for the small molecules and the transmembrane C α atoms for the larger systems). Next, the conformation matrix is constructed from these atoms, and a PCA computed as described in Section 2.2. In direct analogy to block averaging, the BCOM is typically computed for a range of block sizes, up to half the size of the trajectory. Given a trajectory with L frames and a block size k , the trajectory is divided into L/k contiguous blocks. A PCA is then computed for each block. Within an individual block, there is no additional alignment performed since the trajectory as a whole is already in an optimal alignment. However, the average structure from the block is used in the subtraction for the PCA. The covariance overlap (eq 1) is then computed against the PCA for the entire trajectory. In essence, the full trajectory is treated as the gold standard. The average covariance overlap is then reported as a function of block size.

The block covariance overlap is then normalized by the value expected if the trajectory was totally uncorrelated. This value is determined by bootstrapping¹³ the blocks, i.e., each block is created by randomly drawing (with replacement) samples from the entire trajectory, and the covariance overlap is computed between the PCA of the block and the overall trajectory. This

Table 1. Systems Used for Analyzing Convergence

system	length (ns)	conditions
dileucine	1000	500 K, implicit solvent
LfB6 #1	3164	323 K, explicit solvent
LfB6 #2	3127	323 K, explicit solvent
LfB6 #3	4275	323 K, explicit solvent
LfB6 #4	4275	323 K, explicit solvent
LfB6 #5	4285	323 K, explicit solvent
LfB6 #6	4285	323 K, explicit solvent
β_2 AR	1023	310 K, explicit solvent and membrane
rhodopsin	1605	310 K, explicit solvent and membrane
CB2	1882	310 K, explicit solvent and membrane

procedure is repeated a set number of times (typically 50) for each block size, and the average covariance overlap is used to scale the block overlap data.

The inverse of the bootstrap-normalized blocked covariance overlap is a decaying curve that can generally be fit by a 3-exponential function, $f(t) = k_1 e^{-t/t_1} + k_2 e^{-t/t_2} + k_3 e^{-t/t_3} + 1$, where t_1 – t_3 are different correlation time constants. The 3-exponential function was determined to give the best fit to the observed data by fitting 1–4 exponential functions and examining the residual errors.

2.6. Cosine Content. The cosine content for the first mode was calculated based on the method presented by Hess.⁵ In brief, the trajectory is aligned and divided into multiple contiguous blocks of a given size, as described above. The eigendecomposition is computed as previously described. The projection of the conformation matrix along the first eigenvector is then used to calculate the cosine content according to

$$c = \frac{2}{T} \left(\int_0^T \cos(\pi t/T) p(t) dt \right)^2 \left(\int_0^T p^2(t) dt \right)^{-1} \quad (2)$$

where $p(t)$ is the t^{th} element of the projection vector and the average value over all contiguous blocks is reported.

2.7. Lightweight Object Oriented Structure Analysis Package. All analyses were performed using the Lightweight Object Oriented Structure analysis package (LOOS),^{14,15} an object-oriented library for creating new analytical tools for MD that is implemented in C++. LOOS uses Boost¹⁶ and atlas^{17,18} for additional functionality and high-performance linear algebra calculations. LOOS includes a powerful “selection expression” parser that enables tools to easily select which atoms to operate on. In addition, LOOS provides support for reading the native file formats for most major MD packages, including CHARMM,¹⁹ NAMD,²⁰ Amber,²¹ Gromacs,²² and Tinker.²³ Also included with LOOS are over 50 analytical tools, including suites of programs for computing elastic network model solutions and the convergence analyses presented in this work. LOOS is freely available from SourceForge (<http://loos.sourceforge.net>).

2.8. Structural Decorrelation Time and Effective Sample Size. For comparison purposes, we also applied two previously published methods developed by Zuckerman and co-workers for assessing the convergence of MD simulations, which we believe represent the state of the art in the field. The first method uses the “decorrelation time” as described by Lyman et al.¹⁰ Briefly, this method partitions the conformational space of the trajectory using a set of randomly drawn “reference” structures

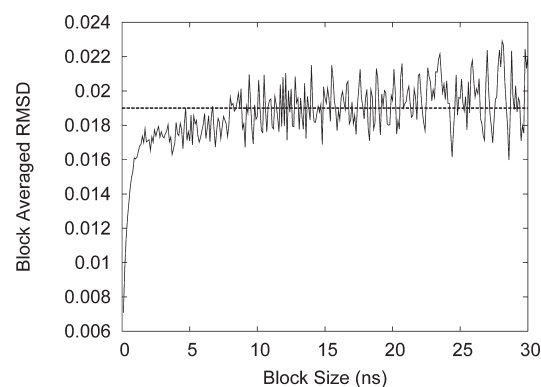


Figure 1. Block averaging of the rmsd between each block and the optimal global average for the 1 μ s dileucine trajectory.

and compares the variance of the histograms of subsets of the ensemble (using the reference structures) with the expected variance if the structures were uncorrelated. The second method, described by Zhang et al.,¹¹ uses the same tessellation of configuration space but then clusters the reference “bins” based on the rates of exchange between the different clusters. This gives an estimate for the number of uncorrelated conformations found in the ensemble from which a decorrelation time can be estimated. Implementations of both tools are available from the Zuckerman lab Web site (<http://www.cccb.pitt.edu/Faculty/zuckerman/software.html>) and are also available in the convergence suite in LOOS.¹⁴

2.9. Model Systems. Several different systems were used to assess convergence, ranging from a dipeptide to a set of G protein-coupled receptors, and are listed in Table 1. The first system was a “toy” system based on Lyman et al.,¹⁰ consisting of a dileucine in implicit water at 500 K simulated in Tinker²³ for 1 μ s, with structures saved every 100 ps. The second system is a short hexapeptide derived from lactoferrin B in explicit water;^{24,25} NAMD²⁰ was used on a BlueGene/P²⁶ to simulate this peptide in the NVT ensemble at 323 K using CHARMM22^{19,27,28} with CMAP²⁸ parameters. Snapshots were saved every 1 ns. We ran multiple simulations, with 2 lasting approximately 3.1 μ s each and 4 other simulations that are each approximately 4.3 μ s long. For both the LfB6 and the dileucine systems, all heavy atoms were used in performing alignments and for computing the PCA.

Three different G protein-coupled receptors (GPCRs) were used as the larger test systems. The details of their construction have been previously described, but in brief, each consisted of the GPCR embedded in a lipid bilayer along with explicit solvent. The first system is a 1.02 μ s all-atom simulation of β_2 AR,²⁹ the second is a 1.6 μ s simulation of dark-state rhodopsin,³⁰ and the final system is an approximately 1.9 μ s simulation of the CB2 cannabinoid receptor.³¹ For these larger systems, only transmembrane α carbons were considered for analysis.

3. RESULTS AND DISCUSSION

3.1. Convergence of the Average Structure. As a trajectory evolves over time, the estimate for the average structure will change. This can be the result of new conformational substates being found or from changes in the relative population of the existing substates. This variation is particularly important for methods such as PCA that depend upon the average structure as a reference point. This dependence can be easily demonstrated

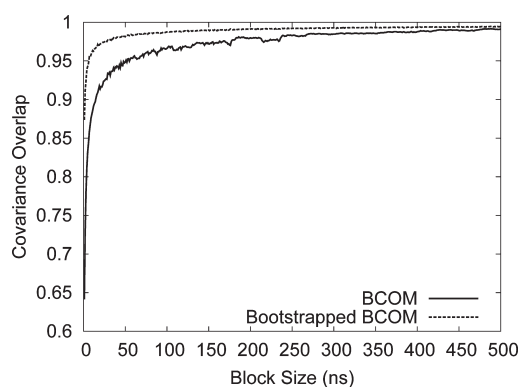


Figure 2. BCOM and BBCOM for the 1 μ s dileucine trajectory.

by randomly perturbing the average structure used in computing a PCA and then calculating the covariance overlap between the perturbed and the unperturbed modes. An rms perturbation as small as 0.1 Å in the β_2 AR average structure reduces the covariance overlap from 1 to 0.92 and an overlap of 0.94 for rhodopsin and CB2. Therefore, we must first ask at what point is it reasonable to believe that the average structure has converged such that its statistical errors no longer perturb the PCA results significantly. For the case of the 1 μ s dileucine simulation, Figure 1 shows the block averaged rmsd. The rmsd error in the average structure does not plateau until around 10 ns with an error of 0.019 Å. This is a higher uncertainty than would be suggested based on a plot of the running average rmsd (≈ 0.002 Å) for the simulation (see Figure S11A, Supporting Information). The running average rmsd for the transmembrane C α 's of rhodopsin is also shown in Figure S11B, Supporting Information. It appears to show a convergence at approximately 1.4 μ s with an error of 0.025 Å. It is important to note however that the rmsd beyond the first several hundred nanoseconds is quite small in both cases, and it would be tempting to determine convergence of the average structure at this point. The rmsd between each structure of the rhodopsin simulation and the starting crystal structure is shown in Figure S12, Supporting Information. Here, there is a short plateau at 0.5 μ s where the rmsd does not increase, followed by an increase and a longer plateau at about 1 μ s. This "classical" measure of convergence therefore suggests that the simulation has "converged" by 1 μ s.

3.2. Convergence and Correlation Times. The PCA of an MD trajectory is defined by both the fluctuations of the system as well as how well these are sampled. To illustrate this, imagine a simple model system that has two distinct states. If the system stays in one well, then the average structure and the fluctuations about that average will be different from the case where the system transitions between the two wells. Similarly, once the simulation samples the true statistical distribution of states (e.g., twice as many samples in state one as in state two), then the average structure and the fluctuations about that average will be different still. Only once the fluctuation subspaces cease to change, can we consider the simulation well sampled.

The covariance overlap (eq 1) is a powerful tool for determining how similar the conformational spaces sampled by two trajectories are. In contrast to the subspace overlap,³² which only considers the similarity of directions, the covariance overlap also includes the relative significance (i.e., power) of each mode. The covariance overlap also considers all modes, rather than an arbitrary subset. This is a more stringent test on whether the

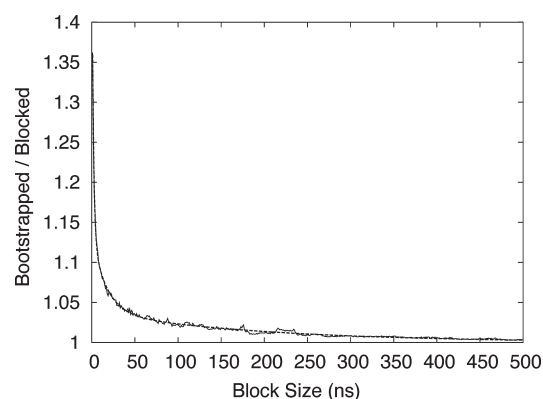


Figure 3. Inverse of bootstrap-normalized BCOM for the 1 μ s dileucine trajectory. A three-exponential function is fit to the curve giving three different correlation times. The correlation times are $t_1 = 1.7$ ns, $t_2 = 16.0$ ns, and $t_3 = 194.3$ ns.

subspaces are similar, since it requires a better sampling of the underlying fluctuations in order for both the directions and the power spectra to match.

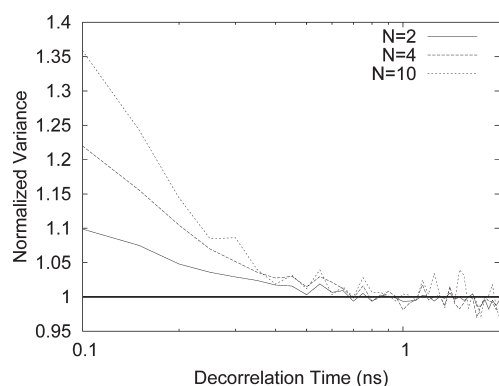
If we wish to say that a trajectory has "converged," then it is reasonable to expect that the fluctuations from different, large subsets of the trajectory should be similar, i.e., have a covariance overlap that approaches 1. The point at which convergence has been achieved can then be determined by choosing successively smaller subsets and determining at what point the covariance overlap diverges from 1. In practice, the subsets are picked as contiguous blocks of a given size spanning the trajectory. This method is what we call the BCOM. It is, in effect, a quantification of the statistical error present in the MD simulation due to finite and discrete sampling. The overlap for each block is then normalized by the overlap value expected were there no correlations via bootstrapping, and the inverse of the resulting curve is fit to a three exponential function. As the block sizes increases, this ratio will decay toward 1, where the blocks are long enough that they are effectively uncorrelated.

Hess⁵ notes that autocorrelation functions of principal components can be fitted with a double exponential function using a fast and a slow correlation time. Indeed, such a hierarchy has been seen previously in PCA analysis for large systems and manifests as a "beads on a string" when visualizing the fluctuation phase space.^{29,33} Within a bead, or conformation state, there is a short correlation time as the local well(s) are explored. The transition between different beads occurs at a much slower time scale. In contrast, our method results in three different correlation times, typically at different scales (i.e., fast, medium, and slow scale). While we could hypothesize that the third time scale is the time to get a sampling of the distribution of states, we have no underlying model to justify a triple exponential; fitting with two exponentials leaves a clear residual, while the residual from triple exponential appears random.

The covariance overlap from the BCOM and the bootstrapped BCOM (BBCOM) for the dileucine system is shown in Figure 2. Both the BCOM and the BBCOM approach 1 after 500 ns. The BCOM curve rapidly increases to 0.95 around 50 ns and then slowly increases thereafter. The bootstrapped curve, in contrast, is very close to its peak value for random sets containing as many points as 25 ns blocks. These curves illustrate the effect of correlation within the blocks. The ratio of the BBCOM to the

Table 2. Fitting of Three-Exponential Curve to BCOM Results for Different Model Systems

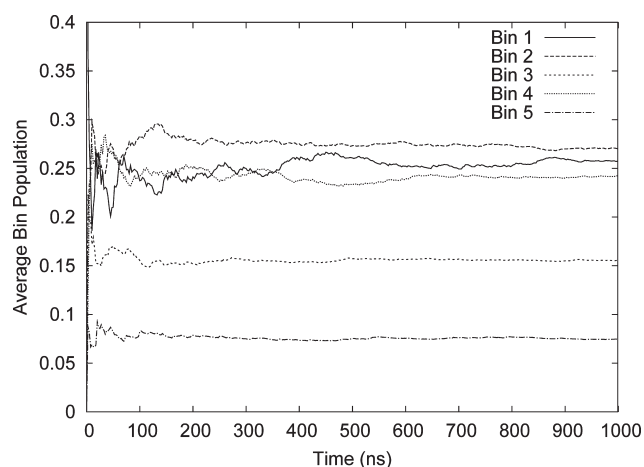
model	k_1	t_1 (ns)	k_2	t_2 (ns)	k_3	t_3 (ns)
dileucine	0.42	1.7	0.10	16.0	0.04	194.3
LfB6	0.19	32.9	0.17	199.9	0.12	1316.3
β_2 AR	0.76	23.3	0.51	248.9	0.19	2481.2
rhodopsin	2.18	10.0	0.89	47.0	1.24	804.5
CB2	1.69	7.3	0.98	39.1	1.15	934.7

**Figure 4.** Decorrelation time estimation by plotting $\sigma_{\text{obs}}^2(t)$ for three different step sizes: 2, 4, and 10. The time point where the curves reach 1 is the approximate decorrelation time.**Table 3. Decorrelation Times As Estimated by Variance Plots (τ_d) and from Automated Effective Sample Size Analysis (τ'_d)^a**

model	τ_d (ns)	τ'_d (ns)
dileucine	0.7–0.9	0.45–0.53
LfB6	40–55	35.6–51.0
β_2 AR ^b	75–90	37.2–46.7
rhodopsin ^b	130–150	60.3–70.9
CB2 ^b	140–175	76.7–91.2
β_2 AR ^c	90–100	60.5–78.2
rhodopsin ^c	140–180	120.0–150.5
CB2 ^c	170–230	160.8–188.7

^aAll analyses used 20 replicates. The τ_d for dileucine and LfB6 used step sizes of 2, 4, and 10, while all other models used 2, 3, and 4. ^b20 bins were used and not all clustering for τ'_d resulted in 2 top-level clusters. ^c10 bins were used and not all clustering for τ'_d resulted in 2 top-level clusters.

BCOM is shown in Figure 3, along with the best three-exponential fit. The correlation times, determined from the three-exponential fit, are shown in Table 2. It is quite clear from this graph that the dileucine system has converged since the ratio decays to 1 by 500 ns. What is striking in this figure is the magnitude of the longer time scale for such a simple system—nearly 200 ns. This is a measure of the simulation time required to accurately determine the relative populations of the different states available to dileucine. It is worth noting that this time is far longer than the average lifetime of any single state. Rather, it reflects the fact that even a molecule as simple as dileucine has a significant number of states available to it.

**Figure 5.** The running average of the population for each bin over time using 5 bins for the dileucine simulation.

In order to further assess the quality of sampling, we also performed a decorrelation time and an effective sample size analysis, as described in Section 2.8. The normalized variance versus decorrelation time is shown in Figure 4 for three different step sizes through the trajectory. Each curve converges to 1 at ≈ 0.7 – 0.9 ns. Similarly, the decorrelation time estimated from the effective sample size (also described in Section 2.8) is 0.45–0.53 ns. The decorrelation times and their estimates for each system are shown in Table 3. It is important to remember that these two methods are not measuring the same thing as the BCOM; both of these methods hinge on the rate of interconversion between individual states, which BCOM measures as the convergence of their populations. Examining the pairwise (all-to-all) rmsd map (Figure SI3A, Supporting Information), it is apparent that dileucine has substates that last approximately 1 ns. Indeed, the pairwise rmsd map is a very simple test that can be used to give a qualitative assessment of how well sampled the trajectory is.^{2,29} However, the off-diagonal blocks, despite having a low rmsd, are not guaranteed to be structurally similar.

While the decorrelation time (or the rate of transitions between states) can be quite rapid, the convergence of the distribution of observables may take much longer. For example, the dileucine trajectory can be partitioned as above but using 5 bins. A running average of the bin populations is computed and graphed over time in Figure 5. The majority of the changes in populations have smoothed out in the first 200 ns or so, with some variation in bins 1 and 4 that continue until nearly 1 μ s.

3.3. LfB6. The LfB6 hexapeptide (RRWQWR-NH₂) is a slightly larger model system that was simulated in explicit solvent. As with the dileucine model, all nonhydrogen atoms were considered in the analysis. The average inverse bootstrap-normalized BCOM plot for all 6 LfB6 simulations is shown in Figure 6 with the fit parameters listed in Table 2. There is a short correlation time of 33 ns with longer correlation times of 200 ns and 1316 ns. When fit individually, half of the simulations have long correlation times exceeding the largest block size. This accounts for the variation in the BCOM plot at 1 μ s and beyond and is evidence of a lack of convergence despite being a small system that was simulated for far longer than is typical for full proteins.

The decorrelation time plots for LfB6 are shown in Figure 7. The curves for the different step sizes are averaged over all six simulations. This figure yields an estimated decorrelation time of

40–55 ns. The average decorrelation time derived from the effective sample size analysis (using 20 replicates) is approximately the same, 35.6–51 ns. This time scale also matches the short correlation time from the BCOM analysis.

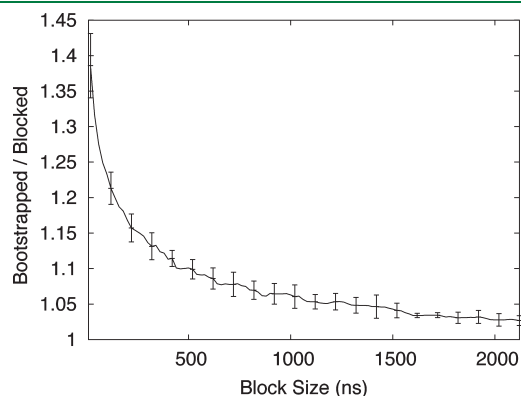


Figure 6. The inverse of the averaged, bootstrap-normalized BCOM plot for all six LfB6 simulations. The error bars are the standard deviation of the normalized BCOM from all simulations. The correlation times are $t_1 = 32.0$ ns, $t_2 = 199.9$ ns, and $t_3 = 1316.3$ ns. The t_1 correlation time closely matches the decorrelation time estimated by other methods.

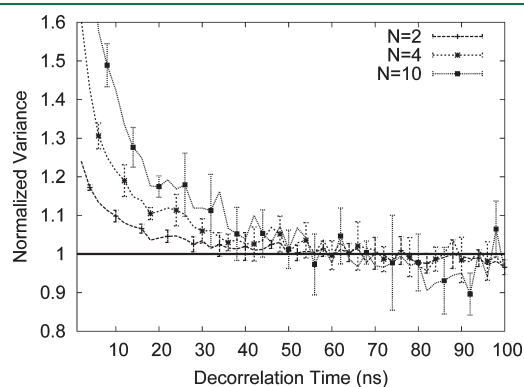


Figure 7. Decorrelation time estimation from $\sigma_{\text{obs}}^2(t)$ for all 6 LfB6 simulations averaged together. The error bars are the standard deviation across all six simulations. The decorrelation time is estimated as being between 40 and 55 ns.

Figure 8A shows the running average of the cluster populations for a 20 bin partitioning of the first LfB6 trajectory. Here, the bins are across the x axis with time along the y axis. Color indicates the deviation of the running average from the overall average population for each column. Qualitatively, what we find is that the cluster populations continue changing until approximately 1–1.5 μs , although there are again variations up to the full 3 μs of the simulation. The time scale of the convergence of these cluster populations corresponds to the long time scale predicted by the BCOM analysis of 1.3 μs .

3.4. GPCRs. The three GPCR simulations represent an application of these methods to large systems of biological interest, with state of the art trajectories ranging from 1 to nearly 2 μs . In order to make the convergence criterion more lenient, only the transmembrane α carbons were considered in the analysis, since incorporating the fluctuations of the flexible loops and the termini would vastly expand the configuration space of the systems. The BCOM curve and fit for $\beta_2\text{AR}$ is shown in Figure 9A. The best exponential fit to the data is again a three-exponential giving three well-separated correlation times of approximately 23 ns, 250 ns, and 2.5 μs . It is important to note that the longest correlation time, 2.5 μs , is longer than the largest block size used in the analysis. Moreover, the final ratio is above 1.2, indicating that the longest blocks still do not appear uncorrelated, which in turn suggests that the system is poorly converged. Interestingly, a previous analysis of the phase space formed by the first three principal components for $\beta_2\text{AR}$ ²⁹ found the presence of “beads”, indicative of conformational substates, with an average duration of 252 ns. This time scale is virtually identical to the medium time scale found by the BCOM.

The decorrelation time and effective sample sizes for $\beta_2\text{AR}$ are shown in Table 3. In this case, the analyses were repeated using two different numbers of bins for the partitioning of configuration space: 10 and 20 bins. There were an insufficient number of frames to support larger step sizes (i.e., $N = 10$) in the decorrelation time plots, so step sizes 2–4 were used instead. The estimated decorrelation times for $\beta_2\text{AR}$ are 90–100 ns for 10 bins and 75–90 ns for 20 bins. The effective sample size hierarchical clustering did not generally result in two top-level states (this can occur in this analysis when there are states that do not interconvert during the trajectory, which is itself a sign that the trajectory has not converged). In addition, the effective sample size (N_{eff}) per bin was only slightly greater than 1, suggesting that the system is not converged and that the resulting

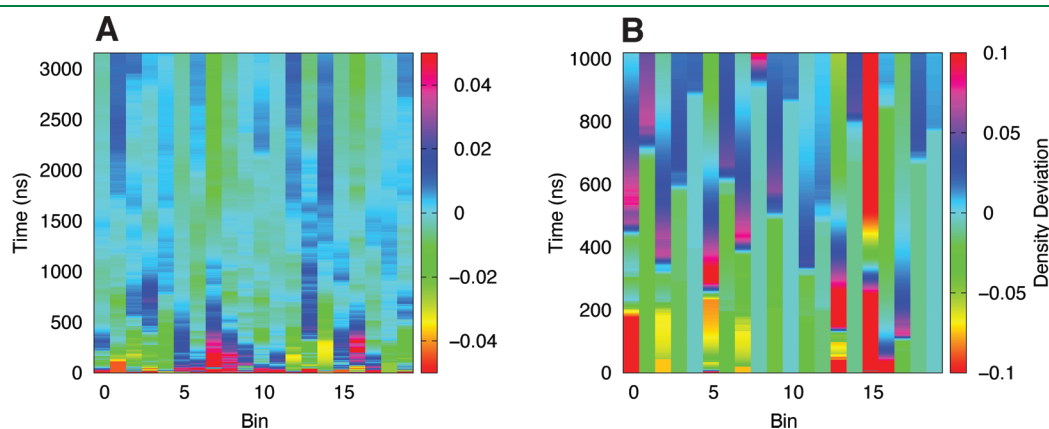


Figure 8. A visualization of the change in cluster populations over time for the first LfB6 trajectory (panel A) and for $\beta_2\text{AR}$ (panel B). Color represents the deviation from the column average.

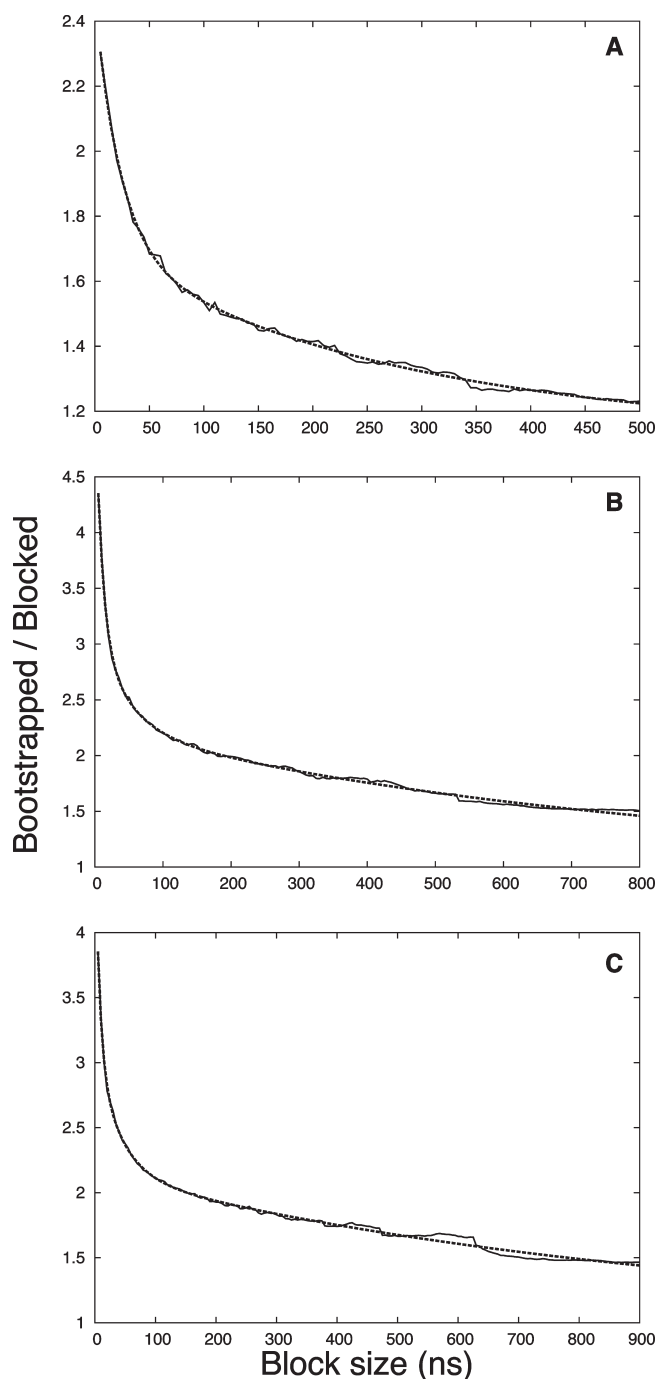


Figure 9. Inverse of the bootstrap-normalized BCOM for the simulation of β_2 AR (panel A), rhodopsin (panel B), and CB2 (panel C). The longest correlation times are 2481, 804, and 935 ns, respectively, and are approximately equal or greater than the largest block size, indicating a lack of convergence.

statistics are suspect. The estimated decorrelation times based on N_{eff} are 60–78 and 37–47 ns for 10 and 20 bins, respectively. The running average of the cluster populations for β_2 AR is shown in Figure 8B; many of the bins show sharp population changes well into the simulation, such as bin 1 at nearly 700 ns, while bin 5 continues to drift until the 1 μ s mark, and bin 8 shows a sudden change at almost 1 μ s. The extent of convergence for CB2 can be seen in the all-to-all rmsd plot in Figure SI3B,

Supporting Information. The only significant cross-peak, indicating revisiting of conformational substates, occurs at around 1.4 μ s, although there are broad regions with some self-similarity, such as the first 0.4 μ s and 0.6–1.1 μ s. In addition, one can qualitatively discern small blocks along the diagonal whose size is approximately 100 ns. This again reinforces the challenges of adequately sampling a large protein system, even with start of the art, microsecond-scale simulations.

The BCOM, decorrelation time, and N_{eff} for rhodopsin and CB2 are also given in Tables 2 and 3. In each case, the BCOM curve never comes close to 1. Rhodopsin plateaus at approximately 1.7 (Figure 9B), and CB2 reaches about 1.6 (Figure 9C), indicating neither system is converged. Moreover, the long correlation time exceeds the largest block size used in the analysis, further suggesting that the systems have not converged. The decorrelation time and N_{eff} analyses are somewhat more complicated given the differences between 10 and 20 bins. While the change in decorrelation time is not that significant, the N_{eff} varies considerably. In the 10 bin case, the number of top clusters found is smaller than in the 20 bin case, and it is likely that the larger numbers are more reflective of the “true” decorrelation time. Nevertheless, the decorrelation times for all of the GPCRs are quite long, approaching 100 ns or longer.

While the time scales found by the BCOM and the decorrelation and N_{eff} analysis diverge for the GPCR systems, all methods indicate that there are quite long time scales involved and that the number of statistically independent configuration samples, even in a multimicrosecond simulation, is small. This divergence is also not entirely unexpected considering that the BCOM is using a very different approach from the decorrelation time and N_{eff} analysis. Moreover, since the per bin N_{eff} is very close to 1, indicating insufficient sampling quality, the estimated decorrelation times are suspect.

3.5. Cosine Content. The cosine content for contiguous blocks along the first mode is shown in Figure 10. The average cosine content across all blocks of a given size is plotted, and the error bars are the standard deviation. Figure 10A illustrates the cosine content for a converged simulation, the first LfB simulation. The cosine content reaches 0 around 1 μ s, suggesting convergence on a slightly longer time-scale than determined by BCOM (\approx 900 ns, data not shown). In contrast, Figure 10B shows the cosine content for the rhodopsin simulation, a nonconverged simulation. The cosine content begins high and increases with larger block sizes. Hess suggests that, in practice, the cosine content can be a useful negative indicator of conformational sampling.⁵ In a simpler form, the cosine content of the first mode for the entire trajectory can be a simple test to determine whether a system is undersampled. For example, the cosine content for β_2 AR, rhodopsin, and CB2 are 0.78, 0.87, and 0.9, respectively. In contrast, the average cosine content for the LfB simulations is 0.003.

3.6. Known Unknowns: Suggestion for Practical Applications of BCOM. There are two tantalizing questions for the BCOM analysis: How reliably can it indicate when a system is not converged, and can it predict how much longer is required to run until convergence is achieved? Philosophically, this is a difficult prospect since it is difficult to know what is not known. It is always possible that increasing the simulation time will reveal a new conformation state, and there is no way to know that this state exists solely by using the previously seen configuration space. In tests where the trajectories were arbitrarily truncated and the BCOM performed, the reliability of using the long

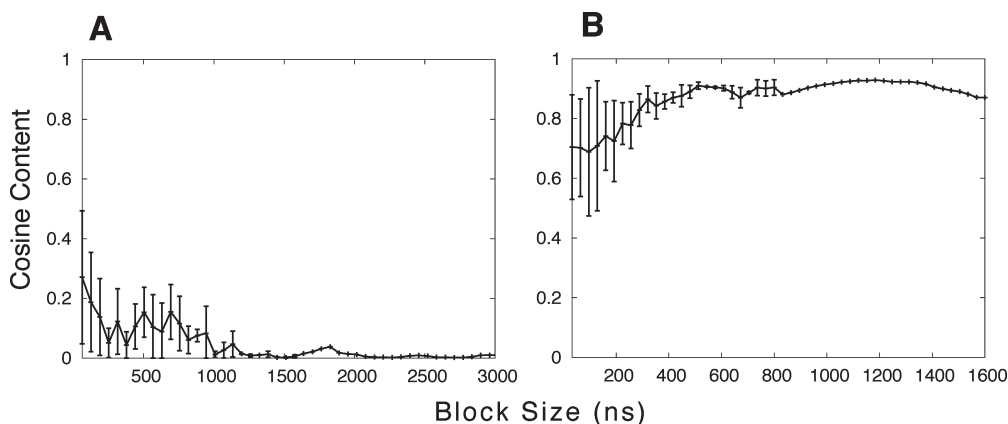


Figure 10. The average cosine content as a function of block size for the first LfB simulation (panel A) and rhodopsin (panel B). The error bars are the standard deviations across all contiguous blocks of a given size.

time-scale coefficient and the BCOM ratio is low until at least the average structure begins to converge. Indeed, when the simulation is so small that it stays within a single well, it may in fact be well sampled within this restricted conformation space—in effect, convergence is in the eye of the beholder. An alternative interpretation is that if the simulation is far too short, then there is no way to know this. Once the trajectories become sufficiently large, however, more credence can be given to these indicators. This is not conclusive proof, however. In some cases, the trajectory was truncated before a new state was discovered, resulting in an apparent convergence (data not shown). In some respects, however, the BCOM was not incorrect in that given the known configuration space, the simulation did appear to be converged. However, once the trajectory was extended and the new state was seen, it was apparent that the simulation was not close to convergence. It bears repeating that there is no way to prove a simulation (or even an ensemble of simulations) has converged to the correct answer. Rather, one can only demonstrate that the system has not converged or that it may be converged.

It is important to bear in mind that none of these methods replace the simpler methods of assessing the simulation's sampling, such as visually inspecting the all-to-all rmsd of the system, examining the convergence of the average structure, and the time series of observables of interest. While it has been repeatedly shown that these measures are not necessarily indicative of good sampling quality, they are a simple test that represents a minimum threshold that any simulation must pass before more sophisticated methods are employed. Similarly, much can be learned by plotting the projection of the system along the first several principal components, and this is typically a computationally inexpensive analysis.

4. CONCLUSIONS

We have devised a new method for assessing both the quality of sampling and the rate of convergence for a molecular dynamics simulation. The method relies on the similarities of the subspace sampled by the simulation and defined by the system's fluctuations. This method differs from other approaches in that it relies on fitting the resulting curves to a three-exponential function rather than a graphical interpretation. In addition, more information is considered in the analysis by using the covariance overlap, utilizing a wide range of block sizes for partitioning the trajectory and normalizing the resultant overlaps by a bootstrapped block

sample. Moreover, the hierarchical nature of the different correlation times indicates a longer time scale for sufficient sampling of the “known” configuration space than is suggested by other measures. Armed with this knowledge of the correlation times in the simulation, we can now make assertions regarding the effective sample size for any observable statistics and hence the statistical errors that are present in those quantities.

All of the methods investigated in this work suffer when the simulation is far too short. That is, they require a minimum sampling quality before one can hope to determine how well converged or sampled the system is. Even in the case where the simulation is converged, there is appreciable variation in the numbers obtained for the correlation and decorrelation times, although the general time scales are similar. Absent a priori knowledge of what the fluctuation space should look like, there is no method known to the authors that can determine an “unknown unknown,” that is, states that should have been seen but have not yet been visited by the system. Moreover, many of these methods for assessing sampling quality are new, and it is not yet clear under what conditions they perform well and those under which they fail. Given this, we strongly recommend that all available methods, ranging from the simple rmsd plots and the cosine content along the first few modes to the structural decorrelation analyses and BCOM, be used to assess sample quality and convergence. These tests are computationally inexpensive (particularly relative to the cost of running an all-atom simulation), and implementations are freely available as part of LOOS. Only in concert, combining N_{eff} with BCOM for example, can we hope to assert what the statistical error is and whether or not a simulation is well sampled. We must emphasize again that despite the improvements in these methods for assessing sampling quality, there is no substitute for visually observing the time series, be it torsions or projections along a principal component to look for multiple transitions.

■ ASSOCIATED CONTENT

S Supporting Information. Three figures: The first shows the rmsd between the running average structure for time t and $t + 1$ for dileucine and rhodopsin. The second shows the rmsd between the structure at time t and the starting crystal structure for rhodopsin. The final figure shows the all-to-all rmsd for each structure in the dileucine simulation with every other structure in

the same simulation as well as for the CB2 simulation. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: alan_grossfield@urmc.rochester.edu.

ACKNOWLEDGMENT

We would like to thank Dan Zuckerman for his insightful discussions and reading of this manuscript. We would also like to thank IBM Watson for the BlueGene supercomputer and Michael Pitman for the GPCR simulations. We are also grateful to the Center for Research Computing at the University of Rochester for providing computing systems and personnel to enable the research presented in this manuscript.

REFERENCES

- (1) Flyvbjerg, H.; Petersen, H. *J. Chem. Phys.* **1989**, *91*, 88–103.
- (2) Grossfield, A.; Zuckerman, D. M. *Ann. Rep. Comp. Chem.* **2009**, *5*, 23–48.
- (3) Balsera, M.; Wriggers, W.; Oono, Y.; Schulten, K. *J. Phys. Chem.* **1996**, *2567–2572*.
- (4) Amadei, A.; Ceruso, M. A.; Nola, A. D. *Proteins* **1999**, *36*, 419–24.
- (5) Hess, B. *Phys. Rev. E* **2002**, *65*, 031910.
- (6) Faraldo-Gómez, J. D.; Forrest, L. R.; Baaden, M.; Bond, P. J.; Domene, C.; Patargias, G.; Cuthbertson, J.; Sansom, M. S. P. *Proteins* **2004**, *57*, 783–91.
- (7) Grossfield, A.; Feller, S. E.; Pitman, M. C. *Proteins* **2007**, *67*, 31–40.
- (8) Sullivan, D. C.; Kuntz, I. D. *Proteins* **2001**, *42*, 495–511.
- (9) Smith, L. J.; Daura, X.; van Gunsteren, W. F. *Proteins* **2002**, *48*, 487–96.
- (10) Lyman, E.; Zuckerman, D. M. *J. Phys. Chem. B* **2007**, *111*, 12876–82.
- (11) Zhang, X.; Bhatt, D.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2010**, *6*, 3048–3057.
- (12) Efron, B. *Ann. Stat.* **1979**, *1*, 1–26.
- (13) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; CRC Press LLC: Boca Raton, FL, 1998; pp 45–57.
- (14) Romo, T. D.; Grossfield, A. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2009**, *1*, 2332–5.
- (15) Romo, T. D.; Grossfield, A. *LOOS: Lightweight Object Oriented Structure analysis*; Grossfield Lab, University of Rochester Medical School: Rochester, NY; <http://loos.sourceforge.net>. Accessed May 20, 2011).
- (16) *BOOST C++ Libraries*; <http://www.boost.org>.
- (17) Whaley, R. C.; Dongarra, J. In *Proceedings from Ninth SIAM Conference on Parallel Processing for Scientific Computing*, San Antonio, TX, March 22–24, 1999; SIAM: Philadelphia, PA, 1999; (in CD-ROM).
- (18) Whaley, R. C.; Petitet, A. *Software: Practice and Experience* **2005**, *35*, 101–121.
- (19) Brooks, B.; Bruccoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (20) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–802.
- (21) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–88.
- (22) Spoel, D. V. D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–18.
- (23) Ponder, J. *Tinker*, version 4.2; Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine: St. Louis, MO, 2007.
- (24) Tomita, M.; Takase, M.; Bellamy, W.; Shimamura, S. *Acta Paediatr. Jpn.* **1994**, *36*, 585–91.
- (25) Romo, T. D.; Bradney, L. A.; Greathouse, D. V.; Grossfield, A. *Biochim. Biophys. Acta, Biomembr.* **2011**, *1808*, 2019–30.
- (26) Allen, F.; et al. *IBM Syst. J.* **2001**, *40*, 310.
- (27) MacKerell, A. D., Jr.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. *CHARMM: The Energy Function and Its Parameterization with an Overview of the Program*; John Wiley and Sons: Chichester, 1998; Vol. 1; pp 271–277.
- (28) A. D. MacKerell, J.; Feig, M.; C. L. Brooks, I. *J. Comput. Chem.* **2004**, *25*, 1400–15.
- (29) Romo, T. D.; Grossfield, A.; Pitman, M. C. *Biophys. J.* **2010**, *98*, 76–84.
- (30) Grossfield, A.; Pitman, M. C.; Feller, S. E.; Soubias, O.; Gawrisch, K. *J. Mol. Biol.* **2008**, *381*, 478–86.
- (31) Hurst, D. P.; Grossfield, A.; Lynch, D. L.; Feller, S.; Romo, T. D.; Gawrisch, K.; Pitman, M. C.; Reggio, P. H. *J. Biol. Chem.* **2010**, *285*, 17954–17964.
- (32) Hess, B. *Phys. Rev. E* **2000**, *62*, 8438–48.
- (33) Clarage, J. B.; Romo, T.; Andrews, B. K.; Pettitt, B. M.; Phillips, G. N. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3288–92.

Benchmarking the Approximate Second-Order Coupled-Cluster Method on Biochromophores

Robert Send,[†] Ville R. I. Kaila,^{‡,§,||} and Dage Sundholm^{*,§}[†]Institut für Physikalische Chemie, Karlsruher Institut für Technologie, Kaiserstrasse 12, 76131 Karlsruhe, Germany[‡]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, 5 Memorial Drive, Bethesda, Maryland, United States[§]Department of Chemistry, P.O. Box 55 (A. I. Virtanens plats 1), FIN-00014 University of Helsinki, Finland^{||}Helsinki Bioenergetics Group, Programme of Structural Biology and Biophysics, Institute of Biotechnology, University of Helsinki, P.O. Box 65, FI-00014 Helsinki, Finland

ABSTRACT: Extensive benchmarking calculations are presented to assess the accuracy of commonly used quantum chemical methods in studying excited state properties of biochromophores. The first few excited states of 12 common model chromophores of photoactive yellow protein, green fluorescent protein, and rhodopsin have been studied using approximate second-order coupled-cluster (CC2) and linear-response time-dependent density functional theory (TDDFT) calculations. The study comprises investigations of basis-set dependences on CC2 excitation energies as well as comparisons of the CC2 results with excitation energies obtained at other computational levels and with experimental data. The basis-set study shows that the accuracy of the two lowest excitation energies is generally sufficient when triple- ζ basis sets augmented with polarization functions are employed, whereas the third and higher excited states were found to require diffuse basis functions in the basis set. Augmenting the basis set with diffuse functions contributes less than 0.15 eV to the excitation energies of low-lying excited states, except for some of the studied anionic states and for Rydberg states. Calculations at the TDDFT level using the B3LYP functional show the necessity of stabilizing anions with point charges or counterions when aiming at reliable electronic excitation spectra. The two lowest excitation energies of the green fluorescent protein and rhodopsin chromophores calculated at the CC2 level agree within 0.15 eV with experimental excitation energies, whereas the B3LYP values are somewhat less accurate, with a maximum deviation of 0.27 eV. The computed excitation energies for the photoactive yellow protein chromophore models deviate from available experimental values by 0.3–0.4 eV and 0.1–0.5 eV, at the CC2 and B3LYP levels of theory, respectively.

INTRODUCTION

Prediction of excitation energies is of profound importance in elucidating molecular mechanisms of photochemical and photo-biological processes, and in characterizing chemical intermediates determined by spectroscopic techniques. For quantitative predictions, this requires high accuracy of the quantum chemical method, i.e., a deviation of 0.1 eV or less. The use of sufficiently large molecular model systems puts additional constraints on the choice of quantum chemical methods. The accuracy of calculated excitation energies depends on the level of the electron correlation treatment and the basis-set size. Accurate computational approaches such as high-order configuration interaction or coupled-cluster-based methods can yield excitation energies in close agreement with experimental results, provided that the employed basis set is sufficiently large.^{1–5} However, such calculations are still limited to very small systems, comprising less than 10 atoms.⁶ The use of small basis sets introduces significant errors, implying that the results are unreliable despite the accurate treatment of the electron correlation. Density functional theory (DFT) and low-order *ab initio* electron correlation methods, such as approximate second-order coupled-cluster (CC2) calculations, can be employed in combination with large basis sets to yield results in the complete basis-set limit. However, even the CC2 calculations are computationally very

expensive for the larger biochromophores. The linear-response time-dependent DFT (TDDFT) method, which has become a very popular tool for investigating excited states, provides accurate excitation energies at much lower computational cost relative to *ab initio* correlation methods.^{7–9} The drawback is a significant number of problematic cases where today's functionals do not provide accurate excitation energies.^{2,5,10–13}

Biological chromophores are an intriguing challenge for quantum chemical methods.^{14,15} The complexity of biological systems such as protein surroundings are part of these challenges. However, isolated chromophore models of well studied proteins suffice to reveal the limitations of any contemporary excited state quantum chemical method. The discrepancies in the excited state descriptions have caused some controversies but also confusion with respect to the reliability and applicability of the concerned methods.^{16–19} There is an increasing amount of extensive experimental gas-phase data for excitation energies of common biological chromophore models.^{20–26} These data have been compiled by Ma et al.²⁷ and used to assess the accuracy of the many-body Green's function approach.²⁸ The compilation of reference data has revealed the need for benchmarking the most

Received: March 30, 2011

Published: June 13, 2011

common excited state methods to obtain an overview of their performance and to resolve the confusion about applicability, basis-set requirements, and further parameter dependencies.

Benchmark sets of vertical excitation energies are numerous, especially for organic chromophores. Most recent efforts comprise the *best estimates* benchmark of the Thiel group, in which reference values are obtained computationally,¹⁰ and the meta studies on various benchmarks by Jacquemin and co-workers.²⁹ Epifanovsky et al.³⁰ recently benchmarked excited states of the anionic green fluorescent protein chromophore, which is one of the molecules studied in this work.

Vertical excitation energies are not readily deduced from experimental results. Calculated vertical excitation energies are therefore usually compared to the maxima of the experimental absorption spectra. A more accurate way of benchmarking would be to compare calculated adiabatic excitation energies with experimental results.^{31,32} However, such comparisons require information about excited state minima as well as experimentally resolved vibrational spectra, both of which are unavailable for the present series of biochromophores. Thus, we rely on comparisons of vertical excitation energies to experimental absorption maxima and assume that the introduced uncertainties are smaller than the error of the computational method.

In this work, we investigate the basis-set dependence of the lowest excitation energies of important biochromophore models by CC2 calculations with basis sets extended to quadruple- ζ quality. The obtained values are compared to excitation energies and oscillator strengths obtained by TDDFT calculations and to experimental data. The benchmark set comprises seven p-hydroxycinnamic acid chromophore models representing the blue-light photoreceptor of the photoactive yellow protein (PYP), three p-hydroxybenzylideneimidazolinone (pHBDI) chromophore models of the green fluorescent protein (GFP), and two protonated Schiff-base (PSB) retinal models, which are the chromophores of rhodopsin proteins. Part of this study aims at identifying problems in the TDDFT calculations by direct comparison to CC2 results.

This article is structured as follows. The computational methods and the studied chromophores are described in the Computational Details section. The basis-set convergence of the CC2 calculations is discussed in the Results section, where the CC2 and DFT excitation energies obtained in the complete-basis-set limit are also compared to previously calculated and measured values. The main conclusions are summarized in the Conclusions section.

COMPUTATIONAL DETAILS

Methods and Basis Sets. This study employs the Karlsruhe basis sets of double-, triple-, and quadruple- ζ quality (def2-SVP, def2-TZVP, and def2-QZVP), as well as the triple- ζ basis set augmented with diffuse functions from Dunning's aug-cc-pVTZ basis set (aug-TZVP).^{33–36} The def2 prefix is omitted in the following. The number of spherical (5d, 7f, 9g) basis functions is 2565 in the QZVP calculations on all-*trans*-retinal. Optimization of the molecular ground state structures was performed at the second-order Møller–Plesset perturbation theory (MP2) level employing the TZVP basis sets and the resolution-of-the-identity (RI) approximation.^{37–39} The excitation energies were calculated at the CC2 level using the RI approximation.^{40–43} The 1s orbitals of C, N, and O were uncorrelated in the *ab initio* calculations. The excitation energies were also calculated at the

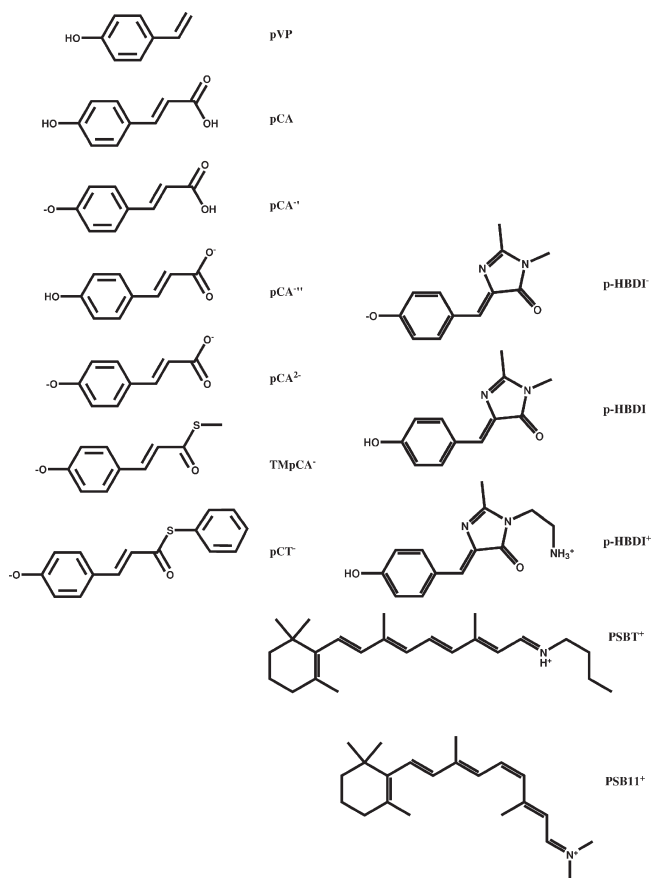


Figure 1. The molecular structures of the studied chromophore models: (a) the photoactive yellow protein chromophore models, (b) the green fluorescent protein chromophore models, (c) the rhodopsin chromophore models.

TDDFT level using Becke's three-parameter functional (B3LYP) and aug-TZVP basis sets.^{44–48}

Extrapolated basis-set limits at the CC2 level were estimated by combining values for the excitation energies obtained with the three largest basis sets (TZVP, aug-TZVP, and QZVP). All calculations were done with Turbomole version 6.1⁴⁹ on a Linux cluster equipped with 2.6 GHz AMD Opteron processors. The calculations are computationally demanding. Typical computational times for the CC2 calculations are 100 CPU hours for calculating the first excited state of HBDI[−] using the QZVP basis set and 150 CPU hours for obtaining the first excited state of PSBT⁺ at the CC2/TZVP level.

Benchmark Set Composition. In this work, we investigate the biochromophores shown in Figure 1, recently studied by Ma et al. at the many-body Green's function theory (MBGFT) level.²⁷ The abbreviations reported in the figure indicate the protonation states with prime (phenolate) and double prime (carboxylate) symbols, as well as the total charge of the molecular system. TDDFT calculations on the anionic chromophores are also performed with external point charges to obtain charge neutrality. The point charges ensure that the DFT ionization threshold is larger than the studied excitation energies.

Photoactive Yellow Protein Chromophore Models. The PYP molecule is a bacterial photoreceptor, responsible for absorbing blue light photons by its p-coumaric acid chromophore.^{50,51} The studied PYP chromophore models comprise *trans*-p-coumaric acid

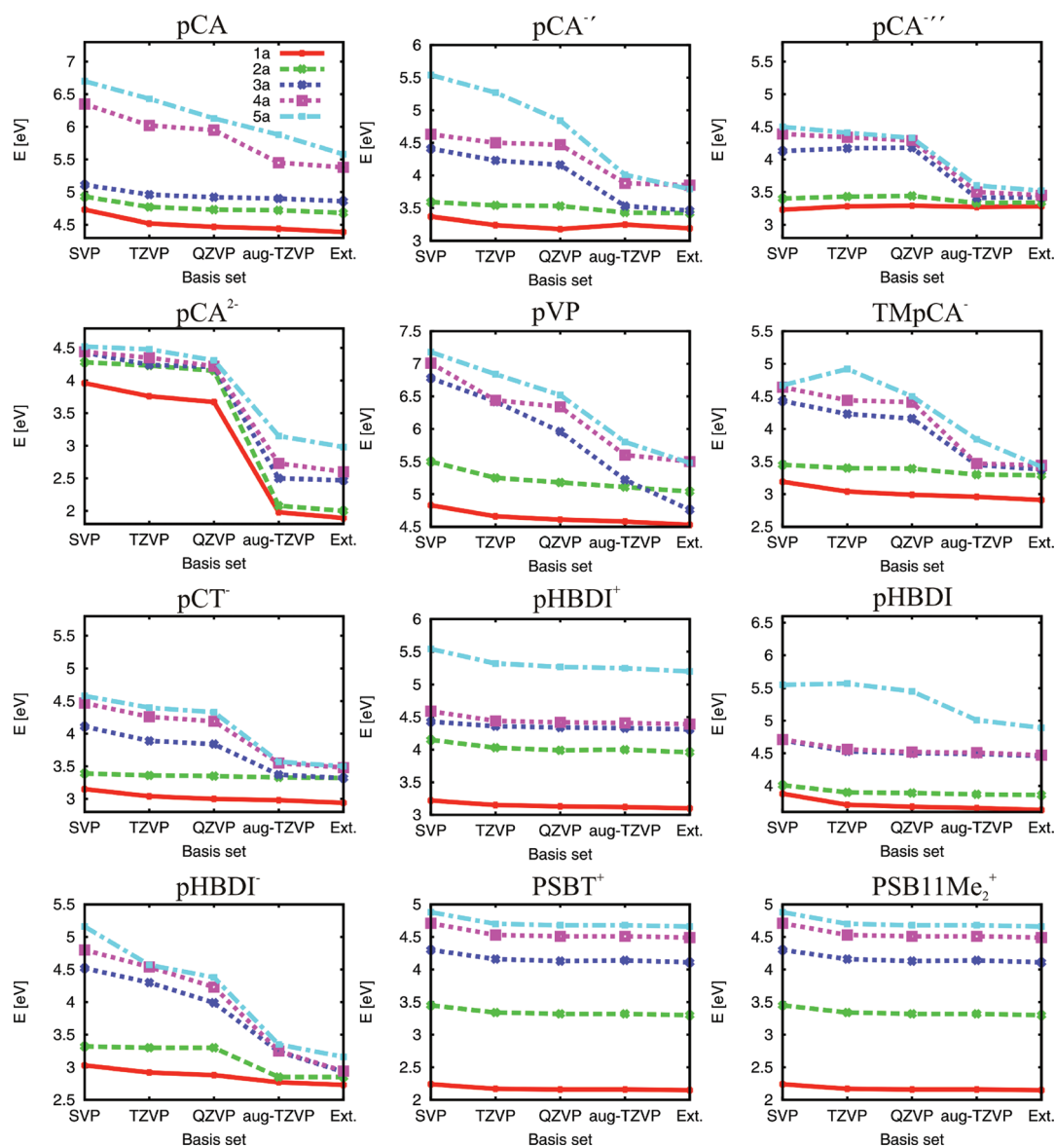


Figure 2. The basis-set convergence of the five lowest singlet excitation energies (in eV) of the investigated chromophore models studied at the CC2 level. The extrapolated aug-QZVP excitation energies are denoted Ext. The legend for the line structures is only given in the first subfigure.

(pCA), deprotonated *trans*-p-coumaric acid (pCA^-), *trans*-p-coumarate (pCA^{--}), deprotonated *trans*-p-coumarate (pCA^{2-}), p-vinyl phenol (pVP), thiomethyl-p-coumarate (TMpCA^-), and thiophenyl p-coumarate (pCT^-), shown in Figure 1a.

Green Fluorescent Protein Chromophore Models. The GFP molecule is responsible for the bioluminescence of many marine organisms.⁵² The luminescence takes place only when its chromophore, pHBDI embedded in the β -barrel structure of the protein, is exposed to blue light but not in solution.⁵² Three GFP model chromophores, shown in Figure 1b, are investigated, namely, pHBDI, pHBDI^- , and pHBDI^+ . The phenol group is deprotonated in the pHBDI^- model, and the methyl group of the imidazolinone nitrogen in the pHBDI model is substituted by a protonated ethylamine group ($-\text{CH}_2\text{CH}_2\text{NH}_3^+$) in the pHBDI^+ model.

Rhodopsin Chromophore Models. Rhodopsin is a G-protein coupled receptor, responsible for light absorption in vertebrate visual pigments.^{53,54} Two models of the light absorbing

chromophore of rhodopsin were studied, 11-*cis* and all-*trans* retinal (Figure 1c). In the former chromophore model, a butyl group is connected to the protonated nitrogen atom of the retinyl chain, simulating the cross-linked Lys-296 in the protein. In the latter chromophore model, the Schiff base nitrogen has two methyl substituents. Both models have a net positive charge of +1.

RESULTS

Basis-Set Convergence of the CC2 Calculations. The basis set convergence of the five lowest singlet excited states is shown in Figure 2 and Tables 1–4. The two lowest excited states of most chromophores are valence states and well described using the TZVP basis sets. The excitation energies obtained at the CC2/TZVP level deviate less than 0.15 eV from the extrapolated CC2/aug-QZVP values. The pHBDI^- chromophore constitutes an exception, where the second excited state obtained at the

Table 1. Basis-Set Dependence of the CC2 Excitation Energies (in eV) for the Five Lowest Excited Singlet States of the Photoactive Yellow Protein Chromophore Models pCA, pCA⁻, pCA⁻, and pCA²⁻^a

molecule	basis	1a	2a	3a	4a	5a
pCA	SVP	4.73	4.93	5.11	6.35	6.70
pCA	TZVP	4.52	4.77	4.96	6.02	6.43
pCA	QZVP	4.47	4.73	4.92	5.95	6.13
pCA	aug-TZVP	4.44	4.72	4.90	5.45	5.88
pCA	extrapolated	4.39	4.68	4.86	<u>5.38</u>	<u>5.58</u>
pCA	oscillator strength	0.54	0.27	0.00	0.22	0.16
pCA ⁻	SVP	3.37	3.59	4.41	4.63	5.54
pCA ⁻	TZVP	3.24	3.54	4.23	4.50	5.27
pCA ⁻	QZVP	3.18	3.53	4.16	4.47	4.84
pCA ⁻	aug-TZVP	3.25	3.43	3.53	3.88	4.01
pCA ⁻	extrapolated	3.19	3.42	<u>3.46</u>	<u>3.85</u>	<u>3.79</u>
pCA ⁻	oscillator strength	1.01	0.00	0.10	0.00	0.05
pCA ⁻	SVP	3.23	3.40	4.13	4.39	4.50
pCA ⁻	TZVP	3.28	3.43	4.17	4.34	4.41
pCA ⁻	QZVP	3.29	3.44	4.18	4.29	4.33
pCA ⁻	aug-TZVP	3.27	3.33	3.41	3.50	3.60
pCA ⁻	extrapolated	3.28	3.34	<u>3.42</u>	<u>3.45</u>	<u>3.52</u>
pCA ⁻	oscillator strength	0.00	0.00	0.02	0.02	0.05
pCA ²⁻	SVP	3.96	4.28	4.43	4.44	4.52
pCA ²⁻	TZVP	3.76	4.23	4.24	4.35	4.48
pCA ²⁻	QZVP	3.67	4.15	4.21	4.22	4.31
pCA ²⁻	aug-TZVP	1.98	2.08	2.50	2.73	3.15
pCA ²⁻	extrapolated	<u>1.89</u>	<u>2.00</u>	<u>2.47</u>	<u>2.60</u>	<u>2.98</u>
pCA ²⁻	oscillator strength	0.13	0.04	0.69	0.00	0.00

^aThe molecular structures were optimized for the ground state at the MP2/TZVP level. Excitation energies for Rydberg and anionic states that demand diffuse basis functions in the basis set are underlined. The oscillator strengths calculated at the CC2/TZVP level are given.

CC2/QZVP level deviates by 0.45 eV from the CC2/aug-TZVP value. Diffuse anionic states, states in the continuum, as well as states with significant Rydberg character can be identified by augmenting the basis set with diffuse functions. They might lower the excitation energy even by more than 1 eV. The higher excited states (states 3–5) of the PYP chromophore models seem to require the use of diffuse basis functions. Diffuse states are not observed for the positively charged chromophores among the five lowest excited states. As previously shown, a second set of diffuse functions is sometimes necessary in order to approach the basis-set limit.⁵ The excitation energies of the diffuse states are underlined in Tables 1–4. For pCA²⁻, the first set of diffuse functions lowers the excitation energies by a factor of 2, implying that the electrons of the doubly charged anion are unbound. Thus, no reliable values for the excitation energies of pCA²⁻ can be obtained without stabilizing the system with counterions, solvent molecules, or by protein residues.

The present benchmark calculations show that the two lowest excited states of the investigated chromophores, which are also the biologically relevant states, are accurately described with TZVP basis sets. Moreover, the calculations show the importance of diffuse basis functions for higher excited states of anionic and neutral chromophores. The calculations also indicate the importance of assessing the basis set requirement individually for such chromophores.

Table 2. Basis-Set Dependence of the CC2 Excitation Energies (in eV) for the Five Lowest Excited Singlet States of the Photoactive Yellow Protein Chromophore Models pVP, TMpCA⁻, and pCT⁻^a

molecule	basis	1a	2a	3a	4a	5a
pVP	SVP	4.83	5.50	6.78	7.01	7.18
pVP	TZVP	4.66	5.25	6.42	6.44	6.84
pVP	QZVP	4.61	5.18	5.96	6.34	6.52
pVP	aug-TZVP	4.58	5.11	5.22	5.60	5.80
pVP	extrapolated	4.53	5.04	<u>4.76</u>	<u>5.50</u>	<u>5.48</u>
pVP	oscillator strength	0.05	0.52	0.00	0.28	0.45
TMpCA ⁻	SVP	3.19	3.45	4.43	4.64	4.67
TMpCA ⁻	TZVP	3.04	3.40	4.23	4.44	4.92
TMpCA ⁻	QZVP	2.99	3.39	4.16	4.41	4.50
TMpCA ⁻	aug-TZVP	2.96	3.30	3.45	3.47	3.84
TMpCA ⁻	extrapolated	2.91	3.29	<u>3.38</u>	<u>3.44</u>	<u>3.42</u>
TMpCA ⁻	oscillator strength	1.21	0.00	0.09	0.00	0.00
pCT ⁻	SVP	3.15	3.39	4.11	4.47	4.58
pCT ⁻	TZVP	3.04	3.36	3.89	4.26	4.40
pCT ⁻	QZVP	3.00	3.35	3.84	4.19	4.33
pCT ⁻	aug-TZVP	2.98	3.33	3.37	3.55	3.57
pCT ⁻	extrapolated	2.94	3.32	<u>3.32</u>	<u>3.48</u>	<u>3.50</u>
pCT ⁻	oscillator strength	1.37	0.00	0.00	0.08	0.00

^aThe molecular structures were optimized for the ground state at the MP2/TZVP level. Excitation energies for Rydberg and anionic states that demand diffuse basis functions in the basis set are underlined. The oscillator strengths calculated at the CC2/TZVP level are given.

Comparison of CC2 and B3LYP Excitation Energies. *General Trends.* The excitation energies and oscillator strengths, obtained at the B3LYP and CC2 levels, are compared in Figure 3 and Table 5. The excitation energies calculated for the cationic and neutral chromophore models at the B3LYP level are systematically smaller than the corresponding CC2 energies, with a deviation of -0.14 to $+0.44$ eV and a maximum deviation of 0.64 eV for some of the higher excited states (e.g., for pCA⁻). At first glance, the B3LYP excitation energies seem to be in better agreement with experimental data than the CC2 ones, particularly for the smaller PYP chromophores. However, the oscillator strengths of the lower excited states of the anions obtained at the B3LYP level indicate that the TDDFT calculations suffer from problems, most likely related to the previously discussed DFT continuum problem.^{5,55} The B3LYP calculations on some of the neutral chromophores also yield oscillator strengths that significantly differ from those obtained at the CC2 level, indicating that the states have different orbital character.

Cationic and Neutral Chromophores. The excitation energies of the PYP chromophore model pVP calculated at the B3LYP/aug-TZVP level are 0.15–0.44 eV smaller than the corresponding CC2 values. The B3LYP calculations yield very small oscillator strengths for the higher excited states, whereas four of the five states are bright at the CC2 level. The basis set convergence at the CC2 level shown in Figure 2 indicates that some of the higher excited states are not fully converged with the aug-TZVP basis set. When comparing the extrapolated values, a smaller difference between B3LYP and CC2 excitation energies of 0.02–0.32 eV is obtained.

For the pCA chromophore, which has an additional protonated carboxyl group relative to pVP, the first and the fifth

Table 3. Basis-Set Dependence of the CC2 Excitation Energies (in eV) for the Five Lowest Excited Singlet States of the Green Fluorescent Protein Chromophore Models pHBDI, pHBDI⁻, and pHBDI⁺

molecule	basis	1a	2a	3a	4a	5a
pHBDI ⁺	SVP	3.22	4.15	4.43	4.59	5.54
pHBDI ⁺	TZVP	3.15	4.03	4.36	4.44	5.32
pHBDI ⁺	QZVP	3.13	3.99	4.34	4.42	5.27
pHBDI ⁺	aug-TZVP	3.12	4.00	4.33	4.41	5.25
pHBDI ⁺	extrapolated	3.10	3.96	4.31	4.39	5.20
pHBDI ⁺	oscillator strength	1.09	0.02	0.00	0.00	0.10
pHBDI	SVP	3.88	4.01	4.70	4.71	5.55
pHBDI	TZVP	3.71	3.90	4.53	4.56	5.57
pHBDI	QZVP	3.68	3.89	4.50	4.52	5.45
pHBDI	aug-TZVP	3.66	3.87	4.49	4.51	5.01
pHBDI	extrapolated	3.63	3.86	4.46	4.47	<u>4.89</u>
pHBDI	oscillator strength	0.77	0.00	0.15	0.00	0.16
pHBDI ⁻	SVP	3.03	3.32	4.52	4.80	5.16
pHBDI ⁻	TZVP	2.92	3.30	4.30	4.54	4.57
pHBDI ⁻	QZVP	2.88	3.30	3.99	4.23	4.38
pHBDI ⁻	aug-TZVP	2.77	2.85	3.24	3.25	3.35
pHBDI ⁻	extrapolated	2.73	<u>2.85</u>	<u>2.93</u>	<u>2.94</u>	<u>3.16</u>
pHBDI ⁻	oscillator strength	1.17	0.00	0.07	0.00	0.00

^a The molecular structures were optimized for the ground state at the MP2/TZVP level. Excitation energies for Rydberg and anionic states that demand diffuse basis functions in the basis set are underlined. The oscillator strengths calculated at the CC2/TZVP level are given.

transitions are strong at the B3LYP level, while the CC2 calculations predict similarly as for the pVP model four strong transitions among the five lowest excited states. The two lowest excitation energies obtained at the CC2 level are 0.30–0.35 eV larger than the experimental values, whereas the discrepancy at the B3LYP level is only 0.09–0.14 eV (cf. Comparison to Experiment section).

For the GFP and rhodopsin models (pHBDI⁺, pHBDI, PSBT⁺, and PSB11Me₂⁺), the B3LYP and CC2 excitation energies and oscillator strengths are in qualitative agreement. The B3LYP excitation energies of pHBDI⁺ and pHBDI are 0.2–0.3 eV smaller than the CC2 excitation energies, whereas the lowest excitation energies differ from the experimental values of the GFP and rhodopsin chromophores by 0.07–0.27 eV. The deviation from experiments at the CC2 level is 0.10–0.14 eV. The differences of 0.14 eV and –0.19 eV between the B3LYP and CC2 excitation energies of the first excited state of pHBDI⁺ and pHBDI affect the conclusions drawn from the B3LYP and CC2 calculations. For the two lowest states of PSBT⁺ and PSB11Me₂⁺, the B3LYP and CC2 energies differ by 0.09–0.23 eV, whereas for the three higher excited states, the B3LYP energies are 0.37–0.63 eV smaller than the CC2 values.

Anionic Chromophores. The DFT ionization potential is given by the absolute orbital energy of the highest occupied molecular orbital (HOMO).⁵⁵ This implies that if the excitation energies exceed the DFT ionization potential (IP), the TDDFT calculations might yield unreliable results. In some cases, the excitation energies obtained using an increasing size of the basis set converge toward the HOMO energy corresponding to the first IP according to Koopmans' theorem.^{5,55} This is especially problematic in the description of anions which have a weakly bound HOMO and

Table 4. Basis-Set Dependence of the CC2 Excitation Energies (in eV) for the Five Lowest Excited Singlet States of the Retinal Protonated Schiff-Bases PSBT⁺ and PSB11Me₂⁺

molecule	basis	1a	2a	3a	4a	5a
PSBT ⁺	SVP	2.23	3.45	4.34	4.72	4.89
PSBT ⁺	TZVP	2.17	3.35	4.22	4.54	4.73
PSBT ⁺	QZVP	2.15	3.33	4.19	4.52	4.70
PSBT ⁺	aug-TZVP ^b	2.16	3.34	4.20	4.52	4.70
PSBT ⁺	extrapolated	2.14	3.32	4.17	4.50	4.67
PSBT ⁺	oscillator strength	1.92	0.50	0.10	0.01	0.05
PSB11Me ₂ ⁺	SVP	2.24	3.45	4.30	4.71	4.88
PSB11Me ₂ ⁺	TZVP	2.17	3.34	4.16	4.53	4.70
PSB11Me ₂ ⁺	QZVP	2.16	3.32	4.13	4.51	4.68
PSB11Me ₂ ⁺	aug-TZVP ^b	2.16	3.32	4.14	4.51	4.68
PSB11Me ₂ ⁺	extrapolated	2.15	3.30	4.11	4.49	4.66
PSB11Me ₂ ⁺	oscillator strength	1.56	0.47	0.15	0.06	0.07

^a The molecular structures were optimized for the ground state at the MP2/TZVP level. The oscillator strengths calculated at the CC2/TZVP level are given. ^b In the aug-TZVP calculations on retinals, the d and f functions were removed from the augmentation due to ground state convergence problems.

thus a very low IP. To avoid the DFT ionization problem, TDDFT calculations were performed with the anions stabilized by external point charges (pc-B3LYP), placed 5 Å from the molecular plane, both above and below the negatively charged substituent. The point charges lift the ionization threshold, ensuring that the lowest excited states are bound states at the DFT level. The dependence of the position and distribution of the point charges were not investigated, since the aim of the point charges is merely to avoid the DFT continuum problem.

The DFT IPs are much smaller than the Hartree–Fock self-consistent field (HF SCF) ones, because of the self-interaction problem, which causes problems for the anions as all excited states lie formally in the DFT continuum above the ionization threshold. The calculations show that the lowest excitation energies are only slightly dependent on the presence of the point charges, indicating that the states could be considered as some kind of metastable resonances in the DFT continuum. The high density of states of the anions becomes more sparse in the DFT calculation when point charges are used, showing the necessity of using point charges. Comparison of the B3LYP and CC2 oscillator strengths for the anionic chromophores shows that the B3LYP calculations often yield low-lying intruder states, which disappear when the anion is stabilized by point charges. This stabilization leads to an increase of the absolute HOMO energy, whereas the gap between the HOMO and the lowest-unoccupied molecular orbital (LUMO) remains nearly unaltered, indicating that excitation energies are less affected by the point charges than the IP. The excitation energies and oscillator strengths calculated with and without point charges are compared in Table 6 and Figure 4.

The external point charges red-shift the first excited state of pCA⁻ from 3.05 to 2.79 eV (Figure 4). The bright transition is blue-shifted by 0.08 eV from 3.34 to 3.42 eV, whereas the oscillator strengths remain nearly unchanged (not shown). The higher excitation energies are blue-shifted, yielding values in good agreement with the CC2 energies. The bright state at the CC2 level is the first excited state, whereas at the pc-B3LYP level, the second excited state has the largest oscillator strength.

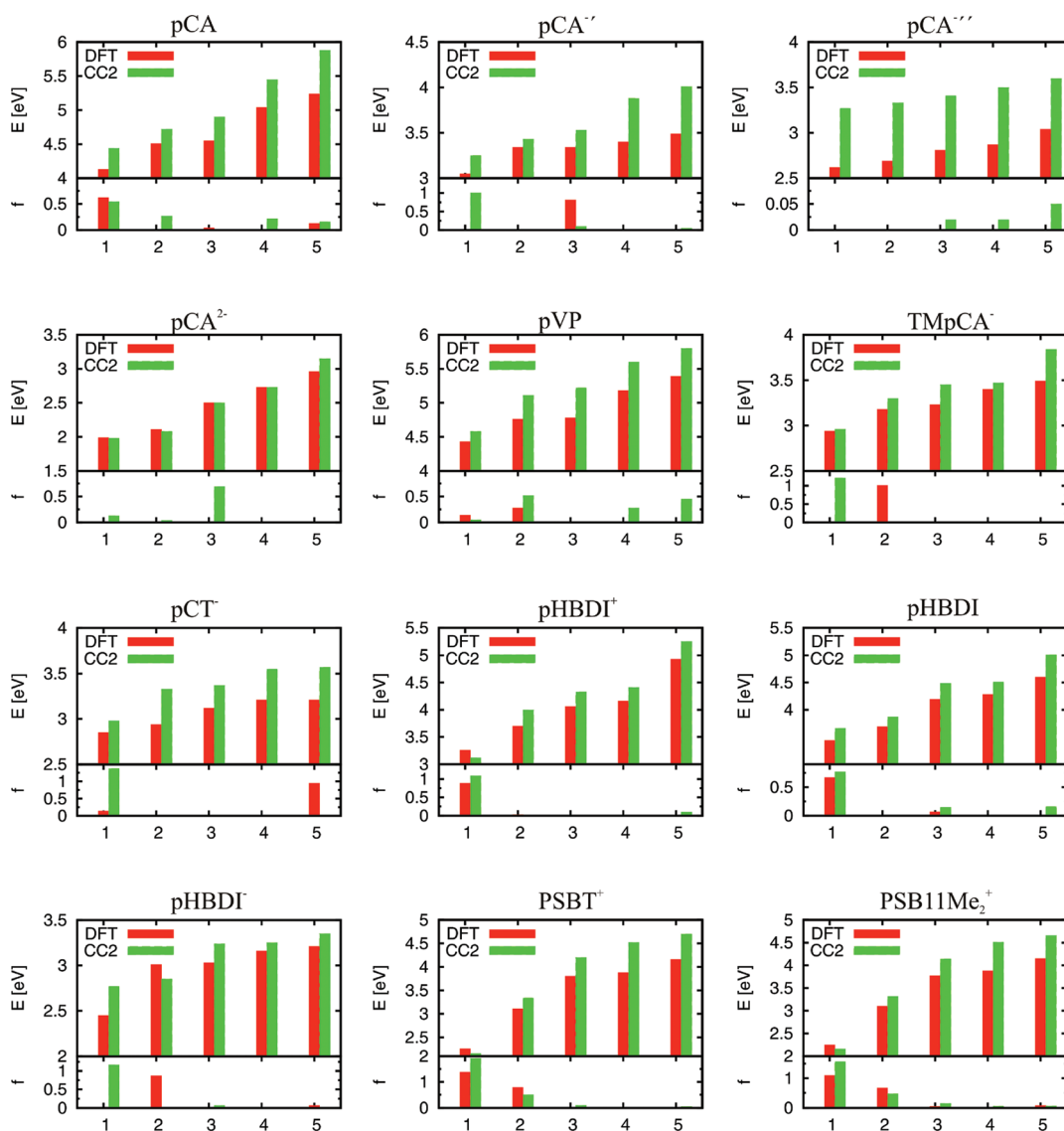


Figure 3. Comparison of the five lowest singlet excitation energies (in eV) and corresponding oscillator strengths calculated at the B3LYP and CC2 levels. All values are obtained using aug-TZVP basis sets. The integers on the x axis enumerate the excited states.

The excitation energies obtained for pCA'' at the pc-B3LYP and CC2 levels are in qualitative agreement. The 10 lowest excited states of pCA'' have a vanishing oscillator strength at the pc-B3LYP level. The external point charges increase the excitation energies of the chromophore in addition to producing a less dense spectrum. Still, at the pc-B3LYP level, the 10 lowest excitation energies are located in an energy interval of less than 1 eV, within 2.98 and 3.83 eV.

Similar results were obtained for the PYP chromophore models $TMpCA^-$ and pCT^- and the GFP chromophore model $pHBDI^-$. Upon point-charge stabilization, the first excitation energy remains nearly unchanged, while the higher excitation energies are blue-shifted. For these anions, the first excited state has the largest oscillator strength at the CC2 level, whereas B3LYP and pc-B3LYP calculations suggest that the second and fifth excited states are the strongly light-absorbing ones.

The first CC2 excitation energy of the studied chromophores is smaller than the orbital energy of the HOMO orbital as obtained in the HF SCF calculations. The only exception is pCA^{2-} , having a

positive HOMO energy of 0.22 eV, whereas pCA^- and $pHBDI^-$ have only one state below the IP obtained at the HF SCF level. For $TMpCA^-$, three excited states lie below the ionization threshold of the HF SCF calculation. The rest of the studied chromophores have all five excited states below the HF SCF ionization threshold. The HOMO energies obtained in the HF SCF calculations are given in Table 5. The ionization potentials obtained at the CC2 level are not expected to deviate much from the HF SCF ones. For pCA^- , the HOMO energy of 3.15 eV is 0.23 eV larger than the ionization threshold obtained at the coupled-cluster singles and doubles (CCSD) level.⁵⁶ The ionization and excitation thresholds of pCA^- (phenolate) and pCA'' (carboxylate) have very recently been discussed by Zuev et al.⁵⁶

pCA^{2-} . The doubly anionic PYP chromophore model pCA^{2-} was neutralized by four positive charges, placed 5 Å from the molecular plane above and below the two negatively charged substituents. At the pc-B3LYP level, the point charges shift the excitation energies by less than ± 0.2 eV, and the oscillator strengths of the five lowest states are very small. At the pc-CC2

Table 5. The Five Lowest Excitation Energies (E in eV) of the Biochromophore Models Calculated at the B3LYP/aug-TZVP Level^a

molecule		1a	2a	3a	4a	5a	HOMO ^b
pCA	<i>E</i>	4.13	4.51	4.55	5.04	5.24	−6.29
pCA	<i>f</i>	0.62	0.00	0.04	0.00	0.13	(−8.40)
pCA [−]	<i>E</i>	3.05	3.34	3.34	3.40	3.49	−1.39
pCA [−]	<i>f</i>	0.00	0.00	0.81	0.00	0.00	(−3.15)
pCA [−]	<i>E</i>	2.62	2.69	2.81	2.87	3.04	−1.49
pCA [−]	<i>f</i>	0.00	0.00	0.00	0.00	0.00	(−4.66)
pCA ^{2−}	<i>E</i>	1.99	2.11	2.50	2.73	2.96	+2.08
pCA ^{2−}	<i>f</i>	0.00	0.00	0.00	0.00	0.00	(+0.22)
pVP	<i>E</i>	4.43	4.76	4.78	5.18	5.39	−5.86
pVP	<i>f</i>	0.14	0.28	0.00	0.01	0.00	(−7.92)
TMpCA [−]	<i>E</i>	2.94	3.18	3.23	3.40	3.49	−1.50
TMpCA [−]	<i>f</i>	0.00	1.01	0.00	0.00	0.00	(−3.23)
pCT [−]	<i>E</i>	2.85	2.94	3.12	3.21	3.21	−1.71
pCT [−]	<i>f</i>	0.14	0.00	0.01	0.01	0.95	(−3.39)
pHBDI ⁺	<i>E</i>	3.26	3.70	4.06	4.16	4.93	−9.05
pHBDI ⁺	<i>f</i>	0.89	0.03	0.00	0.02	0.01	(−10.9)
pHBDI	<i>E</i>	3.44	3.69	4.19	4.28	4.60	−5.62
pHBDI	<i>f</i>	0.67	0.00	0.07	0.01	0.00	(−7.55)
pHBDI [−]	<i>E</i>	2.45	3.01	3.03	3.16	3.21	−1.30
pHBDI [−]	<i>f</i>	0.00	0.87	0.00	0.00	0.07	(−2.89)
PSBT ⁺	<i>E</i>	2.26	3.11	3.80	3.88	4.16	−8.02
PSBT ⁺	<i>f</i>	1.38	0.78	0.01	0.00	0.03	(−9.75)
PSB11Me ₂ ⁺	<i>E</i>	2.25	3.10	3.77	3.88	4.15	−8.10
PSB11Me ₂ ⁺	<i>f</i>	1.10	0.67	0.05	0.01	0.08	(−9.87)

^aThe molecular structures were optimized for the ground state at the MP2/TZVP level. The oscillator strengths (*f*) and the highest occupied molecular orbital energies (HOMO) obtained in the B3LYP and HF SCF calculations are also reported. ^bThe HOMO energies obtained at the HF SCF level are given within parentheses.

level, the lowest excitation energy is red-shifted as compared to the CC2 value, while the higher excited states are blue-shifted. The three lowest lying states have nonvanishing oscillator strengths at the CC2 level, with the third state being the brightest one. The external point charges effectively damp the intensity of the five lowest excited states, at both levels of theory. The absorption spectra for pCA^{2−} calculated at the pc-B3LYP and pc-CC2 levels agree qualitatively despite a difference between some excitation energies by more than 0.4 eV

Comparison to MBGFT Calculations. The chromophores of PYP, GFP, and rhodopsin were recently studied with a MBGFT method by Ma and co-workers.²⁷ Uncontracted, universal, and almost even-tempered basis sets consisting of 40 primitive Gaussian basis functions of the *s*, *p*, and *d* types for the heavier atoms were employed in the calculations. An analogous basis set consisting of 30 basis functions was used for hydrogen. The basis-set convergence was assessed by adding basis functions in the π -electron region centered above and below bond centers.²⁷ Unfortunately, the accuracy and reliability of such nonstandard basis sets are not well established. The MBGFT method might treat molecular excited states more accurately than, e.g., TDDFT with today's functionals, which, however, cannot be assessed prior to a more systematic study.

Ma et al. reported MBGFT vacuum excitation energies for the chromophore models of PYP, GFP, and rhodopsin, with

Table 6. Comparison of Excitation Energies (in eV) and Oscillator Strengths of the Anionic Chromophores Calculated at the B3LYP/aug-TZVP and pc-B3LYP/aug-TZVP Levels^a

molecule	level and point charge		1a	2a	3a	4a	5a
pCA [−]	B3LYP	+1	<i>E</i> 2.79	3.42	3.58	3.66	3.75
pCA [−]	B3LYP	+1	<i>f</i> 0.00	0.77	0.00	0.00	0.00
pCA [−]	B3LYP		<i>E</i> 3.05	3.34	3.34	3.40	3.49
pCA [−]	B3LYP		<i>f</i> 0.00	0.00	0.81	0.00	0.00
pCA [−]	B3LYP	+1	<i>E</i> 2.98	3.14	3.40	3.58	3.62
pCA [−]	B3LYP	+1	<i>f</i> 0.00	0.00	0.00	0.01	0.00
pCA [−]	B3LYP		<i>E</i> 2.62	2.69	2.81	2.87	3.04
pCA [−]	B3LYP		<i>f</i> 0.00	0.00	0.00	0.00	0.00
pCA ^{2−}	B3LYP	+2	<i>E</i> 1.81	2.31	2.68	2.76	2.83
pCA ^{2−}	B3LYP	+2	<i>f</i> 0.00	0.00	0.00	0.00	0.00
pCA ^{2−}	B3LYP		<i>E</i> 1.99	2.11	2.50	2.73	2.96
pCA ^{2−}	B3LYP		<i>f</i> 0.00	0.00	0.00	0.00	0.00
pCA ^{2−}	CC2	+2	<i>E</i> 1.86	2.59	2.83	2.91	3.27
pCA ^{2−}	CC2	+2	<i>f</i> 0.00	0.00	0.00	0.01	0.00
pCA ^{2−}	CC2		<i>E</i> 1.98	2.08	2.50	2.73	3.15
pCA ^{2−}	CC2		<i>f</i> 0.13	0.04	0.69	0.00	0.00
TMpCA [−]	B3LYP	+1	<i>E</i> 2.90	3.26	3.49	3.70	3.75
TMpCA [−]	B3LYP	+1	<i>f</i> 0.00	0.99	0.00	0.00	0.00
TMpCA [−]	B3LYP		<i>E</i> 2.94	3.18	3.23	3.40	3.49
TMpCA [−]	B3LYP		<i>f</i> 0.00	1.01	0.00	0.00	0.00
pCT [−]	B3LYP	+1	<i>E</i> 3.00	3.22	3.46	3.56	3.81
pCT [−]	B3LYP	+1	<i>f</i> 0.00	1.04	0.00	0.00	0.00
pCT [−]	B3LYP		<i>E</i> 2.85	2.94	3.12	3.21	3.21
pCT [−]	B3LYP		<i>f</i> 0.14	0.00	0.01	0.01	0.95
pHBDI [−]	B3LYP	+1	<i>E</i> 2.73	3.06	3.15	3.39	3.56
pHBDI [−]	B3LYP	+1	<i>f</i> 0.00	0.93	0.00	0.00	0.00
pHBDI [−]	B3LYP		<i>E</i> 2.45	3.01	3.03	3.16	3.21
pHBDI [−]	B3LYP		<i>f</i> 0.00	0.87	0.00	0.00	0.07

^aThe CC2/aug-TZVP and pc-CC2/aug-TZVP values for pCA^{2−} are also reported.

deviations from experimental values of only 0.01–0.17 eV,²⁷ which is a much higher precision than obtained at any other computational level. However, some of the available experimental excitation energies are not obtained from gas-phase measurements but determined from measurements in solution or a protein environment. For example, the excitation energies of the GFP chromophore pHBDI measured in solvent (3.51 eV) and in protein (3.12 eV) differ significantly.^{57,58} A recent calculation at the complete-active-space second-order perturbation theory (CASPT2) level using the newer zero-order Hamiltonian yields a lowest excitation energy of 3.58 eV,¹⁹ which agrees well with the present CC2 value of 3.63 eV. This suggests that the experimental gas-phase value should be closer to 3.51 eV as obtained in solution than to 3.12 eV as measured in the protein. Thus, the excitation energy of 3.17 eV obtained in the MBGFT calculation is most likely 0.3–0.4 eV too small. Similar discrepancies might occur for the other chromophores, for which gas-phase measurements are not available, suggesting that the MBGFT approach might be less accurate than reported by Ma and co-workers.²⁷

The B3LYP excitation energies are 0.10–0.33 eV higher than the corresponding MBGFT values, except for the anionic pHBDI[−] chromophore model, where B3LYP is 0.22 eV below

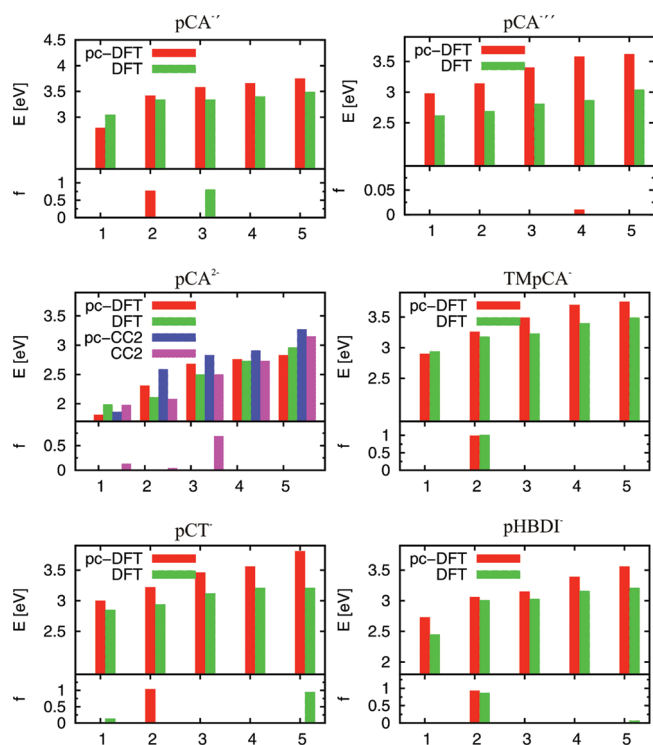


Figure 4. Comparison of the excitation energies (in eV) and the corresponding oscillator strengths of the anionic chromophores calculated at the B3LYP level with those obtained with point-charge stabilization. The corresponding CC2 energies for pCA^{2-} are also shown. All values are obtained using aug-TZVP basis sets. The integers on the x axis enumerate the excited states.

the MBGFT value. However, as discussed above, the anionic chromophore suffers at the TDDFT level from the continuum problem. An external neutralizing point charge increases the TDDFT excitation energy to 2.73 eV, which is 0.06 eV larger than the excitation energy predicted by MBGFT. The pc-B3LYP excitation energies are 0.06–0.25 eV higher than the corresponding MBGFT energies, while the CC2 excitation energies are 0.06–0.45 eV larger than the corresponding MBGFT energies.

Comparison to CASPT2 Calculations. *General Remarks.* Complete-active-space self-consistent-field (CASSCF) calculations combined with CASPT2 energy corrections have for a long time been the prime tool for studying ground and excited states of biochromophores. However, recent studies show that a number of older CASPT2 results are of limited value because the basis-set convergence has not been extensively assessed and the employed zeroth-order Hamiltonian was less accurate than originally anticipated. The recently introduced zeroth-order Hamiltonian in CASPT2 incorporates an empirical parameter and increases average excitation energies by 0.3 eV.^{13,18,59} The CASPT2 excitation energies obtained using the new zeroth-order Hamiltonian seem to be in good agreement with experimental results provided that the employed basis sets are large enough. Valsson and Filippi¹⁸ recently showed that the use of ground-state CASSCF optimized structures in CASPT2 studies results in artifacts when the excitation energies are sensitive to the geometry. The popular combination of CASSCF structures and single-point CASPT2 energies can thus lead to a qualitatively wrong character of the excited states, yielding results in complete disagreement with the consistent use of molecular structures and

excitation energies calculated at the CASPT2 level. Thus, many older CASSCF/CASPT2 studies of photochemical systems might have predicted incorrect reaction mechanisms.

Nevertheless, properly combined CASSCF and CASPT2 calculations are indispensable in studies of the excited state, when the reference state or the excited state has significant multiconfiguration character. Such situations occur quite frequently in photochemical processes, especially when the wave packets are approaching conical intersections. For vertical excitations, the lowest excited states of the investigated chromophores can be rather accurately expressed by expanding them in the configuration state basis of all singly and doubly excited determinants from a single-reference Slater determinant. At the *ab initio* level, the contributions from the double excitations are important as they take correlation effects into account, whereas in the TDDFT calculation the correlation effects are considered by the employed density functional. For many of the studied chromophores, CASPT2 calculations of the vertical excitation energies are not expected to yield more accurate excitation energies than obtained at the CC2 level. The computational costs for CASPT2 calculations increase rapidly with the size of the active space, hampering the feasibility of the CASPT2 calculations on larger molecules. For the GFP and rhodopsin chromophores, the approximations in the CASPT2 model and the inevitable limitations of the CASPT2 calculations introduce uncertainties of at least the same magnitude as the CC2 approximation, whereas for the PYP chromophore high-order correlation effects are more significant.

PYP Chromophores. A larger discrepancy between CASPT2 and CC2 excitation energies was obtained for the methyl ester of the pCA^{-} chromophore, recently studied at the CASPT2 level by Coto et al.⁶⁰ They obtained an excitation energy of 2.54 eV, which is 0.34 eV smaller than the experimental value of 2.88 eV.²⁵ The choice of the zeroth-order Hamiltonian is not mentioned in their work.⁶⁰ Experimentally, the absorption maximum of the methyl ester and the corresponding acid deviate in absorbance by only 3 nm, corresponding to an excitation energy shift of ± 0.02 eV. Thus, the discrepancy between the CASPT2 and experimental results cannot be due to the methyl group. Recent CASPT2 calculations on pCA^{-} using the new standard IPEA shift of 0.25 eV as well as approximate third-order coupled-cluster (CC3) calculations yielded very similar excitation energies of 2.96 and 2.98 eV, respectively.⁵⁶ Depending on the choice of the zeroth-order Hamiltonian in the CASPT2 calculation, the excitation energies differ by about 0.3 eV. The present CC2 value of 3.19 eV calculated for the same chromophore model is 0.31 eV larger than the experimental value. The CC2 calculations overestimate the excitation energies for pCA^{-} and the other PYP chromophore models, whereas the older CASPT2 values for pCA^{-} are too small. The recent CASPT2 and CC3 excitation energies agree within 0.1 eV with experimental results.^{25,56}

GFP Chromophores. The GFP chromophore illustrates another example where CASPT2 and CC2 excitation energies agree well. The present CC2 calculation yields an excitation energy of 3.63 eV for pHBDI, which is in almost perfect agreement with the CASPT2 value of 3.58 eV.¹⁹ Gas-phase measurements of the excitation energies of pHBDI have not been reported, whereas the lowest excitation energy of pHBDI measured in solvents is 3.51 eV.⁵⁸ The corresponding excitation energy obtained for the chromophore in the protein is 3.12 eV,⁵⁷ indicating that the protein shift is much larger than the solvent shift.

Table 7. Comparison of CC2, B3LYP, pc-B3LYP, and Experimental Excitation Energies (in eV) of the Bright State of the Anions^a

molecule	CC2	B3LYP	pc-B3LYP	exp.	reference
pCA ⁻	3.19(1)	3.34(3)	3.42(2)	2.88	21
pCT ⁻	2.94(1)	2.85(1) ^b	3.22(2)	2.70	20
TMpCA ⁻	2.91(1)	3.18(2)	3.26(2)		
pHBDI ⁻	2.73(1)	3.01(2)	3.06(2)	2.59, 2.76	70, 71

^aThe order of the excited state is given within parentheses. The calculated values are extrapolated CC2 values from Tables 1, 2, and 3. The B3LYP and pc-B3LYP data are gathered from Table 6. ^bThe energy of the second bright state (the fifth state) at the B3LYP level is 3.21 eV.

Retinal Chromophores. The first excitation energy of isolated 11-*cis*-retinal at the CASPT2 level converged toward 2.28 eV prior to the gas-phase measurement.⁶¹ After the experimental value was reported, CASPT2 and CC2 calculations yielded excitation energies of 2.05 and 2.10 eV, which are in fair agreement with the experimental value of 2.03 eV.^{22,62,63} However, when the more recent zeroth-order Hamiltonian is employed, a CASPT2 excitation energy of 2.30 eV is obtained.¹⁸ This zeroth-order Hamiltonian was designed to eliminate the mean error in CASPT2 excitation energies, and Valsson and Filippi conclude that the experimental value corresponding to the vertical transition is located around 2.34 eV.¹⁸ The recent storage ring experiments do not render comparisons of experimental and computed retinal spectra easier.²⁶

This illustrates that discrepancies between CASPT2 and CC2 excitation energies are not necessary due to the multireference character of the wave function and must be separately determined by a more careful analysis. In addition, gas-phase measurements of electronic excitation spectra of biochromophores are indispensable²² and need to be interpreted with the help of simulated vibronic spectra. Unfortunately, in the case of retinals, such a simulation is currently not possible, as the excited state potential energy surface does not correspond to a harmonic potential.

Comparison to Experimental Results. General Trends. The B3LYP excitation energies seem to be in better agreement with experimental data than the CC2 ones, particularly for the smaller PYP chromophores. However, the oscillator strengths obtained at the B3LYP level indicate that the TDDFT calculations suffer from problems, most likely related to the previously discussed DFT continuum problem.^{5,55} For pVP, pCA, pCA⁻, and pCT⁻, the excitation energies obtained at the B3LYP level are indeed closer to the experimental values than the CC2 ones, whereas for pHBDI⁺, pHBDI⁻, PBST⁺, and PBSMe₂⁺, the CC2 excitation energies agree well with experimental results. For pCA⁻, TMpCA⁻, and pHBDI⁻, the first excited state is a spurious state at the B3LYP level. The excitation energies of the lowest bright states calculated at the CC2, B3LYP, and pc-B3LYP levels are compared to corresponding gas-phase values in Table 7. The comparison of the excitation energies for the first bright state of the anions shows that CC2 values are with deviations of -0.03 to +0.31 eV in much better agreement with experimental results than values obtained at the B3LYP (0.15–0.46 eV) and pc-B3LYP levels (0.30–0.52 eV). The range of the deviations is given in parentheses. For pCT⁻, the

Table 8. The Extrapolated CC2 Excitation Energies As Well As the Excitation Energies Calculated at the B3LYP/aug-TZVP Level Compared to Experimental Data^a

molecule	state	CC2	B3LYP	MBGFT	experiment	references
pVP	1	4.53	4.43	4.17	4.12	21
	2	4.76	4.76	4.60	4.75	21
pCA	1	4.39	4.13	3.94	4.04 ^b	68
	2	4.68	4.51	4.20	4.37 ^c	69
pCA ⁻	1	3.19	3.05 (2.79)	2.95	2.88	25
pCA ⁻	1	3.28 ^d	2.62 ^d (2.98)	4.37 ^e	4.36 ^f	20
pCA ²⁻	1	1.89 ^g (1.86)	1.99 (1.81)	3.73 ^h	3.69 ⁱ	20
TMpCA ⁻	1	2.91	2.94 (2.90)	2.80	2.78 ⁱ	66
pCT ⁻	1	2.94	2.85 (3.00)	2.75	2.70	20
pHBDI ⁺	1	3.10	3.26	2.93	2.84 ^j , 2.99	24, 58
pHBDI ⁻	1	2.73	2.45 (2.73)	2.67	2.59, 2.76	70, 71
pHBDI	1	3.63	3.44	3.17	3.51 ^j	58
PSBT ⁺	1	2.14	2.26	2.03	2.00	22, 26
	2	3.32	3.11	3.10	3.22	22
PSB11Me ₂ ⁺	1	2.15	2.25	2.04	2.03	22
	2	3.30	3.10	3.01	3.18	22

^aThe pc-B3LYP/aug-TZVP and pc-CC2/aug-TZVP values are given in parentheses. The excitation energies obtained at the many-body Green's function theory (MBGFT) level by Ma et al.²⁷ are also reported. All experimental values that have not been recorded in the gas phase are printed in italics. ^bFor the methyl ester, the experimental and CC2/aug-TZVP values are 4.06²¹ and 4.46 eV. ^cMeasured for the methyl ester, the CC2/aug-TZVP value is 4.73 eV. ^dThe excitation energy of the lowest excited state. The fourth excitation energy is 3.45 eV (CC2), 2.87 eV (B3LYP), and 3.58 eV (pc-B3LYP). ^eThe fourth excitation energy of the MBGFT calculation. ^fMeasured in solvents. The value for pHBDI, measured in the protein, is 3.12 eV.⁵⁷ ^gThe excitation energy of the second state is 2.00 eV (CC2), 2.11 eV (B3LYP), 2.59 eV (pc-CC2), and 2.31 eV (pc-B3LYP). ^hThe second excitation energy of the MBGFT calculation. ⁱMeasured in the protein. ^jVacuum value extrapolated from measurements in solution.

first and bright excited state at the B3LYP level is shifted up in energy when the molecule is stabilized by point charges. The first excited state of the anions is a dark state at the pc-B3LYP level, whereas at the CC2 level, the lowest excited state has large oscillator strengths.

PYP Chromophores. Theoretical prediction of spectra requires accurate determination of both excitation energies and intensities. However, the experimental spectrum is subject to vibrational broadening as well as intensity borrowing, which makes comparison of calculated and experimental spectra challenging.^{23,64,65} The calculated excitation energies of all PYP chromophore models are compared to experimental data in Table 8.

In the case of pVP, the experimental spectrum is obtained using two-photon ionization spectroscopy, implying that the band intensities cannot be compared to the present calculations.²¹ The strong dipole transitions obtained in the CC2 calculations appear at 5.04 and 5.5 eV, corresponding to a valence and a Rydberg transition, respectively. The two lowest experimental excitation energies are 4.12 and 4.75 eV. In the calculation, the two lowest transitions are weak and appear at 4.53 and 4.76 eV. The former is a valence band with a very small transition probability, whereas the latter state has significant Rydberg character, as suggested by the requirement of diffuse basis functions. At the B3LYP level, the lowest excitation energy of 4.43 eV is 0.3 eV larger than the experimental value, whereas the B3LYP and CC2 excitation energies of the second state are identical and in perfect agreement with experimental results.

For pCA, the two lowest excitation energies have large oscillator strengths, making the comparison to experimental results easier. The excitation energy of the second state is measured only for the methyl ester. However, CC2 calculations show that pCA and its methyl ester have similar excitation energies. The CC2 excitation energies of pCA are 4.39 and 4.68 eV as compared to the experimental values of 4.04 and 4.37 eV. The B3LYP excitation energies of 4.13 and 4.51 eV also agree well with experimental values.

For pCA⁻, the CC2 excitation energy of 3.19 eV is 0.3 eV larger than the experimental value. The B3LYP calculation without point charges yields the bright excitation at 3.34 eV. However, this value is larger than the HOMO energy for pCA⁻, indicating that the B3LYP calculation might suffer from ionization problems. The pc-B3LYP calculation yields an excitation energy of 3.42 eV for the bright state, which is the second excited state in the calculation. For pCA^{-''} and pCA²⁻, the pc-B3LYP and pc-CC2 calculations do not yield any bright states among the five lowest excited states. The calculated excitation energies for pCA²⁻ are 1–2 eV smaller than the experimental value of 3.69 eV, measured in the solvent.²⁰ The excitation energy of 3.67 eV obtained for pCA²⁻ at the CC2 level using the QZVP basis sets agrees well with the experimental result, suggesting that the solvent molecules mainly prevent the expansion of the electron density due to Coulomb repulsion. Thus, it seems to be possible to simulate the solvated dianion by omitting very diffuse basis functions in the CC2 calculation. The present study shows that CC2 generally overestimates the gas-phase values for the PYP chromophores by 0.24–0.39 eV.

The excitation energies for pCA^{-''}, pCA²⁻, and pHBDI cannot be directly compared to experimental values because they were obtained from measurements in solvents. The excitation energies for TMpCA⁻ and pHBDI were deduced from UV–vis spectra measured for the chromophores embedded in the protein. Comparison of the calculated excitation energies for TMpCA⁻ with the result from the measurement on the protein indicates that the protein shift is small. The CC2 and pc-B3LYP calculations yielded values of 2.91 and 2.90 eV as compared to the experimental value of 2.78 eV.⁶⁶ However, the first excited state at the B3LYP level has a very small oscillator strength. The first bright transition at the pc-B3LYP level occurs at 3.26 eV.

GFP Chromophores. Lammich et al. measured the electronic excitation spectra for pHBDI⁺ in the gas phase, since neutral molecules cannot be studied in storage ring experiments.²⁴ Their idea was that the cation might have very similar excitation energies as neutral pHBDI, because of the large distance between the positive charge and the light-absorbing part of the chromophore. The present calculations show that this assumption is wrong. The lowest excitation energies of pHBDI⁺ and pHBDI at the CC2 level differ by 0.53 eV, whereas the corresponding difference obtained at the B3LYP level is 0.18 eV. The good agreement between calculated and measured excitation energies for pHBDI⁺ suggests that the CC2 excitation energy obtained for pHBDI is also accurate, whereas the B3LYP value has a larger uncertainty. Furthermore, recent CASPT2 calculations on the pHBDI chromophore¹⁹ yield an excitation energy in close agreement with the present CC2 value. Comparisons of the calculated excitation energy for pHBDI with those measured in the protein and in solution indicate that the solvent shift of the excitation energy is very small, whereas the protein shift is about 0.4 eV.

Retinal Chromophores. The two lowest excitation energies for the retinal models, PSBT⁺ and PSB11Me₂⁺, have recently been measured in the gas phase.^{22,26} The combined experimental and

computational study suggests that the observed spectrum is a mixture of the ones for 6-*cis* and 6-*trans* retinal.²⁶ The present CC2 calculations on the 6-*cis* conformation yield values in very close agreement with the earlier experimental values,²² indicating that the two lowest states are well described by low-order excitations from the ground state. The CC2 calculations also indicate that the recent interpretation of the measured spectrum should be carefully investigated. The photochemical properties of the protonated Schiff base retinals significantly differ from those of polyenes.^{5,67}

The obtained CC2 excitation energies for the retinal models deviate less than 0.15 eV from experimental values. The two lowest excitation energies calculated at the B3LYP level agree within 0.08–0.26 eV with experimental values. The first excitation energy is slightly larger than that obtained experimentally, whereas the second excitation energy is smaller than the experimental value. Comparing gas-phase experiments to our extrapolated CC2/aug-QZVP results yields a good agreement for the rhodopsin and GFP chromophores with a maximum deviation of only 0.14 eV.

CONCLUSIONS

This work gives an overview on the performance of the CC2 and the B3LYP/TDDFT methods applied to excited states of chromophore model systems of the photoactive yellow protein (PYP), the green fluorescent protein (GFP), and rhodopsin. The CC2 method is a black box *ab initio* method that allows the assessment of basis set requirements for the five lowest singlet excited states of the chromophore models up to extrapolated aug-QZVP basis sets, thus yielding accurate benchmark data for comparison with experimental results.

The CC2 basis set requirements are fulfilled up to 0.15 eV at the triple- ζ level, except for the studied anions and for Rydberg states of the neutral chromophores. The excitation energies compare well to the experimental values, especially for the rhodopsin and GFP chromophore models. The excitation energies of the PYP chromophore models are not particularly well described at the CC2 level, with deviations up to 0.4 eV, when experimental gas-phase values are available.

B3LYP/TDDFT calculations show a similar performance with deviations of 0.2–0.3 eV from experimental results for the GFP and rhodopsin chromophore models. However, the anionic PYP chromophores are problematic at the B3LYP level, possibly due to the unphysically low ionization potential predicted at the DFT level. Stabilization of the anions with positive point charges seems to reduce this problem, while it still does not produce completely satisfactory results, when calculated intensities are taken into account.

The present work is the first compilation of accurately benchmarked excitation energies at the black box *ab initio* CC2 level, comprising the most important and experimentally most thoroughly investigated biochromophores. All molecules in our study are challenging for contemporary quantum chemical methods, and we provide an accurate benchmark with respect to all entering parameters. We hope that this work will serve as a reference point in the discussion of other computational methods where the assessment of parameter dependence is limited by computational costs.

AUTHOR INFORMATION

Corresponding Author

*E-mail: sundholm@chem.helsinki.fi.

ACKNOWLEDGMENT

We thank Claudia Filippi for helpful comments on the manuscript. This research has been supported by the Academy of Finland through its Centers of Excellence Programme 2006–2011. This work was also supported by the Center for Functional Nanostructures (CFN) of the Deutsche Forschungsgemeinschaft (DFG) within project C3.9, by the Sigrid Jusélius Foundation, and the HPC-EUROPA2 project (project number: 228398) with the support of the European Commission—Capacities Area—Research Infrastructures. CSC—the Finnish IT Center for Science is acknowledged for computer time and hosting this HPC-EUROPA2 project. V.R.I.K. acknowledges the European Molecular Biology Organization for Long-Term Fellowship, and the Intramural Research Program of the National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases for support.

REFERENCES

- Hellweg, A.; Grün, S. A.; Hättig, C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4119.
- Lehtonen, O.; Sundholm, D.; Vänskä, T. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4535.
- Send, R.; Sundholm, D. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2862.
- Lehtonen, O.; Sundholm, D. *J. Chem. Phys.* **2006**, *125*, 144314.
- Lehtonen, O.; Sundholm, D.; Send, R.; Johansson, M. P. *J. Chem. Phys.* **2009**, *131*, 024301.
- Kállay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 6841.
- Furche, F.; Rappoport, D. In *Computational Photochemistry, Computational and Theoretical Chemistry*; Olivucci, M., Ed.; Elsevier: Amsterdam, 2005; Ch. III, Vol. 16, pp 93–128.
- Rappoport, D.; Furche, F. In *Time-Dependent Density Functional Theory*; Marques, M. A. L., Ullrich, C. A., Nogueira, F., Rubio, A., Burke, K., Gross, E. K. U., Eds.; Springer-Verlag: New York, 2006; no. 706 in *Lecture Notes in Physics*, pp 337–354.
- Castro, A.; Marques, M. A. L.; Varsano, D.; Sottile, F.; Rubio, A. *C. R. Phys.* **2009**, *10*, 469.
- Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103.
- Caricato, M.; Trucks, G. W.; Frisch, M. J.; Wiberg, K. B. *J. Chem. Theory Comput.* **2010**, *6*, 370.
- Fabian, J. *Dyes Pigm.* **2010**, *84*, 36.
- Send, R.; Valsson, O.; Filippi, C. *J. Chem. Theory Comput.* **2011**, *7*, 444.
- Send, R.; Kaila, V. R. I.; Sundholm, D. *J. Chem. Phys.* **2011**, *134*, 214114.
- Kaila, V. R. I.; Send, R.; Sundholm, D., submitted.
- Send, R.; Sundholm, D. *J. Phys. Chem. A* **2007**, *111*, 27.
- Send, R.; Sundholm, D. *J. Mol. Model.* **2008**, *14*, 717.
- Valsson, O.; Filippi, C. *J. Chem. Theory Comput.* **2010**, *6*, 1275.
- Filippi, C.; Zaccheddu, M.; Buda, F. *J. Chem. Theory Comput.* **2009**, *5*, 2074.
- Nielsen, I. B.; Boyé-Péronne, S.; El Ghazaly, M. O. A.; Kristensen, M. B.; Nielsen, S. B.; Andersen, L. H. *Biophys. J.* **2005**, *89*, 2597.
- de Groot, M.; Buma, W. J.; Gromov, E. V.; Burghardt, I.; Köppel, H.; Cederbaum, L. S. *J. Chem. Phys.* **2006**, *125*, 204303.
- Nielsen, I. B.; Lammich, L.; Andersen, L. H. *Phys. Rev. Lett.* **2006**, *96*, 018304.
- Jankowiak, H. C.; Stuber, J. L.; Berger, R. *J. Chem. Phys.* **2007**, *127*, 234101.
- Lammich, L.; Petersen, M. A.; Nielsen, M. B.; Andersen, L. H. *Biophys. J.* **2007**, *92*, 201.
- Rocha-Rinza, T.; Christiansen, O.; Rajput, J.; Gopalan, A.; Rahbek, D. B.; Andersen, L. H.; Bochenkova, A. V.; Granovsky, A. A.; Bravaya, K. B.; Nemukhin, A. V.; Christiansen, K. L.; Nielsen, M. B. *J. Phys. Chem. A* **2009**, *113*, 9442.
- Rajput, J.; Rahbek, D.; Andersen, L.; Hirshfeld, A.; Sheves, M.; Altoé, P.; Orlandi, G.; Garavelli, M. *Angew. Chem., Int. Ed.* **2010**, *49*, 1790.
- Ma, Y.; Rohlfing, M.; Molteni, C. *J. Chem. Theory Comput.* **2010**, *6*, 257.
- Onida, G.; Reining, L.; Rubio, A. *Rev. Mod. Phys.* **2002**, *74*, 601.
- Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420.
- Epifanovsky, E.; Polyakov, I.; Grigorenko, B.; Nemukhin, A.; Krylov, A. I. *J. Chem. Theory Comput.* **2009**, *5*, 1895.
- Grimme, S.; Izgorodina, E. I. *Chem. Phys.* **2004**, *305*, 223.
- Rappoport, D.; Furche, F. *J. Chem. Phys.* **2005**, *122*, 064105.
- Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.
- Weigend, F.; Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *119*, 12753.
- Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- Rappoport, D.; Furche, F. *J. Chem. Phys.* **2010**, *133*, 134105.
- Weigend, F.; Häser, M. *Theor. Chem. Acc.* **1997**, *97*, 331.
- Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143.
- Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
- Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409.
- Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154.
- Köhn, A.; Hättig, C. *J. Chem. Phys.* **2003**, *119*, 5021.
- Hättig, C. *Adv. Quantum Chem.* **2005**, *50*, 37.
- Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454.
- Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2002**, *117*, 7433.
- Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2004**, *121*, 12772.
- Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165. Current version, see: <http://www.turbomole.com>.
- Hellingwerf, K. J.; Hendriks, J.; Gensch, T. *J. Phys. Chem. A* **2003**, *107*, 1082.
- Borgstahl, G. E. O.; Williams, D. R.; Getzoff, E. D. *Biochemistry* **1995**, *34*, 6278.
- Tsien, R. Y. *Annu. Rev. Biochem.* **1998**, *67*, 509.
- Wald, G. *Nature* **1968**, *219*, 800.
- Wald, G. *Science* **1968**, *162*, 230.
- Casida, M. E.; Jamorski, C.; Casida, K. C.; Salahub, D. R. *J. Chem. Phys.* **1998**, *108*, 4439.
- Zuev, D.; Bravaya, K. B.; Crawford, T. D.; Lindh, R.; Krylov, A. I. *J. Chem. Phys.* **2011**, *134*, 034310.
- Creemers, T. M. H.; Lock, A. J.; Subramaniam, V.; Jovin, T. M.; Völker, S. *Nat. Struct. Biol.* **1999**, *6*, 557.
- Dong, J.; Solntsev, K. M.; Tolbert, L. M. *J. Am. Chem. Soc.* **2006**, *128*, 12038.
- Ghigo, G.; Roos, B. O.; Malmqvist, P. Å. *Chem. Phys. Lett.* **2004**, *396*, 142.
- Coto, P. B.; Roca-Sanjuán, D.; Serrano-Andrés, L.; Martín-Pendás, A.; Martí, S.; Andrés, J. *J. Chem. Theory Comput.* **2009**, *5*, 3032.
- Cembran, A.; Bernardi, F.; Olivucci, M.; Garavelli, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6255.
- Send, R.; Sundholm, D. *J. Phys. Chem. A* **2007**, *111*, 8766.
- Sekharan, S.; Weingart, O.; Buss, V. *Biophys. J.* **2006**, *91*, L07.
- Köppel, H.; Domcke, W.; Cederbaum, L. S. *Adv. Chem. Phys.* **1984**, *57*, 59.
- Stanton, J. F. *J. Chem. Phys.* **2007**, *126*, 134309.
- Genick, U. K.; Soltis, S. M.; Kuhn, P.; Canestrelli, I. L.; Getzoff, E. D. *Nature* **1998**, *392*, 206.
- Send, R.; Sundholm, D.; Johansson, M. P.; Pawłowski, F. *J. Chem. Theory Comput.* **2009**, *5*, 2401.

- (68) Smolarek, S.; Vdovin, A.; Perrier, D. L.; Smit, J. P.; Drabbels, M.; Buma, W. J. *J. Am. Chem. Soc.* **2010**, *132*, 6315.
- (69) de Groot, M.; Gromov, E. V.; Köppel, H.; Buma, W. J. *J. Phys. Chem. B* **2008**, *112*, 4427.
- (70) Nielsen, S. B.; Lapierre, A.; Andersen, J. U.; Pedersen, U. V.; Tomita, S.; Andersen, L. H. *Phys. Rev. Lett.* **2001**, *87*, 228102.
- (71) Forbes, M. W.; Jockusch, R. A. *J. Am. Chem. Soc.* **2009**, *131*, 17038.

The Application of Constricted Variational Density Functional Theory to Excitations Involving Electron Transitions from Occupied Lone-Pair Orbitals to Virtual π^* Orbitals

Tom Ziegler,^{*,†} Mykhaylo Krykunov,[†] and John Cullen[‡]

[†]Department of Chemistry, University of Calgary, University Drive 2500, Calgary AB T2N-1N4, Canada

[‡]Department of Chemistry, University of Manitoba, Winnipeg MB, R3T-2N2, Canada

 Supporting Information

ABSTRACT: We have applied the constricted variational density functional method (CV(n)-DFT) to $n \rightarrow \pi^*$ transitions in which an electron is moved from an occupied lone-pair orbital n to a virtual π^* orbital. A total of 34 transitions involving 16 different compounds were considered using the local density approximation (LDA), Becke, three-parameter, Lee–Yang–Parr (B3LYP), and BHLYP functionals. The DFT-based results were compared to the “best estimates” (BE) from high-level ab initio calculations. With energy terms included to second order in the variational parameters (CV(2)-DFT), our theory is equivalent to the adiabatic version of time-dependent density functional theory (DFT). We find that calculated excitation energies for CV(2)-DFT using LDA and BHLYP differ substantially from BE with root-mean-square-deviations (rmsd) of 0.86 and 0.69 eV, respectively, whereas B3LYP affords an excellent fit with BE at rmsd = 0.18 eV. Resorting next to CV(∞)-DFT, where energy terms to all orders in the variational parameters are included, results in all three functionals in too high excitation energies with rmsd = 1.69, 1.14, and 0.93 eV for LDA, B3LYP, and BHLYP, respectively. Adding in orbital relaxation considerably improves the results with rmsd = 0.54, 0.30, and 0.48 eV for LDA, B3LYP, and BHLYP, respectively. It is concluded that CV(∞)-DFT with orbital relaxation is a robust method for which the accuracy is less functionally dependent than that of CV(2)-DFT or adiabatic TDDFT.

1. INTRODUCTION

Time-dependent DFT (TDDFT) in its adiabatic formulation^{1–7} has emerged as an efficient tool for the study of excited states. Extensive benchmarking of adiabatic TDDFT^{8–12} has revealed that the calculated excitation energies are in fair agreement with experiment. It is thus to be expected that adiabatic TD-DFT will be used increasingly as a reasonable compromise between accuracy and computational cost in many applications.^{8–12} However, the extensive benchmarking has revealed some systematic errors^{8–14} in the calculated excitation energies when use is made of the generalized gradient approximation (GGA) as well as the popular approximate hybrid functionals containing fractions of exact Hartree–Fock exchange. The largest deviations^{8–14} are found for transitions where electrons are moved between two separated regions of space [charge-transfer (CT) transitions] or between orbitals of different spatial extent (Rydberg transitions).^{10–14}

We have in some recent studies^{15–18} analyzed the reason for the deviations between experimental CT excitation energies and estimates obtained from TDDFT applications. It was found that the deviations for a large part can be traced back to the simple approach taken in standard TDDFT, where terms depending on the linear orbital response parameter set U only are kept to second order in U for the energy expression.^{15–18} While this simple linear response approach is adequate for the corresponding Hartree–Fock time-dependent formulation where self-interaction is absent,¹⁷ it is inadequate for TDDFT applied to most approximate functionals where self-interaction terms are present.^{15–18} In those cases, higher-order terms in U must be included into the energy expression.^{15–18}

With the intention of including higher-order terms, we have developed a constricted variational density functional approach (CV(n)-DFT) to the calculation of excitation energies and excited-state properties.^{18–20} In its second-order formulation (CV(2)-DFT) this theory coincides with adiabatic TD-DFT when use is made of the popular Tamm–Dancoff approximation²¹ within both theories. However, in the general formulation of CV(n)-DFT terms to any order in U^n are included.

The objective of the current study is to consider another category of excitations in addition to CT where electrons are moved between orbitals of different spatial extent, namely the $n \rightarrow \pi^*$ transitions. In this case an electron is excited from an in-plane doubly occupied lone-pair orbital n to a π^* orbital situated perpendicular to the molecular plane. Obviously there will be only a modest overlap between the densities of n and π^* , although it might be larger than the overlaps for the densities of π_A and π_B^* in the CT transition $\pi_A \rightarrow \pi_B^*$, where A and B might be several angstroms apart.

We shall in the present study apply the pure LDA²² functional as well as the hybrids²³ B3LYP and BHLYP to the $n \rightarrow \pi^*$ transitions. As these transitions in all the considered cases turned out to be true one-orbital transitions where an electron is transferred from $n \rightarrow \pi^*$, we shall also make use of the self-consistent field Δ SCF scheme,²⁴ where one carries out Kohn–Sham SCF calculations on the $(n)^2$ and $n\pi^*$ configurations. For the simple one-orbital $n \rightarrow \pi^*$ transitions studied here the Δ SCF

Received: April 15, 2011

Published: June 28, 2011

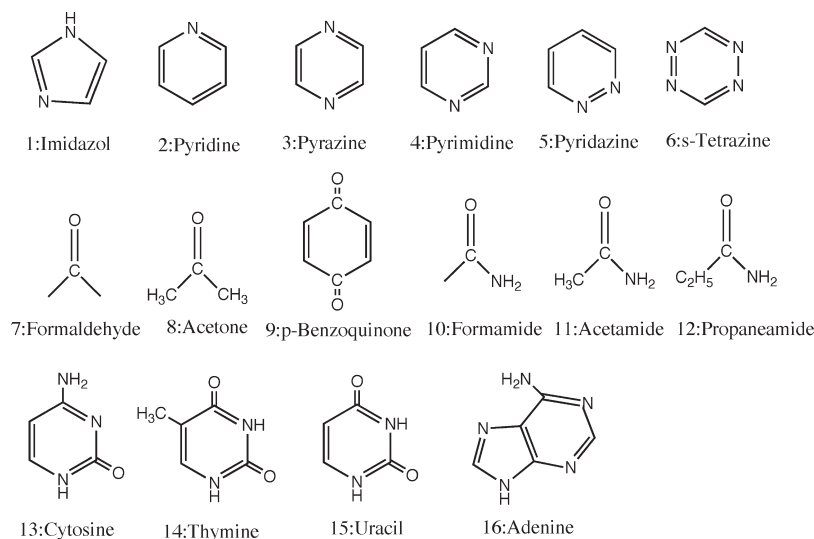


Figure 1. Molecules investigated in this study.

scheme can be considered as equivalent to the SCF-CV(∞)-DFT method²⁰ discussed in Section 3. We shall compare our calculated DFT-based excitation energies to those obtained in a recent benchmark study by Schreiber¹⁰ et al. based on high level ab initio methods¹⁰ and termed “best estimates” (BE) by the authors. The 16 molecules investigated here are shown in Figure 1.

2. COMPUTATIONAL DETAILS

All calculations were based on DFT as implemented in the ADF program version 2010.²⁵ Our calculations employed a standard triple- ζ Slater-type orbital (STO) basis with one set of polarization functions for all atoms.²⁶ Use was made of the local density approximation in the Vosko, Wilk, and Nusair (VWN) parameterization²² as well as the B3LYP and BHLYP hybrid functionals by Becke.²³ All electrons were considered as valence. The parameter (ACINT) for the precision of the numerical integration was set to 5.0. Use was made of a special auxiliary STO basis to fit the electron density in each cycle for an accurate representation of the exchange and the Coulomb potentials. All CV(2)-DFT¹⁸ and CV(∞)-DFT²⁰ calculations were carried out with a developers version of ADF-2010. The Cartesian coordinates of the 16 benchmark molecules shown in Figure 1 were taken from the Supporting Information of ref 10. The ground-state geometries of these molecules were optimized at the MP2/6-31G* level of theory.¹⁰

3. CV(N)-DFT

We have recently introduced a variational approach based on density functional theory for the description of excited states.^{18,20} In CV(n)-DFT we carry out a unitary transformation²⁰ among occupied $\{\varphi_i; i = 1, \text{occ}\}$ and virtual $\{\varphi_a; a = 1, \text{vir}\}$ ground-state orbitals:

$$Y \begin{pmatrix} \varphi_{\text{occ}} \\ \varphi_{\text{vir}} \end{pmatrix} = e^U \begin{pmatrix} \varphi_{\text{occ}} \\ \varphi_{\text{vir}} \end{pmatrix} = \left(\sum_{n=0}^{\infty} \frac{(U^2)^n}{2n!} \right) \begin{pmatrix} \varphi_{\text{occ}} \\ \varphi_{\text{vir}} \end{pmatrix} = \begin{pmatrix} \varphi'_{\text{occ}} \\ \varphi'_{\text{vir}} \end{pmatrix} \quad (1)$$

Here φ_{occ} and φ_{vir} are concatenated column vectors containing the sets $\{\varphi_i; i = 1, \text{occ}\}$ and $\{\varphi_a; a = 1, \text{vir}\}$, whereas φ'_{occ} and φ'_{vir}

are concatenated column vectors containing the resulting sets $\{\varphi'_i; i = 1, \text{occ}\}$ and $\{\varphi'_a; a = 1, \text{vir}\}$ of occupied and virtual excited-state orbitals, respectively. The unitary transformation matrix Y is in eq 1 expressed in terms of a skew symmetric matrix U as

$$Y = e^U = I + U + \frac{U^2}{2} + \cdots = \sum_{n=0}^{\infty} \frac{U^n}{n!} = \sum_{n=0}^{\infty} \frac{(U^2)^n}{2n!} + U \sum_{n=0}^{\infty} \frac{(U^2)^n}{(2n+1)!} \quad (2)$$

Here $U_{ij} = U_{ab} = 0$ where “ ij ” refer to the occupied set $\{\varphi_i; i = 1, \text{occ}\}$, whereas “ ab ” refer to $\{\varphi_a; a = 1, \text{vir}\}$. Further, U_{ai} is the variational mixing matrix elements that combines virtual and occupied ground-state orbitals in the excited state with $U_{ai} = -U_{ia}$. Thus, the entire matrix U is made up of $\text{occ} \times \text{vir}$ independent elements U_{ai} that also can be organized in the column vector \vec{U} . For a given U we can, with the help of eq 2, generate a set of “occupied” excited-state orbitals:

$$\varphi'_i = \sum_p^{\text{occ} + \text{vir}} Y_{pi} \varphi_p = \sum_j^{\text{occ}} Y_{ji} \varphi_j + \sum_a^{\text{vir}} Y_{ai} \varphi_a \quad (3)$$

that are orthonormal to any order in U_{ai} .

In the simple CV(2)-DFT theory,¹⁸ the unitary transformation of eq 2 is carried out to second order in U . We thus obtain the occupied excited-state orbitals to the second order as

$$\varphi'_i = \varphi_i + \sum_a^{\text{vir}} U_{ai} \varphi_a - \frac{1}{2} \sum_j^{\text{occ}} \sum_a^{\text{vir}} U_{ai} U_{aj} \varphi_j \quad (4)$$

from which we can generate the excited state Kohn–Sham density matrix to the second order as

$$\begin{aligned} \rho'(1, 1') &= \rho^{(0)}(1, 1') + \Delta\rho'(1, 1') = \rho^{(0)}(1, 1') \\ &+ \sum_i^{\text{occ}} \sum_a^{\text{vir}} U_{ai} \varphi_a(1') \varphi_i^*(1) + \sum_i^{\text{occ}} \sum_a^{\text{vir}} U_{ai}^* \varphi_a^*(1) \varphi_i(1') \\ &+ \sum_i^{\text{occ}} \sum_a^{\text{vir}} \sum_b^{\text{vir}} U_{ai}^* U_{bi} \varphi_a(1') \varphi_b^*(1) \\ &- \sum_i^{\text{occ}} \sum_j^{\text{occ}} \sum_a^{\text{vir}} U_{ai} U_{aj}^* \varphi_i(1') \varphi_j^*(1) \end{aligned} \quad (5)$$

Next the expression for $\rho'(1,1')$ makes it possible to write down the corresponding excited state Kohn–Sham energy to the second order as

$$E_{KS}[\rho'(1,1')] = E_{KS}[\rho^0] + \sum_{ai} U_{ai} U_{ai}^* (\varepsilon_a^0 - \varepsilon_i^0) + \sum_{ai} \sum_{bj} U_{ai} U_{bj}^* K_{ai,bj} + \frac{1}{2} \sum_{ai} \sum_{bj} U_{ai} U_{bj} K_{ai,jb} + \frac{1}{2} \sum_{ai} \sum_{bj} U_{ai}^* U_{bj}^* K_{ai,jb} + O[U^3] \quad (6)$$

Here $E_{KS}[\rho^0]$ is the ground-state energy and “ a,b ” run over virtual ground-state canonical orbitals, whereas “ i,j ” run over occupied ground-state canonical orbitals. Further

$$K_{nu,tq} = K_{nu,tq}^C + K_{nu,tq}^{XC} \quad (7)$$

where

$$K_{nu,tq}^C = \int \int \varphi_r^*(1) \varphi_u(1) \frac{1}{r_{12}} \varphi_t^*(2) \varphi_q(2) dv_1 dv_2 \quad (8)$$

whereas

$$K_{nu,tq}^{XC(HF)} = - \int \int \varphi_r^*(1) \varphi_q(1) \frac{1}{r_{12}} \varphi_t^*(2) \varphi_u(2) dv_1 dv_2 \quad (9)$$

for Hartree–Fock exchange correlation and

$$K_{nu,tq}^{XC(DFT)} = \delta(m_{sr}, m_{su}) \delta(m_{st}, m_{sq}) \int \varphi_r^*(\vec{r}_1) \varphi_u(\vec{r}_1) \times [f^{(m_{sr}, m_{st})}(\rho^0)] \varphi_t^*(\vec{r}_1) \varphi_q(\vec{r}_1) d\vec{r}_1 \quad (10)$$

for DFT exchange correlation. In eq 10 $m_{sr} = 1/2$ for a spin orbital $\varphi_r(1)$ of α -spin, whereas $m_{sr} = -1/2$ for a spin orbital $\varphi_r(1)$ of β -spin. In addition the kernel $f^{(\tau,v)}(\rho^0)$ is the second functional derivative of E_{XC} with respect to ρ_α and ρ_β :

$$f^{\tau,v}(\rho_\alpha^0, \rho_\beta^0) = \left(\frac{\delta^2 E_{XC}}{\delta \rho_\tau \delta \rho_v} \right)_0 \quad \tau = \alpha, \beta; \quad v = \alpha, \beta \quad (11)$$

Finally ε_i^0 and ε_a^0 in eq 6 are the ground-state orbital energies of respectively $\varphi_i(1)$ and $\varphi_a(1)$.

In CV(2)-DFT¹⁸ we seek points on the energy surface $E_{KS}[\rho']$ such that $\Delta E_{KS}[\Delta\rho'] = E_{KS}[\rho'] - E_{KS}[\rho^0]$ represents a transition energy. Obviously, a direct optimization of $\Delta E_{KS}[\Delta\rho']$ without constraints will result in $\Delta E_{KS}[\Delta\rho'] = 0$ and $U = 0$. We¹⁸ now introduce the constraint that the electron excitation must represent a change in density $\Delta\rho'$, where one electron in eq 5 is transferred from the occupied space represented by $\Delta\rho_{occ} = -\sum_{ija} U_{ai} U_{aj}^* \varphi_i(1') \varphi_j^*(1)$ to the virtual space represented by $\Delta\rho_{vir} = \sum_{iab} U_{ai} U_{bi}^* \varphi_a(1') \varphi_b^*(1)$. An integration of $\Delta\rho_{occ}$ and $\Delta\rho_{vir}$ over all space affords $-\Delta q_{occ} = \Delta q_{vir} = \sum_{ai} U_{ai} U_{ai}^*$. We shall thus introduce the constraint $\sum_{ai} U_{ai} U_{ai}^* = 1$. Constructing next the Lagrangian $L = E_{KS}[\rho'] + \lambda(1 - \sum_{ai} U_{ai} U_{ai}^*)$ with λ being a Lagrange multiplier and demanding that L be stationary to any real variation in U results in the eigenvalue equation:

$$(A^{KS} + B^{KS}) \vec{U}^{(I)} = \lambda_{(I)} \vec{U}^{(I)} \quad (12)$$

where

$$A_{ai,bj}^{KS} = \delta_{ab} \delta_{ij} (\varepsilon_a^0 - \varepsilon_i^0) + K_{ai,bj}^{KS}; \quad B_{ai,bj}^{KS} = K_{ai,jb}^{KS} \quad (13)$$

We can now from eq 12 determine the sets of mixing coefficients $\{\vec{U}^{(I)}; I = 1, occ \times vir\}$ that make L stationary and represent excited states. The corresponding excitation energies are given by

$\lambda_{(I)}$, as it can be seen by substituting $\vec{U}^{(I)}$ into eq 6 and making use of the constraint and normalization condition $\vec{U}^{(I)+} \vec{U}^{(I)} = 1$ after multiplying on both sides with $\vec{U}^{(I)+}$.

Within the Tamm–Dancoff approximation²¹ eq 12 reduces to

$$A^{KS} \vec{U}^{(I)} = \lambda_{(I)} \vec{U}^{(I)} \quad (14)$$

which is identical in form to the equation one obtains from TDDFT in its adiabatic formulation^{1–6} after applying the same Tamm–Dancoff²¹ approximation.

Having determined $\vec{U}^{(I)}$ from either eq 12 or 14 allows us²⁰ now to carry out the unitary transformation of eq 1 to all orders. The resulting occupied excited-state orbitals are given by²⁰

$$\phi_j' = \cos[\eta\gamma_j] \phi_j^o + \sin[\eta\gamma_j] \phi_j^v; \quad j = 1, occ \quad (15)$$

here ϕ_j^o and ϕ_j^v are according to the corresponding orbital theory of Hall and Amos²⁷ eigenvectors to, respectively, D_{occ} and D_{vir} with the same eigenvalues γ_i where $(D_{occ}^2)_{ij} = \sum_a^{vir} U_{ai} U_{aj}$ and $(D_{vir}^2)_{ab} = \sum_i^{occ} U_{ai} U_{bi}$. Here ϕ_j^o is a linear combination of occupied ground-state orbitals, whereas ϕ_j^v is a linear combinations of virtual ground-state orbitals. Thus in the corresponding orbital representation²⁷ only one occupied orbital ϕ_j^o mixes with one corresponding virtual orbital ϕ_j^v for each occupied excited-state orbital ϕ_j' when the unitary transformation is carried out to all orders according to eq 1. Martin²⁸ has used the representation of corresponding orbitals to analyze excitations described by TDDFT and TDHF. In his interesting analysis $\{\varphi_j^o(1), \varphi_j^v(1)\}$ are referred to as natural transition orbitals (NTO).

The change in the density matrix $\Delta\rho^{(\infty)}$ due to a one-electron excitation takes on the compact form of

$$\Delta\rho^{(\infty)}(1,1') = \sum_j^{occ} \sin^2[\eta\gamma_j] [\phi_j^v(1') \phi_j^v(1) - \phi_j^o(1') \phi_j^o(1)] + \sum_j^{occ} \sin[\eta\gamma_j] \cos[\eta\gamma_j] [\phi_j^v(1) \phi_j^o(1') + \phi_j^v(1') \phi_j^o(1)] \quad (16)$$

when the unitary transformation in eq 1 is carried out to all orders. In eqs 15 and 16 the scaling factor η is introduced to ensure that $\Delta\rho^{(\infty)}(1,1')$ represents the transfer of a single electron from the occupied orbital space density, $\sin^2[\eta\gamma_j] \phi_j^o(1') \phi_j^o(1)$ to the virtual orbital space density $\sum_j^{occ} \sin^2[\eta\gamma_j] \phi_j^v(1') \phi_j^v(1)$ or

$$\sum_j^{occ} \sin^2[\eta\gamma_j] = 1 \quad (17)$$

Here the constraint of eq 17 is a generalization of the corresponding second-order constraint $\sum_{ai} U_{ai} U_{ai}^* = 1$ used to derive eqs 13 and 14.

We finally get for the excitation energy including terms to all orders in U :

$$\Delta E^{(\infty)} = E_{KS}^{\infty}[\rho^0 + \Delta\rho^{(\infty)}] - E_{KS}[\rho^0] = - \sum_j^{occ} \sin^2[\eta\gamma_j] F_{j,j}^{KS}[\rho^0 + \frac{1}{2} \Delta\rho^{(\infty)}] + \sum_j^{occ} \sin^2[\eta\gamma_j] F_{j,j}^{KS}[\rho^0 + \frac{1}{2} \Delta\rho^{(\infty)}] + \sum_j^{occ} \cos[\eta\gamma_j] \sin[\eta\gamma_j] F_{j,j} \left[\rho^0 + \frac{1}{2} \Delta\rho^{(\infty)} \right] + \sum_j^{occ} \cos[\eta\gamma_j] \sin[\eta\gamma_j] F_{j,j} \left[\rho^0 + \frac{1}{2} \Delta\rho^{(\infty)} \right] + O^{[3]}(\Delta\rho^{(\infty)}) \quad (18)$$

Here eq 18 is derived by Taylor expanding²⁹ $E_{KS}^{\infty}[\rho^0 + \Delta\rho^{(\infty)}]$ and $E_{KS}[\rho^0]$ from the common intermediate density $\rho^0 + 1/2 \Delta\rho^{(\infty)}$. Further, $F_{j,j}^{KS}[\rho^0 + 1/2 \Delta\rho^{(\infty)}]$ is the Kohn–Sham Fock

Table 1. Vertical $n \rightarrow \pi^*$ Singlet Excitation Energies^a

molecule	state	best ^b	LDA(VWN)			B3LYP		
			CV(2) ^c	CV(∞) ^d	Δ SCF	CV(2) ^c	CV(∞) ^d	Δ SCF ^e
imidazole	A''	6.81	5.79	9.51	6.59	6.53	8.01	6.47
pyridine	B ₁	4.59	4.30	6.30	4.58	4.92	5.91	4.69
	A ₂	5.11	4.35	7.22	4.94	5.17	7.13	5.15
pyrazine	B _{3u}	3.95	3.52	3.68	3.45	4.09	3.94	3.85
	A _u	4.81	3.91	5.01	4.16	4.74	5.40	4.63
	B _{2g}	5.56	5.03	5.59	5.10	5.67	5.72	5.48
	B _{1g}	6.6	5.40	7.40	5.87	6.40	7.84	6.38
pyrimidine	B ₁	4.55	3.73	4.61	3.87	4.37	4.81	4.14
	A ₂	4.91	3.93	5.22	4.21	4.68	5.35	4.54
pyridazine	B ₁	3.78	3.10	4.63	3.29	3.74	4.38	3.55
	A ₂	4.31	3.41	5.49	3.84	4.26	5.62	4.15
	A ₂	5.77	4.97	5.78	5.06	5.55	5.80	5.35
s-tetrazine	B _{3u}	2.29	1.83	2.04	1.75	2.41	2.29	2.15
	A _u	3.51	2.73	3.58	2.95	3.59	3.98	3.49
	B _{1g}	4.73	4.01	4.27	3.91	4.88	4.72	4.56
	A _u	5.5	4.55	4.78	4.50	5.20	5.06	4.97
	B _{2g}	5.2	4.72	5.04	4.82	5.40	5.18	5.17
formaldehyde	A ₂	3.88	3.64	4.65	3.73	3.93	4.48	3.52
acetone	A ₂	4.4	4.16	5.44	4.33	4.41	5.11	4.02
p-benzoquinone	B _{1g}	2.76	1.86	2.18	1.88	2.54	2.55	2.40
	A _u	2.77	2.00	2.62	2.11	2.69	2.90	2.55
	B _{3u}	5.64	4.29	6.36	4.67	5.47	6.69	5.40
formamide	A''	5.63	5.33	7.23	5.66	5.58	6.75	5.28
acetamide	A''	5.69	5.31	7.33	5.67	5.59	6.84	5.31
propanamide	A''	5.72	5.34	7.30	5.67	5.60	6.82	5.34
cytosine	A''	4.87	3.74	10.04	4.64	4.78	7.49	4.83
	A''	5.26	4.41	7.89	5.12	5.17	7.19	
thymine	A''	4.82	4.04	6.95	4.57	4.74	6.64	5.59
	A''	6.16	4.75	9.07	6.06	5.83	8.43	6.15
uracil	A''	4.8	3.91	7.22	4.52	4.66	6.73	4.54
	A''	6.1	4.69	9.26	6.03	5.75	7.98	6.07
	A''	6.56	5.15	8.84	5.58	6.14	6.75	
adenine	A''	5.12	4.22	5.78	4.47	5.01	5.73	4.91
	A''	5.75	5.00	6.01	5.16	5.69	5.86	
rmsd($n \rightarrow \pi^*$)			0.86	1.69	0.54	0.18	1.14	0.30

^aIn eV. ^bTheoretical best estimates from ref 10. ^cCV(2)-DFT singlet transition energy from eq 19 where the Tamm–Dancoff approximation²¹ has been used. ^dCV(∞)-DFT singlet transition energy from eq 22. ^e Δ SCF^{24,30} excitation energies from eq 23.

operator defined with respect to the intermediate Kohn–Sham density matrix $\rho^0 + 1/2\Delta\rho^{(\infty)}$, whereas $F_{pq}^{KS}[\rho^0 + 1/2\Delta\rho^{(\infty)}]$ is a matrix element of this operator involving the two orbitals φ_p, φ_q . The expression in eq 18 is exact to the third order in $\Delta\rho^{(\infty)}$, which is usually enough.²⁹ However its accuracy can be extended to any desired order in $\Delta\rho^{(\infty)}$.²⁹

The energy expression in eq 18 is perturbative in the sense that we make use of a U matrix optimized with respect to the second-order energy expression of eq 6. We refer to this method as CV(∞)-DFT.²⁰ We might alternatively optimize U directly with respect to the energy expression of eq 18 in a self-consistent fashion. Such a procedure termed SCF-CV(∞)-DFT has been formulated²⁰ but not yet implemented. We shall in the present study of $n \rightarrow \pi^*$ transitions make use of the Δ SCF scheme^{24,30} which is equivalent to SCF-CV(∞)-DFT when, as in the current case, the excitation can be described by a single orbital transition.

4. RESULTS AND DISCUSSION

We sample in Table 1 the calculated $n \rightarrow \pi^*$ excitation energies based on LDA and B3LYP for 34 transitions. For each functional, results are given for CV(2)-DFT,¹⁸ CV(∞)-DFT,²⁰ and Δ SCF.^{24,30} We compare further the DFT-based results with

the “best estimate” by Schreiber¹⁰ et al. based on high-level ab initio wave function methods. Please note that the CV(2)-DFT results in Table 1 are identical to those based on TD-DFT after use has been made of the Tamm–Dancoff approximation.²¹

LDA. The LDA results compiled in Table 1 are also shown in Figure 2 where we plot the calculated excitation energies due to CV(2)-DFT, CV(∞)-DFT and Δ SCF relative to BE.¹⁰

In the single orbital transition $n \rightarrow \pi^*$ we can represent the resulting singlet excited states as $\Psi_{n \rightarrow \pi^*}^S = 1/(2)^{1/2}[|n\bar{n}^*| + |\pi^*\bar{n}|]$. In adiabatic TDDFT or CV(2)-DFT the corresponding singlet transition energy is given by^{2,15}

$$\Delta E_S^{(2)} = \varepsilon_{\pi^*} - \varepsilon_n + 2K_{n\pi^*, n\pi^*} - K_{n\pi^*, \pi^*n} \quad (19)$$

within the Tamm–Dancoff²¹ approximation.

We see from Figure 1 where we plot $\Delta E_S^{(2)}$ relative to BE¹⁰ that LDA systematically underestimates the excitation energies at the CV(2)-DFT level with a rmsd of 0.86 eV, see Table 1. This can be attributed to the fact that the highest occupied molecular orbital–lowest unoccupied molecular orbital (HOMO–LUMO) gap $\varepsilon_{\pi^*} - \varepsilon_n$ for pure functionals typically is much smaller than the HOMO to LUMO excitation energy. Further, $K_{n\pi^*, n\pi^*}$, $K_{n\pi^*, \pi^*n}$ although positive are rather small, since n and π^* are in different

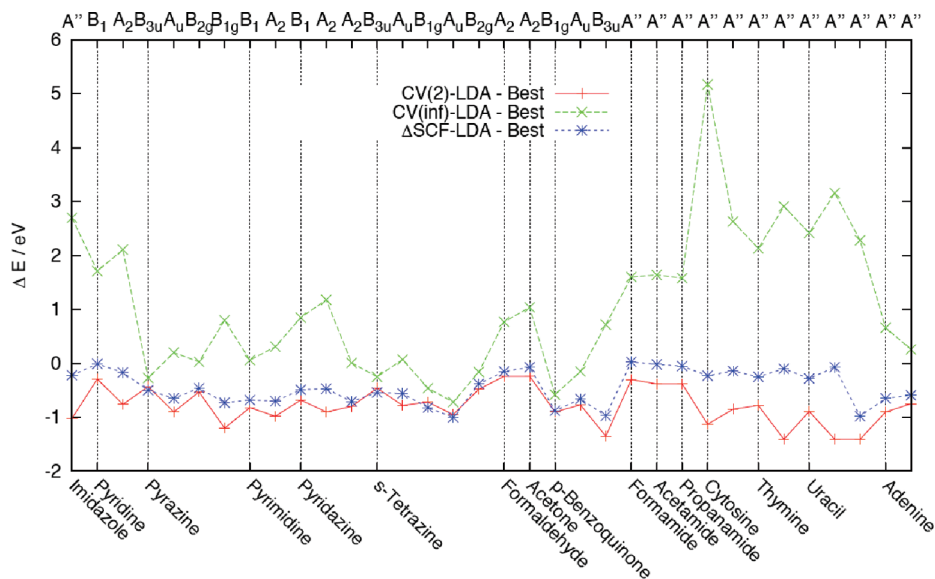


Figure 2. Difference between CV(2)-LDA, CV(∞)-LDA, ΔSCF-LDA, and best estimate¹⁰.

regions. The situation is quite similar to that of the $\pi_A \rightarrow \pi_B^*$ CT transitions, where pure functionals also underestimate the excitation energies at the CV(2)-DFT level.¹⁵ In that case the underestimation is even larger as $K_{\pi_A \pi_B^*, \pi_A \pi_B^*}$ is negligible due to the physical separation of A and B.

Turning next to the CV(∞)-DFT level of theory, we have in the case of the single orbital transition $n_G \rightarrow \pi^*$ one set of corresponding orbitals ($n\pi^*$) for which $\gamma = 1$, whereas $\gamma \approx 0$ for all other corresponding pairs. In this case eq 16 affords $\Delta\rho^{(\infty)} = \pi^*(1)\pi^*(1) - n(1)n(1')$, where we have made use of eq 17. The related excitation energy of eq 18 gives

$$\begin{aligned} \Delta E_{n \rightarrow \pi^*}^{(\infty)} &= E^{KS}[\pi^* \bar{n}] - E^{KS}[n \bar{n}] \\ &= \varepsilon_{\pi^*} - \varepsilon_n + 1/2K_{m, m} + 1/2K_{\pi^* \pi^*, \pi^* \pi^*} - K_{m, \pi^* \pi^*} \end{aligned} \quad (20)$$

where $E^{KS}[n \bar{n}]$ and $E^{KS}[\pi^* \bar{n}]$ are the KS energies of the ground-state determinant $|n \bar{n}\rangle$ and excited-state determinant $|\pi^* \bar{n}\rangle$, respectively. For the singlet excitation energy corresponding to $\Psi_{n \rightarrow \pi^*}^S = 1/(2)^{1/2} [|n \bar{\pi}^* \rangle + | \pi^* \bar{n} \rangle]$, use can be made of the Slater sum rules²⁴ to obtain

$$\begin{aligned} \Delta E_S^{(\infty)} &= 2E^{KS}[n \bar{\pi}^*] - E^{KS}[n \pi^*] - E^{KS}[n \bar{n}] \\ &= 2\Delta E_{n \rightarrow \pi^*}^{(\infty)} - \Delta E_{n \rightarrow \pi^*}^{(\infty)} \end{aligned} \quad (21)$$

We thus get from eq 20 and the corresponding expression for $E_{n \rightarrow \pi^*}^{(\infty)}$ that

$$\begin{aligned} \Delta E_S^{(\infty)} &= \varepsilon_{\pi^*} - \varepsilon_n + 1/2K_{m, m} + 1/2K_{\pi^* \pi^*, \pi^* \pi^*} \\ &\quad + 2K_{m, \pi^* \pi^*} - K_{m, \pi^* \pi^*} \end{aligned} \quad (22)$$

The expression for $\Delta E_S^{(2)}$ of eq 19 appears to be quite different from $\Delta E_S^{(\infty)}$ of eq 22. However, we note that if use is made of pure Hartree–Fock exchange then $K_{\pi^* \pi^*, \pi^* \pi^*}^{HF} = K_{m, m}^{HF} = 0$, whereas $K_{m, \pi^* \pi^*}^{HF} = -K_{\pi^* \pi^*, m}^{HF}$ and $K_{m, \pi^* \pi^*}^{HF} = -K_{\pi^* \pi^*, m}^{HF}$. Thus, in this case $\Delta E_S^{(2)HF} = \Delta E_S^{(\infty)HF}$. As a consequence, for Hartree–Fock the excitation energy of a single orbital transition is fully determined by $\Delta E_S^{(2)HF}$, and higher-order terms are zero. However, this is not the case for LDA, where $K_{\pi^* \pi^*, \pi^* \pi^*}^{LDA} \neq K_{m, m}^{LDA} \neq 0$, whereas $K_{m, \pi^* \pi^*}^{LDA} \neq -K_{\pi^* \pi^*, m}^{LDA}$ and $K_{m, \pi^* \pi^*}^{LDA} \neq -K_{\pi^* \pi^*, m}^{LDA}$. In this case the calculated excitation energies

are quite different depending on whether $\Delta E_S^{(2)}$ or $\Delta E_S^{(\infty)}$ is used. This is also the case for functionals based on the GGA.

The calculated singlet energies $\Delta E_S^{(\infty)LDA}$ are plotted in Figure 2 relative to the BE.¹⁰ It follows from the figure that CV(∞)-DFT in general affords too high excitation energies with a rmsd of 1.69 eV. This is opposite to CV(2)-DFT, where $\Delta E_S^{(2)LDA}$ systematically was too low. The contributions in eq 22 responsible for the high excitation energies are $K_{\pi^* \pi^*, \pi^* \pi^*}^{LDA}$ and especially $K_{m, m}^{LDA}$ which are both positive as the Coulomb contributions $K_{\pi^* \pi^*, \pi^* \pi^*}^C, K_{m, m}^C$ in absolute terms and are larger than the exchange contributions $K_{\pi^* \pi^*, \pi^* \pi^*}^{XC(LDA)}, K_{m, m}^{XC(LDA)}$. The lack of cancellation between Coulomb and exchange terms has been termed “self-interaction error” because the cancellation must be complete for a one-electron system. However, cancellation is not required to apply for many-electron systems where the term self-interaction error might be a misnomer.

It is perhaps not surprising that $\Delta E_S^{(\infty)LDA}$ is too high when we note in eq 20 that the orbital set used is optimized with respect to the ground state $|n \bar{n}\rangle$ but not with respect to $|n \pi^*\rangle$ and $|\pi^* \bar{n}\rangle$. A more balanced description would result if the orbital sets were optimized separately for the ground state and each excited state, as suggested in the SCF-CV(∞)-DFT theory.²⁰ Such an optimization becomes especially simple for one-orbital excitations, such as $n \rightarrow \pi^*$, where we can employ the ΔSCF procedure^{24,30} and write

$$\Delta E_S^{(\Delta SCF)} = 2E_{SCF}^{KS}[n \bar{\pi}^*] - E_{SCF}^{KS}[n \pi^*] - E_{SCF}^{KS}[n \bar{n}] \quad (23)$$

In eq 23 $E_{SCF}^{KS}[n \bar{\pi}^*]$, $E_{SCF}^{KS}[n \pi^*]$ are the energies from SCF KS calculations on the configurations $n \bar{\pi}^*$ and $n \pi^*$, respectively, whereas $E_{SCF}^{KS}[n \bar{n}]$ is the SCF KS ground-state energy. We plot in Figure 2 $\Delta E_S^{(\Delta SCF)LDA}$ relative to BE.¹⁰ It is apparent that $\Delta E_S^{(\Delta SCF)LDA}$ represents a clear improvement over $\Delta E_S^{(2)LDA}$ and especially $\Delta E_S^{(\infty)LDA}$, with a rmsd of 0.54. It is interesting to note that $\Delta E_S^{(\Delta SCF)LDA}$ gives rise to a considerable lowering in energy compared to $\Delta E_S^{(\infty)LDA}$ for those cases (1, 2, and 10–15 of Figure 1), where $\Delta E_S^{(\infty)LDA}$ is much too high, whereas $\Delta E_S^{(\Delta SCF)LDA}$ hardly changes compared to $\Delta E_S^{(\infty)LDA}$ in those cases (3, 6, and 9 of Figure 1), where $\Delta E_S^{(\infty)LDA}$ is close to BE. Obviously for the pure functional LDA the ΔSCF procedure^{24,30} is the more accurate procedure. We note further that (1, 2, and

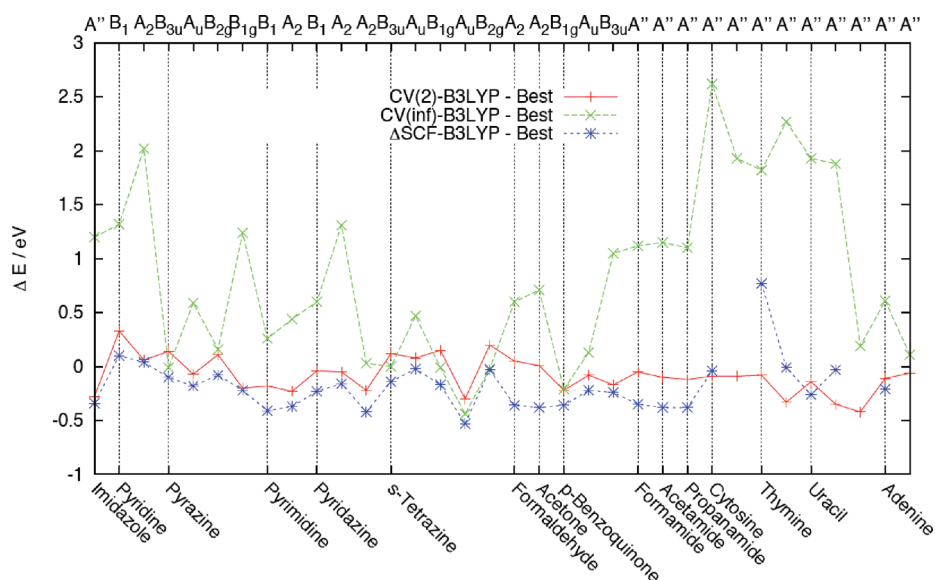


Figure 3. Difference between CV(2)-B3LYP, CV(∞)-B3LYP, Δ SCF-B3LYP, and BE.¹⁰

10–15) correspond to systems where n is localized on a single heteroatom leading to high values for $K_{nn,nn}^{\text{LDA}}$ whereas n in (3, 6, and 9) is delocalized over two or more heteroatoms resulting in much smaller values for $K_{nn,nn}^{\text{LDA}}$.

B3LYP. The B3LYP excitation energies are all given in Table 1. We plot further the calculated excitation energies due to CV(2)-DFT, CV(∞)-DFT, and Δ SCF relative to BE¹⁰ in Figure 3. It is to be noted that $\Delta E_S^{(2)\text{B3LYP}}$ for the various molecules is of a higher energy than $\Delta E_S^{(2)\text{LDA}}$. It is further in much better agreement with BE¹⁰ affording a rmsd of only 0.18 eV, Table 1. We attribute the increase in $\Delta E_S^{(2)\text{B3LYP}}$ compared to $\Delta E_S^{(2)\text{LDA}}$ to the larger gap in $\varepsilon_{\pi^*} - \varepsilon_n$ introduced by the 20% exact HF-exchange that is part of the B3LYP functional. We have in the evaluation of $\Delta E_S^{(2)\text{B3LYP}}$ according to eq 19 employed the Tamm–Dancoff approximation.²¹ Without this approximation one¹⁰ obtains quite similar results with rmsd of 0.22 eV. For CV(∞)-DFT we find again that the calculated excitation energies $\Delta E_S^{(\infty)}$ are larger than for that of $\Delta E_S^{(2)}$. This is especially the case for those systems (1, 2, and 10–15), where n is located on a single heteroatom (N or O). However compared to LDA, the difference $\Delta E_S^{(2)} - \Delta E_S^{(\infty)}$ has decreased in the case of B3LYP, and the rmsd is now 1.14 eV, Table 1. The reductions stem from the fact that the 20% HF-exchange in B3LYP do not contribute to $\Delta E_S^{(2)} - \Delta E_S^{(\infty)}$ as argued previously. Introducing finally Δ SCF leads, as in the case of LDA, to a considerable improvement with a rmsd of 0.30 eV, which is only slightly larger than for $\Delta E_S^{(2)\text{B3LYP}}$ with rmsd of 0.18 eV. For the $n \rightarrow \pi^*$ transitions studied here, the B3LYP proves to afford excitation energies closest to BE¹⁰ for both $\Delta E_S^{(2)}$ and $\Delta E_S^{(\infty)}$. Unfortunately, in a few cases we were unable to obtain SCF convergence for the Δ SCF method. These cases are marked by blank entries in Table 1.

GGA, MO6-L, BHLYP, and MO6. Exploratory calculations with standard functionals based on the GGA revealed that that these functionals afford results quite similar in quality to LDA. A possible exception is the meta-GGA functional MO6-L,^{34,35} which has been applied by Jacquemin et al. for the same sample of excitation shown in Table 1 yielding a rmsd of 0.45 eV. We have also carried out a full investigation employing the BHLYP functional with 50% HF exchange. The results are not reported in details as they follow the same trends as B3LYP. Thus, for CV(2)-

DFT the excitation energies are somewhat overestimated with a rmsd of 0.69 eV using the Tamm–Dancoff approximation²¹ and 0.62 eV without.¹⁰ Introducing CV(∞)-DFT increases the excitation energies even more with a rmsd of 0.93 eV. Finally relaxing the orbitals in Δ SCF leads to a substantial improvement with a rmsd of 0.48 eV. It is apparent from the BHLYP results given in Table S1 of the Supporting Information that a considerable increase in the fraction of exact exchange in B3LYP from 20% leads to a poorer agreement with BE.¹⁰ We should note that the MO6 functional by Zhao and Truhlar⁸ with 27% HF-exchange for the sample of excitations in Table 1 affords a rmsd of 0.24 eV.

5. CONCLUDING REMARKS

We have carried out a theoretical study of $n \rightarrow \pi^*$ transitions based on constricted variational density functional theory (CV(n)-DFT)^{18,20} as a natural extension of a previous study on $\pi_A \rightarrow \pi_B^*$ CT transitions.¹⁵ In both types of excitations an electron is moved between two regions of space with little overlap where the regions are defined by the density of the orbitals involved. In CV(n)-DFT we carry out a unitary transformation $\exp[U]$ of eq 1 between the occupied $\{\varphi_i; i = 1, \text{occ}\}$ and virtual $\{\varphi_a; a = 1, \text{vir}\}$ ground-state orbitals to any desired order n in U to produce a new set of occupied $\{\varphi'_i; i = 1, \text{occ}\}$ and virtual $\{\varphi'_a; a = 1, \text{vir}\}$ excited state orbitals.²⁰ Here the matrix U consists of $\text{occ} \times \text{vir}$ independent variational parameters that are determined in such a way as to minimize the excitation energies under the constraint that the change in density due to an excitation represents the transfer of one electron from the density space spanned by $\{\varphi_i; i = 1, \text{occ}\}$ to the density space spanned by $\{\varphi_a; a = 1, \text{vir}\}$.^{18,20}

For the CV(2)-DFT level of theory we recover within the Tamm–Dancoff approximation²¹ the adiabatic TDDFT theory^{1,2} of eq 14. It was shown in our previous study on $\pi_A \rightarrow \pi_B^*$ CT transitions¹⁵ that the energy expression ($\Delta E_S^{(2)}$) for CV(2)-DFT or adiabatic TDDFT is unable even qualitatively to describe CT transition energies. Further, for such transitions, $\Delta E_S^{(2)}$ leads to a severe underestimation of the excitation energies. In the case of the $n \rightarrow \pi^*$ type of transitions the error for $\Delta E_S^{(2)}$ is much smaller and

for B3LYP, we find that $\Delta E_S^{(2)B3LYP}$ is in close agreement with BE¹⁰ leading to a rmsd of only 0.18 eV.

For the CT transitions $\pi_A \rightarrow \pi_B^*$ the inclusion of energy terms to all orders in U (CV(∞)-DFT) leads to a qualitative correct energy expression $\Delta E_S^{(\infty)}$ which contains the “self-interaction” terms $K_{\pi_A \pi_A \pi_A \pi_A}, K_{\pi_B \pi_B \pi_B \pi_B}, K_{\pi_A \pi_B \pi_B \pi_A}$. Further, combining the energy expression $\Delta E_S^{(\infty)}$ with a relaxation of $\{\varphi_i; i = 1, \text{occ}\}$ and $\{\varphi_a; a = 1, \text{vir}\}$ gives rise to excitation energies ($\Delta E_S^{(\Delta SCF)}$) in excellent agreement with experiment even for LDA where the rmsd was 0.20 eV [15]. In the case of $n \rightarrow \pi^*$ transitions, the inclusion of terms to all orders followed by orbital relaxation as accomplished in the ΔSCF procedure^{24,30} results for LDA and B3LYP in a better agreement with BE¹⁰ than $\Delta E_S^{(2)}$. However, for B3LYP $\Delta E_S^{(2)}$ and $\Delta E_S^{(\Delta SCF)}$ are quite close with rmsd values of 0.18 and 0.30 eV, respectively. The orbital relaxation achieves in many ways the same as the introduction of double displacements through frequency dependent kernels.³¹

The $n \rightarrow \pi^*$ transitions discussed here are of a special kind that can be described by a single orbital transition. This considerably simplifies the self-consistent optimization of U through the use of the ΔSCF procedure.^{24,30} For the general transition involving many orbital transitions the self-consistent optimization of U has been formulated but not yet implemented.²⁰ We finally note that we have modified the name of our approach from constrained variational theory in the previous study¹⁵ to constricted variational theory in the present investigation. This is done in order not to confuse our CV(n)-DFT approach with other constrained variational methods in the literature.^{32,33} In the method by Van Voorhis,³² charge is constricted to certain regions of Cartesian space, whereas our method constricts the charge to certain regions of orbital space.

■ ASSOCIATED CONTENT

S Supporting Information. Results from B3LYP functional given in Table S1. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ziegler@ucalgary.ca.

■ ACKNOWLEDGMENT

T.Z. would like to thank the Canadian government for a Canada research chair in theoretical inorganic chemistry and the NSERC for financial support. J.C. would like to thank University of Manitoba for a sabbatical leave.

■ REFERENCES

- (1) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997.
- (2) Casida, M. E. In *Recent advances in density functional methods*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; pp 155–193.
- (3) van Gisbergen, S. J. A.; Snijders, J. G. *J. Chem. Phys.* **1995**, *103*, 9347.
- (4) Petersilka, M.; Grossmann, U. J.; Gross, E. K. U. *Phys. Rev. Lett.* **1996**, *76*, 12.
- (5) Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454.
- (6) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218.
- (7) Jensen, F. *Introduction to Computational Chemistry*; Wiley: New York, 2006.
- (8) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C.; Valero, R.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, *6*, 2071.

- (9) Goerigk, L.; Grimme, S. *J. Chem. Phys.* **2010**, *132*, 184103.
- (10) Schreiber, M.; Silva-Junior, M.; Sauer, S.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.
- (11) (a) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Phys. Chem.* **2004**, *120*, 8425. (b) Song, J.-W.; Watson, M. A.; Hirao, K. *J. Chem. Phys.* **2009**, *131*, 144108.
- (12) Stein, T.; Kronik, L.; Baer, R. *J. Am. Chem. Soc.* **2009**, *131*, 2818.
- (13) (a) Neugebauer, J.; Gritsenko, O.; Baerends, E. J. *J. Chem. Phys.* **2006**, *124*, 214102. (b) Schipper, P. R. T.; Gritsenko, O. V.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2000**, *112*, 1344–1352.
- (14) Gritsenko, O.; Baerends, E. J. *J. Chem. Phys.* **2004**, *121*, 655–660.
- (15) Ziegler, T.; Krykunov, M. *J. Chem. Phys.* **2010**, *133*, 074104.
- (16) Ziegler, T.; Seth, M.; Krykunov, M.; Autschbach, J.; Wang, F. *J. Mol. Struct. THEOCHEM* **2009**, *914*, 106.
- (17) Ziegler, T.; Seth, M.; Krykunov, M.; Autschbach, J. *J. Chem. Phys.* **2008**, *129*, 184114.
- (18) Ziegler, T.; Seth, M.; Krykunov, M.; Autschbach, J.; Wang, F. *Chem. Phys.* **2009**, *130*, 154102.
- (19) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem.* **2008**, *A 112*, 1095.
- (20) Cullen, J.; Krykunov, M.; Ziegler, T. *Chem. Phys.*, in press, doi:10.1016/j.chemphys.2011.05.021.
- (21) Hirata, S.; Head-Gordon Chem, M. *Phys. Lett.* **1999**, *314*, 291.
- (22) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.
- (23) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (24) Ziegler, T.; Baerends, E. J.; Rauk, A. *Theoret. Chim. Acta (Berlin)* **1976**, *43*, 261.
- (25) te Velde, G.; Bickelhaupt, F. M.; van Gisbergen, S. J. A.; Fonseca Guerra, C.; Baerends, E. J.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931.
- (26) Van Lenthe, E.; Baerends, E. J. *J. Comput. Chem.* **2003**, *24*, 1142.
- (27) Amos, A. T.; Hall, G. G. *Proc. R. Soc.* **1961**, *A263*, 483.
- (28) Martin, R. L. *J. Chem. Phys.* **2003**, *118*, 4775.
- (29) Ziegler, T.; Rauk Theoret, A. *Chim. Acta (Berlin)* **1977**, *46*, 1.
- (30) Slater, J. C.; Johnson, K. H. *Phys. Rev. B* **1972**, *844–853*.
- (31) Maitra, N. T.; Zhang, F.; Cave, R. J.; Burke, R. *J. Chem. Phys.* **2004**, *120*, 5932.
- (32) (a) Cheng, C. L.; Wu, Q.; Van Voorhis, T. *J. Chem. Phys.* **2008**, *129*, 124112. (b) Wu, Q.; Van Voorhis, T. *Phys. Rev. A* **2005**, *72* (2), 024502.
- (33) Artacho, E.; Rohlfling, M.; Côté, M.; Haynes, P. D.; Needs, R. J.; Molteni, C. *Phys. Rev. Lett.* **2004**, *93*, 116401.
- (34) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.
- (35) Zhao, Y.; Truhlar Theor, D. G. *Chem. Acc.* **2008**, *120*, 215.

Critical Examination of Explicitly Time-Dependent Density Functional Theory for Coherent Control of Dipole Switching

Shampa Raghunathan* and Mathias Nest

Theoretische Chemie, Technische Universität München, Lichtenbergstrasse 4, 85747 Garching, Germany

ABSTRACT: We compare the performance of explicitly time-dependent density functional theory (DFT) with time-dependent configuration interaction (TDCI) to achieve the control task of a population inversion in LiCN. We assume that if a given pulse achieves the control task when used in TDCI, then there should be a pulse with similar frequency and intensity that achieves the task in time-dependent DFT (TDDFT). The present investigation indicates that this is not the case, if standard functionals are used in the adiabatic approximation.

1. INTRODUCTION

Time-dependent density functional theory (TDDFT)¹ is one of the most successful and popular approaches for the calculation of excited-state properties of molecules, as well as to describe the real-time (RT) dynamics of many electron systems. Its computational efficiency has made it one of the first electronic structure methods (after time-dependent Hartree–Fock, TDHF²) to be generalized to the real-time domain.^{3,4} Much experience about the strengths and limitations of TDDFT has been gained through applications to a wide range of phenomena. Rubio et al.⁵ studied as diverse processes as high harmonic generation, Coulomb explosion, and laser-induced photodissociation. The same group also developed the real-space code Octopus,^{6–8} one of the most widely used implementations of TDDFT. Other successful applications were to the nonlinear dynamics of electrons in metal clusters^{9,10} and molecules.^{11–14} Less satisfactory were attempts to describe the nonsequential double ionization of atoms such as He or Ne, especially with respect to the famous “knee” structure in the yield as a function of laser fluence.^{15,16} There are various reasons for this mixed success of explicitly time-dependent DFT. First of all, there is the so-called adiabatic approximation;¹⁷ i.e., the time-dependent density $\rho(t)$ is inserted into standard ground-state exchange–correlation (xc) functionals. Obviously, there is no guarantee that these functionals will generate the correct dynamics, although they have been found to be useful in a number of applications mentioned above. Also, some authors have argued that a proper density functional for quantum dynamics should contain a memory term.^{18–21} These problems have not yet been fully solved, so that most work is being done using the adiabatic approximation mentioned above. Recently, some significant progress²² has been made to combine TDDFT with coherent control, or more precisely quantum optimal control theory.^{23–25}

In this paper, we want to study the suitability of TDDFT for a simple control task: a state-to-state transition using a so-called π -pulse. An estimate of the necessary laser pulse parameters can be obtained from a configuration interaction (CI) calculation, described in the following sections. Our test molecule will be LiCN, because this has been used in a similar study before.^{26,46,47} It also has the advantage that the transition is accompanied by a strong change in the dipole moment, which can serve as a

criterion when comparing TDCI^{27–29} and TDDFT calculations. If TDDFT provides a correct description of the correlated electron dynamics, then a pulse similar to the one that worked for TDCI should also work for TDDFT. Similarity between laser pulses is defined here as having a similar carrier frequency and intensity. The pulse duration and carrier-envelope phase will be kept fixed.

In section 2 we give some computational details of our simulations and discuss the extent of deviations of molecular properties of LiCN if different electronic structure methods are used. Section 3 presents the main results of this work, and section 4 concludes the paper. Atomic units will be used if not stated otherwise.

2. COMPUTATIONAL DETAILS

The TD Kohn–Sham (KS) equations are given by

$$i \frac{\partial \varphi_i(\mathbf{r}, t)}{\partial t} = \left[-\frac{\nabla^2}{2} + v_{\text{eff}}[\rho(t)] \right] \varphi_i(\mathbf{r}, t) \quad (1)$$

where the effective potential

$$v_{\text{eff}}[\rho(t)](\mathbf{r}, t) = v_{\text{ext}}(\mathbf{r}, t) + v_{\text{H}}[\rho(t)](\mathbf{r}, t) + v_{\text{xc}}[\rho(t)](\mathbf{r}, t) \quad (2)$$

consists of external v_{ext} , Hartree v_{H} , and exchange–correlation v_{xc} contributions. The external potential contains the electron–nuclear attraction and the laser–electron interaction in the dipole approximation

$$v_{\text{ext}}(\mathbf{r}, t) = v_{\text{Ne}}(\mathbf{r}) - \boldsymbol{\mu} \mathbf{f}(t) \cos[\omega(t - t_p)] \quad (3)$$

with the envelope

$$\mathbf{f}(t) = \begin{cases} \mathbf{f}_0 \cos^2 \left[\frac{\pi}{2\sigma} (t - t_p) \right] & \text{if } |t - t_p| < \sigma \\ \mathbf{0} & \text{else} \end{cases} \quad (4)$$

Received: April 19, 2011

Published: June 24, 2011

Here, t_p is the time at which the pulse is maximal, σ is the full width at half-maximum (fwhm), and f_0 is the polarization and maximum pulse amplitude.

We performed TDDFT calculations using the program packages Octopus^{6,7} and Parsec,^{30,31} both of which employ a real-space uniform grid representation using the finite difference approach. For the real-space grid, a simulation sphere of radius $8 a_0$, and a grid spacing of $0.4 a_0$ are used except for exact-exchange optimized effective potential (OEP-X) calculations, where the radius of simulation sphere was $15 a_0$ and the grid spacing was $0.7 a_0$. In all calculations, we used norm-conserving, nonlocal, ionic Troullier–Martins³² pseudopotentials to model the interactions of valence electrons with core electrons. In this scheme, the 1s shells of all three atoms Li, C, and N were treated as the core. In the density functional calculations using the program Parsec, we employed the xc functionals according to the local-density approximation (LDA) with the Perdew–Zunger parametrization^{33,34} (CA-PZ), and generalized gradient approximation (GGA) of Perdew–Burke–Ernzerhof (PBE).³⁵ We further carried out GGA-PBE, and hybrid-DFT PBE0,³⁶ extended hybrid functional combined with Lee–Yang–Parr correlation functional (X3LYP),³⁷ and OEP-X calculations using the program Octopus.

The TDKS equations (eq 1) are solved by direct numerical integration in real time. For this purpose, Parsec³⁸ uses a fourth-order Taylor approximation, whereas Octopus uses an exponential midpoint rule combined with the Lanczos exponential approximation.³⁹ We have used an electronic time step of 0.24

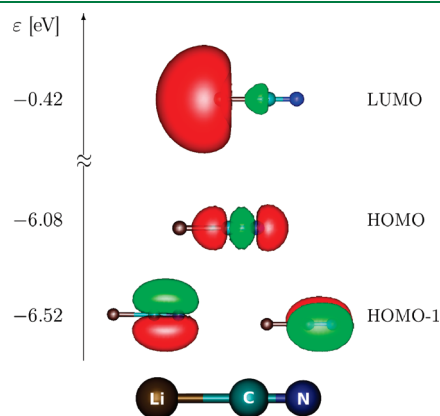


Figure 1. HOMO-1, HOMO, and LUMO molecular orbitals of LiCN along with their orbital energies computed at the GGA-PBE level using Parsec.

as (0.01 au) with a total propagation time of 50 fs, and an x -polarized laser pulse has been employed throughout unless stated otherwise.

Because we want to compare different explicitly time-dependent methods, we will present first some molecular properties from time-independent calculations, like CIS and linear response (LR)-TDDFT and LR-TDHF, in order to assess how much these values change from one method to another. We computed the dipole moments of the ground and selected excited states of LiCN along with the corresponding transition dipole moments and excitation energies. For the LR-TDDFT calculations we employed LSDA,⁴⁰ PBE, and PBE0 xc functionals using Gaussian 09.⁴¹ We also performed more accurate correlated calculations such as CISD (10,15) (all singles, doubles restricted to a (10,15) active space) using our own code,^{42,43} and EOM-CCSD calculation using the program Molpro.⁴⁴ The former, somewhat unusual method is included, because in a later section we report time-dependent results on this level of theory. In all of the above calculations we employed the polarized double- ζ basis set 6-31G*.⁴⁵

From the study of Klamroth et al.,²⁶ we adopted the equilibrium geometry of LiCN; i.e., $r_{\text{Li-C}} = 3.683 a_0$ and $r_{\text{C-N}} = 2.168 a_0$ (with orientation along the z -axis). For the ground-state electronic configuration, the HF method predicted the HOMO to be doubly degenerate $C(2p_{x/y})-N(2p_{x/y})$ π -type MOs and the penultimate MO (i.e., HOMO-1) to be a $C(2p_z)-N(2p_z)$ σ -type MO. On the other hand, in the Kohn–Sham DFT calculations, this order is reversed and the σ -type MO forms the HOMO except for OEP-X calculation, where, the HF ordering of ground-state MOs is preserved. This trend in the ordering of the occupied MOs in DFT calculations has been predicted by both grid-based and basis set approaches. Both in HF and DFT approaches the LUMO is essentially the Li (2s) orbital with a weak σ^* character; see Figure 1.

Table 1 summarizes dipole moments, transition dipole moments, and the excitation energies. The transition we are interested in, $\text{CN}(\pi) \rightarrow \text{Li}(\sigma^*)$, is from the ground state (S_0) to the second or third excited states of LiCN (S_2 or S_3). These degenerate states can be accessed selectively by laser pulses polarized along the x or y direction. A complete population transfer to these excited states will switch the z -component of the dipole moment from $\mu_{0,0;z} \approx -3.5 e a_0$ to $\mu_{2(3),2(3);z} \approx +2 e a_0$, with comparatively small deviations between various quantum chemical methods. In all our LR-TDDFT calculations, done using Gaussian 09,⁴¹ the permanent dipole moments, $\mu_{0,0;z}$ and $\mu_{2(3),2(3);z}$ are predicted to be in good agreement with the EOM-CCSD values; however, the transition dipole moments are

Table 1. Selected Ground- and Excited-State Properties of LiCN Computed Using Linear Response (LR) Theories along with EOM-CCSD^a

methods	$\mu_{0,0;z}$	$\mu_{2(3),2(3);z}$	$\mu_{0,2;x}$	$\mu_{0,2;y}$	$\mu_{0,3;x}$	$\mu_{0,3;y}$	$\Delta E_{0 \rightarrow 2(3)}$
LR-HF ^b	-3.7080	1.8329	0.2682	0.1503	0.1503	-0.2682	6.57
CIS ^b	-3.7080	1.8450	0.3075	0.0258	0.0258	-0.3075	6.58
CIS(D) ^c	-3.7082	2.7952	0.3084	0.0095	0.0095	-0.3084	6.13
CISD(10,15) ^d	-3.4662	1.8338	0.2981	0.0902	0.0902	-0.2981	6.77
EOM-CCSD ^e	-3.5502	2.0315	0.3833	0.0	0.0	-0.3833	6.28
LR-LSDA ^b	-3.4193	2.2172	0.2675	0.0933	0.0933	-0.2675	4.77
LR-PBE ^b	-3.3951	2.1553	0.2442	0.0890	0.0890	-0.2442	4.31
LR-PBE0 ^b	-3.5114	2.0748	0.2628	0.0123	0.0123	-0.2628	5.13

^a Ground- and excited-state dipoles, transition dipole moments ($\mu_{i,j;q}$ $e a_0$) and excitation energies (ΔE , eV) are given. ^b Using Gaussian 09. ^c From ref 26. ^d Own code. ^e Using Molpro.

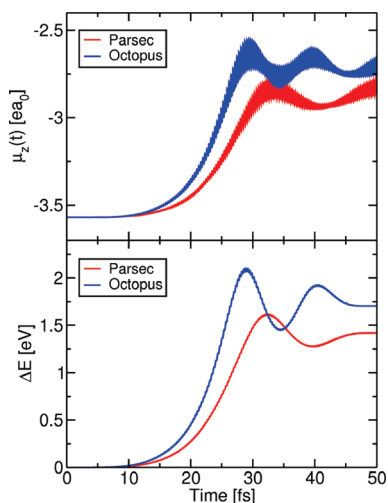


Figure 2. Time evolution of dipole moments and energies computed at the GGA-PBE level using Parsec and Octopus programs: dipole signal of LiCN upon laser excitation (top); time-dependent relative energy (with respect to ground electronic state energy at $t = 0$ fs) (bottom). $\omega = 6.3$ eV, field strength $0.004\,928 E_h/ea_0$ polarized along x , and total propagation time = 50 fs.

slightly underestimated to lie in the range of $0.244\text{--}0.268 ea_0$, close to the LR-TDHF value.

Before we present results of our quantum dynamical calculations, it is worthwhile also to compare the different implementations of TDDFT in Parsec and Octopus. We do so by showing the excitation energy and dipole moment of LiCN computed using these two programs at the GGA-PBE level with a laser frequency (ω) of 6.3 eV and a field strength of $0.004\,928 E_h/ea_0$; see Figure 2. Although the results are qualitatively the same, they are not identical. The size of the grid, the grid spacing, and the pseudopotentials were the same in both programs. We can only speculate that internal differences in either the propagation scheme or the treatment of the Coulomb singularity are responsible for the differences, which become amplified by the non-linearity of the equations of motions. Alternatively, a bug in one of the implementations cannot be ruled out completely.

3. RESULTS

3.1. Dipole Switching. The full population inversion from the ground to the second excited state can be achieved by a π -pulse. For laser pulses with a \cos^2 envelope, the condition for the amplitude is

$$|f_0| = \frac{\pi}{\sigma|\mu_{i,j;q}|} \quad (5)$$

where $\mu_{i,j;q}$ is the transition dipole moment between states i and j , and q indicates the polarization. Also, a π -pulse requires that the laser pulse is long enough to be resonant with the desired transition and that no energetically close states are present which have a significant transition dipole moment, too. Using this condition for TDCIS and TDCISD(10,15), together with the information from Table 1, we achieve a controlled switching of the z -component of the dipole moment and excitation energies; see Figure 4. Because the transition dipole moments are similar for both methods, the pulses are similar too, and they can serve as a starting point when looking for laser pulses

suitable for TDDFT. The total propagation time is 50 fs, or $\sigma = 25$ fs in eq 4.

The π -pulse condition can be readily applied to TDCI calculations, because a time-independent many-body basis exists. In other words, the Slater determinants made of Hartree–Fock orbitals span always the same Hilbert space. It has been shown already for MCTDHF^{48–50} that a time-dependent basis leads to a time-dependent electronic structure, making the concept of a resonant transition questionable. Moreover, the linear-response values of TDDFT given in Table 1 cannot easily be translated to the explicitly time-dependent, nonlinear regime. On the other hand, if TDDFT provides a realistic approximation to the true dynamics of the system, then a pulse similar to those used in TDCI should perform a similar task. Looking for this “similar pulse”, we pursued the following strategy for a variety of density functionals. First, we scanned a range of laser frequencies (ω) between 4.9 and 6.5 eV (4–7 eV for PBE). We identified promising frequencies, to which the system reacts strongly, which indicates proximity to a resonance. For the two TDCI based methods this is simply the excitation energy given in Table 1, whereas TDDFT calculations show the resonances at 6.1, 6.3, 5.3, 5.5, and 6.1 eV using the xc functionals LDA, PBE, PBEO, X3LYP, and OEP-X, respectively, and for TDHF $\omega = 6.9$ eV. Next, an optimal field strength has to be found. This posed a somewhat bigger problem, because several criteria could be applied. On the one hand, the dipole switching as described by TDCI is a one-photon absorption process. On the other hand, one would like to achieve a final state, which is as close to stationary as possible. We have attempted both and report the results in the following.

Figure 3 shows the energy uptake and time-dependent dipole moments of LiCN, if the field strengths are adjusted to a one-photon absorption. For TDCIS ($f_{0,x} = 0.009\,377 E_h/ea_0$) and TDCISD(10,15) ($f_{0,x} = 0.010\,198 E_h/ea_0$), the field strengths follow from the π -pulse equation (eq 5), while for TDDFT and TDHF a scan over pulse amplitudes is necessary. For TDDFT calculations the field strengths are $0.0371, 0.0392, 0.0182, 0.021\,57,$ and $0.0261 E_h/ea_0$ using the xc functionals LDA, PBE, PBEO, X3LYP, and OEP-X, respectively, and for TDHF $f_{0,x} = 0.045 E_h/ea_0$. The density functional based methods and Hartree–Fock obviously require a much larger fluence, than the CI based methods. Clearly a multitude of excited states is populated, leading to oscillating wave packets, that have no similarity with dipole switching.

The results obtained from the second criterion, which focuses on stationarity of the dipole moment at late times, are shown in Figure 4. TDCIS and TDCISD(10,15) use the same field strengths as used in the case of one-photon absorption calculations. TDDFT calculations use the field strengths of $0.004\,928 E_h/ea_0$ for LDA, PBE, and X3LYP functionals and 0.003 and $0.009\,856$ for PBEO and OEP-X functionals, and for TDHF $f_{0,x} = 0.009\,919 E_h/ea_0$. The results for all five functionals and HF are qualitatively the same: the dipole moment changes by about $1 ea_0$, instead of about $5.5 ea_0$ as in TDCI.

How can this failure of real-time TDDFT, using standard functionals in the adiabatic approximation, be explained? It is well-known that linear-response TDDFT has problems describing charge transfer. But, in our case, the charge transfer is only on a small length scale, and is not ameliorated by hybrid functionals. Also the use of exact exchange does not improve the results. To emphasize this point, we included TDHF in the list of methods above. Further, in the LR-TDDFT theory, Tozer and co-workers

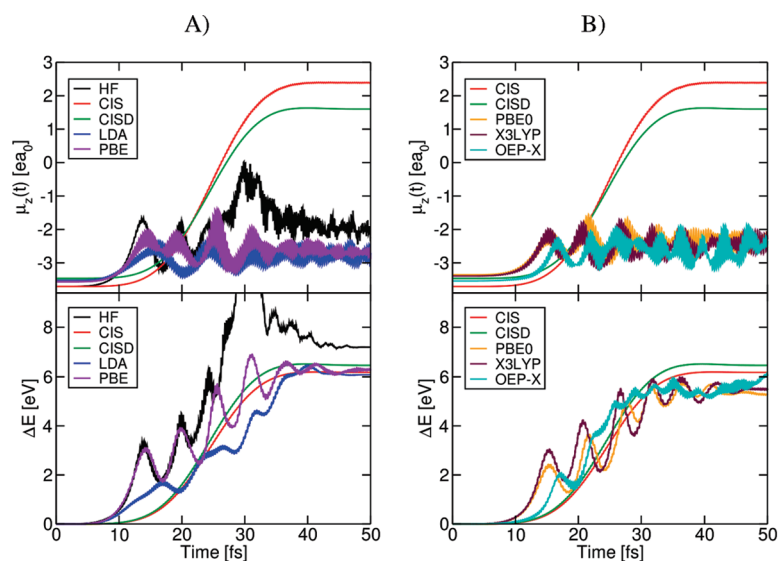


Figure 3. Time evolution of dipole moments and energies for one-photon absorption computed using different theoretical methods: dipole z -component as a function of time (top); time-dependent relative energy (with respect to ground electronic state energy at $t = 0$ fs) (bottom). Laser parameters ($f_{0,x}$ in E_h/ea_0 and ω in eV) for A: (HF) $f_{0,x} = 0.045$, $\omega = 6.9$; (CIS) $f_{0,x} = 0.009377$, $\omega = 6.6$; (CISD) $f_{0,x} = 0.010198$, $\omega = 6.8$; (LDA) $f_{0,x} = 0.0371$, $\omega = 6.1$; (PBE) $f_{0,x} = 0.0392$, $\omega = 6.3$. Laser parameters for B, including the same CIS and CISD values: (PBE0) $f_{0,x} = 0.0182$, $\omega = 5.3$; (X3LYP) $f_{0,x} = 0.02157$, $\omega = 5.5$; (OEP-X) $f_{0,x} = 0.0261$, $\omega = 6.1$. Total propagation time = 50 fs.

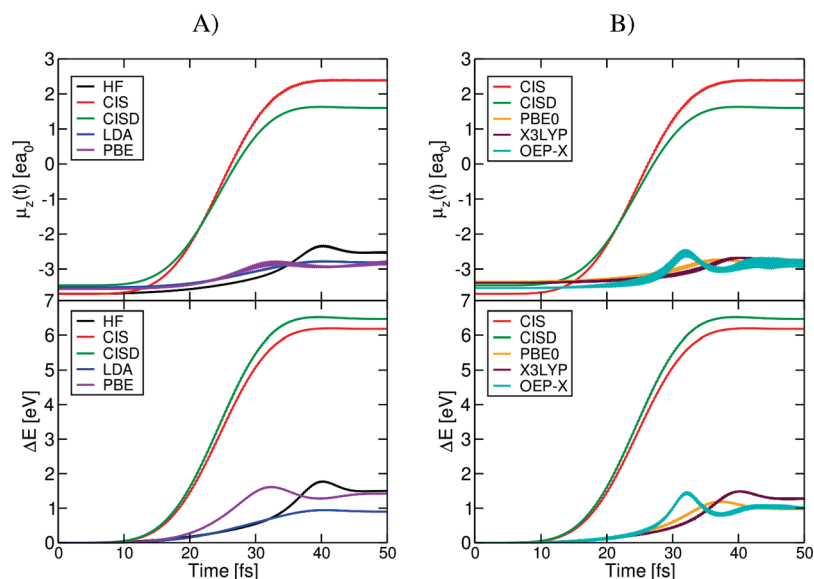


Figure 4. Time evolution of dipole moments and energies computed using different theoretical methods: dipole z -component as a function of time (top); time-dependent relative energy (with respect to ground electronic state energy at $t = 0$ fs) (bottom). Laser parameters ($f_{0,x}$ in E_h/ea_0 and ω in eV) for A: (HF) $f_{0,x} = 0.009919$, $\omega = 6.9$; (CIS) $f_{0,x} = 0.009377$, $\omega = 6.6$; (CISD) $f_{0,x} = 0.010198$, $\omega = 6.8$; (LDA) $f_{0,x} = 0.004928$, $\omega = 6.1$; (PBE) $f_{0,x} = 0.004928$, $\omega = 6.3$. Laser parameters for B, including the same CIS and CISD values: (PBE0) $f_{0,x} = 0.003$, $\omega = 5.3$; (X3LYP) $f_{0,x} = 0.004928$, $\omega = 5.5$; (OEP-X) $f_{0,x} = 0.009856$, $\omega = 6.1$. Total propagation time = 50 fs.

proposed^{51,52} a weighted overlap parameter which can be considered as an indirect measure for the probability of charge transfer. They suggested that a small overlap indicates less sharing of similar regions of space between the corresponding orbitals involved in the excitation, which restricts the CT excitation. In our study, we simply considered the spatial overlap between the π and σ^* states, $O_{\pi\sigma^*}$, computed as

$$O_{\pi\sigma^*} = \int |\varphi_{\pi}(\mathbf{r})| |\varphi_{\sigma^*}(\mathbf{r})| d\mathbf{r} \quad (6)$$

For this quantity, $O_{\pi\sigma^*}$, we found a value of 0.18 from LDA-CA-PZ and GGA-PBE calculations which implies a weak overlap between the fragment orbitals CN(π) and Li(σ^*). This is a small value, but probably not small enough to account for the failure of all functionals. The origin of the problem is probably indeed related to the fact that TDDFT and TDHF (as to some degree MCTDHF) use a time-dependent basis. We assume that this leads to an effective time-dependent excitation energy between S_0 and S_2 so that resonant excitation cannot take place. Similar observations have been made with regard to Rabi oscillations by

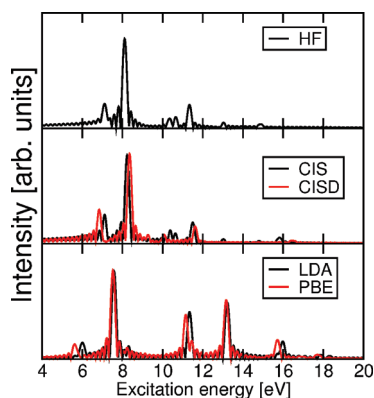


Figure 5. Excitation spectra of LiCN obtained from Fourier transform of $\mu_z(t)$, computed at different explicitly TD theories with a laser pulse of frequency 6.0 eV, $\mathbf{f}_0 = (0.002, 0.002, 0.002) E_h/ea_0$, $\sigma = 0.5$ fs, and total propagation time = 20 fs.

other groups.^{53,54} The paper of Rubio et al.⁵³ suggests a new constraint on standard functionals is required to prevent this detuning. It should also be noted that it is quite difficult to quantify by how far the target state has been missed. In this paper we used the energy and dipole moment as measure, but this provides only restricted information. Simulations for model systems, where difference densities and currents can be evaluated much easier are under way. This will also provide insight whether the quantum systems are taking the same “path” from the initial to the final state. While terms such as $\langle S_0 | \mu | S_2 \rangle$ in TDCI produce the desired population transfer, there is no analogue when time-dependent orbitals are used. The operation of μ on an orbital does not guarantee that the orbitals will change in the physically correct way.

3.2. Ultrashort Laser Pulse Excitation. Finally, we discuss a different aspect of the problem at hand. The task in the previous subsection was to populate a state which was orthogonal to the initial ground state. This turned out to be surprisingly difficult but does not say anything about the performance of TDDFT if the perturbation is smaller. Therefore, we excited the LiCN molecule with an ultrashort laser pulse with $\sigma = 0.5$ fs, $\omega = 6$ eV, and a field strength/polarization of $\mathbf{f}_0 = (0.002, 0.002, 0.002) E_h/ea_0$. This pulse gives a “kick” to the electron system so that they start to oscillate. We recorded the time-dependent dipole moment for 20 fs and calculated the Fourier transform to obtain the absorption spectrum. Figure 5 shows the excitation energy spectra calculated from $\langle \mu_z \rangle(t)$ for various explicitly time-dependent methods. The TDHF spectrum is almost identical to the TDCIS spectrum, and TDCISD(10,15) shifts the results only marginally. The two TDDFT spectra are superficially similar to those from wave function based methods. The peaks at 6 and 8 eV are both shifted somewhat to lower energies but not more than can be expected if the electronic structure method is changed. Even the relative height of these two dominant peaks is semiquantitatively correct. However, the peaks at higher energies come out too strong. Overall, TDDFT performs rather well in this case. This finding might indicate that there is a relation between the performance of real-time TDDFT with standard functionals and in the adiabatic approximation on one side, and the distance between initial and final states, measured in energy or as a distance in Hilbert space, on the other side.

4. CONCLUSIONS

We compared the performance of explicitly time-dependent DFT with TDCI to achieve the control task of a population inversion in LiCN. We found that no pulse similar to the one used for TDCI will perform the task in TDDFT, if standard functionals and the adiabatic approximation are used. Similarity of pulses is defined here as similarity in laser frequency and field strength. Maybe the introduction of a chirp or a pulse sequence would have allowed TDDFT to achieve the task, but probably for a physically wrong reason. The use of hybrid functionals or exact exchange did not improve the results. This indicates that results obtained from a combination of TDDFT in its standard implementation and coherent control schemes have to be taken with care.

AUTHOR INFORMATION

Corresponding Author

*E-mail: shampa@mytum.de.

ACKNOWLEDGMENT

We thank the Munich Centre for Advanced Photonics for financial support and Michael Mundt for the technical help.

REFERENCES

- (1) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997.
- (2) Kulander, K. C. *Phys. Rev. A* **1987**, *36*, 2726.
- (3) Ullrich, C. A.; Erhard, S.; Gross, E. K. U. In *Super Intense Laser Atom Physics IV*; Muller, H. G., Fedorov, M. V., Eds.; Nato Advanced Study Institutes, Ser. 3; Kluwer: Dordrecht, The Netherlands, 1996; Vol. 13, p 267.
- (4) Calvayrac, F.; Reinhard, P.-G.; Suraud, E. *Phys. Rev. B* **1995**, *52*, R17056.
- (5) Castro, A.; Marques, M. A. L.; Alonso, J. A.; Bertsch, G. F.; Rubio, A. *Eur. Phys. J. D* **2004**, *28*, 211.
- (6) Marques, M. A. L.; Castro, A.; Bertsch, G. F.; Rubio, A. *Comput. Phys. Commun.* **2003**, *151*, 60.
- (7) Castro, A.; Appel, H.; Oliveira, M.; Rozzi, C. A.; Andrade, X.; Lorenzen, F.; Marques, M. A. L.; Gross, E. K. U.; Rubio, A. *Phys. Status Solidi B* **2006**, *243*, 2465.
- (8) <http://www.tddft.org/programs/octopus>.
- (9) Calvayrac, F.; Reinhard, P.-G.; Suraud, E.; Ullrich, C. A. *Phys. Rep.* **2000**, *377*, 493.
- (10) Fennel, Th.; Meiwes-Broer, K.-H.; Tiggesbäumker, J.; Reinhard, P.-G.; Dinh, P. M.; Suraud, E. *Rev. Mod. Phys.* **2010**, *82*, 1793.
- (11) Chu, X.; Chu, S. *Phys. Rev. A* **2001**, *64*, 063404.
- (12) Suzuki, Y.; Seideman, T.; Stener, M. *J. Chem. Phys.* **2004**, *120*, 1172.
- (13) Remacle, F.; Levine, R. D. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 6793.
- (14) Remacle, F.; Levine, R. D. *J. Phys. Chem. C* **2007**, *111*, 2301.
- (15) Lappas, D. G.; van Leeuwen, R. *J. Phys. B* **1998**, *31*, L249.
- (16) Bauer, D.; Ceccherini, F. *Opt. Express* **2001**, *8*, 377.
- (17) Gross, E. K. U.; Burke, K. In *Lecture Notes in Physics*; Marques, M. A. L., Ullrich, C. A., Nogueira, F., Rubio, A., Burke, K., Gross, E. K. U., Eds.; Springer-Verlag: Berlin, Heidelberg, Germany, 2006; Vol 706, pp 1–13.
- (18) Maitra, N. T.; Burke, K.; Woodward, Chr. *Phys. Rev. Lett.* **2002**, *89*, 023002.
- (19) Baer, R.; Kurzweil, Y.; Cederbaum, L. S. *Isr. J. Chem.* **2005**, *45*, 161.
- (20) Kurzweil, Y.; Baer, R. *Phys. Rev. B* **2006**, *73*, 075413.
- (21) Wijewardane, H. O.; Ullrich, C. A. *Phys. Rev. Lett.* **2005**, *95*, 086401.

- (22) Castro, A.; Werschnik, J.; Gross, E. K. U. arxiv.org/1009.2241
- (23) Shapiro, M.; Brumer, P. In *Principles of the Quantum Control of Molecular Processes*; Wiley: New York, 2003.
- (24) Krotov, V. F. In *Global Methods in Optimal Control Theory*; Dekker: New York, 1996.
- (25) Zhu, W.; Rabitz, H. *J. Chem. Phys.* **1998**, *109*, 385.
- (26) Krause, P.; Klamroth, T.; Saalfrank, P. *J. Chem. Phys.* **2005**, *123*, 074105.
- (27) Klamroth, T. *Phys. Rev. B* **2003**, *68*, 245421.
- (28) Klamroth, T. *J. Chem. Phys.* **2006**, *124*, 144310.
- (29) Klamroth, T.; Nest, M. *Phys. Chem. Chem. Phys.* **2009**, *11*, 349.
- (30) Kronik, L.; Makmal, A.; Tiago, M. L.; Alemany, M. M. G.; Jain, M.; Huang, X.; Saad, Y.; Chelikowsky, J. R. *Phys. Status Solidi B* **2006**, *243*, 1063.
- (31) Mundt, M. *J. Theor. Comput. Chem.* **2009**, *8*, 561.
- (32) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993 The core radii (in a_0) are as follows: Li, s(2.43), p(2.43); C, s(1.49), p(1.42); N, s(1.50), p(1.50), with Li(s), C(p), and N(p) as local components.
- (33) Ceperley, D. M.; Alder, B. J. *Phys. Rev. Lett.* **1980**, *45*, 566.
- (34) Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048.
- (35) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (36) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029.
- (37) Xu, X.; Goddard, W. A., III *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 2673.
- (38) Mundt, M.; Kümmel, S. *Phys. Rev. B* **2007**, *76*, 035413.
- (39) Castro, A.; Marques, M. A. L.; Rubio, A. *J. Chem. Phys.* **2004**, *121*, 3425.
- (40) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.
- (41) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *GAUSSIAN09*; Gaussian, Wallingford, CT, 2009.
- (42) Nest, M. *Chem. Phys.* **2010**, *370*, 119.
- (43) Nest, M. *Int. J. Quantum Chem.* **2011**, *111*, 505.
- (44) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrenar, T.; Knizia, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, R.; Pflüger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *Molpro*, Version 2009.1, a package of ab initio programs; 2008.
- (45) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.
- (46) Tremblay, J. C.; Klamroth, T.; Saalfrank, P. *J. Chem. Phys.* **2008**, *129*, 084302.
- (47) Klinkusch, S.; Saalfrank, P.; Klamroth, T. *J. Chem. Phys.* **2009**, *131*, 114304.
- (48) Padmanaban, R.; Nest, M. *Chem. Phys. Lett.* **2008**, *463*, 263.
- (49) Mundt, M.; Tannor, D. J. *New J. Phys.* **2009**, *11*, 105038.
- (50) Nest, M.; Remale, F.; Levine, R. D. *New J. Phys.* **2008**, *10*, 025019.
- (51) Plötner, J.; Tozer, D. J.; Dreuw, A. *J. Chem. Theor. Comput.* **2010**, *6*, 2315.
- (52) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.
- (53) Fuks, J. I.; Helbig, A.; Tokatly, I. V.; Rubio, A. arxiv.org/abs/1101.2880
- (54) Ruggenthaler, M.; Bauer, D. *Phys. Rev. Lett.* **2009**, *102*, 233001.

A Qualitative Index of Spatial Extent in Charge-Transfer Excitations

Tangui Le Bahers,* Carlo Adamo, and Ilaria Ciofini*

LECIME, Laboratoire d'Electrochimie, Chimie des Interfaces et Modélisation pour l'Energie, CNRS UMR-7575, Ecole Nationale Supérieure de Chimie de Paris—Chimie ParisTech, 11 rue P. et M. Curie, 75231 Paris Cedex 05, France

S Supporting Information

ABSTRACT: With the aim of defining the spatial extent associated to an electronic transition, of particular relevance in the case of charge-transfer (CT) excitations, a new index, evaluated only from the computed density for the ground and excited state, is here derived and tested on a family of molecules that can be considered as prototypes of push–pull chromophores. The index (D_{CT}) allows to define the spatial extent associated to a given transition as well as the associated fraction of electron transferred. By definition of centroids of charges associated to the density increase and depletion zones upon excitation, a qualitative and easy to visualize measure of the spatial extent of the donor and the acceptor moieties within a given molecular system is also given. Finally, an index (t) allowing to define the presence eventually pathologic CT transitions for time-dependent density functional theory treatment in conjunction with standard generalized gradient approximation or hybrid functional, that is through space CT, is disclosed.

1. INTRODUCTION

Push–pull systems made up of an electron donor (D) and an electron withdrawing (A) groups covalently connected are one of the most pursued families of compounds when aiming at obtaining intense—and solvatochromic—optical transitions in the visible spectral region.¹ The latter are basically associated to the formation of an excited state corresponding to the transfer of an electron from the donor to the acceptor, that is to the formation of a $[D^+ - A^-]^*$ excited state.²

Clearly, this simple representation of the excited state and of the electron transfer process is modulated and strongly depends both on the type of chemical link present between the donor and the acceptor subunits (normally defined as spacer) and on their intrinsic nature.

In most of the real cases, both the length and the magnitude of the electron transfer due to the excitation are far from ideal (that is one electron transferred from the donor and localized on the acceptor).³ Depending on the intramolecular geometrical and electronic coupling, the transferred electron is, in some cases, delocalized from the region of the molecule nearby the donor to one in the vicinity of the acceptor.⁴

This effect can be qualitatively inspected by the analysis of the density distribution at the ground and the excited state.⁵ Nevertheless, even regions of the molecule not presenting a net change in total density can show an alternation of excess and depletion of density as a consequence of the electronic transition, thus making difficult to use directly the density distribution maps to follow and quantify the charge-transfer (CT) phenomena.^{5,6}

Obtaining a more quantitative measure of the length and magnitude of the CT is far from trivial both from the experimental and the theoretical point of view.

At the theoretical level, quantifying the magnitude of the charge transferred on a given spatial region implies to transform the known electron distribution over space computed for the ground and excited states to a condensed, that is localized, index, therefore implying the use of a (arbitrary) localization scheme.

Any measures derived in such a way will thus depend both on the model, level of theory, used to compute the excited and ground state densities and on the localization procedure applied.

The basic idea of the simple model here proposed is to define a measure of the length of a CT excitation solely on the basis of the total electronic density computed for the ground and excited states. This method is thus applicable to any quantum chemical method supplying densities for the ground and excited states, providing in principle a very simple way to qualitatively compare the outcomes of post-Hartree–Fock (HF) and density functional theory (DFT)-based approaches. To this end, the following steps will be considered:

- Definition of the barycenters of the density depletion and the density increment zones associated with the electronic transition and computed over a grid of points.
- Definition of the CT length as distance between the barycenters.
- Definition of the transferred charge by integration of the density depletion function over all space.
- Calculation of the associated dipole moment.

None of these steps is either time or resources demanding so that our method can be applied to screen and analyzed any kind of molecular system for which ground and excited densities on a grid can be computed. Furthermore, none of the previous steps requires defining a subdomain in the molecule, thus it does not introduce any arbitrary localization scheme.

Next, in order to provide a simple way to visualize the spatial extent of the transition, inspired by literature works aimed at giving a mathematical definition to the concept of size of an electron pair,^{7,8} we propose to define two centroids of density corresponding to the density depletion and the density increment regions. A qualitative measure of the overlap between the

Received: May 4, 2011

Published: June 23, 2011

centroids is then used to quantify the through-space character associated to a computed CT excitation.

Overall, the model here developed is finally aimed to be used to qualitatively screen push–pull compounds belonging to different chemical families in terms of length and magnitude of charge transferred in order to provide useful insights for experimental chemists. Besides it can also be shown that a simple qualitative diagnostic index for through-space CT excitations, of interest at time dependent DFT (TD-DFT) level, may be easily derived.

The paper is organized as follows: After a description of the method (Section 2) and the computational details (Section 3), the results obtained at a fixed level of theory for a family of molecules experimentally characterized are discussed in Section 4. Finally, in Section 5 some general insights about the applicability of these indexes to identify pathological behavior of time-dependent (TD)-DFT for the treatment of through-space CT transitions are given.

2. THE MODEL

Defining $\rho_{\text{GS}}(r)$ and $\rho_{\text{EX}}(r)$ as the electronic densities associated to the ground and excited states, the density variation associated to the electronic transition is given by

$$\Delta\rho(r) = \rho_{\text{EX}}(r) - \rho_{\text{GS}}(r) \quad (1)$$

Two functions, $\rho_+(r)$ and $\rho_-(r)$, defining the points in space where an increment or a depletion of the density upon absorption is produced ($\Delta\rho(r)$) can be defined as follows:

$$\rho_+(r) = \begin{cases} \Delta\rho(r) & \text{if } \Delta\rho(r) > 0 \\ 0 & \text{if } \Delta\rho(r) < 0 \end{cases} \quad (2)$$

$$\rho_-(r) = \begin{cases} \Delta\rho(r) & \text{if } \Delta\rho(r) < 0 \\ 0 & \text{if } \Delta\rho(r) > 0 \end{cases} \quad (3)$$

The barycenters of the spatial regions defined by $\rho_+(r)$ and $\rho_-(r)$, referred in the following as R_+ and R_- , can thus be defined [for instance discretizing it on a three-dimensional (3D) grid around the molecule] as

$$R_+ = \frac{\int r\rho_+(r)dr}{\int \rho_+(r)dr} = (x_+, y_+, z_+) \quad (4)$$

$$R_- = \frac{\int r\rho_-(r)dr}{\int \rho_-(r)dr} = (x_-, y_-, z_-) \quad (5)$$

The spatial distance between the two barycenters of density distributions can thus be used to “measure” the CT excitation length D_{CT} as

$$D_{\text{CT}} = |R_+ - R_-| \quad (6)$$

Integrating over all space ρ_+ (ρ_-), the transferred charge (q_{CT}) can be defined. For one electron excitation, q_{CT} can assume values between 0 and 1. Analogously, a variation in dipole moment between the ground and the excited states (μ_{CT})

can be defined, and its norm is computed by the following relation:

$$\begin{aligned} \|\mu_{\text{CT}}\| &= D_{\text{CT}} \int \rho_+(r)dr = -D_{\text{CT}} \int \rho_-(r)dr \\ &= D_{\text{CT}}q_{\text{CT}} \end{aligned} \quad (7)$$

This value has to be identical to the difference between the dipole moments computed for the ground and the excited states, that is $\Delta\mu_{\text{ES-GS}}$. In order to lighten notations, the $\|\mu_{\text{CT}}\|$ term will be replaced in the following by μ_{CT} .

For visualization purposes, it is also interesting to define two centroids of charges associated to the positive and negative density regions. To this end, first the root-mean-square deviations along the three axis (σ_{aj} , $j = x, y, z$; $a = +$ or $-$) are computed as

$$\sigma_{a,j} = \sqrt{\frac{\sum_i \rho_a(r_i)(j_i - j_a)^2}{\sum_i \rho_a(r_i)}} \quad (8)$$

The two centroids (C_+ and C_-) can then simply be defined as

$$C_+(r) = A_+ e \left(-\frac{(x-x_+)^2}{2\sigma_{+x}^2} - \frac{(y-y_+)^2}{2\sigma_{+y}^2} - \frac{(z-z_+)^2}{2\sigma_{+z}^2} \right) \quad (9)$$

$$C_-(r) = A_- e \left(-\frac{(x-x_-)^2}{2\sigma_{-x}^2} - \frac{(y-y_-)^2}{2\sigma_{-y}^2} - \frac{(z-z_-)^2}{2\sigma_{-z}^2} \right) \quad (10)$$

The normalization factors (A_+ and A_-) can be chosen so as to impose the integrated charge on the centroid to be equal to the corresponding density change integrated in the whole space:

$$A_+ = \frac{\int \rho_+(r)dr}{\int e \left(-\frac{(x-x_+)^2}{2\sigma_{+x}^2} - \frac{(y-y_+)^2}{2\sigma_{+y}^2} - \frac{(z-z_+)^2}{2\sigma_{+z}^2} \right) dr} \quad (11)$$

$$A_- = \frac{\int \rho_-(r)dr}{\int e \left(-\frac{(x-x_-)^2}{2\sigma_{-x}^2} - \frac{(y-y_-)^2}{2\sigma_{-y}^2} - \frac{(z-z_-)^2}{2\sigma_{-z}^2} \right) dr} \quad (12)$$

This definition has been used throughout in the present work. By definition this normalization imposes also the physical constrain that $\int \rho_-(r)dr = -\int \rho_+(r)dr$

For purposes of analysis of rod-like dyads, it may be of interest to define an index (named H) as half of the sum of the centroids axis along the D – A direction. For instance, if the D – A direction is along the x axis, H is defined by the relation:

$$H = \frac{\sigma_{+x} + \sigma_{-x}}{2} \quad (13)$$

For $H \geq D_{\text{CT}}$, an overlap between the centroids along this axis is thus expected.

Finally, the t represents the difference between D_{CT} and H :

$$t = D_{\text{CT}} - H \quad (14)$$

3. COMPUTATIONAL DETAILS

At the ground state, the structure of all systems was optimized at DFT level using the hybrid PBE0⁹ functional and the 6-31+G(d) basis set.¹⁰ All calculations were performed using the Gaussian package,¹¹ and bulk solvent (ethanol, EtOH) effects were taken into account using a continuum solvation model (C-PCM).^{12,13} When not differently specified, vertical excitation energies and associated excited state densities were computed at TD-DFT at the same level of theory. Five excited states were computed for all systems analyzed. Using PBE0 ground-state structures, the first five excited states were also computed at CIS level and (using the TD-DFT approach) also using the Perdew, Burke, and Ernzerhof (PBE)¹⁴ and LC-PBE¹⁵ functional, always keeping the 6-31+G(d) basis set.

In this work, densities (for both the ground and the excited states) were computed on a grid of points using the cubegen utility (and associated default options) provided by the Gaussian package.¹¹ All density derived quantities were computed using numerical integration procedure by using internally developed software. The consistency of the size of the box used has been checked computing the net fraction of escaped electrons (δ_{esc}), that is

$$\delta_{\text{esc}} = N - \int_{\text{box}} \rho^{\text{box}}(r) dr \quad (15)$$

where N is the total number of electrons in the molecule and $\rho(r)$ is the total ground-state density.

4. A TEST CASE: A SIMPLE D- π -A FAMILY OF MOLECULES

4.1. Structural and Electronic Properties. In order to test the performance of our index, we considered the family of molecules depicted in Figure 1, which can be considered as prototypes of donor–acceptor systems (dyads) expected to show a significant intramolecular CT character for the first excitation. In these molecules, spectroscopically characterized,¹⁶ the length of CT

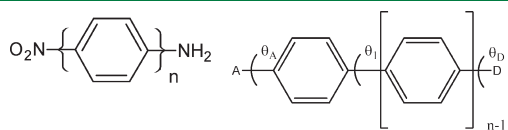


Figure 1. The family of molecules considered in the present work (right) and their labeling scheme (right), $n = 1-5$.

Table 1. Computed (PBE0/6-31+G(d)/PCM-EtOH) Structural Parameters (angles in $^\circ$, distances in \AA) for the P_n Systems^a

n	$dN_D N_A$	$\Theta_A d(N_A C)$	$\Theta_D d(N_D C)$	$\Theta_1 d(C-C)_1$	$\Theta_2 d(C-C)_2$	$\Theta_3 d(C-C)_3$	$\Theta_4 d(C-C)_4$	$\Theta_5 d(C-C)_5$
1	5.579	0.05 1.429	-12.15 1.355	—	—	—	—	—
2	9.933	0.37 1.447	-23.82 1.380	-31.23 1.471	—	—	—	—
3	14.262	0.47 1.451	-25.27 1.386	-33.83 1.475	33.31 1.476	—	—	—
4	18.585	0.28 1.452	-25.98 1.388	-35.14 1.476	35.01 1.478	-34.40 1.477	—	—
5	22.906	0.22 1.452	-25.91 1.388	-35.03 1.477	35.23 1.478	-35.24 1.478	34.07 1.477	—

^a For labeling refer to Figure 1. The phenyl spacers are numbered going from A to D, $d(CC)_x$ distances represent the distance between carbon atoms belonging to the x and $x + 1$ phenyl groups.

from the donor (amino, NH_2 noted D) to the acceptor (nitro, NO_2 noted A) group is modulated by the presence of an increasing number of conjugated phenyl spacers (π_n , $n = 1-5$). In the following these dyads will therefore be labeled as P_n , where n represents the number of phenyl spacers present in the molecule.

The main structural parameters corresponding to the optimized structures at PBE0/6-31+G(d) level are reported in Table 1, while the computed excitation energies and oscillator strengths associated to the first five transitions are collected in Table 2. As can be noted all P_n molecules present a staggered conformation of the phenyl rings, with an inter-ring dihedral angle going from ca. 31° for the shortest molecules to ca. 35° for P_n systems with $n \geq 4$ and a practically constant carbon to carbon inter-ring bond (between 1.471 and 1.478 \AA). The coupling of the donor and acceptor groups to the phenyl spacer, which is ruled both by the N_x-C distances and the Θ_x dihedral angles ($x = A, D$), is also comparable for all the systems analyzed (Table 1). Therefore we can reasonably assume that from a purely geometrical point of view the same kind of coupling is present in all P_n systems and that the only relevant geometrical parameter is the increase in distance between the donor and acceptor units. In this respect, they constitute an ideal benchmark for our purposes.

The CT excitation for all systems corresponds basically to an one-electron highest occupied molecular orbital–lowest unoccupied molecular orbital (HOMO–LUMO) excitation. The computed vertical excitation energies (Table 2) are strongly underestimated at PBE0 level for all systems but P1. Only for P1, the error on computed transition energies (0.10 eV) is in line with what was expected at this level of theory when neglecting direct

Table 2. Computed (PBE0/6-31+G(d)/PCM-EtOH) and Experimental Transition Energies (in nm) Associated to the First Excitation for the P_n Systems along with the Corresponding Computed Oscillator Strengths and Experimental ϵ Values (in $\text{M}^{-1} \cdot \text{cm}^{-1}$ from ref 16)

N	λ_{calc}	f_{calc}	λ_{exp}	$\epsilon_{\text{exp}} (10^{-4})$
1	362	0.65	372 (EtOH)	1.68
2	460	0.69	377 (EtOH)	1.38
3	483	0.54	358 (EtOH)	1.65
4	478	0.38	340 (DMF)	3.7
5	470	0.27	—	—

solute solvent interactions (hydrogen bonding) that may be of importance in the case of EtOH.¹⁷ Indeed, if a good spatial proximity between the donor (HOMO) and the acceptor (LUMO) orbitals is computed for P1 (Figure 2) starting from the P2 system, only a weak overlap can be evidenced between the donor and the acceptor units (Figure 2).

The very large errors on computed transition energies obtained for P_n systems ($n = 2$ to 5, Table 2) are thus not surprising, since the underestimation and the collapse of through-space CT transition energies when using TD-DFT with nonasymptotically corrected exchange correlation functionals or in absence of correction to the linear response approach is very well documented in literature (see for instance refs 18–31). Fixing to the standard LR-TD-DFT, the use of range separated functionals, such as LC-PBE should cure this problem although vertical excitation energies computed at this level of theory may not be quantitative agreement with experimental data.^{32–36}

4.2. CT Indexes at PBE0 Level. In this paragraph the indexes obtained at PBE0 level will be discussed in order to show their potential interest for the description of the nature of the CT states. Indeed, the real nature of the excited states, which are better described using range-separated hybrids, will be discussed in the following sections.

Using the total density computed for the ground and excited states, it is possible to evaluate $\Delta\rho$, ρ_+ , and ρ_- on a grid of points around the molecule as defined by eqs 1–3. The density variation upon excitation ($\Delta\rho$), computed for the first electronic transition at PBE0 level, is graphically depicted in Figure 3, the green and red zones corresponding to ρ_+ and ρ_- , respectively.

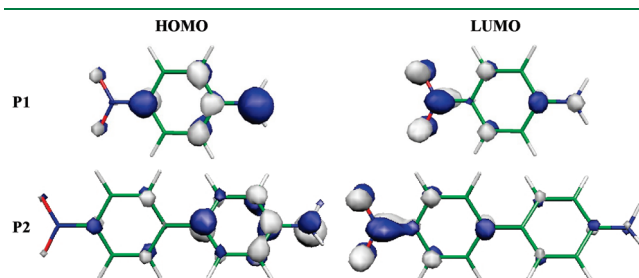


Figure 2. Computed (PBE0/6-31+G(d)/PCM) HOMO and LUMO orbitals for the P1 (top) and P2 (bottom) systems. Isocontour value 0.005 au.

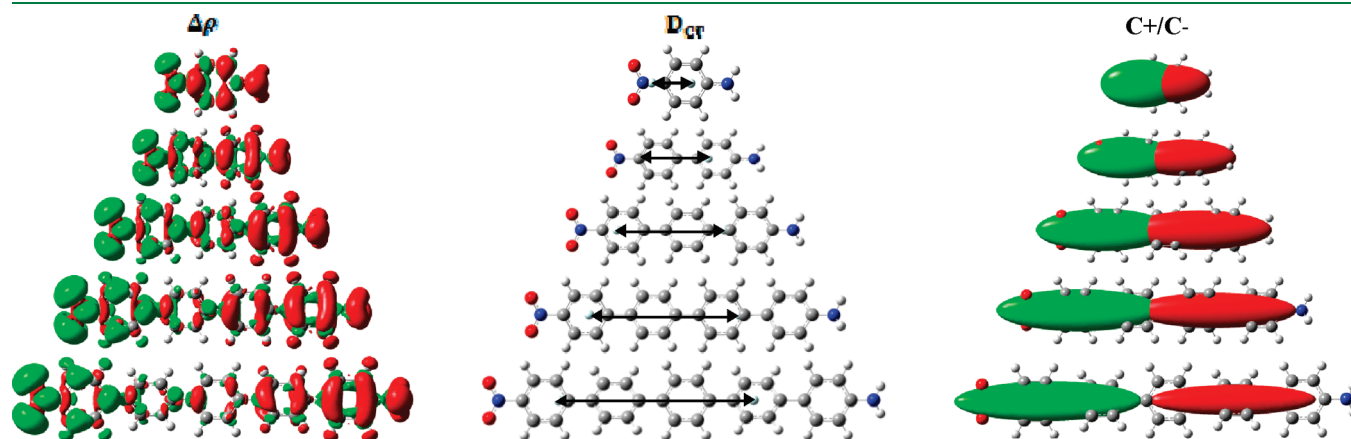


Figure 3. Computed (PBE0/6-31+G(d)/PCM) difference in total density computed for the ground and excited states ($\Delta\rho(r) = \rho_{\text{EX}}(r) - \rho_{\text{GS}}(r)$, isocontour value 0.001 au), graphical representation of D_{CT} , and centroids of charge ($C_+(r)/C_-(r)$, isocontour value 0.001 au).

As expected the density depletion zones (red, Figure 3) are mostly located on the NH_2 donor group but actually quite delocalized on the phenyl(s) directly connected to it. This finding is consistent with a weak donor character of the phenyl ring. On the other hand the regions of density increment (green, Figure 3) look more localized on the acceptor moiety, formally corresponding to the nitro (NO_2) group and, only to a minor extent, to the phenyl ring directly connected to it.

The use of the barycenter of charge (R_+ and R_- , eqs 4 and 5) allows to quantify these qualitative observations. In fact, while the R_- barycenter is very close to the NO_2 group (Figure 3) in the case of R_+ a displacement toward the phenyl rings closer to the amino group is clearly computed, thus highlighting their donor character.

As a consequence, the length of CT (D_{CT}) does not correspond to the distance between the donor and the acceptor groups ($d(\text{N}_\text{A}\text{N}_\text{D})$) as clearly evident from the data collected in Table 3 and in Figure 4. In practice, due to the delocalization of mainly the donor group, while the physical distance between A and D increases ca. 4.33 Å for each phenyl unit added, the corresponding increase in D_{CT} is only of ca. 2.36 Å. Indeed, the length of the CT, D_{CT} , increases more slowly than the distance $d(\text{N}_\text{A}\text{N}_\text{D})$ when phenyl spacers are added. This results in an A coefficient lower than 1 in Figure 4.

Table 3. Computed (PBE0/6-31+G(d)/PCM-EtOH) Distances between Donor and Acceptor Groups ($d(\text{N}_\text{D} - \text{N}_\text{A})$, in Å), Length, Charge and Dipole Moment (D_{CT} , q_{CT} and μ_{CT} in Å, $|e^-|$ and Debye, respectively) Associated to the CT Excitation Together with the Computed Fraction of Escaped Electron (δ_{esc} , in $|e^-|$) and the Difference between Ground- and Excited-State Dipole Moments ($\Delta\mu_{\text{GS-ES}}$, in Debye)^a

N	PBE0						H
	$d(\text{N}_\text{D} - \text{N}_\text{A})$	D_{CT}	q_{CT}	μ_{CT}	$\Delta\mu_{\text{GS-ES}}$	δ_{esc}	
1	5.58	2.6	0.5	6.6	6.6	0.07	2.0
2	9.93	4.9	0.8 _s	20.1	20.1	0.06	3.1
3	14.26	7.2	1.0	36.4	36.4	0.00	4.3
4	18.59	9.5	1.2	52.9	53.0	0.04	5.4
5	22.91	11.9	1.2	69.9	69.9	0.17	6.6

^aThe last column (H , in Å) represents the half sum of the centroids axis length along the D – A direction.

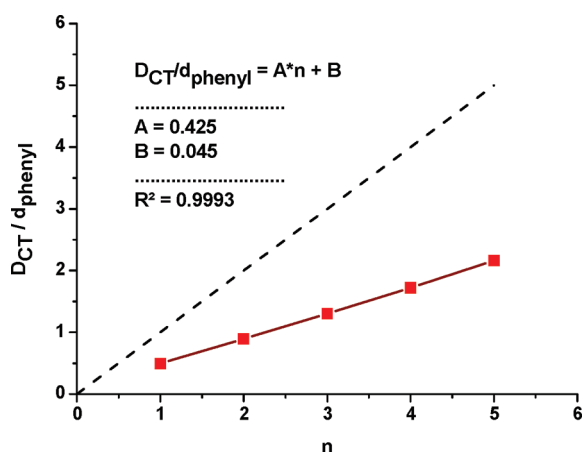


Figure 4. Normalized D_{CT} (D_{CT}/d_{phenyl} , d_{phenyl} being the average phenyl length set to 4.33 Å) as a function of the number of phenyl spacer, n .

By integration of ρ_+ (ρ_-) over all space, the transferred charge can be evaluated, and the numerical values are reported in Table 3. Clearly, if for the P1 system the transition has still a local character (D_{CT} and q_{CT} being only 2.5 Å and 0.5 $|e^-|$, respectively) the transition computed for the P2 dyad shows a net CT character, the computed q_{CT} value being 0.8 $|e^-|$. Of note, for the larger systems (P4 and P5) unphysical, that is larger than 1, values of q_{CT} (of the order of 1.2–1.3 $|e^-|$) are computed. This finding is related to the accuracy of the sampling of the density over a grid. As a matter of fact, the integration volume (box) chosen for these systems is too small, as demonstrated by the computed escaped density, which leads to 0.04–0.22 $|e^-|$ computed to be out of our integration grid (δ_{esc} in Table 3). The value of the escaped charge for the excited state is, in the present case, strictly equivalent to that computed for the ground state. The computed change in dipole moment between the ground and excited states is evaluated using either D_{CT} and q_{CT} by eq 7 (μ_{CT} , Table 3) or as expectation values over the density of the ground and excited states ($\Delta\mu_{GS-ES}$, Table 3) that are thus numerically equivalent. Nevertheless, in general cases particular care should be taken when using diffuse orbitals which may contribute more to the description of the excited states, thus determining a larger value for the escaped charge at the excited state.

It is worth noting that none of the quantities discussed up to now (D_{CT} , q_{CT} , and μ_{CT}) depend on an arbitrary localization scheme but only on the quality of the density used for the ground and the excited state. Therefore, they could be used to fairly compare different levels of theory used to evaluate the ground- and excited-state densities. This point will be addressed in the next section.

In order to visualize the spatial extent and the overlap between the regions of density depletion and increment, two centroids of charge have been defined using eq 9 and 10. The computed centroids are reported in Figure 3. Using this arbitrarily condensed function, on one hand, it is much easier to identify the spatial extent of the “real” donor and acceptor moieties and, on the other hand, to visualize the presence (or not) of overlap between the ρ_+ and ρ_- regions. In particular, it is clear that starting from P2, a very weak spatial proximity between the donor and the acceptor units is computed at PBE0 level.

More quantitatively, starting from P2, D_{CT} becomes much larger than H thus pointing out the presence of a through-space CT, which, as reported in literature, could introduce artifacts in computed transitions when using TD-DFT in conjunction with nonasymptotically corrected functionals.^{19–27}

5. CT INDEXES AND PERFORMANCE OF DIFFERENT FUNCTIONALS: A DIAGNOSTIC TOOL?

In order to clarify if the previously defined indexes can also be used as a diagnostic tool for the functional behavior in the reproduction of CT excitations, the first five transitions for systems P1–P4 were computed at different levels of theory, namely at TD-DFT (using the PBE, PBE0, LC-PBE functionals) and post-HF (CIS) levels.

In this context, it should be noted that although the origin of the failure of standard functionals in the description of CT excited state has been largely debated, explained, and pointed out, for the time being, practically only one efficient diagnostic index can be found in literature. This index (Λ), introduced by Tozer in 2008³⁷ and varying between 0 and 1, defines the spatial overlap between orbitals involved in a given electronic excitation computed using a single reference formalism (TD-DFT or CIS, for instance). Whenever the overlap is small, a through-space CT is predicted, and thus a significant error is expected when using standard GGA or hybrid functionals. Several papers by Tozer and collaborators^{37–39} have confirmed the validity of such a diagnostic index, applying it to different molecular systems. As a general conclusion it seems that for transitions characterized by a small Λ , both standard GGA and hybrid functionals will yield extremely large errors. Indeed, Λ alone, contrary to D_{CT} , cannot be used to identify if a given transition is a CT one or not, since CT excitations with substantial overlap will both be correctly predicted at the TD-DFT level and give a large Λ .^{37–39}

Although our index is not aimed at being a quantitatively diagnostic tool, its first aim being related to the definition of the spatial extent and magnitude of a CT transition, here we would like to test if D_{CT} and the centroids already both could be used to screen the CT character and as a first indicator of TD-DFT failures, at no extra computational cost.

To this end, the transition energies, oscillator strengths, and character of the first five transitions computed at different levels of theory are compared and collected in Table 4. The CT indexes of selected transitions, discussed in the text and graphically depicted in Figure 5, are reported in Table 5.

From the data reported in Table 4 several points concerning functional performances could be clearly derived. In particular, for P1 all methods provide a reasonable description of both the transition energies and their character. At all levels of theory, the most intense transition, the first one, corresponds to a CT excitation from the NH_2 to NO_2 groups (in the following noted as NN), with computed energies ranging from 406 (PBE) to 306 nm (CIS), the best agreement with experiment being obtained at PBE0 level (360 versus 372 nm experimentally determined).

From the analysis of the CT indexes reported in Table 5, it is also clear that for P1 all levels of theory predict a CT transition with substantial overlap between the centroids of charge representing the zones of increase and decrease of electron density upon excitation, with the computed value of D_{CT} being very close to the H value.

This means that for P1, this intense transition, although having a CT character, is indeed not a through-space one, and it is

Table 4. Computed Transition Energies (in nm), Oscillator Strengths, and Character for the P_n Class of Compounds (*n* = 1–4)^a

<i>n</i>	PBE			<i>n</i>	PBE0			<i>n</i>	LC-PBE			<i>n</i>	CIS		
	λ	<i>f</i>	type		λ	<i>f</i>	type		λ	<i>f</i>	type		λ	<i>f</i>	type
P1															
1	406	0.55	NN	1	362	0.65	NN	1	312	0.75	NN	1	306	1.10	NN
2	349	0.00	NO ₂	2	315	0.00	NO ₂	2	303	0.00	NO ₂	2	237	0.00	NO ₂
3	323	0.01	PN	3	282	0.01	PN	3	259	0.01	NO ₂	3	226	0.01	PN
4	288	0.00	NO ₂	4	267	0.00	NO ₂	4	248	0.01	–	4	218	0.00	–
5	280	0.06	NP	5	253	0.11	NP	5	215	0.14	–	5	208	1.14	–
P2															
1	641	0.49	NN	1	460	0.69	NN	1	312	1.09	NN	1	312	1.57	NN
2	403	0.01	PN	2	319	0.00	NO ₂	2	305	0.09	NO ₂	2	241	0.00	NO ₂
3	356	0.00	NO ₂	3	303	0.01	PN	3	265	0.00	NO ₂	3	240	0.71	–
4	348	0.01	–	4	298	0.03	PN	4	253	0.08	–	4	234	0.06	–
5	342	0.01	NN	5	279	0.01	NN	5	250	0.01	–	5	226	0.01	–
P3															
1	839	0.28	NN	1	483	0.54	NN	1	306	0.08	NO ₂	1	303	2.03	PN/NN
2	440	0.00	PN	2	334	0.52	PN	2	299	1.64	NN/PN	2	255	0.06	–
3	439	0.42	PN	3	319	0.03	NO ₂	3	266	0.01	NO ₂	3	242	0.00	NO ₂
4	414	0.02	PN	4	308	0.56	NP	4	253	0.07	NN/PN	4	242	1.16	–
5	384	0.44	NN	5	306	0.00	PN	5	253	0.05	–	5	240	0.01	–
P4															
1	940	0.13	NN	1	478	0.38	NN	1	305	0.02	NO ₂	1	298	2.44	PN
2	526	0.35	PN	2	371	0.66	PN	2	295	2.25	PN	2	271	0.18	–
3	456	0.00	PN	3	332	0.88	NP	3	266	0.00	NO ₂	3	242	0.00	NO ₂
4	437	0.00	PN	4	319	0.00	NO ₂	4	265	0.17	NN	4	242	1.12	–
5	436	0.54	NP	5	304	0.00	PN	5	253	0.07	–	5	240	0.27	–

^aThe following notation is used: NN = NH₂ to NO₂ CT; NO₂ = NO₂ centered excitation; PN = phenyl to NO₂ excitation; NP = NH₂ to phenyl excitation; – = π - π^* excitation.

Table 5. Computed CT indexes for P_n systems (*n* = 1 to 5) for selected transitions at different level of theory. Experimental values from ref.¹⁶ λ is nm and other indexes are in Å

	P1				P2				P3				P4			
	λ	<i>D</i> _{CT}	<i>H</i>	<i>t</i>	λ	<i>D</i> _{CT}	<i>H</i>	<i>t</i>	λ	<i>D</i> _{CT}	<i>H</i>	<i>t</i>	λ	<i>D</i> _{CT}	<i>H</i>	<i>t</i>
PBE	406	2.5	2.0	0.5	641	5.0	3.2	2.8	839	7.7	4.3	3.4	940	10.2	5.5	4.7
									439	7.1	4.0	3.1	526	7.8	4.8	3.0
PBE0	362	2.6	2.0	0.6	460	5.0	3.1	2.9	483	5.9	4.0	1.9	478	9.5	5.5	4.0
									334	4.6	3.0	1.6	371	7.6	5.2	2.4
LC-PBE	312	2.5	2.0	0.5	312	3.7	3.2	0.5	299	3.8	4.2	–0.4	295	3.4	5.0	–1.6
CIS	306	2.1	2.0	0.1	312	3.0	3.0	0.0	303	3.0	5.1	–2.1	298	2.6	5.4 ₅	2.8 ₅
expt	372				377				358				340			

therefore expected to be well predicted using standard GGA or hybrid functional, as it is in the present case. The computed centroids (graphically depicted in Figure 5) provide a simple pictorial representation of this concept.

By analysis of Table 4 it could also be noted that, starting from P2, a clearly different behavior of PBE and PBE0 with respect to LC-PBE and CIS is observed. In particular, for P2, although all the approaches predict a NN type excitation, while PBE and PBE0 predict a significant decrease in transition energies going from P1 to

P2 (of the order of 1.12 and 0.73 eV for PBE and PBE0, respectively), this energy is computed as almost constant in the case of CIS and LC-PBE, in better agreement with the experimental trend.

The absolute error in transition energy computed for P2 at both PBE and PBE0 level is indeed extremely large (1.36 and 0.59 eV, respectively) when compared to their standard performances for organic molecules.¹⁷

The situation gets even worse when going to P3 and P4. Again, not only the transition energy computed at PBE and PBE0 levels

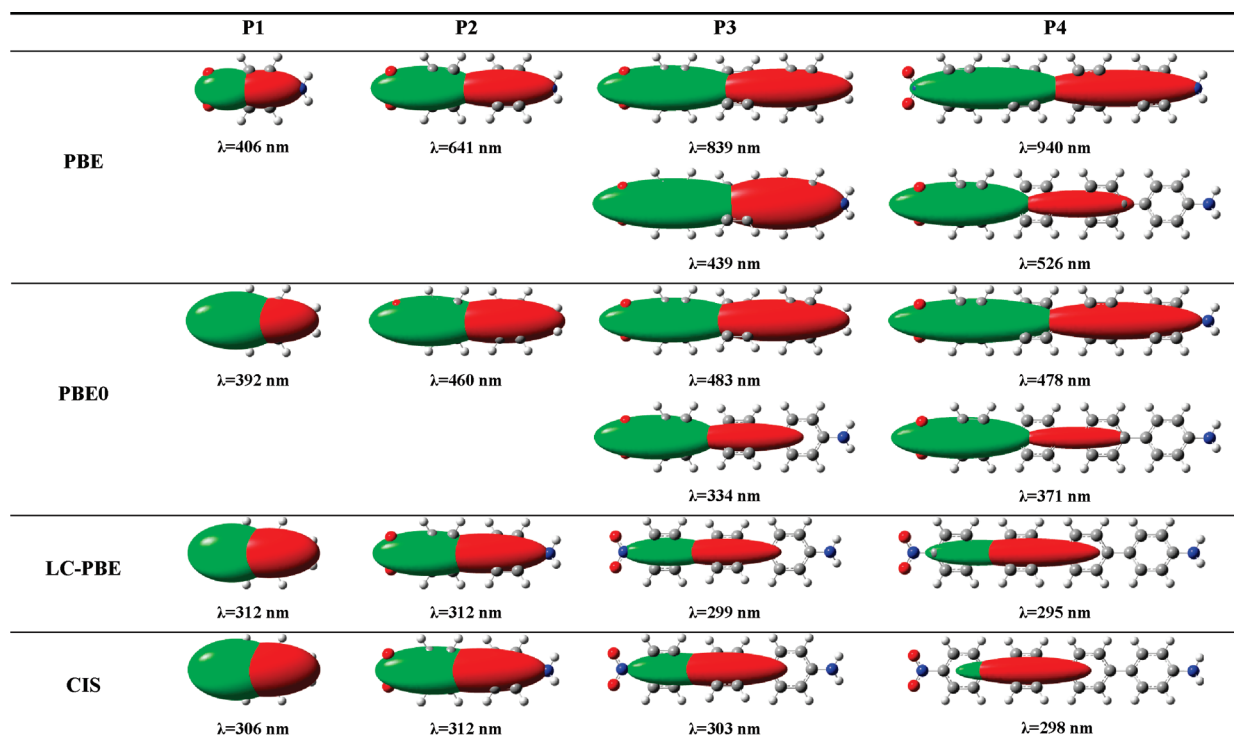


Figure 5. Computed transition wavelength and corresponding centroids for the highlighted transitions of Table 4 (isocontour value 0.001 au).

decreases with the increase of the length of the bridge, contrary to LC-PBE and CIS, which show a constant or slightly increasing energy in qualitative agreement with experimental data, but also and more severely the character of the most intense transition changes as a function of the method considered. In particular, while at CIS and LC-PBE levels, a CT excitation showing a significant contribution on the phenyl group linked to the NO₂ group is found (PN/NN), PBE and PBE0 still predict a full CT excitation from the donor to the acceptor (NN-type) at very low energies (839 and 483 nm for P3 and 940 and 478 nm for P4) and a second excitation, of PN-type, occurring at higher energies (439 and 334 nm for P3 and 526 and 371 nm for P4). These two transitions are both intense although, increasing *n*, the PN transition gains in intensity with respect to the, lower in energy, NN one.

As a consequence, going from P1 to P4, the CT length (D_{CT}) associated to the first transition always increases for PBE and PBE0 (and linearly correlates with the NO₂ to NH₂ distance, as discussed in Section 4.2), while it does not for CIS and LC-PBE (refer to Table 5).

In particular, for these latter two methods, a ‘saturation’ of the CT length is computed starting from P2, and even a slight contraction of D_{CT} is computed when going from P3 to P4. This finding qualitatively correlates with the experimentally observed transition energies only slightly increasing going from P1 to P3, although neither LC-PBE nor CIS provides transition energies in quantitative agreement with the experiment, as expected.^{17,32}

Physically, it also corresponds to the fact, while the acceptor keeps localized nearby the NO₂ group for all systems, the donor, primarily represented by the NH₂ group, actually delocalizes to the phenyl groups when increasing the bridge length and when starting from a three phenyl units bridge, is actually represented only by phenyl groups, as evident from the analysis of the centroids of charge (Figure 5).

Only at LC-PBE and CIS levels, going from P2 to P4, also the acceptor moiety shifts from the NO₂ group to the phenyl linked to it, thus finally giving rise to a very short-range CT transition (Figure 5). Analysis of the centroid allows also to easily visualize the tendency toward delocalization of GGA functional (larger centroids) with respect to the more localized character of hybrid and range-separated functionals.

As a consequence, analyzing the CT indexes computed at different levels of theory (Table 5), it could be noted that at both the CIS and LC-PBE levels, the CT length (D_{CT}) only slightly increases going from P1 to P2, keeping constant from P2 to P3 and slightly contracting while passing from P3 to P4. This behavior closely parallels the experimentally observed transition energies. As a consequence, for all systems the CT length is computed between 2.5 Å (P1) and 3.8 Å (P3) at LC-PBE level and between 2.1 Å (P1) and 3.0 Å (P3) at CIS level.

On the other hand, PBE and PBE0 predict a much larger D_{CT} associated to the first intense transitions for all systems, except P1 for which all methods yield the same picture.

Indeed, it could be noticed that already starting from P2, both at PBE and PBE0 level for the first transition $D_{CT} \gg H$, thus implying the presence of a through-space CT. Not surprisingly starting from P2 the computed transition energies associated with the first electronic transition do not follow both the experimental and the LC-PBE or CIS trends.

The difference between D_{CT} and H , noted t (eq 14 and Table 5), seems thus a reasonable diagnostic index for TD-DFT transitions. From the data collected for the family of molecules currently investigated, a $t > 1.6$ Å points out a potentially problematic transition for standard GGA and hybrid functionals.

Of note this rule of thumb, derived from the analysis of the first transition, also applies to the second intense transition computed

at PBE and PBE0 level for P3 and P4, which indeed has the same PN character as that predicted at LC-PBE or CIS. Indeed, only in the case of P3 computed at PBE0 level, the t criteria is satisfied, thus pointing out that for P4, both PBE and PBE0 should in principle provide very large errors also for the second transition, which is actually the case.

6. CONCLUSIONS

A new simple index (D_{CT}) aimed at giving a measure of the length of the electron transfer associated to an electronic transition has been introduced and tested on a family of push–pull molecules derived from the 1-amino-4-nitrobenzene.

The index, based only on the computed electronic density for the ground and excited states, quantifies the charge-transfer (CT) length as the distance between the barycenters of the density increment and depletion regions upon electronic excitation. As a consequence, it could be computed at any level of theory providing both ground- and excited-state densities. Clearly this index will be exactly zero for any centrosymmetric system. In such a case the index should be evaluated on the corresponding symmetry irreducible subunits.

The definition of centroids of charges associated to the density increase and depletion zones centered on the barycenters of charge previously defined allows to visualize and quantify the spatial extent of the donor and acceptor groups within a given molecular system. For this reason both D_{CT} and the centroids can be of help for the design and the description of new push–pull systems at both the theoretical and the experimental levels.

Finally the possibility of using a quantity related to the overlap between the centroids (namely, t) has been tested as a diagnostic tool to point out the pathological behavior of TD-DFT for the description of through-space CT transitions. From the results obtained for the classes of molecule analyzed it seems that t can be used to highlight the through-space character of a given electronic transitions and, thus, as a first warning for possible TD-DFT failures.

As a consequence, the combined uses of D_{CT} and t will allow to define the character (CT or not) and the magnitude (in term of spatial extent, i.e., the length of transition) of a given transition and to be a first indicator of eventual CT pathologic cases for DFT. Further work is in progress in this direction in our group.

■ ASSOCIATED CONTENT

S Supporting Information. Optimized structures of P_n ($n = 1–5$) systems. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: tangui-le-bahers@chimie-paristech.fr; ilaria-ciofini@chimie-paristech.fr.

■ ACKNOWLEDGMENT

Ciro A. Guido (SNS, Pisa Italy) is thanked for fruitful discussions on CT indexes. Members of the COST Action CODECS are gratefully acknowledged for constructive comments. The French National Agency for Research (ANR) is acknowledged for financial support to I.C. and C.A. in the

framework of the “NEXUS project (Programme Blanc 2007, BLAN07-1-196405).

■ REFERENCES

- (1) (a) Bureš, F.; Pytela, O.; Kivala, M.; Diederich, F. *J. Phys. Org. Chem.* **2011**, *24*, 274–281. (b) Blanchard-Desce, M.; Wortmann, R.; Lebus, S.; Lehn, J.-M.; Krämer, P. *Chem. Phys. Lett.* **1995**, *243*, 526–532. (c) Pocker, Y.; Spyridis, G. T. *J. Am. Chem. Soc.* **2002**, *124*, 7390–7394.
- (2) Juris, A.; Balzani, V.; Barigelletti, F.; Belser, P.; Von Zelewsky, A. *Coord. Chem. Rev.* **1988**, *84*, 85–277.
- (3) Le Bahers, T.; Pauporté, T.; Scalmani, G.; Adamo, C.; Ciofini, I. *Phys. Chem. Chem. Phys.* **2009**, *11*, 11276–11284.
- (4) Ciofini, I.; Lainé, P. P.; Bedioui, F.; Adamo, C. *J. Am. Chem. Soc.* **2004**, *126*, 10763–10777.
- (5) Fortage, J.; Peltier, C.; Nastasi, F.; Puntoriero, F.; Tuyères, F.; Griveau, S.; Bedioui, F.; Campagna, S.; Lainé, P. P. *J. Am. Chem. Soc.* **2010**, *132*, 16700–16713.
- (6) Peltier, C.; Adamo, C.; Lainé, P. P.; Campagna, S.; Punteriero, F.; Ciofini, I. *J. Phys. Chem. A* **2010**, *114*, 8434–8443.
- (7) Robb, M. A.; Haines, W. J.; Csizmadia, I. G. *J. Am. Chem. Soc.* **1973**, *95*, 42–48.
- (8) Amhadi, G. R.; Røeggen, I. *Theor. Chem. Acc.* **1997**, *97*, 41–46.
- (9) Barone, V.; Adamo, C. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (10) Krishnan, R.; Binkley, J. S.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654.
- (11) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.
- (12) Miertuš, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117–129.
- (13) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995–2001.
- (14) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (15) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.
- (16) Berry, R. W. H.; Brocklehurst, P.; Burawoy, A. *Tetrahedron* **1960**, *10*, 109–117.
- (17) Jacquemin, D.; Perpète, E.; Ciofini, I.; Adamo, C. *Acc. Chem. Res.* **2009**, *42*, 326–334.
- (18) Tozer, D. *J. Chem. Phys.* **2003**, *119*, 12697–12699.
- (19) Tozer, D.; Amos, R. D.; Handy, N. C.; Roos, B. O.; Serran-Andres, L. *Mol. Phys.* **1999**, *97*, 859–868.
- (20) Fabian, J. *Theor. Chem. Acc.* **2001**, *106*, 199–217.
- (21) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007–4016.
- (22) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009–4037.
- (23) Hieringer, W.; Görling, A. *Chem. Phys. Lett.* **2006**, *419*, 557–562.
- (24) Dreuw, A.; Head-Gordon, M. *Chem. Phys. Lett.* **2006**, *426*, 231–233.
- (25) Kummel, S.; Kronik, S. *Rev. Mod. Phys.* **2008**, *80*, 3–60.
- (26) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. T. *Science* **2008**, *321*, 792–794.
- (27) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943–2946.

- (28) Ziegler, T.; Krykunov, M. *J. Chem. Phys.* **2010**, *133*, 074104.
- (29) Ziegler, T.; Seth, M.; Krykunov, M.; Autschbach, J.; Wang, F. *J. Mol. Struct. THEOCHEM* **2009**, *914*, 106–109.
- (30) Ziegler, T.; Seth, M.; Krykunov, M.; Autschbach, J. *J. Chem. Phys.* **2008**, *129*, 184114.
- (31) Görling, A.; Ipatov, A.; Götz, A. W.; Hesselmann, A. *Z. Phys. Chem.* **2010**, *224*, 03–04.
- (32) Jacquemin, D.; Perpète, E. A.; Scuseria, G. E.; Ciofini, L.; Adamo, C. *J. Chem. Theory Comput.* **2008**, *4*, 123–135.
- (33) Stein, T.; Kronik, L.; Baer, R. *J. Am. Chem. Soc.* **2009**, *131*, 2818–2820.
- (34) Livshits, E.; Baer, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932–2941.
- (35) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425–8433.
- (36) Rohrdanz, M. A.; Herbert, J. M. *J. Chem. Phys.* **2008**, *129*, 034107.
- (37) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.
- (38) Peach, M. J. G.; Tozer, D. J. *J. Mol. Struct.: THEOCHEM* **2009**, *914*, 110–114.
- (39) Peach, M. J. G.; Le Sueur, C. R.; Ruud, K.; Guillaume, M.; Tozer, D. J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4465–4470.

Ab Initio Trajectory Study on Triplet Ketene Photodissociation via Statistical Sampling of the Crossing Seam

Yusuke Ogihara, Takeshi Yamamoto,* and Shigeki Kato

Department of Chemistry, Kyoto University, Kyoto 606-8502, Japan

ABSTRACT: Triplet ketene exhibits a steplike structure in the experimentally observed dissociation rate, but its mechanism is still unclear despite many theoretical efforts. A previous surface-hopping simulation at the CASSCF level suggests that nonadiabatic transition from the S_0 to T_1 states creates the T_1 species in a highly nonstatistical manner, which raises the question of whether the use of statistical rate theory is valid in itself for the T_1 state. Here, we study this problem by performing ab initio trajectory simulation at the multireference second-order Möller–Plesset perturbation (MRMP) level of theory. Since the MRMP theory is too expensive for such a trajectory calculation, we first construct dual-level potential energy surfaces (PESs) for the S_0 and T_1 states by calibrating the PESs at the B3LYP level with a limited set of MRMP energies. We then introduce the assumption of vibrational equilibrium on the S_0 surface and characterize the $S_0 \rightarrow T_1$ crossing points using the conditional microcanonical distribution on the S_0/T_1 seam surface. The latter distribution is obtained by running a constrained trajectory on the seam surface by use of an efficient SHAKE-like method. Subsequently, we propagate a number of T_1 trajectories from the seam surface to obtain the dissociation rate. The result shows that (i) the $S_0 \rightarrow T_1$ crossing points are localized mainly in the T_1 reactant region; (ii) the lifetime on the T_1 surface is about 30 ps at the MRMP level, which is 2 orders of magnitude greater than the previous estimate obtained from the surface-hopping simulation at the CASSCF level (~ 100 fs); and (iii) the calculated T_1 dissociation rate agrees reasonably well with classical transition state theory. These results suggest that the T_1 dissociation is rather statistical, given that the T_1 trajectories are initiated from the conditional microcanonical distribution on the seam surface.

I. INTRODUCTION

In the early 1990s, Kim et al.^{1,2} performed an experimental study on the photodissociation of triplet ketene, $\text{CH}_2\text{CO} \rightarrow \text{CH}_2 + \text{CO}$. In their experiment, the ketene molecule in the S_0 state was first excited to the S_1 state via a UV laser, and the excited ketene underwent nonadiabatic transition to lower electronic states, followed by intersystem crossing from S_0 to T_1 states; see Figure 1 for a schematic of relevant potential energy surfaces (PESs). An interesting observation here is that the dissociation rate of triplet ketene, $k(E)$, exhibits a steplike structure as a function of total energy E near the dissociation threshold. The steplike structure in $k(E)$ was interpreted as arising from vibrationally quantized states at the transition state (TS). On the basis of this interpretation, the observed steps in $k(E)$ were discussed in terms of energy levels of vibrational modes orthogonal to the reaction coordinate, and semiquantitative agreement was obtained between the width of the observed steps and the level spacing of orthogonal modes at the transition state.²

However, a subsequent quantum dynamical study by Gezelter and Miller³ demonstrated that if one takes into account tunneling effects along the reaction coordinate, the steps in $k(E)$ should be eliminated (or “washed out”) by tunneling, which results in a nearly structureless profile of $k(E)$. The situation can be described more clearly by considering the microcanonical rate expression for unimolecular dissociation:^{4,5}

$$k(E) = \frac{N(E)}{2\pi\hbar\rho(E)} \quad (1)$$

where $N(E)$ is the cumulative reaction probability (CRP) and $\rho(E)$ is the density of reactant states. For the present system, the CRP can be approximated accurately as follows:⁶

$$N(E) \approx \sum_i \kappa(E - E_i^\ddagger) \quad (2)$$

where E_i^\ddagger is the energy of the i th vibrational level at the TS and $\kappa(E)$ is the one-dimensional transmission probability calculated along the reaction coordinate. In the classical limit, $\kappa(E)$ reverts to the Heaviside step function, $\kappa(E - E_i^\ddagger) \rightarrow \Theta(E - E_i^\ddagger)$, and the $N(E)$ always exhibits a steplike structure. In the quantum case, $\kappa(E - E_i^\ddagger)$ increases gradually from 0 to 1 when passing $E = E_i^\ddagger$, and if the width of $\kappa(E - E_i^\ddagger)$ is greater than the level spacing $E_i^\ddagger - E_{i-1}^\ddagger$, one can no longer observe a steplike structure in $N(E)$. By utilizing the accurate PES at the CCSD level obtained by Schaefer and co-workers,⁷ Gezelter et al. demonstrated that the width of $\kappa(E - E_i^\ddagger)$ is actually much greater than the spacing of energy levels $E_i^\ddagger - E_{i-1}^\ddagger$, resulting in a structureless profile of $k(E)$. They also suggested that in order for the steps to appear in $k(E)$, the imaginary frequency ω^\ddagger at the TS needs to be less than $\sim 100 \text{ cm}^{-1}$, which is 3–4 times smaller than the result at the CCSD level (379 cm^{-1}).⁷ Recently, we have also carried out an ab initio calculation of triplet ketene at the multireference levels of theory,^{6,8} but the obtained ω^\ddagger was about 300 cm^{-1} , and thus the steps in $k(E)$ were not reproduced. Here, it is worth noting that a similar discrepancy exists between theory and experiment for the photodissociation of triplet acetaldehyde.^{9,10} In the latter

Received: June 1, 2011

Published: July 06, 2011

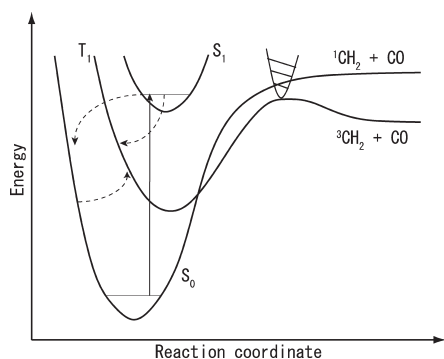


Figure 1. A schematic of S_0 , S_1 , and T_1 potential energy surfaces of triplet ketene. Dashed arrows indicate possible pathways of nonadiabatic transition. In this paper, we assume that the molecule reaches a vibrational equilibrium on the S_0 surface prior to the intersystem crossing to the T_1 surface. The direct $S_1 \rightarrow T_1$ pathway is neglected throughout this paper.

case, the imaginary frequency estimated from the experiment is 60 cm^{-1} , whereas accurate quantum-chemical calculations predict $300\text{--}500 \text{ cm}^{-1}$.

As such, possible mechanisms other than vibrational quantization at the TS were also explored for explaining the observed steps. Cui and Morokuma performed an extensive *ab initio* study on nonadiabatic interactions among the relevant PESs.¹¹ Their study suggested that $S_1 \rightarrow S_0$ internal conversion is very efficient, while the direct $S_1 \rightarrow T_1$ pathway is negligible due to the extremely small spin-orbit coupling as well as the high location of the S_1/T_1 minimum seam crossing (MSX) point. As a result, they concluded that $S_1 \rightarrow S_0 \rightarrow T_1$ is the dominant pathway for the triplet dissociation. Subsequently, Kaledin et al.¹² performed a direct surface-hopping trajectory calculation at the state-averaged (SA) CASSCF level. They propagated a trajectory, starting at the Franck-Condon geometry on S_1 , and branched it out into many child trajectories every time the propagating PES crossed with another PES. Their calculation suggested that (1) the $S_0\text{--}T_1$ nonadiabatic transition creates the T_1 species nonstatistically at restricted regions of phase space and (2) a large fraction of the T_1 species thus created dissociates almost immediately (within 120 fs), thus leaving no time for equilibration on the T_1 surface. Those observations suggest that the dissociation dynamics on the T_1 surface are highly nonstatistical, which raises the question of whether equilibrium-based rate theories [such as eq 1] are valid in themselves for describing the T_1 dissociation.

It should be pointed out, however, that the surface-hopping simulation by Kaledin et al.¹² was not without limitation (which is essentially due to the limited computational power available in 2001). First, they utilized the CASSCF method to describe the S_0 and T_1 surfaces, but in fact the latter method is not very accurate for describing the dissociation profile (see section II.A). Specifically, the dissociation barrier height on the T_1 surface is predicted to be 16 kcal/mol at the CASSCF level, while that at the multireference level of theory is 29 kcal/mol. The significantly smaller barrier height at the CASSCF level may be partly responsible for the fast dissociation observed in ref 12. Second, the surface-hopping simulation was performed with the “ant” method.^{13,14} However, the relatively small number of propagated trajectories (about 10) makes it difficult to assess the statistical significance of the results, particularly considering that the $S_0 \rightarrow T_1$ transition

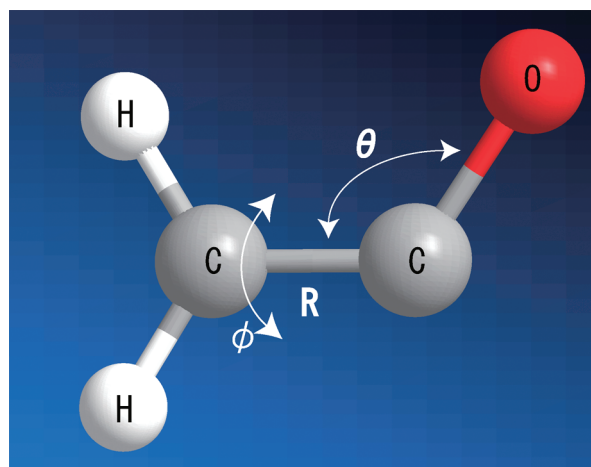


Figure 2. Transition state geometry of triplet ketene. R is the CC bond length, θ is the CCO bending angle, and ϕ is the torsional angle.

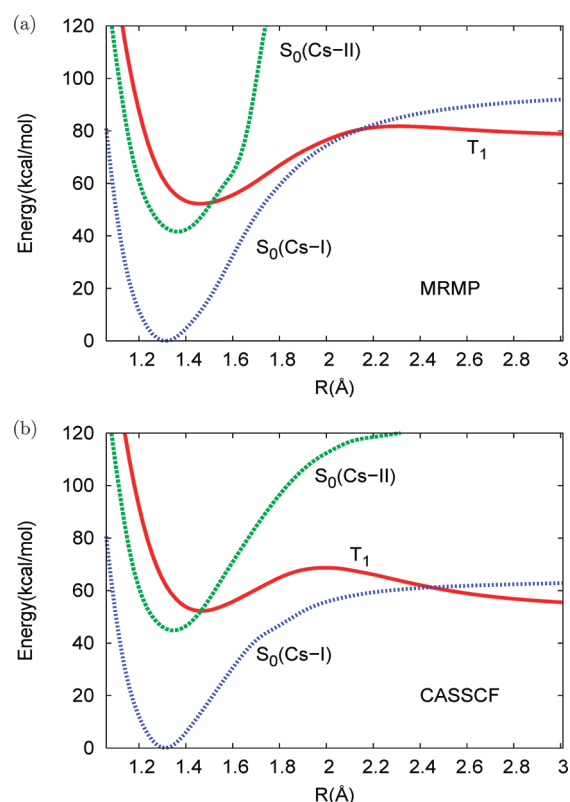


Figure 3. Minimum energy profiles of the S_0 and T_1 surfaces obtained with (a) MRMP and (b) CASSCF methods. Internal coordinates other than R are optimized at each value of R . The minimum energy path on the S_0 surface is of nonplanar $Cs\text{--}I$ symmetry, while that on the T_1 surface is of planar $Cs\text{--}II$ symmetry. The green line with label “ $S_0\text{--}(Cs\text{--}II)$ ” represents the S_0 potential energy curve calculated along the T_1 minimum energy path.

probability is very small (on the order of 10^{-5}).¹² Third, the trajectory calculation was performed in planar $Cs\text{--}II$ symmetry with the torsional angle ϕ set at zero (see Figure 2). This is based on the observation that all of the important geometries (including the S_0/T_1 MSX point) belong to planar $Cs\text{--}II$ symmetry. However, the minimum energy profile on the S_0 surface is known to be of

nonplanar Cs–I symmetry,⁷ and therefore it is desirable that no restriction on ϕ is applied in the trajectory calculation.

The purpose of this paper is thus to perform an ab initio trajectory calculation of triplet ketene by lifting some of the limitations mentioned above, while describing the relevant PESs more accurately with the multireference second-order Möller–Plesset perturbation (MRMP) theory. Specifically, we first construct dual-level PESs for the S_0 and T_1 states by combining the accuracy of the MRMP theory and the efficiency of the B3LYP method (section II.A). We then introduce the assumption of vibrational equilibrium on the S_0 surface and characterize the $S_0 \rightarrow T_1$ crossing points using the conditional microcanonical distribution on the S_0/T_1 seam surface (section II.B). The latter distribution is obtained by running a constrained trajectory on the seam surface via use of an efficient SHAKE-like method.^{15,16} Subsequently, we propagate a number of T_1 trajectories starting from the seam distribution thus obtained and calculate the dissociation rate on the T_1 surface. We emphasize that the present study does not make the assumption of vibrational equilibrium on the T_1 surface; rather, we assume that the molecule reaches a vibrational equilibrium on the S_0 surface prior to $S_0 \rightarrow T_1$ intersystem crossing. As such, comparison of the present trajectory result with statistical rate theory (based on a vibrational equilibrium on the T_1 surface) provides useful insight into the statistical degree of the T_1 dissociation (section III.B).

II. METHOD

II.A. Construction of Dual-Level Potential Energy Surfaces.

To perform an ab initio trajectory calculation, one needs to select an appropriate level of electronic structure theory that is not too expensive yet provides a sufficient accuracy for the relevant PESs. To do so, we first compare in Figure 3 several one-dimensional potential profiles calculated at the CASSCF and MRMP levels. Here, both the CASSCF and MRMP calculations were performed with the cc-pVTZ basis set,¹⁷ and the active space was constructed by distributing 14 electrons in 11 molecular orbitals. The minimum energy profiles of the S_0 and T_1 states in Figure 3 are obtained from energy minimization at fixed R . As noted above, the reaction path on the T_1 surface is of planar Cs–II symmetry, while that on the S_0 surface is of nonplanar Cs–I symmetry. Figure 3 also displays the S_0 potential energy profile calculated along the T_1 reaction path, which crosses the T_1 minimum energy profile at around $R = 1.5$ Å. This is consistent with the fact the S_0/T_1 MSX point is located near the T_1 minimum geometry.¹¹ The dissociation barrier height on the T_1 surface is calculated to be 16.3 and 29.3 kcal/mol at the CASSCF and MRMP levels, respectively. The MRMP result is in excellent agreement with the T_1 barrier height at the UCCSD(T) level⁷ (29.8 kcal/mol), while the CASSCF result is significantly smaller than the MRMP or UCCSD(T) result. This fact underlines the importance of dynamic electron correlation in the TS region.⁶ Furthermore, we compare in Figure 4 the minimum energy profiles obtained at the (U)B3LYP/6-31G* level with the CASSCF and MRMP results. We see that the B3LYP method gives a rather accurate PES for the S_0 state, whereas it is less accurate for the T_1 state.

In the following, we construct an approximate dual-level PES in order to combine the accuracy of the MRMP theory and the efficiency of the B3LYP method. (For previous applications of this type of method, see refs 18 and 19, for example.) Here, we choose the CC bond length R as the reaction coordinate and

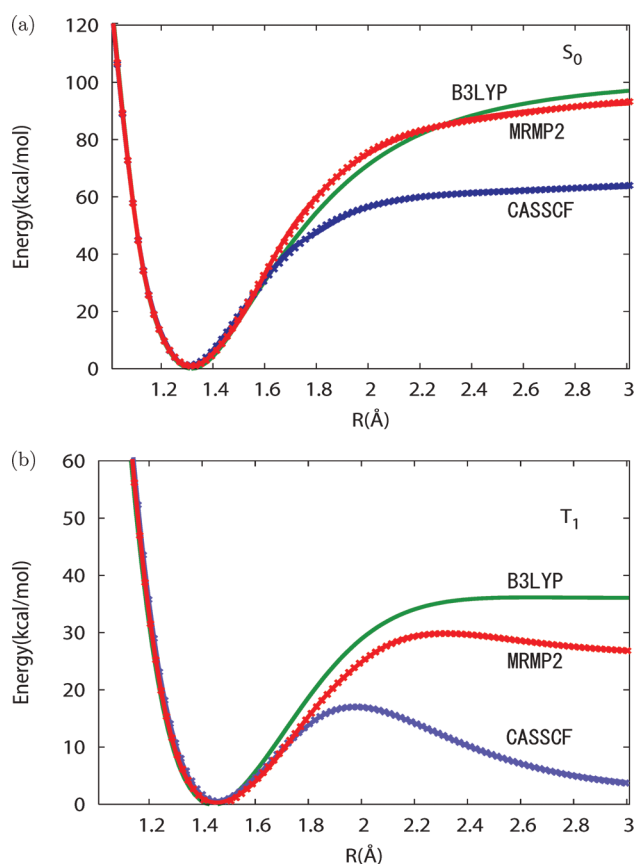


Figure 4. Minimum energy profiles calculated at the MRMP, CASSCF, and B3LYP levels (solid line): (a) S_0 state, (b) T_1 state. Minimum energy profiles of the MRMP::B3LYP and CASSCF::B3LYP potential functions are plotted with cross symbols, which agree almost exactly with the underlying MRMP and CASSCF ones.

consider the functional form

$$U_{DL}(\mathbf{r}) = U_{LL}(\mathbf{r}) + U_{HL}^{\min}(R) - U_{LL}^{\min}(R) \quad (3)$$

Here, \mathbf{r} is the Cartesian coordinates of the molecule and $U_{HL}^{\min}(R)$ and $U_{LL}^{\min}(R)$ are the minimum energy profiles of high- and low-level potential functions denoted as $U_{HL}(\mathbf{r})$ and $U_{LL}(\mathbf{r})$. We suppose that the latter are calculated with the MRMP and B3LYP methods, respectively. The minimum energy profiles of $U_{HL}(\mathbf{r})$ and $U_{LL}(\mathbf{r})$ are defined here as

$$U_{LL}^{\min}(R) = \min_{\mathbf{q}} U_{LL}(R, \mathbf{q}) \quad (4a)$$

$$U_{HL}^{\min}(R) = \min_{\mathbf{q}} U_{HL}(R, \mathbf{q}) \quad (4b)$$

where \mathbf{q} denotes the internal coordinates other than R . $U_{LL}^{\min}(R)$ and $U_{HL}^{\min}(R)$ are obtained from energy minimization at fixed R . The minimum energy profiles are then fit analytically with the following function:

$$U_{HL}^{\min}(R) - U_{LL}^{\min}(R) \approx \Delta U_{\infty} \left\{ 1 - \sum_{k=1}^4 c_k \exp[-(R - R_k)^k] \right\} \quad (5)$$

where $\{c_k\}$ and $\{R_k\}$ are adjustable parameters obtained from a least-squares fit. ΔU_{∞} represents the energy difference of the

products between the high- and low-level theories. Due to the definition in eq 3, the minimum energy profile of the dual-level PES coincides exactly with that of the high-level PES, i.e.,

$$\min_{\mathbf{q}} U_{\text{DL}}(\mathbf{r}) = U_{\text{HL}}^{\text{min}}(R) \quad (6)$$

This property ensures the good accuracy of the dual-level PES in the vicinity of the minimum energy path. On the other hand, the accuracy of $U_{\text{DL}}(\mathbf{r})$ in the direction of \mathbf{q} is determined by the low-level theory [see eq 3]. Since the B3LYP method often provides a good description of bound degrees of freedom, we expect that $U_{\text{DL}}(\mathbf{r})$ also behaves reasonably well as a function of \mathbf{q} (see below for numerical tests). The stationary points on the dual-level PES are determined by the condition

$$\frac{\partial U_{\text{DL}}(R, \mathbf{q})}{\partial R} = 0 \quad (7a)$$

$$\frac{\partial U_{\text{DL}}(R, \mathbf{q})}{\partial \mathbf{q}} = \frac{\partial U_{\text{LL}}(R, \mathbf{q})}{\partial \mathbf{q}} = 0 \quad (7b)$$

The second equation gives the optimized value of \mathbf{q} as a function of R , which will be denoted as $\mathbf{q}^*(R)$. The latter is obtained by minimizing the low-level potential at fixed R . Insertion of $\mathbf{q}^*(R)$ into the first equation gives

$$\frac{\partial U_{\text{DL}}(R, \mathbf{q}^*(R))}{\partial R} = \frac{dU_{\text{HL}}^{\text{min}}(R)}{dR} = 0 \quad (8)$$

which states that the stationary value of R coincides with that of the minimum energy profile of the high-level PES. As for the Hessian matrix of the dual-level PES, we have

$$\frac{\partial^2 U_{\text{DL}}(R, \mathbf{q})}{\partial R^2} = \frac{\partial^2 U_{\text{LL}}(R, \mathbf{q})}{\partial R^2} + \frac{\partial^2 U_{\text{c}}(R)}{\partial R^2} \quad (9a)$$

$$\frac{\partial^2 U_{\text{DL}}(R, \mathbf{q})}{\partial R \partial \mathbf{q}} = \frac{\partial^2 U_{\text{LL}}(R, \mathbf{q})}{\partial R \partial \mathbf{q}} \quad (9b)$$

$$\frac{\partial^2 U_{\text{DL}}(R, \mathbf{q})}{\partial \mathbf{q} \partial \mathbf{q}} = \frac{\partial^2 U_{\text{LL}}(R, \mathbf{q})}{\partial \mathbf{q} \partial \mathbf{q}} \quad (9c)$$

with $U_{\text{c}}(R) = U_{\text{HL}}^{\text{min}}(R) - U_{\text{LL}}^{\text{min}}(R)$. The Hessian matrix is thus largely determined by the low-level theory, while the element in the direction of R is determined by the high-level theory.

Hereafter, we will denote the dual-level potential function thus obtained as MRMP2::B3LYP, where a pair of high- and low-level theories are denoted with double colons. For comparison, we also constructed a dual-level CASSCF::B3LYP potential function. We calculated high-level potential energies using the GAMESS program package,²⁰ while low-level energies are calculated with the Firefly package.²¹ This is because the Firefly package provides a significant acceleration of computational speed (typically by a factor of 4) over the standard GAMESS program for the B3LYP method. To construct the MRMP2::B3LYP PES, we calculated the minimum energy profile $U_{\text{HL}}^{\text{min}}(R)$ approximately using the minimum energy path at the CASSCF level. This is because analytical gradients for the MRMP method are not available in the GAMESS program.²⁰ We emphasize, however, that the property in eq 6 holds even if we calculate $U_{\text{HL}}^{\text{min}}(R)$ approximately.

We now assess the accuracy of the dual-level potential functions thus obtained. First, we display in Figure 4 the

Table 1. Equilibrium Geometry on the S_0 Surface^a

	MRMP(DL)	CASSCF(DL)	CASSCF
CC	1.312	1.307	1.309
CH ₁	1.084	1.084	1.073
CH ₂	1.084	1.084	1.073
CO	1.176	1.176	1.164
∠HCH	120.0	120.0	120.0
∠HCC	120.0	120.0	120.0
∠CCO	180.0	180.0	180.0

^aMRMP(DL) and CASSCF(DL) stand for the dual-level MRMP::B3LYP and CASSCF::B3LYP potential functions, respectively. Bond lengths are in Ångstroms, and angles are in degrees.

Table 2. Equilibrium and TS Geometries on the T_1 Surface^a

	MRMP(DL)	MRMP ^a	CASSCF(DL)	CASSCF
equilibrium geometry				
CC	1.463	1.461	1.470	1.466
CH ₁	1.089	1.083	1.089	1.075
CH ₂	1.083	1.077	1.083	1.069
CO	1.203	1.203	1.202	1.194
∠HCH	119.3	120.1	119.5	120.0
∠HCC	120.4	120.0	120.5	120.2
∠CCO	126.6	126.2	126.7	128.6
transition state				
CC	2.311	2.312	2.000	2.000
CH ₁	1.084	1.079	1.085	1.075
CH ₂	1.082	1.077	1.082	1.069
CO	1.147	1.148	1.162	1.150
∠HCH	132.8	132.3	126.4	126.1
∠HCC	115.1	113.6	117.1	117.7
∠CCO	115.3	113.8	118.5	118.4

^aReference 6. ^aSee the caption of Table 1 for the notation.

minimum energy profiles obtained with the MRMP2::B3LYP and CASSCF::B3LYP functions for the S_0 and T_1 states (shown as cross symbols). It is seen that the latter agrees almost exactly with the minimum energy profiles of the high-level theory (solid line), which is ensured by the property in eq 6. As another test, we calculated the stationary points and normal frequencies on the dual-level PES, which are summarized in Tables 1–3. Here, the stationary points on the (original) CASSCF PES are determined by using analytical gradients, while the stationary points on the MRMP PES are obtained from a local least-squares fit.⁶ Tables 1 and 2 show that the agreement between the dual-level PESs and the underlying CASSCF and MRMP PESs is satisfactory considering the limited accuracy of the B3LYP method. As for normal frequencies, we see that the agreement is better for high-frequency modes while it is slightly worse for low-frequency modes. To improve the accuracy, one could employ a more complicated functional form for the dual-level potential. However, we did not enter this complexity and instead utilized the simple functional form in eq 3.

II.B. Propagation of *ab Initio* Trajectory on the Seam Surface. As mentioned in the Introduction, we assume a dynamical model in which the molecule stays for a long time

Table 3. Normal Mode Frequencies (in cm^{-1}) at the Equilibrium and TS Geometries on the T_1 Surface^a

	MRMP(DL)	MRMP ^a	CASSCF(DL)	CASSCF
equilibrium geometry				
CC-str	992	911	1009	950
asym-CH-str	3268	3355	3268	3419
sym-CH-str	3123	3188	3123	3258
CO-str	1785	1679	1789	1832
CH ₂ -scissor	1464	1418	1468	1548
CH ₂ -rock	1067	1019	1101	1098
CH ₂ -wag	786	727	785	687
CCO-bend	453	455	451	492
torsion	396	337	394	386
transition state				
CC-str	339i	328i	513i	556i
asym-CH-str	3369	3431	3350	3428
sym-CH-str	3144	3162	3146	3235
CO-str	2119	2060	2009	2014
CH ₂ -scissor	1136	1123	1219	1348
CH ₂ -rock	390	344	634	555
CH ₂ -wag	264	308	403	427
CCO-bend	203	184	305	286
torsion	147	108	191	175

^aReference 6. ^a See the caption of Table 1 for the notation.

on the S_0 surface prior to the $S_0 \rightarrow T_1$ intersystem crossing. This is based on the observation^{11,12} that the nonadiabatic transition probability from the S_1 to S_0 surface is on the order of 10^{-2} , while that from the S_0 to T_1 surface (via spin-orbit coupling) is on the order of 10^{-5} . We thus expect that the molecule undergoes a large number of vibrations on the S_0 surface, and thus intramolecular vibrational energy redistribution (IVR) proceeds. On the basis of this consideration, we assume that the S_0 species is described by the microcanonical distribution

$$f_{S_0}(\mathbf{r}, \mathbf{p}) = \delta[E - H_{S_0}(\mathbf{r}, \mathbf{p})] \quad (10)$$

where (\mathbf{r}, \mathbf{p}) are the coordinates and momenta of the molecule, $H_{S_0}(\mathbf{r}, \mathbf{p})$ is the Hamiltonian of the S_0 species, and E is the total energy. The S_0 species crosses to the T_1 surface via the S_0/T_1 seam surface, which is given by $\sigma(\mathbf{r}) = 0$ with

$$\sigma(\mathbf{r}) = U_{S_0}(\mathbf{r}) - U_{T_1}(\mathbf{r}) \quad (11)$$

In this paper, we assume that the surface hopping occurs precisely on the seam surface (i.e., the weak-coupling limit of the intersystem crossing). This indicates that the phase space points crossing to the T_1 surface are characterized by the conditional microcanonical distribution on the seam surface

$$f_{seam}(\mathbf{r}, \mathbf{p}) \propto \delta[E - H_{S_0}(\mathbf{r}, \mathbf{p})] \delta[U_{S_0}(\mathbf{r}) - U_{T_1}(\mathbf{r})] \quad (12)$$

(see Appendix A for more rigorous definitions). Our strategy here is to first statistically sample the $S_0 \rightarrow T_1$ crossing points from the seam distribution in eq 12 and then propagate them on the T_1 surface. In this paper, we perform the first task by running a constrained trajectory on the seam surface via the use of a SHAKE-like method. In usual molecular-dynamics

(MD) simulations, the SHAKE method is used to constrain geometric coordinates such as bond lengths. In the present case, we wish to constrain the energy gap between the S_0 and T_1 surfaces. However, if the original SHAKE method is used for this purpose, the calculation becomes quite expensive because it involves iterative solution of $\sigma(\mathbf{r}) = 0$ at every MD step [note that the calculation of $\sigma(\mathbf{r})$ in eq 11 requires an initial calculation of S_0 and T_1 potential energies]. To avoid this, we consider developing an approximate yet more efficient version of the SHAKE method. In the absence of the constraint, the equation of motion is given by

$$m_i \frac{d^2}{dt^2} \mathbf{r}_i(t) = -\nabla_i U_{S_0}[\mathbf{r}(t)] \equiv \mathbf{F}_i[\mathbf{r}(t)] \quad (13)$$

where $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_N)$ and $\mathbf{F}_i[\mathbf{r}(t)]$ is the potential force for the S_0 surface. One can discretize eq 13 using the velocity Verlet (VV) integrator

$$\mathbf{v}_i(t + \Delta t/2) = \mathbf{v}_i(t) + \frac{\Delta t}{2m_i} \mathbf{F}_i[\mathbf{r}(t)] \quad (14a)$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t + \Delta t/2)\Delta t \quad (14b)$$

$$\mathbf{v}_i(t + \Delta t) = \mathbf{v}_i(t + \Delta t/2) + \frac{\Delta t}{2m_i} \mathbf{F}_i[\mathbf{r}(t + \Delta t)] \quad (14c)$$

where the three steps are evaluated sequentially. Given that the system is ergodic at total energy E , the VV integrator in eq 14 produces the microcanonical distribution in eq 10. In the presence of the constraint $\sigma(\mathbf{r}) = 0$, the equation of motion is given by

$$m_i \frac{d^2}{dt^2} \mathbf{r}_i(t) = \mathbf{F}_i[\mathbf{r}(t)] - \lambda(t) \nabla_i \sigma[\mathbf{r}(t)] \quad (15)$$

where the second term represents the constraint force and $\lambda(t)$ is the Lagrange multiplier. The corresponding VV integrator reads^{15,16}

$$\mathbf{v}_i(t + \Delta t/2) = \mathbf{v}_i(t) + \frac{\Delta t}{2m_i} \{\mathbf{F}_i[\mathbf{r}(t)] - \lambda_R \nabla_i \sigma[\mathbf{r}(t)]\} \quad (16a)$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t + \Delta t/2)\Delta t \quad (16b)$$

$$\mathbf{v}_i(t + \Delta t) = \mathbf{v}_i(t + \Delta t/2) + \frac{\Delta t}{2m_i} \{\mathbf{F}_i[\mathbf{r}(t + \Delta t)] - \lambda_V \nabla_i \sigma[\mathbf{r}(t + \Delta t)]\} \quad (16c)$$

where λ_R and λ_V are discretized analogs of $\lambda(t)$. The original SHAKE method determines the λ_R by requiring that the constraint condition is satisfied exactly at time $t + \Delta t$, i.e.,

$$\sigma[\mathbf{r}(t + \Delta t; \lambda_R)] = 0 \quad (17)$$

This is an implicit equation for the unknown λ_R , which is solved via iteration in the original SHAKE method. This iterative solution requires repeated evaluation of σ , which is

expensive in the present case. We thus solve eq 17 approximately by expanding $\sigma[\mathbf{r}(t + \Delta t)]$ up to the first order in terms of $\mathbf{r}(t + \Delta t) - \mathbf{r}(t)$ about $\mathbf{r}(t)$, namely

$$\begin{aligned} \sigma[\mathbf{r}(t + \Delta t)] \approx & \sigma[\mathbf{r}(t)] \\ & + \sum_i \nabla_i \sigma[\mathbf{r}(t)] \cdot \Delta t \{ \mathbf{v}_i^u(t + \Delta t/2) \\ & - \frac{\Delta t}{2m_i} \lambda_R \nabla_i \sigma[\mathbf{r}(t)] \} \equiv 0 \end{aligned} \quad (18)$$

where $\mathbf{v}_i^u(t + \Delta t/2)$ is the unconstrained velocity at $t + \Delta t/2$

$$\mathbf{v}_i^u(t + \Delta t/2) = \mathbf{v}_i(t) + \frac{\Delta t}{2m_i} \mathbf{F}_i[\mathbf{r}(t)] \quad (19)$$

The above approximation gives

$$\lambda_R = \frac{\sigma[\mathbf{r}(t)] + \Delta t \sum_i \nabla_i \sigma[\mathbf{r}(t)] \cdot \mathbf{v}_i^u(t + \Delta t/2)}{(\Delta t^2/2) \sum_i \{ \nabla_i \sigma[\mathbf{r}(t)] \}^2 / m_i} \quad (20)$$

It is important to note that $\sigma[\mathbf{r}(t)]$ in eq 20 should be retained explicitly in order to allow small fluctuations of $\sigma[\mathbf{r}(t)]$ while suppressing its systematic drift from zero. On the other hand, the second Lagrange multiplier λ_V in eq 16 can be determined exactly by requiring the velocity constraint at $t + \Delta t$:

$$\begin{aligned} \dot{\sigma}[\mathbf{r}(t + \Delta t)] = & \sum_i \nabla_i \sigma[\mathbf{r}(t + \Delta t)] \cdot \{ \mathbf{v}_i^u(t + \Delta t) \\ & - \frac{\Delta t}{2m_i} \lambda_V \nabla_i \sigma[\mathbf{r}(t + \Delta t)] \} \equiv 0 \end{aligned} \quad (21)$$

where $\mathbf{v}_i^u(t + \Delta t)$ is the unconstrained velocity at $t + \Delta t$

$$\mathbf{v}_i^u(t + \Delta t) = \mathbf{v}_i(t + \Delta t/2) + \frac{\Delta t}{2m_i} \mathbf{F}_i[\mathbf{r}(t + \Delta t)] \quad (22)$$

The obtained λ_V is

$$\lambda_V = \frac{\sum_i \nabla_i \sigma[\mathbf{r}(t + \Delta t)] \cdot \mathbf{v}_i^u(t + \Delta t)}{(\Delta t/2) \sum_i \{ \nabla_i \sigma[\mathbf{r}(t + \Delta t)] \}^2 / m_i} \quad (23)$$

With the above procedure, one can propagate a constrained trajectory with only a single evaluation of σ per MD step. Since σ in eq 11 is obtained from the S_0 and T_1 potential energies, the computational cost for propagating a constrained trajectory is identical to that of an unconstrained trajectory.

Before proceeding, it is useful to make some comments on the above integrator. First, because λ_R is evaluated only approximately, the discretization error is slightly greater than in the case where λ_R is evaluated exactly (see below for comparison). To reduce the error, we additionally perform a velocity scaling at each MD step so that the energy is conserved exactly. Second, we note that the constrained trajectory gives the following phase-space distribution:^{22,23}

$$f_{shake}(\mathbf{r}, \mathbf{p}) \propto \delta[E - H_{S_0}] \delta[\sigma] \delta[\dot{\sigma}] \quad (24)$$

The $f_{shake}(\mathbf{r}, \mathbf{p})$ above differs from $f_{seam}(\mathbf{r}, \mathbf{p})$ in eq 12 in that the velocity constraint factor $\delta[\dot{\sigma}]$ is present. The latter factor reflects the fact that the velocity of a constrained trajectory is

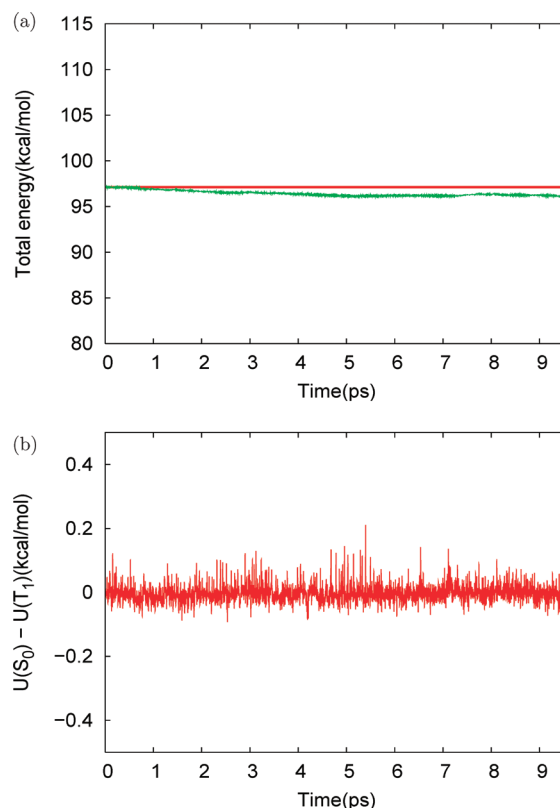


Figure 5. (a) Total energy of the seam trajectory as a function of time. When the velocity scaling is not applied, the total energy exhibits a slight drift on the order of 0.1 kcal/mol/ps (green line). With velocity scaling applied, the total energy is conserved exactly (red line). (b) The difference between the S_0 and T_1 potential energies, $\sigma = U_{S_0} - U_{T_1}$, calculated along the seam trajectory.

always tangent to the constraint surface, $\sigma(\mathbf{r}) = 0$. Since we are interested in generating $f_{seam}(\mathbf{r}, \mathbf{p})$ rather than $f_{shake}(\mathbf{r}, \mathbf{p})$, we need to correct for the above difference. This correction can be made in a manner similar to the bluemoon sampling method²³ (see Appendix A). Third, since we are running a classical trajectory for statistical sampling, the so-called quasi-ergodicity²⁴ may become a problem. To examine this, we performed periodic resampling of velocities during the trajectory propagation. We find that the results obtained with velocity resampling are very similar to those without velocity resampling, which indicates that the quasi-ergodicity problem is not of serious concern for the present system (we note, however, that for low-energy systems such as van der Waals clusters, statistical sampling via classical trajectories can be more problematic; see ref 24).

Using the SHAKE-like method described above, we propagated a constrained trajectory on the seam surface with a time step of 0.5 fs up to a maximum time of 10 ps. Figure 5a displays the time evolution of the total energy with or without velocity scaling applied at each MD step. We see that the energy drift is as small as 0.1 kcal/ps even without velocity scaling applied (green line). Nevertheless, in the following calculations we always apply the velocity scaling to conserve the total energy exactly. Figure 5b displays the time evolution of the energy gap $\sigma(\mathbf{r}) = U_{S_0} - U_{T_1}$. As seen, the $\sigma(\mathbf{r})$ remains close to zero with a maximum deviation of 0.2 kcal/mol. This indicates that the first-order expansion of $\sigma(\mathbf{r})$ in eq 18 does not introduce a significant error into the trajectory propagation. In the next section, we will utilize

the above approximate integrator to obtain the conditional microcanonical distribution on the seam surface.

III. RESULTS AND DISCUSSION

In the following, we consider three simulation conditions regarding the type of dual-level PES and the total energy of the system. First, under the “MRMP” simulation condition, we study the dynamics using the MRMP::B3LYP PES at a total energy of 97.1 kcal/mol (as measured from the S_0 minimum). The latter total energy is equal to the dissociation threshold on the T_1 surface at the MRMP2 level (81.8 kcal/mol) plus the sum of zero-point energies at the T_1 TS geometry (15.3 kcal/mol). The above condition is intended to match most closely with the experimental condition² (we recall that the steplike structure in $k(E)$ was observed in the T_1 dissociation threshold region). In the following, we will refer to the total energy minus the potential energy of the TS geometry as the TS excess energy. With this definition, the excess energy in the MRMP condition is simply equal to the ZPE at the TS geometry. Second, in the “CASSCF” simulation condition we study the dynamics using the CASSCF::B3LYP PES at a total energy of 97.1 kcal/mol (which is identical to that in the MRMP condition). The TS excess energy under the CASSCF condition (28.5 kcal/mol) is significantly larger than that under the MRMP condition (15.3 kcal/mol). This is because the dissociation barrier height on the T_1 surface is considerably small at the CASSCF level (16.3 kcal/mol) compared to that at the MRMP level (29.3 kcal/mol). Third, under the “CASSCF(K)” simulation condition, we study the dynamics using the CASSCF::B3LYP PES at a total energy of 82.8 kcal/mol. This condition is intended to match most closely with the surface-hopping study by Kaledin et al.,¹² where the total energy is set equal to the $S_0 \rightarrow S_1$ vertical excitation at the SA-CASSCF level. The TS excess energy under the CASSCF(K) condition is 14.2 kcal/mol, which is again close to the experimental condition.

III.A. Statistical Distribution on the Seam Surface. Figure 6 displays the statistical distribution of (R, U_{T_1}) on the seam surface obtained from the constrained trajectory calculation (section II. B). Here, R is the CC bond length and U_{T_1} denotes the T_1 potential energy as measured from the T_1 equilibrium geometry. Under the MRMP simulation condition [Figure 6a], the value of U_{T_1} is broadly distributed up to 40 kcal/mol, which roughly corresponds to the total energy of the trajectory (as measured from the T_1 minimum). The CC bond length is distributed in the range $1.25 < R < 1.85 \text{ \AA}$, which is essentially localized in the T_1 reactant region. The sampled values of U_{T_1} are generally greater than the corresponding value of the T_1 minimum energy profile [Figure 4b] at the same value of R . This is because the internal coordinates other than R may be excited vibrationally on the seam surface. At the MRMP level, the stationary values of R for the T_1 minimum, S_0/T_1 MSX, and T_1 TS geometries are 1.46, 1.52, and 2.31 \AA , respectively. Since the value of R is distributed only up to 1.85 \AA , it follows that the phase-space points on the seam surface are located in the vicinity of the S_0/T_1 MSX point ($R = 1.5 \text{ \AA}$), and the seam trajectory hardly accesses the TS region ($R = 2.3 \text{ \AA}$). Figure 6b displays the distribution of (R, U_{T_1}) obtained with the CASSCF simulation condition. As seen, the overall shape of the distribution is rather close to that in the MRMP condition. This is probably because the topography of the T_1 PES in the reactant region ($R < 1.75$) is very similar between the CASSCF and MRMP results, despite the fact that they are significantly different in the TS and product region

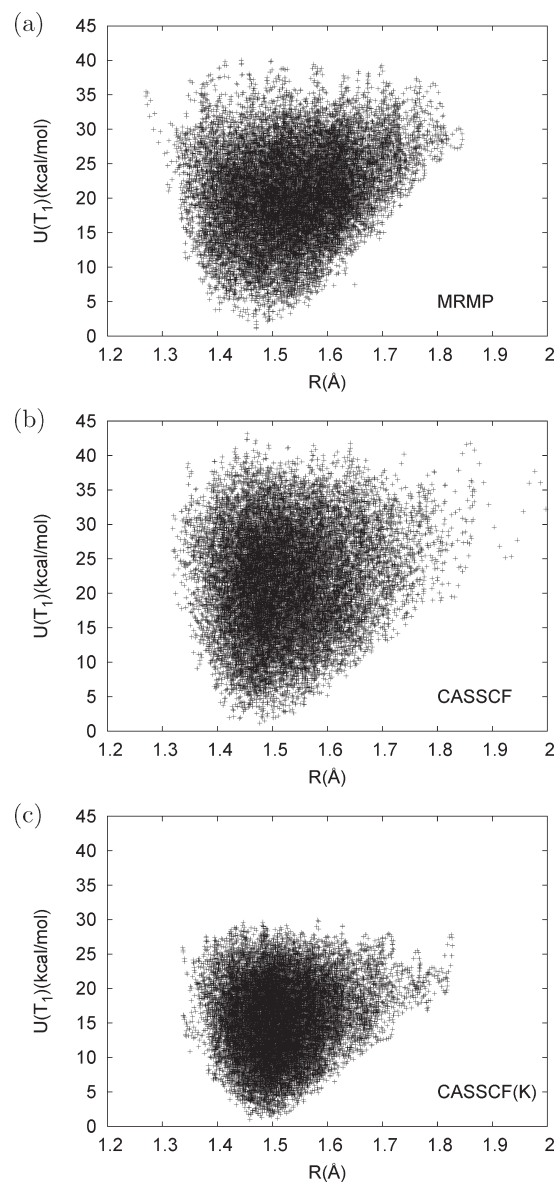


Figure 6. Statistical distribution of (R, U_{T_1}) on the S_0/T_1 seam surface calculated under the MRMP, CASSCF, and CASSCF(K) simulation conditions. R is the CC bond length, and U_{T_1} is the T_1 potential energy measured from the T_1 minimum.

[see Figure 4b]. Under the CASSCF condition, we frequently observed that the seam trajectory escaped from the reactant region and dissociated to the product region. An example of such a trajectory is shown in Figure 4b with $R > 1.9 \text{ \AA}$. This is to be expected because the TS excess energy under the CASSCF condition is very large (28.5 kcal/mol) compared to that under the MRMP condition (15.3 kcal/mol). Another factor that facilitates the dissociation is the closeness of the TS geometry at the CASSCF level to the reactant region. This is seen from the stationary values of R at the CASSCF level for the T_1 minimum, S_0/T_1 MSX, and T_1 TS points, which are 1.47, 1.53, and 2.00 \AA , respectively. Figure 6c displays the result for the CASSCF(K) condition. Since the total energy under the latter condition is smaller by 14 kcal/mol than that in the MRMP condition, the value of U_{T_1} is distributed only up to 30 kcal/mol (as measured from the T_1 minimum). Figure 6c shows that the seam distribution

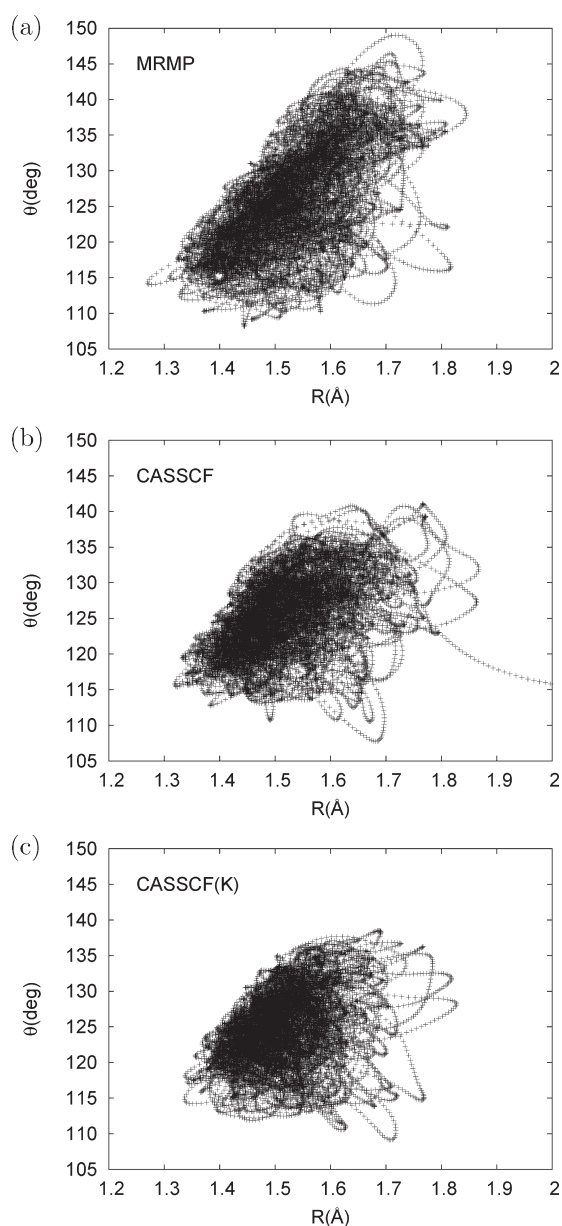


Figure 7. Statistical distribution of (R, θ) on the S_0/T_1 seam surface calculated under the MRMP, CASSCF, and CASSCF(K) simulation conditions. R is the CC bond length, and θ is the CCO bending angle.

is again localized in the reactant region, exhibiting no dissociating trajectory. This indicates that the region $R > 1.9$ Å has little contribution to the conditional microcanonical distribution on the seam and that the $S_0 \rightarrow T_1$ surface crossing occurs mainly in the T_1 reactant region.

Figure 7 displays the distribution of (R, θ) obtained under the individual simulation conditions. Here, θ denotes the CCO bending angle (see Figure 2). Under the MRMP condition, the θ is widely distributed in the range $110^\circ < \theta < 145^\circ$ and roughly centered at the S_0/T_1 MSX point ($R = 1.490$ Å and $\theta = 124.5^\circ$). Under the CASSCF and CASSCF(K) conditions, the distribution of θ is somewhat more localized compared to the MRMP condition. In Figure 7, the values of θ tend to increase with R , which indicates that the molecule assumes a more linear geometry of the CCO moiety as R is increased. This trend is in

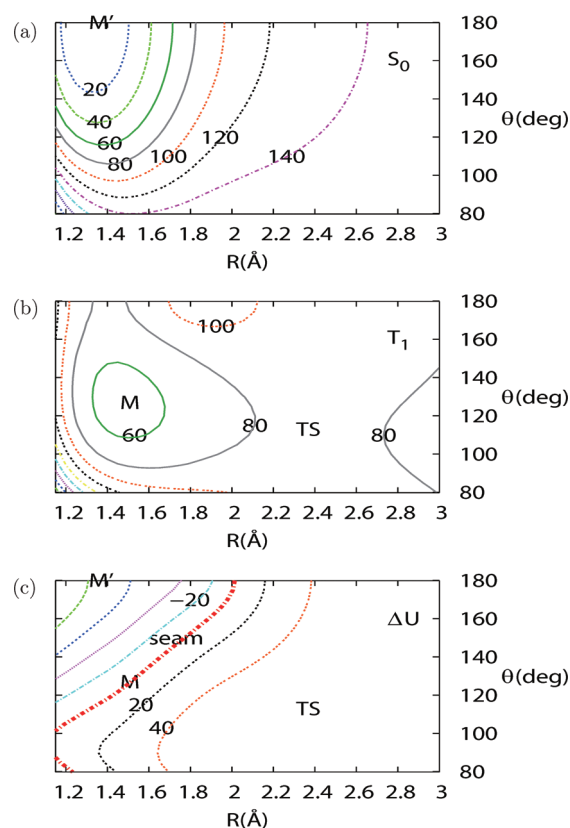


Figure 8. Two-dimensional potential map as a function of (R, θ) calculated with the MRMP::B3LYP potential functions. Panel b is obtained by optimizing the internal coordinates other than (R, θ) on the T_1 PES. Panel a displays the S_0 potential energy calculated with the same coordinate set as used in panel a. Panel c displays the energy difference between panels a and b. The red, thick line in panel c is the seam line defined by $U(S_0) = U(T_1)$. The symbols M' , M , and TS indicate the location of the S_0 minimum, T_1 minimum, and T_1 TS geometries, respectively. Energies are in kcal/mol and measured from the S_0 minimum.

qualitative agreement with the previous study by Cui and Morokuma,¹¹ which obtained partially optimized geometries on the seam surface at several values of R . To obtain further insight, we display in Figure 8 the two-dimensional map of the S_0 and T_1 surfaces at the MRMP level. Here, the T_1 potential map in panel b is obtained by adiabatically optimizing the internal coordinates other than (R, θ) on the T_1 PES, whereas the S_0 potential map in panel a was obtained by using the same coordinate set as that used in panel b. Those potential maps indicate that the S_0 surface favors a more linear geometry with $\theta \approx 180^\circ$, while the T_1 state favors a more bent geometry with $\theta \approx 130^\circ$. The energy difference between the two potential maps is plotted in panel c, which exhibits a seam line running from $R \approx 1.2$ Å and $\theta \approx 100^\circ$ to $R \approx 2.0$ Å and $\theta \approx 180^\circ$. A comparison between Figures 7 and Figure 8c shows that the phase-space points are roughly centered around the seam line thus obtained.

Figure 9 displays the distribution of (R, ϕ) on the seam surface. We recall that the torsional angle ϕ is 0° for planar $Cs-II$ symmetry and 90° for nonplanar $Cs-I$ symmetry. Previous studies have shown^{7,11} that the important geometries such as the T_1 minimum and the S_0/T_1 MSX point are of planar $Cs-II$ symmetry. Nevertheless, we see from Figure 9 that ϕ is widely

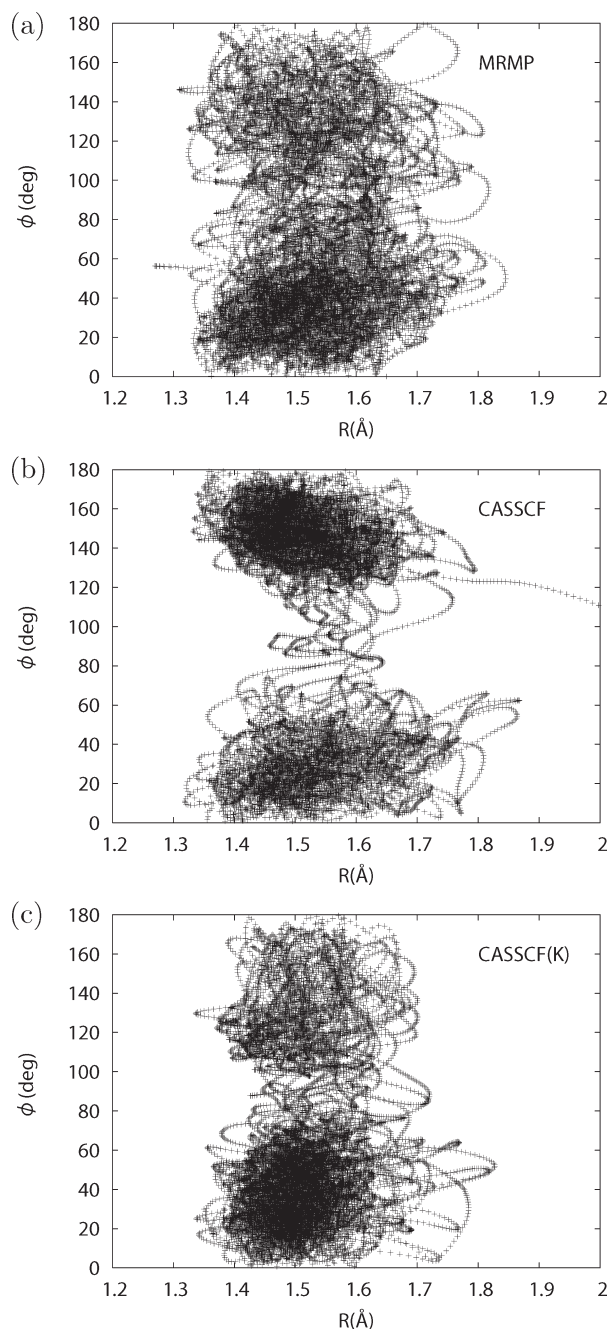


Figure 9. Statistical distribution of (R, ϕ) on the S_0/T_1 seam surface calculated with the MRMP, CASSCF, and CASSCF(K) simulation conditions. R is the CC bond length, and ϕ is the torsional angle.

distributed over $0 < \phi < 180^\circ$, indicating that the molecule is not limited to planar $Cs-II$ geometries upon surface crossing. A closer look at Figure 9 reveals that the distribution of ϕ is bimodal and centered at around 40° and 140° , which are of neither $Cs-I$ nor $Cs-II$ symmetry. Several factors may be responsible for the above observation. First, the distribution on the seam tends to get delocalized by the effect of kinetic energy. Second, since Figure 9 is obtained by projecting out the internal coordinates other than (R, ϕ) , the two-dimensional distribution reflects the accessible number of states (or phase-space volume) of the projected coordinates at individual values of (R, ϕ) . Third, the ϕ dependences of the S_0 and T_1 PESs are rather different. To illustrate this,

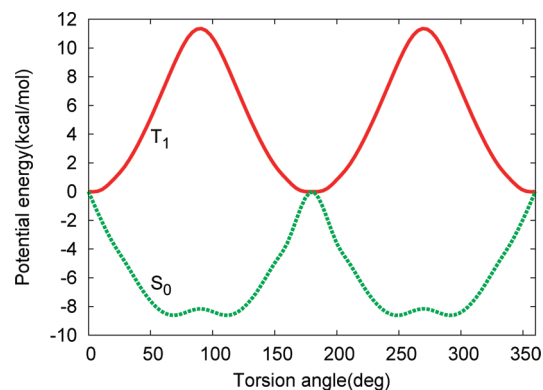


Figure 10. Potential energy profiles of the S_0 and T_1 surfaces as a function of the torsional angle ϕ obtained with the MRMP::B3LYP potential functions. Internal coordinates other than (R, θ) are optimized at each value of ϕ . The values of R and θ are fixed at the S_0/T_1 MSX geometry ($R = 1.491 \text{ \AA}$ and $\theta = 124.5^\circ$). Note that the minima of the S_0 potential are slightly off the $Cs-I$ symmetry because θ is frozen at the MSX geometry.

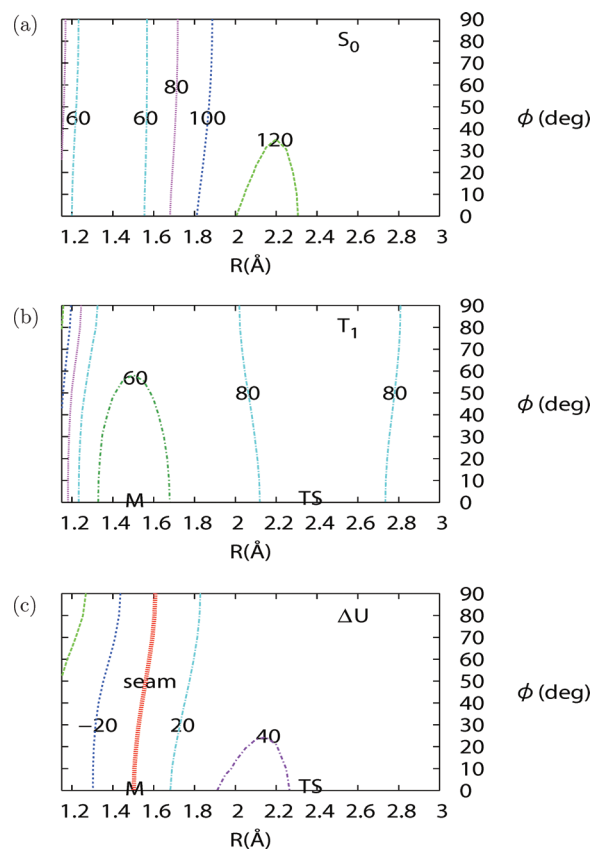


Figure 11. Two-dimensional potential map as a function of (R, ϕ) calculated with the MRMP::B3LYP potential functions. See the caption of Figure 8 for computational details.

we plot in Figure 10 the adiabatically optimized potential curves as a function of ϕ for the S_0 and T_1 states [here, the values of (R, θ) are fixed at the S_0/T_1 MSX geometry]. As seen, the equilibrium values of ϕ are different between the S_0 and T_1 states, which seems most responsible for the broad distribution of ϕ in Figure 9. For further information, Figure 11 displays the

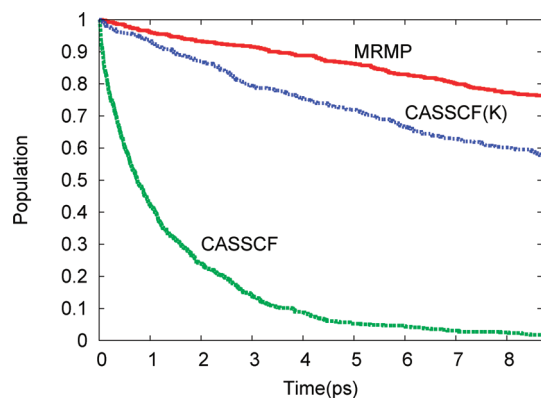


Figure 12. Time-dependent survival probability $P_{\text{surv}}(t)$ on the T_1 surface calculated under the MRMP, CASSCF, and CASSCF(K) simulation conditions.

two-dimensional map of the S_0 and T_1 potentials as a function of (R, ϕ) . The seam line runs from around $R = 1.5 \text{ \AA}$ and $\phi = 0^\circ$ to $R = 1.6 \text{ \AA}$ and $\phi \approx 90^\circ$, and it is again seen that the phase-space points are distributed roughly around the seam line.

III.B. Dissociation Rate on the T_1 Surface. We now calculate the dissociation rate by propagating a number of trajectories on the T_1 surface. To do so, we assume the weak-coupling limit of intersystem crossing, and accordingly a trajectory is propagated adiabatically on the T_1 surface.²⁵ To prepare the initial state of the trajectory, we picked up 500 points randomly from the seam distribution obtained in section III.A, followed by an appropriate resampling of velocities (see Appendix A). The initial phase-space points thus obtained were propagated on the T_1 surface until they dissociated or the integration time exceeded 9 ps. The progress of the reaction was monitored by calculating the time-dependent survival probability defined by

$$P_{\text{surv}}(t) = \frac{\langle |v_\perp| P_{\text{NA}}(v_\perp) \Theta(R^\ddagger - R_t) \rangle_{\text{seam}}}{\langle |v_\perp| P_{\text{NA}}(v_\perp) \rangle_{\text{seam}}} \quad (25)$$

Here, $\langle \dots \rangle_{\text{seam}}$ denotes an ensemble average over the initial phase-space points on the seam surface, $\Theta(R^\ddagger - R_t)$ is the Heaviside step function, and v_\perp is the component of the velocity vector normal to the seam surface (see Appendix A for more details). The $P_{\text{NA}}(v_\perp)$ in eq 25 is the nonadiabatic transition probability calculated with the Landau–Zener formula:^{26–28}

$$P_{\text{NA}}(v_\perp) = 1 - \exp \left[-\frac{2\pi}{\hbar} \frac{H_{\text{SO}}^2}{|\Delta F v_\perp|} \right] \quad (26)$$

Here, H_{SO} is the S_0 – T_1 spin–orbit coupling and ΔF is the normal component of the difference gradient vector $\nabla(U_{S_0} - U_{T_1})$. In this paper, we approximated H_{SO} to be constant and set it to 43 cm^{-1} (calculated at the CASSCF level). By fitting the survival probability to a single exponential form

$$P_{\text{surv}}(t) \approx \exp(-kt) \quad (27)$$

we estimated the dissociation rate k and the associated lifetime $\tau = 1/k$ on the T_1 surface.

Figure 12 displays the survival probability thus obtained for individual simulation conditions. Table 4 lists the corresponding dissociation rate k and lifetime τ . Under the CASSCF conditions, the dissociation is very fast, with $\tau = 1.4 \text{ ps}$, which is clearly due to the large excess energy (28.5 kcal/mol) at the TS geometry.

Table 4. Dissociation Rate k on the T_1 Surface (in ps^{-1}) Obtained from the Trajectory Calculations^a

	MRMP	CASSCF	CASSCF(K)
RRKM ($\hbar = 1.00$)	0.0017 (590)	2.59 (0.39)	0.057 (17.5)
RRKM ($\hbar = 0.50$)	0.0462 (21.6)	3.48 (0.29)	0.223 (4.5)
RRKM ($\hbar = 0.25$)	0.0656 (15.2)	3.69 (0.27)	0.353 (2.8)
RRKM ($\hbar = 0.10$)	0.0715 (14.0)	3.75 (0.26)	0.391 (2.6)
trajectory calcd	0.0276 (36.2)	0.710 (1.4)	0.0677 (14.8)

^a The RRKM rates obtained with various values of \hbar (in atomic units) are also shown. The corresponding lifetime $\tau = 1/k$ (in ps) is given in parentheses.

Under the MRMP and CASSCF(K) conditions, the reaction proceeds much slower than under the CASSCF condition. The calculated values of τ are 36.2 ps for the MRMP condition and 14.8 ps for the CASSCF(K) condition. A critical observation here is that the latter values are significantly greater than that obtained in the previous trajectory study by Kaledin et al.¹² In the latter study, half the trajectories dissociated within $\sim 120 \text{ fs}$ after making a transition to the T_1 surface (or within $\sim 260 \text{ fs}$ after $S_0 \rightarrow S_1$ vertical excitation). This means that the dissociation rate under the MRMP and CASSCF(K) conditions are about 2 orders of magnitude smaller than obtained in ref 12. This is somewhat surprising because the CASSCF(K) condition has the same total energy and level of electronic structure theory as employed in ref 12. The marked difference in the dissociation rates is thus attributed to the different ways of preparing the T_1 trajectories. Specifically, in ref 12, the parent trajectory was first evolved on the S_1 surface (after initial vertical excitation), and it was allowed to branch every time the S_1 surface crossed the S_0 surface. At the branching point, the parent trajectory continued to be propagated on the S_1 surface, whereas the child trajectory started to evolve on the S_0 surface. A similar procedure was applied to obtain child trajectories on the T_1 surface. A critical point here is that the propagation time of the S_0 trajectory was rather short (up to $\sim 300 \text{ fs}$), and the number of the S_0 and T_1 trajectories was relatively small (a total of 10). This indicates that the T_1 trajectories may be affected considerably by the short-time dynamics of the S_0 trajectories after vertical excitation. Since the molecule is vibrationally excited in the CC stretching and CCO bending modes upon vertical excitation, the dissociation may be facilitated to some extent by selective excitation of those vibrational modes (we recall here that the CC stretch plays the role of reaction coordinate for this system). By contrast, in the present calculation, we have assumed a vibrational equilibrium on the S_0 surface and propagated T_1 trajectories starting from the conditional microcanonical distribution on the seam surface. We expect that the above different ways of preparing the T_1 trajectories are mainly responsible for the different rates obtained in the present study and in ref 12.

Another question to be asked here is whether the dissociation dynamics on the T_1 surface may be regarded as statistical or not. This is because the previous rate calculations^{2,3,6} relied on statistical rate theories that assume a vibrational equilibrium on the T_1 surface. Those calculations are thus reliable only when the IVR on the T_1 surface proceeds considerably prior to the dissociation. To examine this, we have calculated the RRKM rate for the T_1 dissociation. Here, it should be noted that the usual RRKM rate refers to a *quantum mechanical* unimolecular reaction rate, which cannot be compared directly to the present classical

trajectory results. Rather, we need to calculate a *classical mechanical* reaction rate assuming a vibrational equilibrium on the T_1 surface. In this paper, we performed this task by calculating the RRKM rate with decreasing values of \hbar (see Appendix B for details), the results of which are summarized in Table 4. We see from this table that the RRKM rate is fairly converged with $\hbar = 0.1$ (in atomic units), which we regard as the “classical” transition state theory (TST) rate, and it is seen that the latter is in reasonable agreement with the present classical trajectory result. The deviation between the trajectory and classical RRKM results is within a factor of 3 for the MRMP condition and 6 for the CASSCF(K) condition. A possible reason for the deviation is the so-called recrossing effect of trajectories over the potential barrier. However, we expect that the latter effect is minor for the present system because the T_1 surface is rather separable in the TS region.⁶ A more plausible reason for the above deviation is an inefficient IVR on the T_1 surface. Figure 11 indicates that the vibrational coupling between R and ϕ coordinates is fairly small, which may lead to inefficient energy transfer from ϕ to R and hence the underestimation of the rate in the trajectory calculation (recall that R plays the role of reaction coordinate, and thus its activation is essential for the dissociation). Nevertheless, since the discrepancy between the trajectory and classical RRKM results is not particularly large (within a factor of 3 in the MRMP condition), the T_1 dissociation seems to be rather statistical (at least not highly nonstatistical), given that the trajectory is initiated from the conditional microcanonical distribution on the seam surface.

IV. CONCLUDING REMARKS

In this paper, we have revisited the still unresolved problem of triplet ketene regarding the stepwise structure in the observed rate.² In particular, we focused on whether the T_1 dissociation may be regarded as statistical, because this has been the fundamental assumption of the previous rate calculations.^{2,3,6} To examine this, we have introduced the assumption of a vibrational equilibrium on the S_0 surface and calculated the dissociation rate by running T_1 trajectories from the seam surface. The main results obtained are the following: (1) The $S_0 \rightarrow T_1$ crossing points are located mainly in the T_1 reactant region. (2) The T_1 dissociation rate is significantly smaller than the previous surface-hopping result by Kaledin et al.¹² (3) The present trajectory result agrees reasonably well with the classical RRKM result. As noted above, the discrepancy between the present study and ref 12 is probably due to the different ways of preparing the T_1 trajectories, rather than the different levels of electronic structure theory used. This is based on the observation that even under the CASSCF(K) condition there remains a significant difference in the rates between the two calculations. We expect that if a surface-hopping simulation is performed with a very long S_0 trajectory and a sufficient number of T_1 trajectories, the discrepancy between the two studies would probably be reduced.

While we have studied the T_1 dissociation from a statistical point of view, we cannot make further arguments on the possible nonadiabatic effect on the steplike structure in $k(E)$. To investigate the latter, it is desirable that one performs a quantum dynamical nonadiabatic simulation involving S_0 and T_1 states. (Here, we emphasize that the present study as well as the previous one by Kaledin et al.¹² is performed wholly in terms of classical mechanics, except that ZPE was taken into account to define the total energy.) An approach to that end is to introduce a reduced-dimensional model that involves S_0 and T_1 states, which

is the strategy used previously for the adiabatic dynamics on the T_1 surface.^{3,6} The multiconfiguration time-dependent Hartree method^{29–32} may also be useful for this purpose. An alternative approach is to use a trajectory-based simulation method that includes nuclear quantum effects approximately.^{33–36} Application of those methods to the nonadiabatic dynamics of ketene may provide more mechanistic insights into the observed steps in $k(E)$.

■ APPENDIX A: CONDITIONAL MICROCANONICAL DISTRIBUTION ON THE SEAM SURFACE

In this paper, we define the conditional microcanonical distribution on the seam surface as follows:

$$f_{seam}(\mathbf{r}, \mathbf{p}) = \delta(E - T - U_0) \delta(\sigma) |\nabla \sigma| \quad (\text{A1})$$

where $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_N)$ and $\mathbf{p} = (\mathbf{p}_1, \dots, \mathbf{p}_N)$ are the coordinates and momenta in mass-weighted coordinates with $\nabla = \partial/\partial \mathbf{r}$, and $T = \mathbf{p}^2/2$ is the kinetic energy of the molecule. Hereafter, we denote the potential energies of the S_0 and T_1 states as $U_0(\mathbf{r})$ and $U_1(\mathbf{r})$, respectively, and σ represents the energy gap

$$\sigma(\mathbf{r}) = U_0(\mathbf{r}) - U_1(\mathbf{r}) \quad (\text{A2})$$

The $|\nabla \sigma|$ in eq A1 is a geometric factor that makes $f_{seam}(\mathbf{r}, \mathbf{p})$ invariant to an arbitrary scaling of $\sigma(\mathbf{r})$, namely, $\sigma(\mathbf{r}) \rightarrow \kappa(\mathbf{r}) \sigma(\mathbf{r})$. For simplicity of discussion, we first consider the following form of the survival probability on the T_1 surface

$$P_{surv}(t) = \langle \Theta(R^\ddagger - R_t) \rangle_{seam} \quad (\text{A3})$$

where $\langle \dots \rangle_{seam}$ is an ensemble average defined by

$$\langle \dots \rangle_{seam} = \frac{\int d\mathbf{r} \int d\mathbf{p} f_{seam}(\mathbf{r}, \mathbf{p}) (\dots)}{\int d\mathbf{r} \int d\mathbf{p} f_{seam}(\mathbf{r}, \mathbf{p})} \quad (\text{A4})$$

We evaluate the above average by propagating a constrained trajectory on the seam surface. To do so, we recall that a microcanonical system with a holonomic constraint $\sigma(\mathbf{r}) = 0$ has the following phase-space distribution:²²

$$f_{shake}(\mathbf{r}, \mathbf{p}) = \delta(E - T - U_0) \delta(\sigma) \delta(\dot{\sigma}) |\nabla \sigma|^2 \quad (\text{A5})$$

A critical difference between $f_{shake}(\mathbf{r}, \mathbf{p})$ and $f_{seam}(\mathbf{r}, \mathbf{p})$ is that the former includes the velocity constraint, $\dot{\sigma} = \nabla \sigma \cdot \mathbf{v} = 0$, which states that the velocity of the constrained system is always tangential to the constraint surface, $\sigma(\mathbf{r}) = 0$. Note that such a velocity constraint is absent in the ensemble of phase-space points associated with S_0/T_1 intersystem crossing. To account for this difference, it is useful to introduce the position distribution function obtained by integrating over the momenta³⁷

$$\begin{aligned} \tilde{f}_{seam}(\mathbf{r}) &= \int d\mathbf{p} f_{seam}(\mathbf{r}, \mathbf{p}) \\ &= const \times [E - U_0]^{(n-2)/2} \delta(\sigma) |\nabla \sigma| \end{aligned} \quad (\text{A6})$$

$$\begin{aligned} \tilde{f}_{shake}(\mathbf{r}) &= \int d\mathbf{p} f_{shake}(\mathbf{r}, \mathbf{p}) \\ &= const \times [E - U_0]^{(n-3)/2} \delta(\sigma) |\nabla \sigma| \end{aligned} \quad (\text{A7})$$

where n is the number of degrees of freedom and $const$ represents a constant independent of (\mathbf{r}, \mathbf{p}) . Note that the exponent of $[E - U_0]$ is different by 1/2 between eqs A6 and A7, which suggests

$$\tilde{f}_{seam}(\mathbf{r}) = const \times [E - U_0]^{1/2} \tilde{f}_{shake}(\mathbf{r}) \quad (\text{A8})$$

The above relation can be used to calculate an ensemble average over $f_{seam}(\mathbf{r}, \mathbf{p})$ by sampling $f_{shake}(\mathbf{r}, \mathbf{p})$. To do so, we first rewrite the ensemble average in eq A3 as follows:

$$\langle \Theta_t \rangle_{seam} = \frac{\int d\mathbf{r} \tilde{f}_{seam}(\mathbf{r}) \langle \Theta_t \rangle_p}{\int d\mathbf{r} \tilde{f}_{seam}(\mathbf{r})} \quad (\text{A9})$$

where $\Theta_t \equiv \Theta(R^\ddagger - R_t)$, and $\langle \dots \rangle_p$ represents an average over the momenta defined by

$$\langle \dots \rangle_p = \frac{\int d\mathbf{p} f_{seam}(\mathbf{r}, \mathbf{p}) (\dots)}{\int d\mathbf{p} f_{seam}(\mathbf{r}, \mathbf{p})} \quad (\text{A10})$$

By inserting eq A8 into eq A9, we have

$$\langle \Theta_t \rangle_{seam} = \frac{\langle [E - U_0]^{1/2} \langle \Theta_t \rangle_p \rangle_{shake}}{\langle [E - U_0]^{1/2} \rangle_{shake}} \quad (\text{A11})$$

where $\langle \dots \rangle_{shake}$ is defined by

$$\begin{aligned} \langle A(\mathbf{r}) \rangle_{shake} &= \frac{\int d\mathbf{r} \tilde{f}_{shake}(\mathbf{r}) A(\mathbf{r})}{\int d\mathbf{r} \tilde{f}_{shake}(\mathbf{r})} \\ &= \frac{\int d\mathbf{r} \int d\mathbf{p} f_{shake}(\mathbf{r}, \mathbf{p}) A(\mathbf{r})}{\int d\mathbf{r} \int d\mathbf{p} f_{shake}(\mathbf{r}, \mathbf{p})} \end{aligned} \quad (\text{A12})$$

where $A(\mathbf{r})$ is an arbitrary function of \mathbf{r} . The above average can be evaluated by running a constrained trajectory $(\mathbf{r}_t, \mathbf{p}_t)$ on the seam surface via use of the SHAKE method.²²

$$\langle A(\mathbf{r}) \rangle_{shake} \approx \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T dt A(\mathbf{r}_t) \quad (\text{A13})$$

In the present case, $A(\mathbf{r})$ is given by $[E - U_0]^{1/2} \langle \Theta_t \rangle_p$ in eq A12. The factor $\langle \Theta_t \rangle_p$ is obtained from an (internal) average over the momenta sampled randomly from a sphere of radius $|\mathbf{p}| = (2(E - U_0))^{1/2}$. (In practice, the latter average can be merged with the trajectory propagation by evaluating Θ_t with random momenta rather than with the current momenta of the trajectory.) We also note that the above sampling procedure is analogous to the bluemoon sampling method.²³ In the latter case, one samples a conditional canonical distribution, $\exp(-\beta H) \delta(\sigma)$, while in the present case we sample a conditional microcanonical distribution, $\delta(E - H) \delta(\sigma)$.

In section III.B, we defined the survival probability on the T_1 surface by taking account of the transition probability, i.e.

$$P_{surv}(t) = \frac{\langle \langle |v_\perp| P_{NA}(v_\perp) \Theta(R^\ddagger - R_t) \rangle \rangle_{seam}}{\langle \langle |v_\perp| P_{NA}(v_\perp) \rangle \rangle_{seam}} \quad (\text{A14})$$

The latter expression can be evaluated analogously by rewriting it to

$$P_{surv}(t) = \frac{\langle [E - U_0]^{1/2} \langle |v_\perp| P_{NA}(v_\perp) \Theta(R^\ddagger - R_t) \rangle_p \rangle_{shake}}{\langle [E - U_0]^{1/2} \langle |v_\perp| P_{NA}(v_\perp) \rangle_p \rangle_{shake}} \quad (\text{A15})$$

and calculating the average $\langle \dots \rangle_{shake}$ with the SHAKE method.

■ APPENDIX B: RRKM RATE CONSTANT AND ITS CLASSICAL LIMIT

The RRKM rate for the T_1 dissociation, $k(E) = N^\ddagger(E)/(2\pi\hbar\rho(E))$, was evaluated via direct count of anharmonic vibrational states. Specifically, we first approximated $\rho(E)$ as

$$\rho(E) \approx \frac{N_R(E + \Delta E) - N_R(E - \Delta E)}{2\Delta E} \quad (\text{B1})$$

where $N_R(E)$ is the number of vibrational states in the T_1 reactant region. The vibrational states at the TS and in the reactant region were obtained approximately as the direct product of one-dimensional eigenstates of the Hamiltonian for individual normal coordinates $\{Q_i\}$:

$$\hat{H}_i = -\frac{\hbar^2}{2} \frac{\partial^2}{\partial Q_i^2} + V_i(Q_i) \quad (\text{B2})$$

where $V_i(Q_i)$ is the one-dimensional section of the T_1 PES in the direction of Q_i . The diagonalization of \hat{H}_i was performed using the discrete variable representation (DVR). For the torsional mode ϕ , we used a hindered-rotor Hamiltonian given by

$$\hat{H}_{tors} = -\frac{\hbar^2}{2I_{eff}} \frac{\partial^2}{\partial \phi^2} + \frac{V_b}{2} [1 - \cos(2\phi)] \quad (\text{B3})$$

where I_{eff} is the effective inertia moment and V_b is the torsional barrier height. Thus, the anharmonicity of the multidimensional PES is approximately taken into account in $\rho(E)$ and $N^\ddagger(E)$. The above RRKM calculation was repeated for several different values of \hbar in order to obtain an approximate classical limit ($\hbar \rightarrow 0$). It should be noted that the number of DVR grid points needs to be increased with decreasing \hbar . This is because the eigenfunctions becomes more and more oscillatory for small \hbar and also the density of states increases with decreasing \hbar . It is also crucial to optimize the way of calculating $N_R(E)$ such that high energy states are not actually included in multidimensional loops, because failure to do so makes the calculation too time-consuming for small \hbar .

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: yamamoto@kuchem.kyoto-u.ac.jp.

■ ACKNOWLEDGMENT

We are grateful to Dr. T. Yonehara and Dr. K. Fujimoto for helpful discussions on the present calculation. This work was

supported by the Grant-in-Aid for the Priority Area “Molecular Theory for Real Systems” and the Global COE Program “International Center for Integrated Research and Advanced Education in Materials Science,” from the Ministry of Education, Culture, Sports, Science and Technology of Japan. During the preparation of this paper, our coauthor Professor Shigeki Kato passed away on March 31, 2010. Y.O and T.Y. are deeply saddened to lose him and would like to dedicate this paper in his memory.

REFERENCES

- (1) Lovejoy, E. R.; Kim, S. K.; Moore, C. B. *Science* **1992**, *256*, 1541.
- (2) Kim, S. K.; Lovejoy, E. R.; Moore, C. B. *J. Chem. Phys.* **1995**, *102*, 3202.
- (3) Gezelter, J. G.; Miller, W. H. *J. Chem. Phys.* **1996**, *104*, 3546.
- (4) Steinfeld, J. I.; Francisco, J. S.; Hase, W. L. *Chemical Kinetics and Dynamics*; Prentice-Hall: Englewood Cliffs, NJ, 1989.
- (5) Henriksen, N. E.; Hansen, F. Y. *Theories of Molecular Reaction Dynamics*; Oxford University Press: New York, 2008.
- (6) Ogihara, Y.; Yamamoto, T.; Kato, S. *J. Phys. Chem. A* **2010**, *114*, 9981.
- (7) King, R. A.; Allen, W. D.; Ma, B.; Schaefer, H. F., III. *Faraday Discuss.* **1998**, *110*, 23.
- (8) Ogihara, Y.; Yamamoto, T.; Kato, S. *Chem. Phys. Lett.* **2011**, *511*, 28.
- (9) Leu, G.; Huang, C.; Lee, S.; Lee, Y.; Chen, I. *J. Chem. Phys.* **1998**, *109*, 9340.
- (10) King, R. A.; Allen, W. D.; Schaefer, H. F., III. *J. Chem. Phys.* **2000**, *112*, 5585.
- (11) Cui, Q.; Morokuma, K. *J. Chem. Phys.* **1997**, *107*, 4951.
- (12) Kaledin, A. L.; Seong, J.; Morokuma, K. *J. Phys. Chem. A* **2001**, *105*, 2731.
- (13) Preston, R. K.; Tully, J. C. *J. Chem. Phys.* **1971**, *54*, 4297.
- (14) Tully, J. C.; Preston, R. K. *J. Chem. Phys.* **1971**, *55*, 562.
- (15) Ryckaert, J.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
- (16) Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24.
- (17) Dunning, J. T. H. *J. Chem. Phys.* **1989**, *90*, 1007.
- (18) Nguyen, K. A.; Rossi, I.; Truhlar, D. G. *J. Chem. Phys.* **1995**, *103*, 5522.
- (19) Ruiz-Pernia, J. J.; Silla, E.; Tunon, I.; Marti, S.; Moliner, V. *J. Phys. Chem. B* **2004**, *108*, 8427.
- (20) 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.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.
- (21) Granovsky, A. A. Firefly version 7.1.G. <http://classic.chem.msu.su/gran/firefly/index.html> (accessed July 2011).
- (22) Ryckaert, J. P.; Ciccotti, G. *J. Chem. Phys.* **1983**, *78*, 7368.
- (23) Carter, E. A.; Ciccotti, G.; Hynes, J. T.; Kapral, R. *Chem. Phys. Lett.* **1989**, *156*, 472.
- (24) Calvo, F.; Galindez, J.; Gadea, F. X. *J. Phys. Chem. A* **2002**, *106*, 4145.
- (25) Note that the back transfer from the T_1 to the S_0 state is thus totally neglected in this paper. Inclusion of the latter may slightly modify the T_1 dissociation rate because a small fraction of the T_1 trajectories may return to the S_0 state prior to the triplet dissociation.
- (26) Landau, L. D. *Phys. Z.* **1932**, *2*, 46.
- (27) Zener, C. *Proc. R. Soc. London, Ser. A* **1932**, *137*, 696.
- (28) Wittig, C. *J. Phys. Chem. B* **2005**, *109*, 8428.
- (29) Manthe, U.; Meyer, H. D.; Cederbaum, L. S. *J. Chem. Phys.* **1992**, *97*, 3199.
- (30) Beck, M. H.; Jackle, A.; Worth, G. A.; Meyer, H. D. *Phys. Rep.* **2000**, *324*, 1.
- (31) Worth, G. A.; Robb, M. A.; Burghardt, I. *Faraday Discuss.* **2004**, *127*, 307.
- (32) Meyer, H.-D.; Gatti, F.; Worth, G. A. *Multidimensional Quantum Dynamics: MCTDH Theory and Applications*; Wiley-VCH: Berlin, 2009.
- (33) Garashchuk, S.; Vazhappilly, T. *J. Chem. Phys.* **2009**, *131*, 164108.
- (34) Czako, G.; Kaledin, A. L.; Bowman, J. M. *J. Chem. Phys.* **2010**, *132*, 164103.
- (35) Czako, G.; Kaledin, A. L.; Bowman, J. M. *Chem. Phys. Lett.* **2010**, *500*, 217.
- (36) Yamamoto, T.; Miller, W. H. *J. Chem. Phys.* **2003**, *118*, 2135.
- (37) Schranz, H. W.; Nordholm, S.; Nyman, G. *J. Chem. Phys.* **1991**, *94*, 1487.

Beyond Microscopic Reversibility: Are Observable Nonequilibrium Processes Precisely Reversible?

Divesh Bhatt and Daniel M. Zuckerman*

Department of Computational and Systems Biology, University of Pittsburgh, 3501 Fifth Ave, Biomedical Sciences Tower 3, Pittsburgh, Pennsylvania 15260, United States

S Supporting Information

ABSTRACT: Although the principle of microscopic reversibility has been studied for many decades, there remain ambiguities in its application to nonequilibrium processes of importance to chemistry, physics, and biology. Examples include whether protein unfolding should follow the same pathways and in the same proportions as folding and whether unbinding should likewise mirror binding. Using continuum-space calculations which extend previous kinetic analyses, we demonstrate formally that the precise symmetry of forward and reverse processes is expected only under certain special conditions. Approximate symmetry will be exhibited under a separate set of conditions. Exact, approximate, and broken symmetry scenarios are verified in several ways: using numerical calculations on toy and molecular systems; using exact calculations on kinetic models of induced fit in protein–ligand binding; and based on reported experimental results. The analysis highlights intrinsic challenges and ambiguities in the design and the analysis of both experiments and simulations.

1. INTRODUCTION

What does unbinding indicate about binding? Unfolding about folding?^{1,2} Do the reverse steps of a motor simply reverse the mechanism of the forward steps? In some systems, experiments have verified an overall type of reversibility, such as for ATP synthase which can either produce ATP driven by a proton gradient or hydrolyze ATP to pump protons, depending on conditions.^{3–5} Similarly, a tethered kinesin motor protein has been shown to hydrolyze or synthesize ATP depending upon concentrations of the reactants and products.⁶ Closer examination of some reversible processes suggests, however, that the forward and reverse mechanisms may not always coincide.^{1,7}

In fact, for processes that can occur via multiple mechanisms/pathways, reversibility entails a fundamental question about symmetry: Does a pathway occur in the same fraction in the forward and reverse directions? For systems at equilibrium, the answer must be in the affirmative.^{8,9} However, most physiological processes take place under nonequilibrium conditions that can resemble steady states. Experiments which study only a single direction of a process (at a time), such as folding or binding, are also out of equilibrium by definition. For such systems, as we will show, the general principle of “microscopic reversibility” is not sufficient to determine whether symmetry should hold. Molecular simulations of protein processes foreshadow this point, having provided ambiguous or conflicting conclusions about forward–reverse symmetry.^{10,11}

Symmetry issues not only are of fundamental interest, but they could have broad practical implications for the design and the analysis of experiments and computer simulations. For example, reverse steps in molecular motors are much rarer than forward motion; study of the reverse process would be greatly facilitated if the forward process could be used as a model. Likewise, in a computational setting, typically it will be much easier to observe the unbinding of two molecules, compared to the binding

(e.g., refs 12 and 13). Similarly, unfolding is more readily simulated than folding.^{1,14,15}

The framework underpinning the question of forward–reverse symmetry can be made precise. We will consider processes which take a system from some “state” A to another state B as well as the reverse processes. A state will be an arbitrary region of configuration space (presumably connected although this is not required by the formalism below). For example, in a conformational transition in a biomolecule, a state could consist of all configurations within a specified root-mean-squared deviation from a reference structure or an ‘inherent structure’ basin.^{16,17} For a process involving multiple molecules (e.g., binding, catalysis), states are regions in the full system configuration space; a bound state might be defined by a cutoff distance between molecular centers of mass, possibly augmented by conformational conditions. Importantly, in our analysis, states need not correspond to a single energy basin nor even to a set of rapidly interconverting basins.

More than one dynamical pathway, or “mechanism,” may connect a pair states A and B as has been suggested for molecular systems, such as kinesin^{18–20} and adenylate kinase^{10,11,21,22} and as in protein folding.^{1,23–25} Colloquially, a pathway Γ is a time-ordered sequence of states through which a trajectory passes going from A to B. For example, if there are two intermediate states, I_1 and I_2 , possible pathways include: (i) $A-I_1-B$, (ii) $A-I_2-B$, and (iii) $A-I_1-I_2-B$. Based on the statistical mechanics of dynamical trajectories, these different pathways will have different probabilities²⁶ under a given set of defined conditions, such as equilibrium or another steady state.

In greater detail, pathways can be understood in terms of trajectories and projection operators. A trajectory is a sequence ζ

Received: February 6, 2011

Published: June 22, 2011

$= \{\mathbf{r}^N(t_0), \mathbf{r}^N(t_0 + \Delta t), \mathbf{r}^N(t_0 + 2\Delta t, \dots)\}$, in the limit $\Delta t \rightarrow 0$, of full-system configurations \mathbf{r}^N through which a system passes at times t_i as it evolves in response to all forces and conditions. Such a trajectory can immediately be “transcribed” as a sequence of states, e.g., $\zeta \rightarrow \{A, I_1, I_1, I_2, I_1, B\}$, given predefined states. The state sequence, in turn, can be queried by a projection operator as to whether it fulfills a set of conditions specific to a pathway Γ . The preceding pathway examples (i–iii) correspond to the conditions: (i) the state sequence includes I_1 at least once but not I_2 ; (ii) it includes I_2 but not I_1 ; and (iii) it includes both I_1 and I_2 such that all occurrences of I_1 precede all occurrences of I_2 .

The question of symmetry now can be posed precisely if we let Γ' be the reverse of the path Γ . We want to study when the following symmetry of path probabilities P holds

$$\frac{P(\Gamma_i)}{P(\Gamma_j)} = \frac{P(\Gamma'_i)}{P(\Gamma'_j)} \quad (1)$$

for every pair of pathways i and j . Previous work has investigated this issue of forward–reverse symmetry. For systems with discrete states, a rather complete treatment was provided by Krupka et al.⁸ showing that symmetry in the sense of eq 1 must hold when A and B are single discrete states. Other work implies the possibility of symmetric continuum systems, for instance in the study of protein unfolding processes to obtain potential folding intermediates.^{14,15} Some computational studies have suggested the presence of symmetry,^{2,11,22} while others appear to show asymmetry.¹⁰ We note that Onsager’s reciprocity relations do not address the issue of forward and reverse processes investigated here; moreover, those relations are based on a linearized theory,²⁷ but no assumption of linear behavior is made in our general derivation.

Here we report a formal derivation of a symmetry relation for dynamical processes, along with explicit conditions required for the symmetry to hold. We find that, in important cases of interest, such as when states are ill-defined or contain multiple basins, symmetry may not hold and should not be assumed. On the other hand, when states can be well-defined, the reverse paths can indeed be studied using the forward process. This fact can be exploited when there is one ‘side’ of a process which is intrinsically better defined—such as the folded state of a protein.

The derivation is based on a novel decomposition of an equilibrium ensemble of trajectories into two special steady states. Based on the derived conditions, several symmetric examples and nonsymmetric counter examples are investigated in toy and molecular (numerical) systems. Examples culminate in the detailed analysis of the balance among “induced fit” and “conformational selection” mechanisms in protein–ligand binding, revealing results which appear to be new and fundamental. We also discuss the case of driven systems, such as when binding is coupled to catalysis of a regulated substrate, such as ATP. Some recent experimental results bearing on symmetry are discussed.

2. DERIVATION OF THE CONDITIONS FOR SYMMETRY

This section presents a trajectory-based derivation of the conditions required for an equilibrium-like symmetry to hold. An alternative derivation based on a mathematical statement of microscopic reversibility is given in the Supporting Information. We highlight the trajectory-based derivation because it yields additional information about retaining symmetry outside of equilibrium.

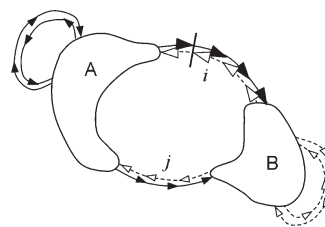


Figure 1. Schematic depiction of a system at equilibrium with two states, A and B. Transitions between the two states occur via two distinct paths, i and j . Directed lines are used to classify possible paths: transitioning “AB” trajectories start from A and reach B before coming back to A, whereas looping “AA” trajectories start in A but come back to A before reaching B (both shown as solid lines). Oppositely directed “BA” and “BB” trajectories which start from B are also shown (dashed lines). At equilibrium, the net flux across any surface, such as the solid bar across path i , is zero.

2.1. Preliminaries: Symmetry in Equilibrium. We begin by considering a situation of equilibrium as sketched in Figure 1, constructed from a large ensemble of systems undergoing natural dynamics, following a long equilibration period. For simplicity, we assume there are two states (A and B) and two distinct pathways or channels (i and j) connecting the states as shown in Figure 1, but our discussion is more generally applicable. In a system obeying detailed balance in equilibrium, both the probability density and the probability flows are unchanging in time, reflecting averages over the large ensemble. By invoking detailed balance, we are assuming that the Hamiltonian and the dynamics are time invariant.

The ensemble can be usefully decomposed in several ways. First, if we consider a single point in time, each system in the ensemble either is in one of the states (A or B) or not. We will focus on the fraction of systems not in either state, which can be further classified if we assume complete knowledge of the past and the future of each system.²⁸ In particular, all systems which were most recently in state A either will proceed back to state A or make a transition to state B; call these AA and AB trajectories, respectively. A similar classification into BB and BA trajectories applies to systems most recently in B, leading to the schematic flows shown in Figure 1. These classifications apply for arbitrary definitions of states A and B—whether physically reasonable or not.

We now want to consider the relative probabilities of two pathways or “channels” (which can be arbitrarily defined) such those schematized as i and j in Figure 1. In equilibrium, there is no net flow anywhere in configuration space because detailed balance is obeyed. For instance, along the surface shown as a straight bar across channel i in Figure 1, there will be an equal number of forward- (AB) and reverse-moving (BA) trajectories in the ensemble. (As shown by Crooks,²⁹ detailed balance requires zero net flow among the set of AA trajectories, and so these do not enter the discussion—and similarly for BB trajectories.) The AB–BA balance within a pathway, in turn, requires that the relative probability of the i and j channels in the A to B direction must be matched exactly by that in the reverse B to A direction. If the relative probabilities were different, then a net flow would occur, violating equilibrium. Thus, the symmetry relation (eq 1) is established for arbitrary definitions of A and B, in equilibrium.

2.2. Exact Symmetry in Specially Constructed Steady States. The first step in understanding implications of equilibrium

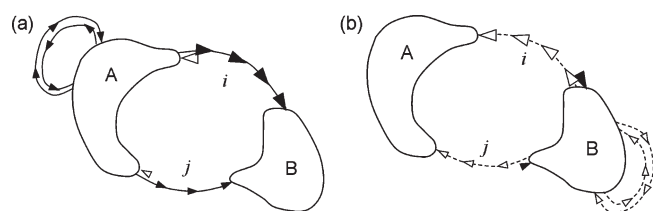


Figure 2. An exact decomposition of equilibrium (Figure 1) into opposing steady states. Panel (a) shows the A to B steady state consisting of trajectories in state A as well as the “AA” and “AB” trajectories which were most recently in state A. The open arrowheads in (a) indicate that trajectories are fed back to the surface of A exactly as they would reach A from B in equilibrium (see Figure 1). This is termed “EqSurf” feedback; see Section 2.2. A similar description applies for the reverse transition in panel b.

symmetry is to consider two unidirectional steady states which exactly “sum” to equilibrium. These steady states are schematized in Figure 2. The “forward” or A to B steady state consists of trajectories in A, along with AA and AB trajectories, as defined above. To “complete the circuit” of this steady state, trajectories arriving in B must be fed back into A. Although there are many ways to feed trajectories back into A, there is a unique prescription which precisely mimics equilibrium behavior in the A to B direction; specifically, as schematized in Figure 2a, trajectories should be fed back at the surface of A according to the distribution which would occur in the equilibrium set of BA (i.e., reverse) trajectories, i.e., according to the distribution of entry points, and momenta if applicable, to state A of BA trajectories. This fully defines the A to B steady state we wish to consider. If the reverse or B to A steady is now prepared with a mirrored prescription (see Figure 2b), then the two steady states together represent an exact decomposition of equilibrium. To see this in terms of the ensemble picture, we can say that the original equilibrium ensemble of systems has been classified into two groups constituting the forward and reverse steady states. Over time, when a successful transition occurs (either A to B or B to A), then an individual system switches to the opposite group—which is a concrete way of visualizing the feedback mechanism prescribed for the steady states.

To make further discussion more precise, we term the feedback procedure just described as “equilibrium-surface-based” (EqSurf) feedback. In EqSurf feedback, trajectories are initiated at the surface of a given state (e.g., A) according to the distribution realized in equilibrium for trajectories entering that state (A) which last visited the other state of interest (B). Although most of the present discussion concerns a feedback procedure in a steady state, we note that the EqSurf distribution (if known) could be used to construct an initial condition for a system not in a steady state.

Conditions for Exact Symmetry. We are now in a position to understand the conditions necessary for equilibrium-like symmetry to hold when unidirectional processes (forward and reverse) are studied separately. Equilibrium-like symmetry implies that path ratios, such as in eq 1, exhibit the same values as in equilibrium.

The conditions required for equilibrium-like symmetry are embodied in the steady states described above. By construction, these nonequilibrium steady states preserve the same pathway symmetry as equilibrium, i.e., eq 1 with equilibrium probabilities. Each “initial” state (A in the AB steady state or B in the BA steady

state) experiences trajectory inflow and outflow according to the EqSurf mechanism, exactly as if equilibrium held. Therefore, exact equilibrium-like symmetry cannot generally be expected unless the EqSurf distribution is used to initiate trajectories. Furthermore, a distribution of paths different from that exhibited in equilibrium will generally be expected if EqSurf-initiated trajectories are not used.

The special EqSurf-feedback mechanism, it should be emphasized, is not what might be expected. The most natural first thought would be to feed back trajectories according to the equilibrium distribution internal to the initial state, which actually would destroy the equilibrium-like behavior. A specific example of this is seen below in the study of induced-fit binding. It is also noteworthy that the alternative derivation based on microscopic reversibility (Supporting Information) does not explicitly yield the feedback mechanism required to maintain symmetry in a steady state.

2.3. Approximate Symmetry Requires Idealized States.

We can inquire whether the conditions for equilibrium-like symmetry might hold approximately when feedback schemes for establishing steady states do not exactly replicate equilibrium or when transition trajectories are generated outside the rubric of a steady state. In other words, under what conditions are the precise details of the feedback scheme (or the scheme for initializing trajectories) unimportant? Such insensitivity should occur if the A and B states are “reasonably deep” physical basins of attraction. Here, “deep” means that trajectories which enter the state are likely to remain there long enough to explore the basin fully and emerge in a quasi-Markovian way, i.e., to emerge the way trajectories would in equilibrium regardless of where they entered. Said another way, approximate symmetry is expected when intrastate time scales are much less than interstate transition times, as might be expected. Indeed, continuum models effectively revert to discrete models in this limit.

The preceding discussion implies care is required for complex systems where it may be difficult to define states obeying the quasi-Markovian property just described. In such cases, the forward–reverse symmetry relation of channel probabilities should be viewed as an approximate guideline or reference point. Interestingly, the presence or the absence of the symmetry (in experimental or computational observations) can be used as a means to validate physically meaningful state definitions.

3. EXPLORATION OF SYMMETRY: EXAMPLES AND COUNTER-EXAMPLES

Cases of symmetric and nonsymmetric processes can be carefully evaluated in several model systems, which serve to illustrate principles governing more complex systems. We examine toy models, a molecular example (alanine dipeptide), as well as the balance of induced fit and conformational selection pathways in a kinetic model of binding.

The parameters for the toy models, the forcefield parameters for alanine dipeptide (AD), and the dynamics used to establish steady state are given in the Supporting Information.

3.1. Two-Dimensional Continuum Models. Continuum toy models can exhibit a diversity of trajectories which permits illustration of symmetry principles. At the same time, the simplicity of toy models allows full sampling and enables statistically confident conclusions. We will use the two-dimensional continuum models of Figure 3a–c to illustrate three key cases: (i) exact symmetry; (ii) symmetry violation; and (iii) approximate symmetry. Each of

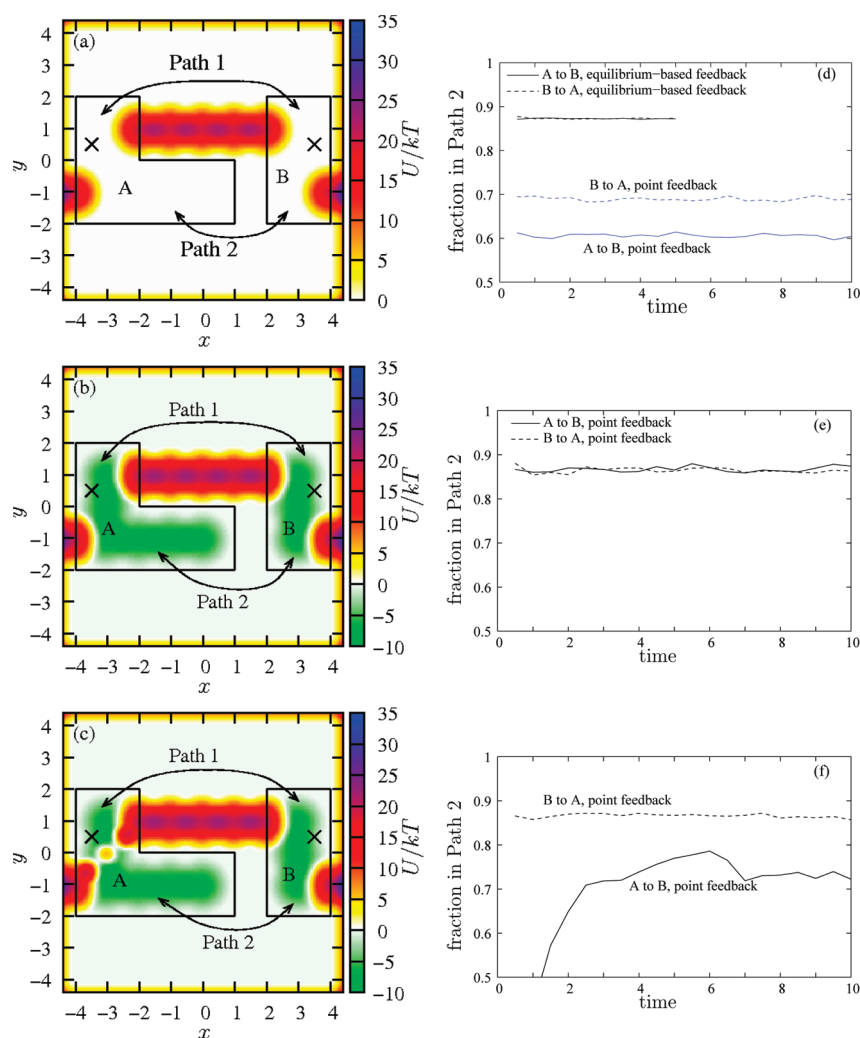


Figure 3. Two dimensional potential energy surfaces (panels a–c), and corresponding fractions of trajectories in a particular transition path (panels d–f). Each potential energy surface shows two states A and B as well as two transition pathways between the two states. Each potential exhibits characteristic features: (a) states are not well-defined basins; (b) states are well-defined basins; and (c) states are well-defined basins but state A has an internal barrier. The positions for point feedback in the respective steady states are denoted by crosses. The corresponding fractions of trajectories in path 2 along both the directions are shown in the right panels. For the potential in panel a, results for equilibrium-based (EqSurf) feedback are also shown.

the potentials is studied in a forward (A to B) and a reverse (B to A) steady state, with different feedback mechanisms designed to probe symmetry issues. Overdamped Langevin (“Brownian”) dynamics are used in all cases; details are given in the Supporting Information along with functional forms for the potentials.

Trajectories based on the potentials and states of Figure 3 can exhibit or violate symmetry depending on how they are initiated. If a steady state is established using EqSurf feedback (Section 2.2), then Figure 3d shows that symmetry between forward and reverse directions is indeed obtained, as expected. That is, the fraction of trajectories taking a given path is the same in the A to B and B to A directions. However, when trajectories are initiated from a single point after feedback, symmetry is violated as indicated in the blue traces of Figure 3d. The violation is particularly acute in the potential of Figure 3a because the states are rather arbitrary and do not correspond to physical basins.

Although EqSurf feedback leads to exact symmetry, approximate symmetry can be achieved without precise initial conditions or feedback, if suitable states can be defined. In the potential

of Figure 3b, the states correspond to well-defined, single energy basins. Figure 3e shows that symmetry can be exhibited for such a system even when the EqSurf feedback procedure is not used. In the present example, a point feedback scheme is used to initiate trajectories. For a sufficiently deep basin, the exit point of a trajectory is uncorrelated with the entry point as might be expected (Section 2.3).

In a nonsymmetric example, the barrier which is internal to state A in Figure 3c breaks the quasi-Markovian property in which a trajectory’s exit point from a state will be uncorrelated with its entry point. When trajectories are initiated in state A exclusively on one side of the barrier, Figure 3f shows that symmetry is violated. In this case, time scales internal to a state have become significant compared to transitions times, which may indeed model complex biomolecules that can be expected to possess significant barriers internal to (hypothesized) states.

3.2. A Molecular System: Alanine Dipeptide. We now show that a molecular system with well-defined states obeys the symmetry relation (eq 1) to a good approximation. Figure 4 shows two stable states of an atomistic model of AD. Although

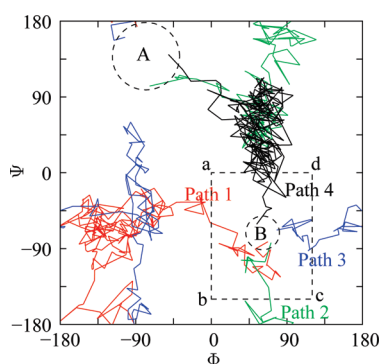


Figure 4. Ψ – Φ plane of alanine dipeptide, with two stable states, labeled A and B, shown via circles. Due to periodicity of the system, there are multiple transitions paths. We divide the paths into four types, as shown. The dashed rectangle quantifies the path definitions, e.g., for A to B transition, if the transition trajectory enters the rectangle from the left of segment *ab*, the path is classified as path 1.

Table 1. Percentage of Transition Trajectories in Four Different Paths (described in Figure 4) at Steady States in the Two Directions for AD at 500 K

	path 1	path 2	path 3	path 4
A to B	46.0 ± 6.1	17.6 ± 1.7	2.7 ± 1.0	33.7 ± 4.4
B to A	47.6 ± 3.8	15.5 ± 2.9	5.5 ± 3.9	31.4 ± 3.0

AD is a relatively simple biomolecule, transitions between the two stable states can follow multiple pathways.^{30–33} We categorize the transition paths into four types, as shown in Figure 4. Steady-state trajectories were generated using steady-state weighted ensemble path sampling as described elsewhere.³³

Table 1 shows the percentages of trajectories at a temperature of 500 K in the four paths between A and B, as shown in Figure 4. The error bars represent two standard errors of the mean and are computed from eight independent simulations. Within errors, each path occurs with the same fraction in the two directions, i.e., symmetry is observed. Trajectories were initiated, after feedback, from configurations at the centers of the states indicated in Figure 4.

AD illustrates a practical consequence of the symmetry relation because overall transition rates in the two directions differ by almost 2 orders of magnitude: 1.5/ns in the A to B direction, compared to 123/ns in reverse. Thus the “easy” direction can be used to reveal pathways of the reverse process.

3.3. Symmetry of Induced Fit Processes. The possible coexistence of both “induced fit” (IF) and “conformational selection” (CS) pathways in protein–ligand binding has recently received attention.^{34,35} The IF mechanism (upper pathway in Figure 5) entails initial complexation of a ligand *L* to a weak-binding conformation *W* of a receptor, leading to state *WL*, followed by fitting to a tightly bound complex *TL*. In the conformational selection pathway (lower pathway in Figure 5), by contrast, the ligand binds directly to the tight-binding conformation *T*, which is in equilibrium with *W*. We analyze the balance among these pathways using the kinetic model shown in Figure 5, which was previously employed by others.^{34,35} The use of nontraditional symbols for rate constants reduces the use of subscripts and superscripts. We note that *k*, *v*, and ω are first-order rate constants, while σ denotes a second-order rate constant.

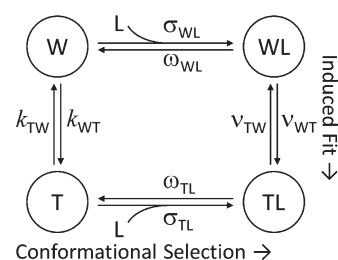


Figure 5. A standard kinetic model including both IF and CS pathways. All rate constants are specified.

Exact calculation of the ratio of IF and CS pathways confirms the symmetry relation fully, as described below, and also demonstrates the predicted dependence of path probabilities on state definitions. To our knowledge, this latter sensitivity has not been previously noted. Symmetry can be violated, moreover, if one of the end states consists of multiple substates and if feedback (or initialization) is not performed suitably.

We show that the symmetry is preserved for different formulations of the problem (via different state definitions), the first of which is schematized in Figure 6. In this case, the initial state is chosen as state *W* alone, with *TL* as the final state, yielding the ratio of pathways:³⁵

$$\frac{P(\Gamma_{CS})}{P(\Gamma_{IF})} = \frac{\sigma_{TL}k_{WT}}{\sigma_{WL}v_{WT}} \frac{\omega_{WL} + v_{WT}}{k_{TW} + [L]\sigma_{TL}} \quad (\text{state A} = \text{W}) \quad (2)$$

where $[L]$ is the free ligand concentration. The same result is found whether the forward or reverse process is considered; see full details in the Supporting Information.

As an alternative formulation of the problem, it is also natural to have the initial state *A* consist of the overall unbound state (i.e., both *W* and *T*), again with *TL* as the final state. This scheme, which we emphasize employs precisely the same kinetic model, is illustrated in Figure S1 in the Supporting Information. It yields a different ratio of path probabilities, now concentration independent:

$$\frac{P(\Gamma_{CS})}{P(\Gamma_{IF})} = \frac{\sigma_{TL}k_{WT}}{\sigma_{WL}v_{WT}} \frac{\omega_{WL} + v_{WT}}{k_{TW}} \quad (\text{state A} = \text{W} + \text{T}) \quad (3)$$

Again, the ratio is the same for forward and reverse process, but the EqSurf feedback described in our derivation (Section 2.2) must be used because the initial state consists of two “substates,” *W* and *T*. Importantly, the symmetry-preserving EqSurf feedback uses the ratio of fluxes from (eq 3), which is not the equilibrium population ratio $[T]_{\text{eq}}/[W]_{\text{eq}} = k_{WT}/k_{TW}$. Details are given in Supporting Information.

Although the symmetry we find here is not surprising, two observations appear to be novel: First, the ratio of the two mechanisms is sensitive to the (subjective) choice of the initial state and is concentration dependent in one case but not the other. Second, when the initial state contains more than one substate (e.g., *W* and *T*), symmetry is only obtained exactly if EqSurf feedback is used. In a discrete system, EqSurf corresponds to initiating trajectories according to the ratio of fluxes which enter the substates in equilibrium, which is different from the equilibrium ratio of substate populations. If EqSurf initiation is not used, symmetry will hold approximately when the transition rates among state *A*’s substates (the rates k_{WT} and k_{TW}) are much faster than other processes, leading to quasi-Markovian behavior

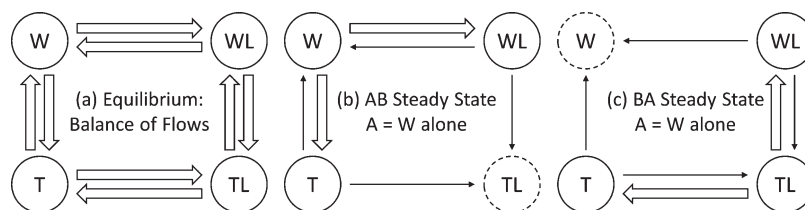


Figure 6. Equilibrium is divided into forward (AB) and reverse (BA) steady states to examine the induced fit question. Steady states are shown for the case when state A is chosen to be W only. (a) In equilibrium, there is an exact balance of flows. (b) In the AB steady state, the feedback from state B (i.e., TL in this model) to state A (i.e., W) reduces some flows. Open arrows denote flows of probability occurring in equilibrium, and single-line arrows denote reduced flows occurring in a steady state. (c) In the BA steady state, there is a similar reduction of flows due to feedback from W to TL. The flows in the two steady states exactly sum to equilibrium flows by construction. Arrow sizes are not to scale.

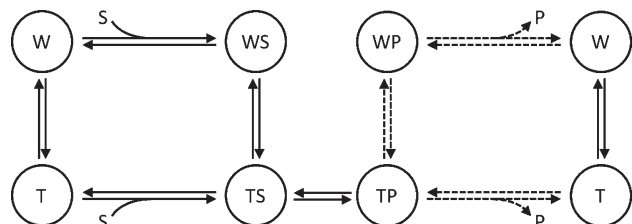


Figure 7. Coupling of IF binding to catalysis. The binding model of Figure 5 is now coupled to catalysis, with ligand labeled “S” for substrate, which is catalyzed to product “P.” The postcatalytic processes shown with dashed arrows will have different rate constants from the corresponding precatalytic processes. Although equilibrium-based symmetry arguments can be applied to this system as shown, an equilibrium basis is invalid when there is an external source of substrate S.

as discussed earlier. The presence of second-order kinetics (i.e., depending on concentrations) does not impact the symmetry arguments developed previously, but details of the calculation are more complicated, not surprisingly; see Supporting Information.

4. DISCUSSION

4.1. Symmetry Considerations in Catalysis-Coupled and Driven Systems. While symmetry arguments appear to have broad application, the underlying equilibrium-based assumptions used above break down for driven systems. Many biological systems are driven by coupling to a secondary reaction, such as ATP hydrolysis.³⁶ Interestingly, it is not the hydrolysis per se that limits the symmetry argument, but rather the fact that the cell “drives” the process by synthesizing ATP through an unrelated mechanism. As can be deduced from Figure 7, symmetry arguments could be applied to a system which couples binding and catalysis, so long as the entire system is analyzed. However, if part of the system is excluded (e.g., ATP synthesis through a mechanism other than reversal of the process shown), then the fact that the rate constants will differ for pre- and postcatalytic processes destroys the symmetry. To see this, note that when the system of Figure 7 is driven to the right, the off-rates from the WP and TP states can be seen as the rates of feedback to the W and T states. These postcatalytic off-rates generally will not fulfill the requirements for preserving symmetry; see the analysis of case 2 in the Supporting Information.

Driven systems in other contexts should also not be expected to show symmetry, and indeed the term “hysteresis” describes this textbook fact in the physical sciences. Hysteresis, such as in magnetic systems, occurs when there is driving (out of equilibrium) by a magnetic field, often followed by reverse

driving along a different pathway in the “plane” of average magnetization vs the applied field. Loads applied to motor proteins can be construed similarly, which will contribute further to the asymmetry already expected from coupling to ATP hydrolysis.

4.2. Could Nonequilibrium-Like Symmetry Occur? On a more fundamental level, one can ask whether it is possible to construct a process which is the symmetric reversal of an arbitrary nonequilibrium process. That is, given a “forward” process (e.g., protein folding following a quench) in which trajectories are initiated from some nonequilibrium distribution and in which the resulting path distribution differs from that in equilibrium, can we construct a reverse process so that the symmetry of eq 1 is realized? In principle, one can imagine time-reversing the entire process by initiating momentum-reversed trajectories from the target state. However, in practical cases of interest where a system’s degrees of freedom are coupled stochastically to a thermal environment, achieving symmetry for a nonequilibrium process does not seem possible in general. This can be seen by constructing a counter example: If the reverse process is initiated from a sufficiently deep basin, emerging trajectories will be effectively Markovian and will exhibit an equilibrium distribution of paths regardless of the nature of the forward process. Nevertheless, we cannot rule out the possibility in principle of processes obeying eq 1 but with ratios different from equilibrium.

4.3. Symmetry Aspects of Related Experimental Results. The preceding discussion hints at the complexities which confront the analysis of experimental results. The apparent clarity of the “microscopic reversibility” concept ultimately offers little insight, as our report demonstrates. Most basically it can be expected that different methods initializing a nonequilibrium study, which is akin to performing steady-state feedback in different ways, will lead to different results, as has been noted in the case of protein folding.^{1,37,38} A folding experiment can be understood as a “quench” to a nonequilibrium condition that will depend on whether the protocol used altered pH, chemical denaturant, or elevated temperature.^{1,39,37,40}

Asymmetric findings for forward and reverse processes can be understood in the context of this study. For example, it was recently reported that the mechanism of a peptide’s insertion into a lipid bilayer differed dramatically from its exit mechanism.⁷ This is hardly a violation of microscopic reversibility. Rather, it reflects the altered driving “force” applied in order to observe the insertion and exit processes separately, namely, two different pH values.

4.4. Can Equilibrium-Like Symmetry Be Observed in Simulations? It seems possible, and even practical, to construct a simulation protocol for a “forward” process which is guaranteed

to mirror the reverse process. Specifically, if one can achieve transitions from state A to B (e.g., using a steady-state path sampling approach)³³ then trajectories reaching B can be fed back to the surface of A using the EqSurf prescription of Section 2.2. The necessary distribution of entry points to state A can be obtained, using a detailed balance argument, by reversing (stored) phase-space coordinates at the surface of A of trajectories which later successfully transitioned to B.

4.5. Overview of Symmetry Issues in Observable Systems.

Although systems in equilibrium exhibit forward–reverse symmetry by virtue of detailed balance, experimental observations of pathways are almost always made out of equilibrium (e.g., for protein folding). In other words, the question of symmetry becomes interesting in cases where standard equilibrium principles, such as detailed balance, cannot be brought to bear.

This study suggests that for systems not in equilibrium, one generally should not expect forward and reverse processes to mirror one another in the sense of eq 1. Exact, equilibrium-like symmetry (i.e., where the ratios of eq 1 are that of equilibrium) only can be generally guaranteed using the EqSurf process for initiating systems, as described in Section 2.2. Approximate symmetry may be observed in cases where both end states are deep basins exhibiting the quasi-Markovian property (Section 2.3), but it is not clear that such states can typically be identified for complex systems. It does appear possible, interestingly, to construct exactly reversed equilibrium-like systems in a simulation context, as described in Section 4.4.

This report has not systematically investigated the issue of whether an arbitrary nonequilibrium process (exhibiting path ratios that differ from equilibrium) can be symmetrically reversed. However, we speculate in Section 4.2 that such symmetry is unlikely to be observed for typically complex systems.

5. CONCLUSIONS

We have addressed the question of symmetry for forward and reverse directions of a wide class of nonequilibrium processes important in biomolecular contexts, including conformational transitions, binding, and folding. To what extent does a reverse process mirror forward events, in terms of the distribution of pathways? Although the symmetry issue previously has been addressed for discrete systems,⁸ the present work contributes a more encompassing theoretical view applicable to continuum systems and also establishes a basis for understanding asymmetry in driven systems. In our trajectory-based derivation of equilibrium-like symmetry conditions, the equilibrium state is exactly decomposed into forward and reverse steady states. The formal symmetry in the two steady states occurs only under special conditions which are made explicit: Processes must be initiated according to a precise prescription (termed “EqSurf”, Section 2.2) in order for exact symmetry to hold and for the pathway distribution to recapitulate that found in equilibrium. Nevertheless, for systems with well-defined physical states characterized by rapid intrastate relaxation, approximate symmetry can be observed even when the symmetry conditions are violated. It appears to be an open question, however, whether biomolecular systems of interest tend to exhibit suitably well-defined states.

From a broader perspective, we have seen that a microscopic law does not have a direct “macroscopic” corollary. That is, microscopic reversibility does not directly translate into symmetric, reversible, observable processes. Unidirectional processes observable in experiments cannot be initiated from a precisely

defined phase-space distribution and hence are subject to the uncertainties described here. The general sensitivity to initial conditions in complex systems has been noted previously for the unfolded state of proteins.^{1,37,38}

The implications of the symmetry relation and its conditions could be wide ranging for both experimental and theoretical/computational studies. For instance, the symmetry conditions offer a prescription for when unfolding pathways will mirror folding, namely, if the folded state is characterized by fast relaxation processes and if folding is initiated from the EqSurf distribution described above. Furthermore, the folding and the unfolding must occur under the same conditions (e.g., temperature, pH), as has been noted before.^{1,2} A similar characterization can be applied to the use of unbinding studies to explore binding. We have also described a possible computational procedure for achieving symmetric reversals. Thus, the symmetry discussion presented here may open new avenues for analysis and production of biophysical data.

■ ASSOCIATED CONTENT

S Supporting Information. An alternative derivation of the symmetry relation under equilibrium conditions; further details for the induced-fit kinetic calculations; and additional information regarding models and calculation methods. This information is available free of charge via the Internet at <http://pubs.acs.org>

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ddmmzz@pitt.edu. Telephone: 412-648-3335.

■ ACKNOWLEDGMENT

We thank Chris Jarzynski and Carlos Camacho for helpful discussions and Terry Oas for useful correspondence. Support for this work was provided by the NIH (grants GM076569 and GM070987) and the NSF (grant MCB-0643456).

■ REFERENCES

- (1) Dinner, A. R.; Karplus, M. *J. Mol. Biol.* **1999**, *292*, 403–419.
- (2) McCully, M. E.; Beck, D. A. C.; Daggett, V. *Biochemistry* **2008**, *47*, 7079–8089.
- (3) Oster, G.; Wang, H. Y. *Structure (Cambridge, MA, U. S.)* **1999**, *7*, R67–R72.
- (4) Kinoshita, K.; Yasuda, R.; Noji, H.; Adachi, K. *Philos. Trans. R. Soc., B* **2000**, *355*, 473–489.
- (5) Schultz, B. E.; Chan, S. I. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 23–65.
- (6) Hackney, D. D. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 18338–18343.
- (7) Andreev, O. A.; Karabadzha, A. G.; Weerakkody, D.; Andreev, G. O.; Engelman, D. M.; Reshetnyak, Y. K. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 4081–4086.
- (8) Krupka, R. M.; Kaplan, H.; Laidler, K. J. *Trans. Farad. Soc.* **1966**, *62*, 2754–2759.
- (9) Go, N.; Abe, H. *Biopolymers* **1981**, *20*, 991–1011.
- (10) Whitford, P. C.; Miyashita, O.; Levy, Y.; Onuchic, J. N. *J. Mol. Biol.* **2007**, *366*, 1661–1671.
- (11) Lu, Q.; Wang, J. *J. Am. Chem. Soc.* **2008**, *130*, 4772–4783.
- (12) Izrailev, S.; Stepaniants, S.; Balsara, M.; Oono, Y.; Schulten, K. *Biophys. J.* **1997**, *72*, 1568–1581.
- (13) Burendahl, S.; Danculescu, C.; Nilsson, L. *Proteins* **2009**, *77*, 842–856.

- (14) Day, R.; Daggett, V. *Protein Sci.* **2005**, *14*, 1242–1252.
- (15) Fersht, A. R.; Daggett, V. *Cell* **2002**, *108*, 573–582.
- (16) Stillinger, F. H.; Weber, T. A. *Phys. Rev. A* **1982**, *25*, 978–989.
- (17) Rao, F.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 9152–9157.
- (18) Liepelt, S.; Lipowsky, R. *Phys. Rev. Lett.* **2007**, *98*, 258102.
- (19) Liepelt, S.; Lipowsky, R. *Phys. Rev. E* **2009**, *79*, 011917.
- (20) Tsygankov, D.; Fisher, M. E. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19321–19326.
- (21) Chu, J. W.; Voth, G. A. *Biophys. J.* **2007**, *93*, 3860–3871.
- (22) Bhatt, D.; Zuckerman, D. *J. Chem. Theory Comput.* **2010**, *6*, 3527–3539.
- (23) Shea, J.-E.; Onuchic, J. N.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 16064–16068.
- (24) Klimov, D. K.; Thirumalai, D. *J. Mol. Biol.* **2005**, *353*, 1171–1186.
- (25) Noe, F.; Schutte, C.; vanden Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–19016.
- (26) Pratt, L. R. *J. Chem. Phys.* **1986**, *85*, 5045–5048.
- (27) Onsager, L. *Phys. Rev.* **1931**, *37*, 405.
- (28) vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *131*, 044120.
- (29) Crooks, G. E. *J. Stat. Phys.* **1998**, *90*, 1481–1487.
- (30) van der Vaart, A.; Karplus, M. *J. Chem. Phys.* **2007**, *126*, 164106.
- (31) Jang, H.; Woolf, T. B. *J. Comput. Chem.* **2006**, *27*, 1136–1141.
- (32) Ren, W.; vanden Eijnden, E.; Maragakis, P.; E, W. *J. Chem. Phys.* **2005**, *123*, 134109.
- (33) Bhatt, D.; Zhang, B. W.; Zuckerman, D. *J. Chem. Phys.* **2010**, *133*, 014110.
- (34) Weikl, T. R.; von Deuster, C. *Proteins* **2009**, *75*, 104–110.
- (35) Hammes, G. G.; Chang, Y.-C.; Oas, T. G. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 13737–13741.
- (36) Hill, T. L. *Free Energy Transduction and Biochemical Cycle Kinetics*; Dover Publications: New York, 1989.
- (37) Hyeon, C.; Morrison, G.; Pincus, D. L.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 20288–20293.
- (38) Ensign, D. L.; Kasson, P. M.; Pande, V. S. *J. Mol. Biol.* **2007**, *374*, 806–816.
- (39) Krishna, M. M. G.; Hoang, L.; Lin, Y.; Englander, S. W. *Methods* **2004**, *34*, 51–64.
- (40) Zuckerman, D. M. *Statistical Physics of Biomolecules: An Introduction*; CRC Press: New York, 2010.

Searching for Computational Strategies to Accurately Predict pK_a s of Large Phenolic Derivatives

Aida Mariana Rebollar-Zepeda, Tania Campos-Hernández, María Teresa Ramírez-Silva, Alberto Rojas-Hernández, and Annia Galano*

Departamento de Química, División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana-Iztapalapa, Av San Rafael Atlixco No.186, Col.Vicentina C.P.09340, México D.F.

S Supporting Information

ABSTRACT: Twenty-two reaction schemes have been tested, within the cluster-continuum model including up to seven explicit water molecules. They have been used in conjunction with nine different methods, within the density functional theory and with second-order Møller–Plesset. The quality of the pK_a predictions was found to be strongly dependent on the chosen scheme, while only moderately influenced by the method of calculation. We recommend the E1 reaction scheme $[HA + OH^- (3H_2O) \leftrightarrow A^- (H_2O) + 3H_2O]$, since it yields mean unsigned errors (MUE) lower than 1 unit of pK_a for most of the tested functionals. The best pK_a values obtained from this reaction scheme are those involving calculations with PBE0 (MUE = 0.77), TPSS (MUE = 0.82), BHandHLYP (MUE = 0.82), and B3LYP (MUE = 0.86) functionals. This scheme has the additional advantage, compared to the proton exchange method, which also gives very small values of MUE, of being experiment independent. It should be kept in mind, however, that these recommendations are valid within the cluster-continuum model, using the polarizable continuum model in conjunction with the united atom Hartree–Fock cavity and the strategy based on thermodynamic cycles. Changes in any of these aspects of the used methodology may lead to different outcomes.

INTRODUCTION

There is a huge amount of substances that behave as Brønsted acids or bases in aqueous solution. Proton-transfer equilibria are therefore very important for a large variety of chemical compounds and in particular for pharmaceuticals, which frequently are weak acids or bases.¹ Acid dissociation constants (K_a) not only characterize the acidity of these compounds but also influence their reactivity. They are commonly reported as pK_a s, and their values are related to numerous properties of drugs, such as solubility and rate of absorption.² They are also taken into account to decide dosage forms and regimes of drugs.³ Therefore the accurate knowledge of pK_a s is highly important for practical purposes as well as for understanding the behavior of chemicals under different conditions.

There are several experimental techniques that have been successfully applied to accurately determine pK_a s. However sometimes this becomes a challenging task,⁴ for example, for short-living intermediates and for very weak or very strong acids.⁵ Thus a large amount of works have been devoted to obtain pK_a values using theoretical methods.^{2,5–26} Different calculation strategies and current trends have been recently and thoroughly reviewed by Ho and Coote.²⁷

Considering all the information gathered so far it becomes evident that estimating accurate pK_a s using computational methodologies remains a very complicated problem. The difficulties are numerous. To start with, deprotonation processes do not conserve the number of charged species on both sides of the equilibrium. Therefore there is no cancellation of errors when computing energies of reaction. As a result, the accuracy of these relative energies is not as predictable as for processes where the

number of charged species is conserved. A second problem is that even though high-level composite methods, such as Gaussian-n (Gn)²⁸ and complete basis set (CBS)²⁹ variations, are now available for producing accurate gas phase energies, the calculations in solution have not reached such level of accuracy yet. Moreover, Gn and CBS methods are still computationally unfeasible for moderately large systems, which is the case of most chemicals. An additional problem, particularly important for calculating pK_a s in aqueous solution, is that specific short-range solute–solvent interactions are not included when continuum models are used. On the other hand, using discrete models, including a large enough number of solvent molecules, within ab initio or density functional theory (DFT) frameworks, would be so computationally demanding that it is currently an unattainable option. If we take into account that all the above-mentioned problems represent sources of error, that do not necessarily cancel out, and that an error of 1.36 kcal/mol in the Gibbs free energy of the deprotonation reactions (ΔG) represents an error of 1 pK_a unit, then it becomes evident that producing accurate pK_a values from theoretical calculations is a very challenging task. In fact it is currently accepted that mean absolute deviations smaller than 2 units of pK_a are reasonably accurate.²⁷

Albeit a wide variety of strategies that have been developed to overcome such difficulties, in this work we will focus on those involving thermodynamic cycles. This approach allows a large flexibility in designing reaction schemes that maximize systematic error cancellations.²⁷ Some of them can be designed in such a way

Received: March 17, 2011

Published: June 29, 2011

that the number of charged species is conserved on both sides of the deprotonation equilibrium. Thermodynamic cycles also allow obtaining Gibbs free energies in solution ΔG_s from Gibbs free energies in gas phase (ΔG_g) and Gibbs free energies of solvation (ΔG_{solv}). This seems to be an efficient strategy since continuum solvation models are parametrized to produce accurate ΔG_{solv} , but the levels of theory at which they are usually implemented are not sufficient to accurately reproduce ΔG_s .²⁷ Therefore it is possible to take advantage of using higher levels of theory to calculate ΔG_g and therefore to improve the accuracy of the results. However, as mentioned before, bulk polarization effects are not necessarily enough to accurately reproduce ΔG_{solv} because of the short-range solute–solvent interactions. One way to overcome this issue, without making the calculations computationally prohibited, is to include a few solvent molecules in close proximity to the solute, in addition to using a dielectric continuum model. This hybrid explicit/implicit solvation model is known as the discrete-continuum,³⁰ the supermolecule-reaction field,³¹ and the cluster-continuum model.³²

Taking into account all the possible levels of theory, solvation methods, reaction schemes and number of explicit solvent molecules that can be used for calculating $\text{p}K_a$ s with thermodynamic cycles, the amount of variations becomes almost infinite. Therefore in this work we will focus on levels of theory that are computationally feasible for relatively large molecules, mainly within the DFT framework. They are used in combination with the polarizable continuum model (PCM), which is currently one of the most widely used for calculating $\text{p}K_a$ s in aqueous solution. Twenty-two reaction schemes have been used, some based on those previously reported by other authors and some proposed for the first time. In addition the cluster-continuum model has been used, including up to seven explicit water molecules. To test the resulting strategies we have chosen four nonsteroidal anti-inflammatory drugs, which are phenolic derivatives: acetaminophen, profadol, tapentadol, and ketobemidone. Then the study has been extended to a larger series of phenols.

Phenolic compounds are the focus of our study since they are ubiquitous and versatile substances. They have been identified to have multiple biological activities, including antioxidant³³ and cardioprotective³⁴ effects as well as anti-inflammatory,³⁵ antimicrobial, and antiviral³⁶ activities. They are also used to prevent and treat cancer³⁷ and neurodegenerative diseases³⁸ and to prevent skin damage³⁹ and osteoporosis.⁴⁰ Phenols are even proposed to have inhibitory effects against obesity.⁴¹ All these biological activities take place within the human body, i.e., at different pHs depending on the specific site of action. Since their effects vary depending of their $\text{p}K_a$, which rules the fractions of their acid/base forms under specific conditions, it is vital to know this data.^{33f} In addition, from a computational point of view, phenols represent an extra challenge for estimating acid constants, due to the influence of solvation in the proper description of the phenoxide anions' geometry.¹⁰ Even though highly accurate strategies have been proposed for calculating the $\text{p}K_a$ s of phenols, they involve CBS-QB3 calculations,¹⁰ which are computationally unfeasible for most phenolic compounds with biological activity.

Based on the systematic analysis of the extensive data obtained, a few strategies are recommended for accurate calculations of aqueous $\text{p}K_a$ s of relatively large-sized molecules with phenolic deprotonation sites, at reasonable computational cost.

COMPUTATIONAL DETAILS

Full geometry optimizations and frequency calculations were performed with the package of programs Gaussian 03.⁴² Different levels of theory have been used: BLYP, B3LYP, BHandHLYP, PBE, PBE0, PW91, BMK, TPSS, and M05-2x DFT methods and the MP2 wave function method for some of the smallest systems. All of them in conjunction with the 6-311++G(d,p) basis set. Local minima were identified by the absence of imaginary frequencies. The stationary points were first modeled in gas phase (vacuum), and solvent effects were included a posteriori by single point calculations using a PCM, specifically the integral equation formalism (IEF-PCM).⁴³ In PCM calculations, the choice of the solute cavity is important because the computed energies and properties strongly depend on the cavity size. In the present study the cavity has been built using the united atom model for Hartree–Fock (UAHF) method,⁴⁴ at HF/6-31+g(d), which is the recommended approach for predicting free energies of solvation according to the Gaussian 03 User's Reference.⁴⁵

Relative Gibbs free energies in solution ΔG_s have been computed using the Hess law and thermodynamic cycles, as the sum of the corresponding gas-phase free energy ΔG_g and the free energy of solvation (ΔG_{solv}). They include standard thermal corrections at 298.15 K. In all the cases the used reference state is 1 M. The aqueous Gibbs free energies for the deprotonation reactions are in turn used to compute the acid equilibrium constant (K_a), according to

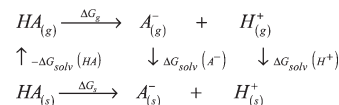
$$K_a = e^{-\Delta G_s/RT} \quad (1)$$

Then, the $\text{p}K_a$ values are obtained using its definition:

$$\text{p}K_a = -\log(K_a) \quad (2)$$

which can also be calculated directly from ΔG_s , from expressions that depend on the particular scheme of reaction used to model the deprotonation process.

Reaction Schemes. The reaction schemes tested in this work are reported in Table 1, together with the expressions corresponding to direct calculations of $\text{p}K_a$ from ΔG_s . The latter have been calculated using the strategy based on thermodynamic cycles described above. For example for the reaction scheme A, the cycle is



and the Gibbs free energy of reaction in solution ΔG_s is obtained as the sum of the Gibbs free energy of reaction in vacuum (ΔG_g) and the difference in solvation free energies ($\Delta\Delta G_{\text{solv}}$):

$$\Delta G_s = \Delta G_g + \Delta\Delta G_{\text{solv}} \quad (3)$$

where $\Delta\Delta G_{\text{solv}}$ and ΔG_g are calculated as

$$\Delta\Delta G_{\text{solv}} = \Delta G_{\text{solv}}(A^-) + \Delta G_{\text{solv}}(H^+) - \Delta G_{\text{solv}}(HA) \quad (4)$$

$$\Delta G_g = G_g(A^-) + G_g(H^+) - \Delta G_g(HA) \quad (5)$$

with ΔG_{solv} representing the free energies of solvation of each species.

Scheme A is by far the most frequently used, probably because of its simplicity. However it involves the proton, and it is known that computational methods poorly reproduce the solvation energies of this particular species. Therefore the $\Delta G_g(H^+)$ and

Table 1. Reaction Schemes and the Corresponding Expressions to Directly Calculate pK_a Values From ΔG_s

scheme	equilibrium	$pK_a =$
A	$HA \leftrightarrow H^+ + A^-$	$\Delta G_s/RT \ln(10)$
B	$HA + H_2O \leftrightarrow H_3O^+ + A^-$	$\Delta G_s/RT \ln(10) - \log[H_2O]$
C1	$HA + 2H_2O \leftrightarrow H_3O^+ + A^-(H_2O)$	$\Delta G_s/RT \ln(10) - 2\log[H_2O]$
C2	$HA + 3H_2O \leftrightarrow H_3O^+ + A^-(2H_2O)$	$\Delta G_s/RT \ln(10) - 3\log[H_2O]$
C3	$HA + 4H_2O \leftrightarrow H_3O^+ + A^-(3H_2O)$	$\Delta G_s/RT \ln(10) - 4\log[H_2O]$
D	$HA + \text{ref}^- \leftrightarrow A^- + \text{HRef}$	$\Delta G_s/RT \ln(10) + pK_a(\text{HRef})$
E1	$HA + OH^-(3H_2O) \leftrightarrow A^-(H_2O) + 3H_2O$	$(\Delta G_s)/RT \ln(10) + 14 + 3\log[H_2O]$
E2	$HA + OH^-(3H_2O) \leftrightarrow A^-(2H_2O) + 2H_2O$	$\Delta G_s/RT \ln(10) - 2\log[H_2O]$
E3	$HA + OH^-(3H_2O) \leftrightarrow A^-(3H_2O) + H_2O$	$\Delta G_s/RT \ln(10) + 14 + \log[H_2O]$
F1	$HA + 4H_2O \leftrightarrow H_3O^+(3H_2O) + A^-$	$\Delta G_s/RT \ln(10) - 4\log[H_2O]$
F2	$HA + 5H_2O \leftrightarrow H_3O^+(3H_2O) + A^-(H_2O)$	$\Delta G_s/RT \ln(10) - 5\log[H_2O]$
F3	$HA + 6H_2O \leftrightarrow H_3O^+(3H_2O) + A^-(2H_2O)$	$\Delta G_s/RT \ln(10) - 6\log[H_2O]$
F4	$HA + 7H_2O \leftrightarrow H_3O^+(3H_2O) + A^-(3H_2O)$	$\Delta G_s/RT \ln(10) - 7\log[H_2O]$
CN1	$HA(H_2O) + H_2O \leftrightarrow H_3O^+ + A^-(H_2O)$	$\Delta G_s/RT \ln(10) - \log[H_2O]$
CN2	$HA(H_2O) + 2H_2O \leftrightarrow H_3O^+ + A^-(2H_2O)$	$\Delta G_s/RT \ln(10) - 2\log[H_2O]$
CN3	$HA(H_2O) + 3H_2O \leftrightarrow H_3O^+ + A^-(3H_2O)$	$\Delta G_s/RT \ln(10) - 3\log[H_2O]$
FN1	$HA(H_2O) + 4H_2O \leftrightarrow H_3O^+(3H_2O) + A^-(H_2O)$	$\Delta G_s/RT \ln(10) - 4\log[H_2O]$
FN2	$HA(H_2O) + 5H_2O \leftrightarrow H_3O^+(3H_2O) + A^-(3H_2O)$	$\Delta G_s/RT \ln(10) - 5\log[H_2O]$
FN3	$HA(H_2O) + 6H_2O \leftrightarrow H_3O^+(3H_2O) + A^-(3H_2O)$	$\Delta G_s/RT \ln(10) - 6\log[H_2O]$
EN1	$HA(H_2O) + OH^-(3H_2O) \leftrightarrow A^-(H_2O) + 4H_2O$	$\Delta G_s/RT \ln(10) + 14 + 4\log[H_2O]$
EN2	$HA(H_2O) + OH^-(3H_2O) \leftrightarrow A^-(2H_2O) + 3H_2O$	$\Delta G_s/RT \ln(10) + 14 + 3\log[H_2O]$
EN3	$HA(H_2O) + OH^-(3H_2O) \leftrightarrow A^-(3H_2O) + 2H_2O$	$\Delta G_s/RT \ln(10) + 14 + 2\log[H_2O]$

$\Delta G_{\text{solv}}(H^+)$ values have been derived from experiments. This constitutes the main disadvantage of using scheme A because the reported experimental values of the solvation free energy of the proton range from -259 to -264 kcal/mol.¹⁹ Such variation is an important source of error in the calculation of pK_a s, i.e., it alone represents about 3 pK_a units. In this work we have used $\Delta G_g(H^+) = -4.39$ kcal/mol and $\Delta G_{\text{solv}}(H^+) = -265.89$ kcal/mol, based on the recommendation of Camaioni and Schwerdtfeger.⁴⁶

A simple strategy to avoid using experimental values of ΔG_{solv} is to use a water molecule as a coreactant (scheme B). This strategy can be extended, within the cluster-continuum model, by including more than one water molecule and modeling the proton and/or the anionic conjugated base explicitly solvated (schemes C and F). The difference between these two kinds of schemes is that in C schemes, the proton is solvated with only one water molecule, while in F schemes, it is solvated by four of them. Moreover the explicit solvation can also be applied to the acid (schemes CN and FN). However the number of charged species on both sides of the equilibrium is not conserved for any of these schemes.

Scheme D, on the other hand, usually lead to more reliable values of pK_a s since the number and the kind, of charged species is conserved on both sides of the chemical equation. This particular approach is known as the proton exchange method, the isodesmic method, or the relative method to calculate pK_a s. The main problem with this approach is that its outcome strongly depends on the choice of the reference acid (HRef). Therefore, the quality of the results would be determined by the structural similarity between HRef and HA, and also by the accuracy of the experimental value of $pK_a(\text{HRef})$. In this work we have used phenol as HRef, and its pK_a value equal to 10.09.⁴⁷ From a theoretical point of view schemes D and A both have the same drawback, they directly depend on experimental data. Such

experiment dependency weakens the predictive character of any computational methodology aiming for predicting pK_a s or any other chemical property. Thus schemes E and EN have the advantage of being experiment independent, and at the same time, they conserve the number and kind of charged species on both sides of the equilibrium. The variations among them (E1–E3 and EN1–EN3) differentiate only on the number of the explicit water molecules and on the explicit solvation of the different species involved in the equilibrium.

It should be noticed that for those reaction schemes that explicitly includes water molecules, it is necessary to correct for the standard state of liquid water, i.e., 55.55 mol/L. In fact it has been previously demonstrated that ignoring such correction leads to systematic errors.^{48,49}

RESULTS AND DISCUSSION

General Considerations. Three of the nonsteroidal anti-inflammatory drugs studied in this work (profadol, tapentadol, and ketobemidone, Figure 1) have a tertiary amine site. Therefore, they may have more than one pK_a , and it is necessary to assess the relative acidity of these sites. To that purpose, we have modeled the protonated species and estimated the deprotonation energies from both acid sites. The formed products are the nonionic (deprotonation from the amino site) and the zwitterion (formed by deprotonation from the phenol site). The zwitterionic forms were found to be 10.63, 11.08, and 20.57 kcal/mol higher in Gibbs free energy than the nonionic form for ketobemidone, profadol, and tapentadol, respectively, in aqueous solution. Therefore the zwitterionic forms of these compounds have been ruled out. In addition, for acetaminophen, we have estimated the relative ease of deprotonation from the $-NH$ and $-OH$ groups and confirmed that the Gibbs free energy of

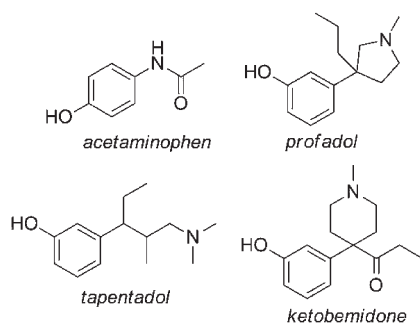


Figure 1. Studied nonsteroidal anti-inflammatory drugs.

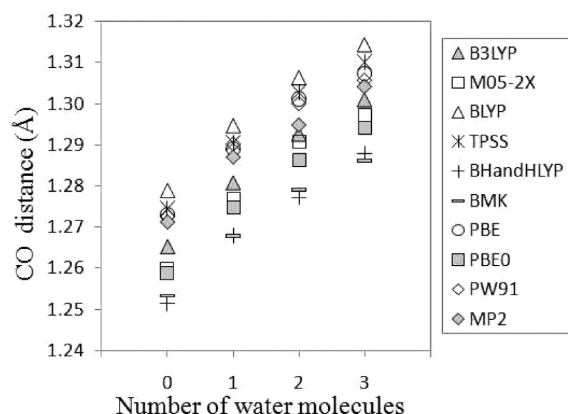


Figure 2. Variation of the CO distance in the phenoxide anions, with the number of explicit water molecules.

the anion formed by deprotonation of the $-\text{OH}$ is 10.97 kcal/mol lower than that of the anion formed by deprotonation of the $-\text{NH}$, in aqueous solution. After these analyses it can be concluded that the processes modeled in the present work correspond to the actual phenolic deprotonations involved in their observed acid/base equilibrium.

In addition, it seems worthwhile to call attention on a previous report, by Liptak et al.,¹⁰ describing that while the inclusion of the solvent has little effect on the phenols' geometries, it has a significant effect on the phenoxide anions. Since the geometrical parameter that changes the most is the CO distance, we have used it to analyze this point. In ref 10, dCO was found to be 1.251 Å in gas phase and 1.298 Å in solution, modeled using the continuum approach CPCM. In both cases the geometry optimizations were performed at HF/6-31+G(d) level of theory. In our case the solvent was modeled by including up to three explicit water molecules in the vicinity of the CO group. The dCO values obtained for each studied phenol, at every level of calculation, are reported in Table 1S, Supporting Information. The average value of dCO, including all methods of calculation and all the studied phenols, is found to be 1.266, 1.282, 1.293, and 1.301 Å for the anions with 0–3 explicit water molecules, respectively. The dCO average values, per method, have been plotted in Figure 2. The geometry optimizations performed with the BLYP functional systematically lead to the longest CO distances, while the shortest ones arise from optimizations with BHandHLYP. To our best knowledge there is no experimental data reported in the CO distance of phenoxides, but taking the value reported by Liptak

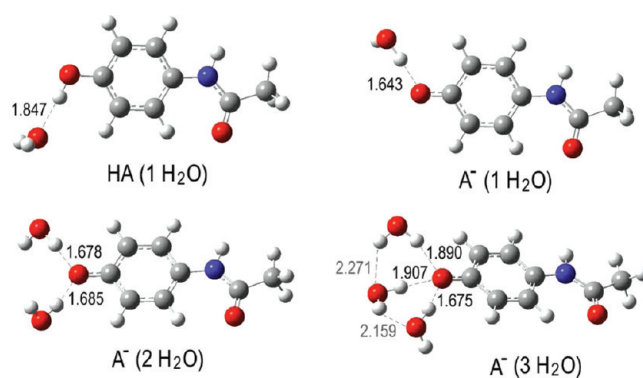


Figure 3. Geometries of the solute–water clusters, for the acetaminophen system, optimized at TPSS/6-311++G(d,p) level of theory.

et al.¹⁰ as reference, it seems that the inclusion of one or two water molecules in the vicinity of the anions is enough to obtain good geometrical descriptions of these species.

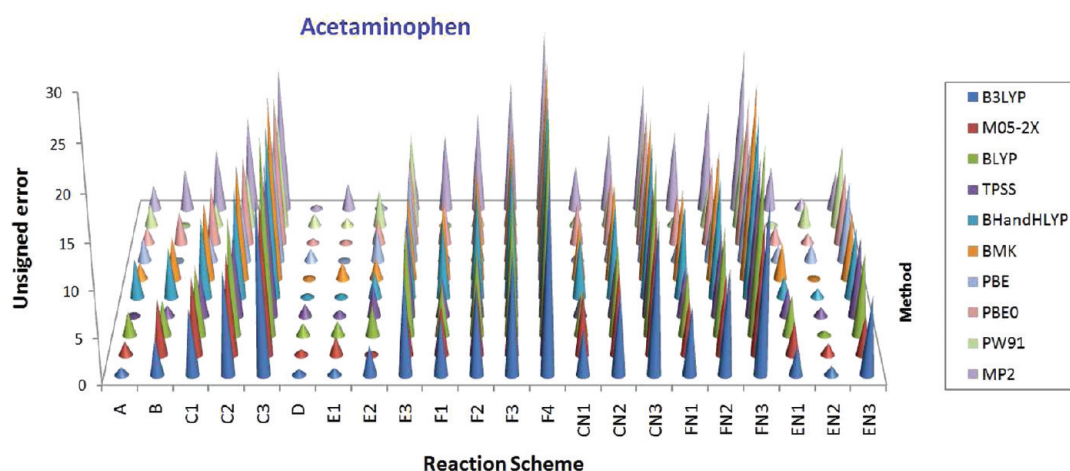
Due to the large amount of calculations involved in the present work, it was unfeasible to perform exhaustive conformational analyses for the solute–water clusters. We have used chemical intuition and previous experience instead to construct the starting geometries in each case. To provide information on the optimized structures of the modeled clusters, the geometries corresponding to the acetaminophen system are shown in Figure 3. Since only small variations were found from changing the calculation method, we have chosen only one of them to show the geometrical distribution of the clusters. Equivalent configurations were located for all the studied phenolic systems.

pK_a Estimations. The pK_a values calculated for the studied nonsteroidal anti-inflammatory drugs (Figure 1) using the reaction schemes from Table 1 and the different levels of theory are provided as Supporting Information (Tables 2S–5S). The signed errors (SE) arising from comparison with the experimental values are reported in Tables 2–5. The SE values smaller than ± 2 units of pK_a have been highlighted in bold letters, since this is the limit of accuracy currently accepted for calculated pK_as.²⁷ To facilitate rapid comparisons among all the reported data, the unsigned errors (UE) have been plotted in Figures 4–7.

Acetaminophen. For the phenolic deprotonation of acetaminophen, it was found that the outcomes from most of the tested reaction schemes lead to overestimated pK_a values (Table 2). The exceptions are schemes D, E1, and EN1, which produce pK_a values lower than the experimental one. In general the quality of the pK_a predictions was found to be strongly dependent on the chosen scheme, while it is only slightly influenced by the method of calculation (Figure 4). The errors arising from MP2 calculations are larger than those obtained within the DFT framework for E1 and EN1 schemes, and smaller when the E2 scheme is used. For all the other reaction schemes, the MP2 deviation from the experimental value was found to be intermediate, with respect to those arising from different DFT functionals. Therefore there is no use to additional computational cost arising from using MP2 instead of DFT. Reaction schemes C, CN, F, and FN lead to very large errors, regardless of the method of calculation and of the number of explicit water molecules included in the modeling. Curiously increasing the number of water molecules when using these reaction schemes does not improve but worsens the results for acetaminophen. In general the reaction schemes leading to the smallest errors are A,

Table 2. Signed Errors for the Calculated Values of the pK_a of Acetaminophen (vs $pK_a = 9.5$)⁵⁰

scheme	B3LYP	M05-2X	BLYP	TPSS	BHandHLYP	BMK	PBE	PBE0	PW91	MP2
A	0.88	1.47	-2.61	-0.11	4.28	1.80	-2.55	1.84	-2.49	2.40
B	3.87	6.06	3.83	1.20	5.58	4.78	-0.01	3.64	0.15	4.24
C1	7.24	8.67	7.77	4.75	8.74	8.52	2.85	6.39	2.66	6.92
C2	11.32	11.08	12.69	8.80	12.09	12.71	6.66	9.58	5.95	10.53
C3	18.90	16.63	21.24	16.49	18.75	20.01	13.59	16.50	12.51	16.10
D	-0.27	-0.24	-0.97	-0.97	-0.01	0.10	-1.03	-0.15	-1.04	0.80
E1	-0.80	-1.98	-1.59	-0.70	-0.42	-1.91	-0.11	-0.37	0.87	2.95
E2	3.28	0.43	3.33	3.35	2.93	2.28	3.70	2.82	4.17	0.66
E3	10.86	5.98	11.88	11.04	9.59	9.58	10.63	9.74	10.72	6.24
F1	6.27	5.40	5.83	4.17	8.77	8.43	1.31	4.81	0.10	8.26
F2	9.63	8.00	9.76	7.72	11.92	12.17	4.16	7.56	2.60	10.93
F3	13.72	10.41	14.69	11.77	15.28	16.37	7.98	10.76	5.90	14.54
F4	21.30	15.96	23.24	19.46	21.94	23.67	14.91	17.68	12.45	20.12
CN1	4.97	6.96	4.94	2.38	6.81	6.36	1.12	4.72	1.26	5.13
CN2	9.05	9.37	9.87	6.44	10.17	10.56	4.94	7.92	4.56	8.74
CN3	16.62	14.92	18.41	14.11	16.82	17.85	11.86	14.83	11.11	14.31
FN1	7.36	6.29	6.94	5.35	10.00	10.02	2.44	5.90	1.21	9.14
FN2	11.44	8.70	11.86	9.40	13.35	14.21	6.25	9.09	4.50	12.75
FN3	19.02	14.25	20.41	17.08	20.02	21.51	13.18	16.01	11.06	18.33
EN1	-3.07	-3.69	-4.42	-3.07	-2.35	-4.07	-1.84	-2.04	-0.53	4.74
EN2	1.01	-1.27	0.50	0.98	1.01	0.13	1.98	1.15	2.77	1.12
EN3	8.58	4.27	9.04	8.66	7.67	7.42	8.90	8.07	9.32	4.45

Figure 4. Unsigned errors for the calculated values of the phenolic pK_a of acetaminophen.

D, E1, and EN2. With the exception of scheme A, these schemes are just those that conserve the number and kind of charged species at both sides of the equilibrium, and among them scheme A is just the one leading to largest errors.

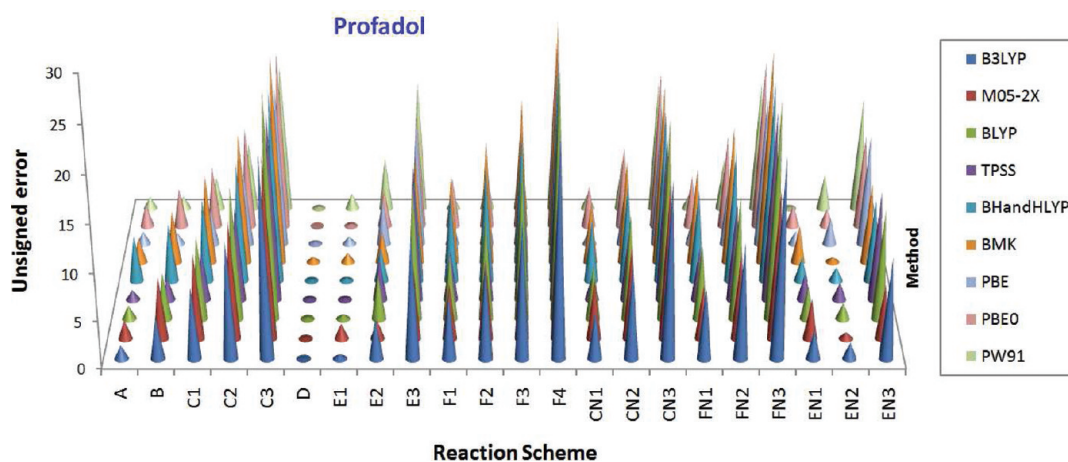
Analyzing in more detail the results reported in Table 2, some peculiarities become evident. For the reaction scheme A, functionals BLYP, BHandHLYP, PBE, and PW91 produce a SE larger than 2 units of pK_a . PW91 also yields a SE larger than 2 within scheme EN2. On the other hand the SE, obtained from M05-2X calculations in conjunction with scheme E2 and from TPSS using the scheme B, is significantly lower than this limit. PBE and PW91 functionals also produce very low errors when used in conjunction with the B scheme. In fact these are the only two

functional for which the UE values arising from using scheme B are smaller than those obtained from scheme A. UE values smaller than 2 were also obtained for acetaminophen from PBE calculations using reaction schemes F1, CN1, and EN1 and from PW91 calculations using react F1, CN1, EN1, and FN1.

Profadol. The general behavior of the calculated data for the phenolic deprotonation of profadol is very similar to that of acetaminophen. The reaction schemes leading to the smallest errors are A, D, E1, and EN2. Most of the tested reaction schemes yield overestimated values of pK_a values (Table 3), while schemes E1 and EN1 lead to negative values of UE. The reaction scheme choice was found to have a much stronger influence on the quality of the results than on the particular method of calculation

Table 3. Signed Errors for the Calculated Values of the pK_a of Profadol (vs $pK_a = 10.27$)⁵¹

scheme	B3LYP	M05-2X	BLYP	TPSS	BHandHLYP	BMK	PBE	PBE0	PW91
A	1.45	1.98	-1.48	1.04	4.86	2.67	-1.37	2.34	-1.29
B	4.45	6.57	4.97	2.35	6.15	5.65	1.18	4.14	1.35
C1	7.57	9.02	8.89	5.26	8.82	9.42	3.77	6.51	3.45
C2	12.34	12.58	14.12	9.96	12.80	13.96	7.90	10.76	7.34
C3	21.21	19.06	23.99	19.64	20.18	21.95	16.83	18.82	15.78
D	0.31	0.26	0.17	0.19	0.57	0.98	0.16	0.36	0.16
E1	-0.46	-1.62	-0.48	-0.19	-0.33	-1.01	0.82	-0.26	1.67
E2	4.30	1.93	4.76	4.51	3.64	3.53	4.94	3.99	5.55
E3	13.17	8.41	14.63	14.19	11.02	11.52	13.87	12.06	14.00
F1	6.84	5.90	6.96	5.32	9.34	9.31	2.49	5.32	1.30
F2	9.97	8.35	10.88	8.22	12.01	13.07	5.09	7.68	3.39
F3	14.74	11.91	16.12	12.93	15.99	17.62	9.22	11.93	7.28
F4	23.61	18.39	25.99	22.61	23.37	25.61	18.15	20.00	15.73
CN1	5.02	6.39	5.85	2.58	6.61	6.40	1.61	4.40	1.64
CN2	9.79	9.95	11.08	7.29	10.59	10.95	5.75	8.65	5.54
CN3	18.65	16.43	20.95	16.96	17.96	18.94	14.67	16.71	13.97
FN1	7.42	5.73	7.85	5.55	9.81	10.06	2.93	5.58	1.59
FN2	12.18	9.28	13.08	10.25	13.78	14.60	7.06	9.82	5.47
FN3	21.05	15.76	22.95	19.93	21.16	22.60	15.99	17.89	13.92
EN1	-3.02	-4.26	-3.52	-2.87	-2.54	-4.03	-1.35	-2.37	-0.14
EN2	1.75	-0.69	1.72	1.83	1.44	0.52	2.79	1.88	3.75
EN3	10.61	5.78	11.59	11.51	8.81	8.51	11.71	9.94	12.19

Figure 5. Unsigned errors for the calculated values of the phenolic pK_a of profadol.

(Figure 5). The largest deviations from the experimental value arise from using reaction schemes C, CN, F and FN, and the inclusion of more water molecules in these schemes increases the UE values.

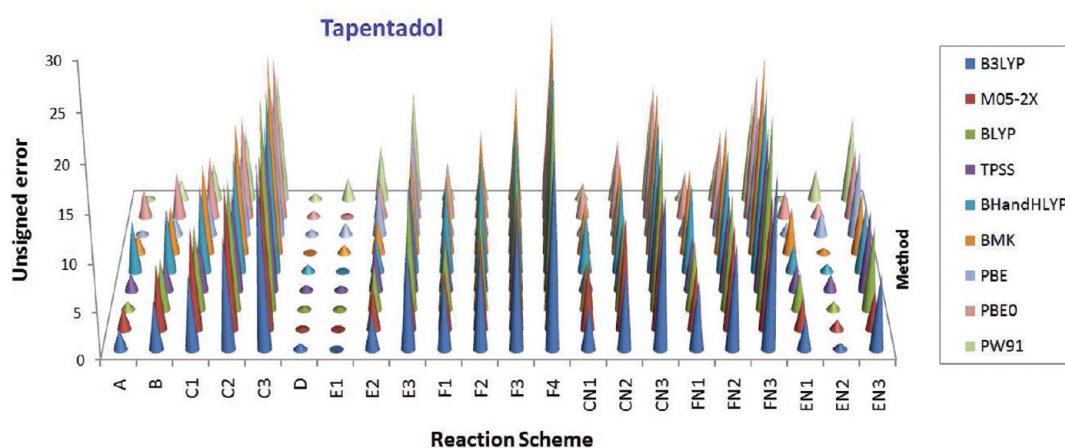
Some deviations from these general trends were found. DFT functionals BHandHLYP, BMK, and PBE0 produce a SE larger than 2 units of pK_a when used in conjunction with scheme A. This also happens when computing scheme EN2 with the PBE functional. As it was the case for acetaminophen, M05-2X calculations using the scheme E2 lead to an UE < 2. However an UE value is significantly larger for profadol than for acetaminophen (1.92 vs 0.43). PBE and PW91 functionals produce UE < 2 when used in conjunction with the B scheme. However, in this case, their values are similar in magnitude but opposite in sign

than those obtained from scheme A. UE values smaller than 2 were also obtained for profadol from PBE calculations using reaction schemes CN1 and EN1 and from PW91 calculations using reactions F1, CN1, FN1, and EN1.

Tapentadol. For this compound the general trends are similar to those of acetaminophen and profadol. It was found that most of the calculated values of pK_a are overestimated (Table 4). The exceptions arise mainly from using schemes E1 and EN1, but they are fewer than those found for acetaminophen and profadol. The quality of the pK_a predictions was found to be more influenced by the chosen scheme than by the method of calculation (Figure 6). Reaction schemes C, CN, F, and FN lead to very large errors, regardless of the method of calculation and the number of explicit water molecules included in the modeling.

Table 4. Signed Errors for the Calculated Values of the pK_a of Tapentadol (vs $pK_a = 10.09$)⁵¹

scheme	B3LYP	M05-2X	BLYP	TPSS	BHandHLYP	BMK	PBE	PBE0	PW91
A	2.15	2.36	-0.88	1.83	5.45	2.11	-0.52	3.04	-0.41
B	5.15	6.95	5.56	3.14	6.74	5.09	2.03	4.84	2.23
C1	8.03	10.72	9.05	5.97	9.09	9.53	4.29	6.56	4.14
C2	11.98	15.09	14.00	10.46	12.84	13.98	8.39	11.05	7.78
C3	19.28	18.19	22.03	17.35	19.11	21.12	14.60	17.10	13.60
D	1.01	0.65	0.77	0.97	1.16	0.42	1.01	1.06	1.04
E1	-0.01	0.07	-0.32	0.52	-0.07	-0.90	1.33	-0.21	2.35
E2	3.94	4.44	4.64	5.00	3.68	3.55	5.43	4.29	5.99
E3	11.24	7.55	12.67	11.90	9.95	10.69	11.65	10.33	11.81
F1	7.54	6.29	7.56	6.11	9.93	8.75	3.35	6.01	2.17
F2	10.42	10.05	11.04	8.93	12.28	13.19	5.60	7.73	4.08
F3	14.38	14.42	16.00	13.43	16.03	17.64	9.70	12.23	7.73
F4	21.67	17.53	24.03	20.32	22.30	24.77	15.92	18.27	13.55
CN1	4.76	7.33	5.31	2.69	6.24	5.46	1.59	3.84	1.42
CN2	8.72	11.71	10.27	7.18	10.00	9.92	5.69	8.34	5.07
CN3	16.01	14.81	18.30	14.07	16.26	17.04	11.90	14.38	10.88
FN1	7.16	6.66	7.31	5.66	9.44	9.12	2.91	5.02	1.37
FN2	11.11	11.03	12.27	10.14	13.19	13.57	7.00	9.51	5.01
FN3	18.41	14.14	20.30	17.04	19.46	20.70	13.22	15.55	10.83
EN1	-3.28	-3.32	-4.05	-2.76	-2.91	-4.97	-1.37	-2.93	-0.36
EN2	0.68	1.06	0.91	1.73	0.85	-0.51	2.73	1.57	3.28
EN3	7.97	4.16	8.93	8.62	7.11	6.61	8.95	7.61	9.10

Figure 6. Unsigned errors for the calculated values of the phenolic pK_a of tapentadol.

Moreover the inclusion of more water molecules in these schemes increases the UE values. The reaction schemes leading to the smallest errors were found to be **A**, **D**, **E1**, and **EN2**.

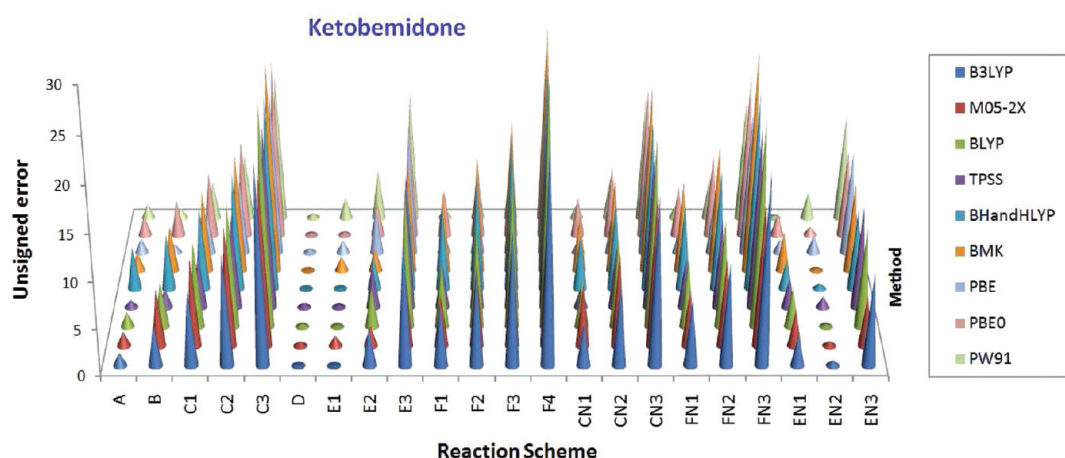
In contrast to these general trends, some peculiarities were found. DFT functionals BHandHLYP, BMK, and PBE0 produce a SE larger than 2 units of pK_a when used in conjunction with scheme **A**, as it was the case for profadol. Computing scheme **EN2** with the PW91 functional also lead $UE > 2$. PBE and PW91 functionals produce $UE < 2$, when used in conjunction with the **B** scheme. However, in this case their values are similar in magnitude but opposite in sign than those obtained from scheme **A**, also in line with the results for profadol. UE values smaller than 2 were also obtained for profadol from PBE calculations using reaction schemes **CN1** and **EN1** and from PW91 calculations

using reactions **F1**, **CN1**, and **EN1**. The outcomes from PBE and PW91 calculations are similar to what was described for acetaminophen and profadol.

Ketobemidone. In this case the general behavior of the calculated data is congruent with what was described for the other compounds in the tested set. The reaction schemes leading to the smallest errors are **A**, **D**, **E1**, and **EN2**. Most of the tested reaction schemes yield overestimated values of pK_a values (Table 5), while schemes **D**, **E1**, and **EN1** lead to negative values of UE. The reaction scheme choice was found to have a much stronger influence on the quality of the results than the particular method of calculation (Figure 7). The largest deviations from the experimental value arise from using reaction schemes **C**, **CN**, **F** and **FN**, and the inclusion of more water molecules in these schemes increases the UE values.

Table 5. Signed Errors for the Calculated Values of the pK_a of Ketobemidone (vs $pK_a = 9.96$)⁵¹

scheme	B3LYP	M05-2X	BLYP	TPSS	BH&HLYP	BMK	PBE	PBE0	PW91
A	1.38	1.59	-1.60	0.77	4.65	2.00	-1.60	2.13	-1.57
B	4.37	6.17	4.84	2.08	5.94	4.98	0.95	3.93	1.07
C1	8.12	9.48	9.19	5.76	9.04	8.86	4.34	7.10	4.14
C2	12.13	12.86	13.62	10.05	12.66	12.93	7.82	10.55	7.25
C3	21.62	18.75	24.09	19.85	20.86	22.87	17.10	19.13	15.88
D	0.23	-0.13	0.05	-0.08	0.36	0.31	-0.07	0.15	-0.12
E1	0.08	-1.16	-0.18	0.31	-0.12	-1.57	1.38	0.34	2.35
E2	4.09	2.22	4.25	4.59	3.51	2.50	4.87	3.78	5.46
E3	13.58	8.11	14.73	14.40	11.71	12.44	14.14	12.36	14.09
F1	6.77	5.51	6.84	5.05	9.14	8.64	2.27	5.11	1.01
F2	10.51	8.81	11.18	8.73	12.23	12.51	5.65	8.27	4.08
F3	14.52	12.20	15.62	13.02	15.86	16.58	9.14	11.73	7.19
F4	24.02	18.09	26.09	22.82	24.06	26.53	18.42	20.30	15.82
CN1	4.62	6.74	5.19	2.34	6.10	6.15	1.38	4.29	1.53
CN2	8.63	10.12	9.63	6.63	9.73	10.22	4.87	7.75	4.65
CN3	18.12	16.01	20.10	16.43	17.92	20.16	14.13	16.32	13.27
FN1	7.01	6.07	7.19	5.31	9.30	9.80	2.70	5.47	1.48
FN2	11.02	9.45	11.63	9.60	12.92	13.87	6.18	8.92	4.59
FN3	20.52	15.34	22.10	19.40	21.12	23.81	15.45	17.49	13.22
EN1	-3.42	-3.91	-4.17	-3.11	-3.05	-4.29	-1.58	-2.47	-0.25
EN2	0.59	-0.52	0.27	1.18	0.58	-0.21	1.91	0.98	2.86
EN3	10.08	5.36	10.74	10.98	8.77	9.73	11.18	9.55	11.48

Figure 7. Unsigned errors for the calculated values of the phenolic pK_a of ketobemidone.

The deviations from these general trends that were found for ketobemidone are described next. It is interesting to notice that even deviations are similar, though not identical, for all the tested cases. DFT functionals BHandHLYP, BMK, and PBE0 produce a SE larger than 2 units of pK_a , when used in conjunction with scheme A. This also happens when computing scheme EN2 with the PW91 functional. PBE and PW91 functionals produce UE < 2 when used in conjunction with the B scheme. The UE values are similar in magnitude but opposite in sign than those obtained from scheme A. UE values smaller than 2 were also obtained for ketobemidone from PBE calculations using reaction schemes CN1 and EN1 and from PW91 calculations using reaction F1, CN1, FN1, and EN1.

Other Phenols. To further analyze the performance of the tested protocols, they have been applied to the pK_a calculations of other phenols. They are: 3-methoxyphenol (*m*-OCH₃), 4-methylphenol (*p*-CH₃), 3-cyanophenol (*m*-CN), 4-methylthiophenol (*p*-SCH₃), 3-hydroxybenzaldehyde (*m*-CHO), and 2,4-dimethylphenol (*op*-2CH₃). The calculated pK_a s are provided as Supporting Information (Tables 6S–11S).

The average UEs, including these six compounds, are shown in Figure 8. The general behavior of the calculated pK_a s is in agreement with what was found for the nonsteroidal anti-inflammatory drugs. The reaction schemes leading to the smallest errors are A, D, E1, and EN2, which supports the good performance of these schemes. The choice of the method of calculation was found

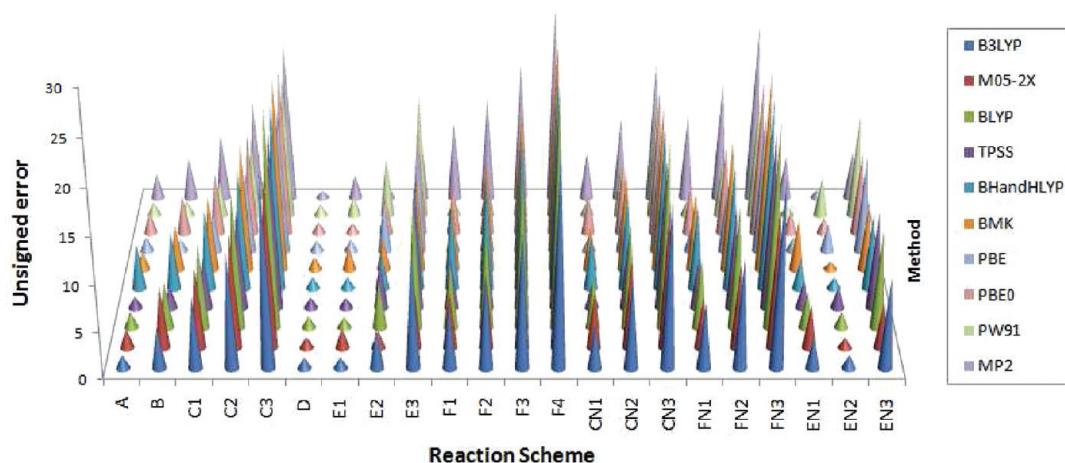


Figure 8. Unsigned errors for the calculated pK_a values of phenolics (average, including *m*-OCH₃, *p*-CH₃, *m*-CN, *p*-SCH₃, *m*-CHO, and *op*-2CH₃).

Table 6. Mean Unsigned Errors for the Calculated Values of Phenolic pK_a s within the Different Approaches

scheme	B3LYP	M05-2X	BLYP	TPSS	BHandHLYP	BMK	PBE	PBE0	PW91	MP2
A	1.46	1.97	1.69	1.11	4.88	2.03	1.58	2.35	1.47	2.67
B	4.45	6.55	4.89	2.33	6.17	5.01	1.24	4.14	1.43	4.51
C1	7.67	9.00	8.88	5.50	8.75	8.52	4.01	6.67	3.81	7.18
C2	12.18	12.55	14.19	10.34	12.88	13.65	8.22	10.71	7.78	10.91
C3	20.48	18.18	23.16	18.69	19.90	21.22	15.90	18.12	14.90	16.96
D	0.88	0.82	0.94	0.90	0.96	1.00	0.91	0.90	0.94	0.80
E1	0.86	1.77	0.95	0.82	0.82	2.02	1.15	0.77	2.02	2.69
E2	4.14	2.01	4.83	4.89	3.72	3.21	5.26	3.94	5.99	1.04
E3	12.44	7.54	13.79	13.24	10.74	10.79	12.94	11.35	13.11	7.09
F1	6.85	5.89	6.89	5.30	9.36	8.67	2.55	5.32	1.38	8.53
F2	10.06	8.33	10.87	8.46	11.94	12.17	5.32	7.84	3.75	11.19
F3	14.58	11.88	16.19	13.31	16.07	17.30	9.54	11.88	7.72	14.92
F4	22.88	17.52	25.16	21.66	23.09	24.88	17.22	19.30	14.84	20.98
CN1	4.73	6.35	5.42	2.55	6.12	5.40	1.57	4.29	1.67	5.23
CN2	9.25	9.90	10.74	7.40	10.25	10.52	5.73	8.34	5.61	8.96
CN3	17.54	15.52	19.70	15.73	17.26	18.09	13.40	15.75	12.72	15.00
FN1	7.13	5.69	7.42	5.52	9.31	9.05	2.83	5.47	1.63	9.24
FN2	11.64	9.23	12.73	10.36	13.43	14.18	7.04	9.51	5.55	12.97
FN3	19.94	14.87	21.70	18.70	20.46	21.75	14.72	16.92	12.66	19.02
EN1	3.31	4.29	3.94	2.91	3.04	5.04	1.53	2.47	0.80	4.64
EN2	1.27	1.01	1.40	1.94	1.12	0.65	2.77	1.59	3.82	0.91
EN3	9.51	4.88	10.33	10.28	8.11	7.66	10.44	8.98	10.93	5.14

to have less influence on the quality of the result than the reaction scheme, also for these six phenols.

Generalizations. According to all the gathered data, several generalizations and recommendations can be made regarding the accuracy of phenolic pK_a s obtained from calculations. Table 6 shows the mean unsigned errors (MUE) for each scheme–method pair, including the 10 tested systems. The MUE values <2 units of pK_a have been highlighted in bold letters. This limit has been chosen based on the accuracy currently accepted for calculated pK_a s.²⁷

It should be kept in mind that these generalizations and recommendations are valid within the cluster-continuum model, using the PCM model and the strategy based on thermodynamic

cycles. Changes in any of these aspects to the used methodology can lead to different outcomes.

The quality of the pK_a predictions was found to be strongly dependent on the chosen scheme while only moderately influenced by the method of calculation. The reaction schemes leading to the smallest errors are A, D, E1, and EN2, while those producing the largest errors are C, CN, F, and FN. The MUE obtained when using reaction schemes A, D, E1, and EN2 is within or below the currently accepted accuracy. In particular, schemes D and E1 produce MUE values significantly lower than 2 units of pK_a with most of the used methods of calculation (Figure 9). Therefore we recommend these approaches to predict reliable pK_a values, for phenolic deprotonations of relative large systems, within the DFT

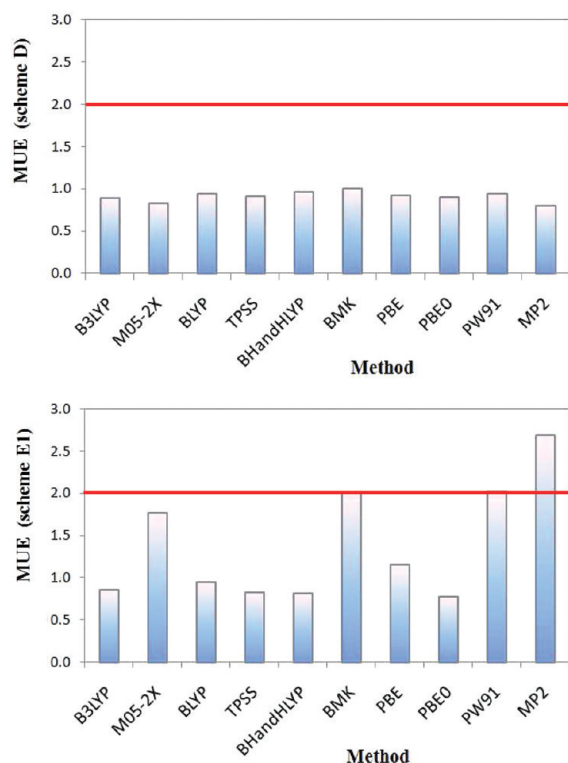


Figure 9. Mean unsigned errors for reaction schemes D and E1, including the 10 tested systems.

framework and using relatively modest basis sets, i.e., at very reasonable computational costs. We particularly recommend the E1 reaction scheme, since it has the additional advantage of being experiment independent.

According to the above discussion, it can be stated that the schemes conserving the number and kind of charged species at both sides of the equilibrium lead to more accurate values of pK_a , since these approaches promote the cancelation of errors. So this seems to be the main problem to address for succeeding in the challenging task of predicting reliable acid constants. Among the reaction schemes leading to the smallest errors (A, D, E1, and EN2), all but scheme A conserve the number and the kind of charges at both sides of the equilibrium, and among them scheme A is just the one leading to largest errors. The reaction schemes producing the largest MUEs (C, CN, F, and FN) have all the charged species at the products side.

Even though the reaction scheme A is among those leading to the smallest errors, it is not recommended to be used in conjunction with BHandHLYP, MP2, or PBE0 methods (Table 6). On the contrary, reaction schemes B, CN1, and EN1 seem to be good approaches for predicting phenolic pK_a s, provided that the calculations are performed with PBE or PW91 functionals. In particular the combination EN1–PW91 produced a very small value of MUE.

CONCLUSIONS

Twenty-two reaction schemes have been tested, within the cluster-continuum model and including up to seven explicit water molecules, using the PCM and the strategy based on thermodynamic cycles [Gibbs free energy of reaction in solution ΔG_s is

obtained as the sum of the Gibbs free energy of reaction in vacuum (ΔG_g) and the solvation free energies ($\Delta\Delta G_{\text{solv}}$)].

The quality of the pK_a predictions was found to be strongly dependent on the chosen scheme, while is only moderately influenced by the method of calculation. The reaction schemes leading to the smallest errors are A, D, E1, and EN2, while those producing the largest errors are C, CN, F, and FN.

In particular, schemes D and E1 produce MUE values significantly lower than the currently accepted accuracy for pK_a calculations. Therefore we recommend these approaches to predict reliable pK_a values for phenolic deprotonations of relative large systems, within the DFT framework and using relatively modest basis sets, i.e., at very reasonable computational costs. We particularly recommend the E1 reaction scheme, since it has the additional advantage of being experiment independent.

The best pK_a values obtained from the E1 reaction scheme (MUE < 1 units of pK_a) are those involving calculations with PBE0 (MUE = 0.77), TPSS (MUE = 0.82), BHandHLYP (MUE = 0.82), and B3LYP (MUE = 0.86) functionals. The best results when using the D reaction schemes correspond to calculations performed with MP2 (MUE = 0.80), M05-2X (MUE = 0.82), B3LYP (MUE = 0.88), PBE0 and TPSS (MUE = 0.90), and PBE (MUE = 0.91).

It is important to notice that these conclusions are valid within the cluster-continuum model, using the PCM model in conjunction with the UAHF cavity, and the strategy based on thermodynamic cycles. Changes in any of these aspects to the used methodology can lead to different results.

ASSOCIATED CONTENT

S Supporting Information. CO distances in the phenoxide anions, pK_a values calculated using all the tested reaction schemes and levels of theory. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: agalano@xanum.uam.mx.

ACKNOWLEDGMENT

The authors thank Laboratorio de Visualización y Cómputo Paralelo at UAM - Iztapalapa for the access to its computer facilities. T.C.-H. thanks CONACyT for Master fellowship and M.T.R.-S. for partial financial support through project 82932. We also acknowledge PROMEP for partial financial support through SIATA network and for the stipend to A.M.R.-Z.

REFERENCES

- (1) Riccardi, D.; Schaefer, P.; Cui, Q. *J. Phys. Chem. B* **2005**, *109*, 17715.
- (2) Brown, T. N.; Mora-Diez, N. *J. Phys. Chem. B* **2006**, *110*, 9270.
- (3) Thomas, G. *Medicinal Chemistry: An Introduction*; John Wiley & Sons: West Sussex, U.K., 2000.
- (4) Albert, A.; Serjeant, E. P. *The Determination of Ionization Constants*; Chapman and Hall: New York, 1984.
- (5) Pliego, J. R., Jr.; Riveros, J. M. *J. Phys. Chem. A* **2002**, *106*, 7434.
- (6) Schuurmann, G. S.; Cossi, M.; Barone, V.; Tomasi, J. *J. Phys. Chem. A* **1998**, *102*, 6706.
- (7) Silva, C. O.; Silva, M. A.; Nascimento, M. A. *C. J. Phys. Chem. A* **1999**, *103*, 11194.

- (8) Toth, A. M.; Liptak, M. D.; Phillips, D. L.; Shields, G. C. *J. Chem. Phys.* **2001**, *114*, 4595.
- (9) Liptak, M. D.; Shields, G. C. *J. Am. Chem. Soc.* **2001**, *123*, 7314.
- (10) Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. C. *J. Am. Chem. Soc.* **2002**, *124*, 6421.
- (11) Chipman, D. M. *J. Phys. Chem. A* **2002**, *106*, 7413.
- (12) Adam, K. R. *J. Phys. Chem. A* **2002**, *106*, 11963.
- (13) Saracino, G. A. A.; Improta, R.; Barone, V. *Chem. Phys. Lett.* **2003**, *373*, 411.
- (14) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. *J. Phys. Chem. A* **2003**, *107*, 9380.
- (15) Almerindo, G. I.; Tondo, D. W.; Pliego, J. R., Jr. *J. Phys. Chem. A* **2004**, *108*, 166.
- (16) Magill, A. M.; Cavell, K. J.; Yates, B. F. *J. Am. Chem. Soc.* **2004**, *126*, 8717.
- (17) Vianello, R.; Maksic, Z. B. *Eur. J. Org. Chem.* **2005**, *16*, 3571.
- (18) Namazian, M.; Zakery, M.; Noorbala, M. R.; Coote, M. L. *Chem. Phys. Lett.* **2008**, *451*, 163.
- (19) Ho, J.; Coote, M. L. *J. Chem. Theory Comput.* **2009**, *5*, 295.
- (20) Song, Y.; Mao, J.; Gunner, M. R. *J. Comput. Chem.* **2009**, *30*, 2231.
- (21) Delgado, E. J. *Chem. Phys. Lett.* **2009**, *471*, 133.
- (22) Casanovas, R.; Frau, J.; Ortega-Castro, J.; Salvà, A.; Donoso, J.; Muñoz, F. *J. Mol. Struct. (THEOCHEM)* **2009**, *912*, 5.
- (23) Dissanayake, D. P.; Senthilnithy, R. *J. Mol. Struct. (THEOCHEM)* **2009**, *910*, 93.
- (24) Uudsemaa, M.; Kanger, T.; Lopp, M.; Tamm, T. *Chem. Phys. Lett.* **2010**, *485*, 83.
- (25) Rayne, S.; Forest, K. *J. Mol. Struct. (THEOCHEM)* **2010**, *949*, 60.
- (26) Casanovas, R.; Fernandez, D.; Ortega-Castro, J.; Frau, J.; Donoso, J.; Muñoz, F. *Theor. Chem. Acc.* **2011**, in press DOI 10.1007/s00214-011-0945-5.
- (27) Ho, J.; Coote, M. L. *Theor. Chem. Acc.* **2010**, *125*, 3.
- (28) (a) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764. (b) Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **1999**, *110*, 7650. (c) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703. (d) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *127*, 124105. (e) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *126*, 84108.
- (29) (a) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A. *J. Chem. Phys.* **1996**, *104*, 2598. (b) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822. (c) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **2000**, *112*, 6532.
- (30) Claverie, P.; Daudey, J. P.; Langlet, J.; Pullman, B.; Piazzola, D.; Huron, M. J. *J. Phys. Chem.* **1978**, *82*, 405.
- (31) Freitas, L. C. G.; Longo, R. L.; Simas, A. M. *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 189.
- (32) Pliego, J. R., Jr.; Riveros, J. M. *J. Phys. Chem. A* **2001**, *105*, 7241.
- (33) (a) Rice-Evans, C. A.; Miller, N. J.; Bolwell, P. G.; Bramley, P. M.; Pridham, J. B. *Free Radical Res.* **1995**, *22*, 375. (b) Cook, N. C.; Samman, S. *J. Nutr. Biochem.* **1996**, *7*, 66. (c) Rice-Evans, C.; Miller, N. J.; Paganga, G. *Free Radical Biol. Med.* **1996**, *20*, 933. (d) Cao, G.; Sofic, E.; Prior, R. L. *Free Radical Biol. Med.* **1997**, *22*, 749. (e) Kahkonen, M. P.; Hopia, A. I.; Vuorela, H. J.; Rauha, J. P.; Pihlaja, K.; Kujala, T. S.; Heinonen, M. *J. Agric. Food Chem.* **1999**, *47*, 3954. (f) Lemanska, K.; Szymusiak, H.; Tyrakowska, B.; Zielinski, R.; Soffers, A. E. M.; Rietens, I. M. C. M. *Free Radical Biol. Med.* **2001**, *31*, 869. (g) Materska, M.; Perucka, I. *J. Agric. Food Chem.* **2005**, *53*, 1750. (h) Fernandez-Pancho, M. S.; Villano, D.; Troncoso, A. M.; Garcia-Parrilla, M. C. *Crit. Rev. Food Sci.* **2008**, *48*, 649.
- (34) (a) Renaud, S.; de Lorgeril, M. *Lancet* **1992**, *339*, 1523. (b) Morton, L. W.; Caccetta, R. A. A.; Puddey, I. B.; Croft, K. D. *Clin. Exp. Pharmacol. Physiol.* **2000**, *27*, 152. (c) Corder, R.; Mullen, W.; Khan, K. Q.; Marks, S. C.; Wood, E. G.; Carrier, M. J.; Crozier, A. *Nature* **2006**, *30*, 444. (d) Manach, C.; Mazur, A.; Scalbert, A. *Curr. Opin. Lipidol.* **2005**, *16*, 77.
- (35) (a) Saito, M.; Hosoyama, H.; Ariga, T. *J. Agric. Food Chem.* **1998**, *46*, 1460. (b) Jiang, F.; Dusting, G. J. *Curr. Vasc. Pharmacol.* **2003**, *1*, 135.
- (36) Sokmen, M.; Serkedjieva, J.; Daferera, D.; Gulluce, M.; Polissiou, M.; Tepe, B.; Akpulat, H. A.; Sahin, F.; Sokmen, A. *J. Agric. Food Chem.* **2004**, *52*, 3309.
- (37) (a) Stavric, B. *Food Chem. Toxicol.* **1994**, *32*, 79. (b) Brown, D. M.; Kelly, G. E.; Husband, A. J. *Mol. Biotechnol.* **2005**, *30*, 253.
- (38) (a) Ramassamy, C. *Eur. J. Pharmacol.* **2006**, *545*, 51. (b) Hamaguchi, T.; Ono, K.; Murase, A.; Yamada, M. *Am. J. Pathol.* **2009**, *175*, 2557.
- (39) Hsu, S. J. *Am. Acad. Dermatol.* **2005**, *52*, 1049.
- (40) Cornwell, T.; Cohick, W.; Raskin, I. *Phytochemistry* **2004**, *65*, 995.
- (41) Hsu, C. L.; Yen, G. C. *Mol. Nutr. Food Res.* **2008**, *52*, 53.
- (42) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision E.01; Gaussian, Inc.: Wallingford, CT, 2004.
- (43) (a) Cancès, M. T.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032. (b) Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151. (c) Mennucci, B.; Cancès, E.; Tomasi, J. *J. Phys. Chem. B* **1997**, *101*, 10506. (d) Tomasi, J.; Mennucci, B.; Cancès, E. *J. Mol. Struct. (THEOCHEM)* **1999**, *464*, 211.
- (44) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.
- (45) Frisch, A. E.; Frisch, M.; Trucks, G. W. *Gaussian 03 User's Reference*; Gaussian, Inc.: Wallingford, CT, 2003.
- (46) Camaioni, D. M.; Schwerdtfeger, C. A. *J. Phys. Chem. A* **2005**, *109*, 10795.
- (47) Torres-Lapasio, J. R.; Garcia-Alvarez-Coque, M. C.; Bosch, E.; Roses, M. *J. Chromatogr. A* **2005**, *1089*, 170.
- (48) Pliego, J. R. *Chem. Phys. Lett.* **2003**, *367*, 145.
- (49) Bryantsev, V. S.; Diallo, M. S.; Goddard, W. A., III. *J. Phys. Chem. B* **2008**, *112*, 9709.
- (50) Hasemann, P.; Balk, M.; Preu, L.; Wätzig, H. *Electrophoresis* **2007**, *28*, 1779.
- (51) Calculated using Advanced Chemistry Development (ACD/Laboratories) Software V11.02 (1994–2011 ACD/Laboratories).

A Solvent-Free Coarse Grain Model for Crystalline and Amorphous Cellulose Fibrils

Goundla Srinivas, Xiaolin Cheng, and Jeremy C. Smith*

UT/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, Tennessee 37831, United States

S Supporting Information

ABSTRACT: Understanding biomass structure and dynamics on a range of time and length scales is important for the development of cellulosic biofuels. Here, to enable length and time scale extension, we develop a coarse grain (CG) model for molecular dynamics (MD) simulations of cellulose. For this purpose, we use distribution functions from fully atomistic MD simulations as target observables. A single bead per monomer level coarse graining is found to be sufficient to successfully reproduce structural features of crystalline cellulose. Without the use of constraints the CG crystalline fibril is found to remain stable over the maximum simulation length explored in this study ($>1 \mu\text{s}$). We also extend the CG representation to model fully amorphous cellulose fibrils. This is done by using an atomistic MD simulation of fully solvated individual cellulose chains as a target for developing the corresponding fully amorphous CG force field. Fibril structures with different degrees of crystallinity are obtained using force fields derived using a parameter coupling the crystalline and amorphous potentials. The method provides an accurate and constraint-free approach to derive CG models for cellulose with a wide range of crystallinity, suitable for incorporation into large-scale models of lignocellulosic biomass.

I. INTRODUCTION

Polysaccharides are the most abundant form of biomaterial on Earth. A polysaccharide of particular importance is cellulose, which is widely present in plants as cell wall or extra-cellular material.¹ Cellulose plays a crucial role in plant growth and cell wall function.^{2,3}

Cellulose has recently gained increasing attention due to its potential application in generating biomass-based renewable energy.^{4,5} Various methods have been proposed for converting cellulose-based biomass to ethanol and other transportation fuels,⁶ typically involving three steps: (i) biomass pretreatment, (ii) conversion of extracted cellulose into sugars, typically through enzymatic hydrolysis, and (iii) fermentation of sugars to obtain ethanol. However, the natural recalcitrance of biomass to hydrolysis has led to the need to improve pretreatment and hydrolysis methods.

A characterization of cellulose structure and assembly at the molecular level is important for understanding the recalcitrance of biomass to hydrolysis. There have been substantial experimental efforts to understand cellulose structure. For example, a combination of X-ray diffraction analysis and computational modeling has revealed that native cellulose (cellulose I) occurs in two forms: $I\alpha$ and $I\beta$.⁷ The $I\alpha$ phase contains a single cellulose chain in a triclinic cell, while a two-chain monoclinic cell is found for $I\beta$. The relative ratio of the two forms depends on the origins of the cellulose. Recently, extensive synchrotron X-ray and neutron diffraction measurements on $I\alpha$ and $I\beta$ structures have provided detailed crystal and molecular structures together with the hydrogen-bonding network within each phase.^{8,9} Cellulose also occurs in amorphous phases.⁸

Although extensive experimental studies have been undertaken, simulation studies of cellulose microfibrils have been relatively rare. This is not surprising, as the complexity and the

size of cellulose fibril structures have rendered them difficult to study with traditional molecular simulation approaches. The smallest cellulose microfibrils consist of approximately 36 individual chains with a large degree of polymerization. These large system sizes, together with the associated long-time dynamics often needed to examine cell wall processes, put simulation studies of cell wall decomposition beyond the scope of a complete atomistic representation.^{10–12} Hence, recent simulation studies have focused on using alternative methodologies based on lower resolution models, such as coarse grain (CG) molecular dynamics (MD),^{13–15} that have been proven to be useful in exploring long time and length scale processes involving proteins,^{15,16} lipids,^{17–20} polymers,²¹ and other materials.²²

In one recent study, a three-site CG model for monosaccharides was developed based on chemical information on the underlying monomers,²³ and the model was employed to study water dynamics in the glassy state formed by sugar–water complexes. Subsequently, a similar but improved three-site CG model was proposed that successfully reproduces the atomistic behavior of a single chain cellulose molecule of 14 monomer units.²⁴ In later work the model of ref 24 was applied to study interactions between cellulose (40 monomer long) and carbohydrate-binding modules.²⁵ However, in ref 25, in order to preserve the fibril structure of the cellulose assembly, cylindrical reflecting boundary conditions needed to be applied. Most recently, another three-site CG model for native cellulose was introduced based on the principles of Martini force field.²⁶ However, this force field was unable to reproduce stable crystalline cellulose fibril structures (additional interaction terms and backbone restraints were needed

Received: March 15, 2011

Published: July 01, 2011

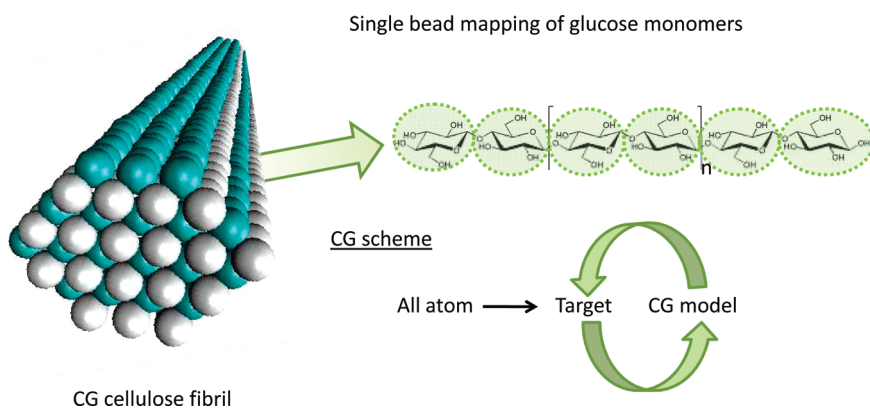


Figure 1. Single bead CG model of a cellulose chain is shown along with the underlying chemical structure. A CG cellulose fibril is also shown. Within the cellulose fibril, chains belong to origin (cyan) and center (white) sheets. A schematic representation of the CG methodology used in this work is also depicted (please see text for details).

to stabilize the crystalline fibril). Hence, to the best of our knowledge, there has hitherto been no CG force field able to model the crystalline cellulose fibril without constraints.

Here, we propose an unconstrained CG model of the cellulose fibril. The polysaccharide is modeled as a flexible chain consisting of single-site monomer units. In other words, each monosaccharide is replaced with a single CG unit, as compared to three sites in the previous studies.^{23–26} The CG model is developed using information obtained from a 20 ns long atomistic trajectory of a cellulose fibril in water.²⁷ In lignocellulosic biomass cellulose exists in varying degrees of crystallinity,^{5,6} and the crystallinity can change after pretreatment.²⁸ Hence, it is important to be able to generate models of cellulose in amorphous and fully crystalline forms as well as various states of intermediate crystallinity. The present CG model also addresses this need. We use a system of individual cellulose chains dissolved in water as a model for the fully amorphous/noncrystalline form. Further, we obtain several noncrystalline forms of cellulose fibril by introducing a coupling parameter between CG force fields of fully crystalline and fully amorphous cellulose systems.

The rest of the paper is organized as follows: In the next section we describe the CG methodology and its application to crystalline cellulose. The simulation details are presented in Section III. The results and discussion for the crystalline models are presented in Section IV along with parameterization and CG results for noncrystalline fibrils. We conclude the paper with closing remarks in Section V.

II. CG METHODOLOGY

A schematic representation of the present CG method is shown in Figure 1. For development of the CG model we make use of the chemical information from all-atom simulations of a representative cellulose system, the details of which are described elsewhere.²⁷ We summarize important aspects here. A cellulose fibril was constructed with 36 chains based on the $I\beta$ crystal structure⁸ with a degree of polymerization of 80 (containing 80 glucose monomers or 40 cellobiose units). The cellulose fibril was represented with the CHARMM force field²⁹ and solvated with more than 200 000 TIP3P water molecules³⁰ in a rectangular simulation box. Atomistic simulations of this system were performed for 20 ns using the NAMD simulation package.³¹ Five such independent 20 ns simulations using a 2 fs time step were performed for the analysis. The convergence of individual simulations

was monitored by calculating the total dipole moment of the cellulose fibril, which converged in less than 5 ns in all the simulations.²⁷ Since our parametrization is completely based on atomistic crystalline fibril data, the present CG potential is expected to be less transferrable but accurately represent crystalline fibril structure.

A. CG Model for Crystalline Cellulose. In the $I\beta$ fibril structure the cellulose chains are categorized as belonging to origin or center planes based on the location of the chain in the crystalline unit cell. Macroscopically, these two different chains constitute alternate layers or sheets of the fibril, as represented in Figure 1. In the CG model, the origin and center chains are distinguished and assigned separate parameter sets: the monomers in the origin chains are denoted as “OR”, while those in the center chains are denoted as “CE”. The need for separate parameter sets for crystalline and noncrystalline structures is examined in detail in the sections relevant to amorphous fibrils.

Bonded Parameters. A CG cellulose chain is constructed using the center of mass of each monomer derived from the corresponding atomistic representation. Hence, each glucose monomer is represented by a single CG bead, as shown in Figure 1. Clearly then, a single cellulose chain with a degree of polymerization 80 would be represented as an 80-bead chain.

The complete CG interaction potential is given by

$$V_{cg} = V_{bond} + V_{angle} + V_{torsion} + V_{nonbond} \quad (1)$$

In eq 1, any two consecutive monomers in a chain are connected by harmonic pseudobond potentials V_{bond} , and similarly, harmonic pseudobond angle potentials are employed for three consecutively connected beads. A torsional potential, $V_{torsion}$, is applied to any four consecutive CG beads. $V_{nonbond}$ represents the nonbonded interactions. Since we represent each glucose monomer as a single CG site, there is no net charge associated with any of the CG units. Hence, there are no explicit electrostatic interaction terms.

In the following we describe each of the terms in eq 1. The harmonic bond interactions are given by

$$V_{bond} = \sum_{bonds} k_b (r - r_0)^2 \quad (2)$$

where k_b and r_0 represent the force constant and equilibrium bond distance for the pseudobond connecting monomers; k_b and r_0 were adjusted until the CG-bond distance distributions reproduced the corresponding distributions in the atomistic MD to the

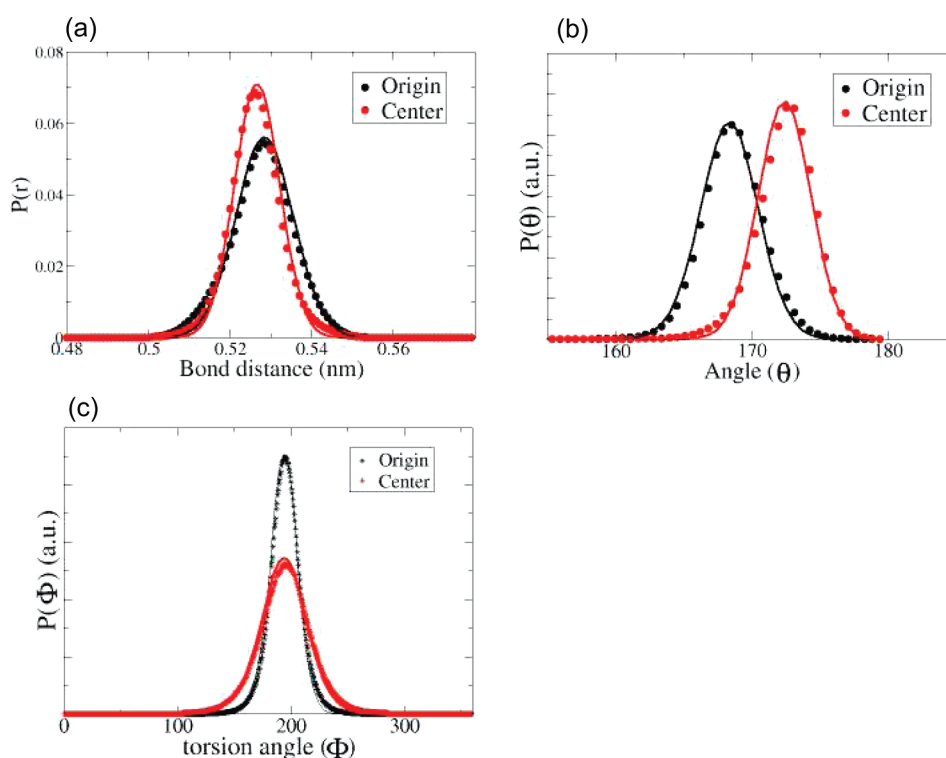


Figure 2. Comparison of distributions obtained from all-atom simulations (symbols) and CG simulations (lines): (a) Bond distance (OR–OR and CE–CE); (b) bond angle (OR–OR–OR and CE–CE–CE); and (c) torsion angle (OR–OR–OR–OR and CE–CE–CE–CE) distribution comparisons. In all the figures, comparisons for origin and center sheets are shown in black and red, respectively.

desired accuracy. The bond distance distributions obtained from the CG simulations using the final parameters for origin and center chains are shown along with the corresponding atomistic distributions in Figure 2a.

The bond angle potential is given by

$$V_{\Theta} = \sum_{\text{angles}} k_{\Theta} (\Theta - \Theta_0)^2 \quad (3)$$

where the equilibrium virtual bond angle Θ_0 and the corresponding force constant k_{Θ} were determined in a similar fashion, as described above. The bond angle distribution comparisons for origin and center chains are shown in Figure 2b. As can be seen from the figures, the chosen set of virtual bond and angle CG parameters reproduces underlying atomistic distributions quite well.

To represent the long-range order associated with the cellulose crystalline structure, it was found necessary to include a torsional potential. A harmonic potential for the torsion angles was found to reproduce the atomistic torsional distributions reasonably well, as shown in Figure 2c for both origin and center chains. The torsional potential is given by

$$V_{\Phi} = \sum_{\text{dihedrals}} k_{\Phi} (\Phi - \Phi_0)^2 \quad (4)$$

where Φ_0 and k_{Φ} represent the equilibrium torsion angle and the corresponding force constant for four consecutively connected monomers. CG parameters for the bond, angle, and torsional potentials are listed in Table 1.

Nonbonded Parameters. While the development of bonded CG parameters is relatively straightforward, treating nonbonded interactions is more complex. There exist different approaches in developing the nonbonded CG parameters for any given system,^{14,17,19} and the specific need of the intended application of the developed CG force field also influences the approach

Table 1. CG-Bonded Parameters for Crystalline and Amorphous Cellulose Fibrils

bonds	k_b (kJ mol ⁻¹ nm ⁻²)	r_0 (nm)
	Crystalline	
OR–OR	4.588×10^4	0.5283
CE–CE	7.700×10^4	0.5250
	Amorphous	
AM–AM	2.420×10^4	0.5228
angles	k_{Θ} (kJ mol ⁻¹ rad ⁻²)	Θ_0 (°)
	Crystalline	
OR–OR–OR	840	168.7
CE–CE–CE	1080	173.2
	Amorphous	
AM–AM–AM	266.8	163.5
torsions	k_{Φ} (kJ mol ⁻¹ rad ⁻²)	Φ_0 (°)
	Crystalline	
OR–OR–OR–OR	23	191.6
CE–CE–CE–CE	8	187.2
	Amorphous	
AM–AM–AM–AM	5.6	224.0

chosen. If the model needs to be developed for a particular application, one may choose to parametrize the system based solely on the information obtained from corresponding, specific

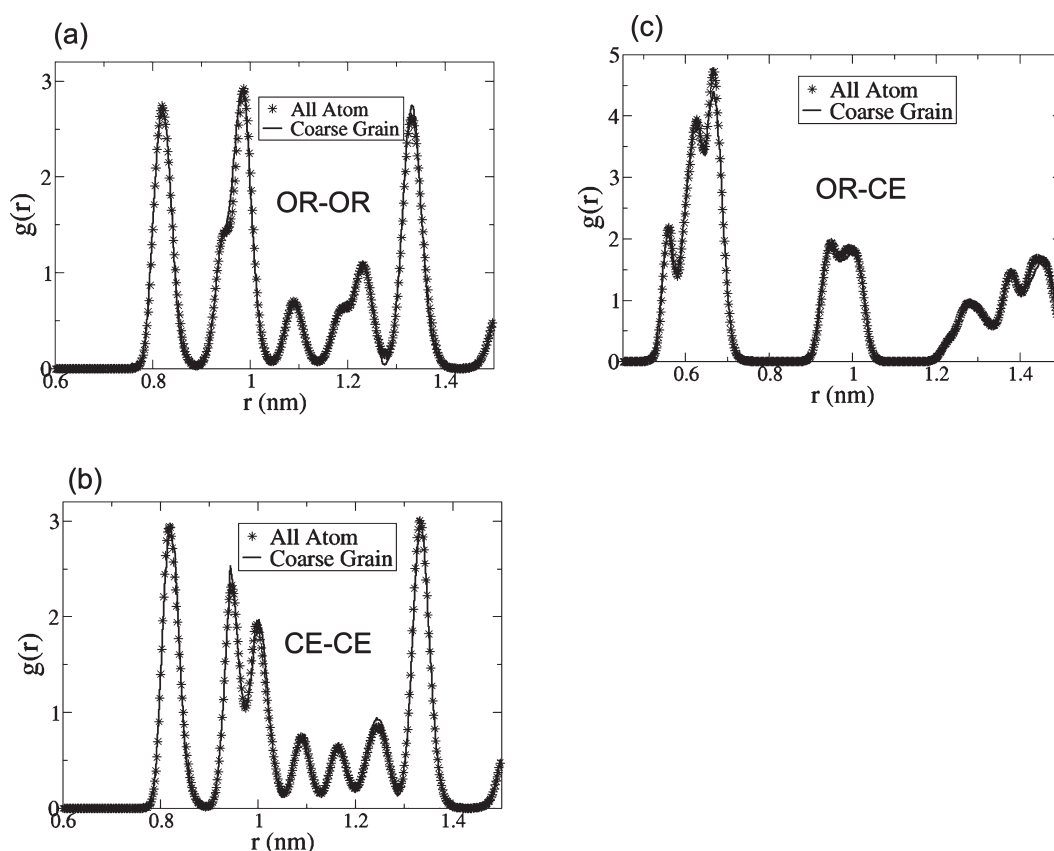


Figure 3. Comparison of all-atom radial distribution functions with those of the CG simulations: (a) $g(r)$ between the CG monomers in origin sheets; (b) $g(r)$ for the CG monomers in center sheets; and (c) $g(r)$ obtained for the cross-interactions (between origin and center sheets). In all the figures, CG distributions (lines) agree with the underlying atomistic distributions (symbols) remarkably well.

higher-resolution simulations. In one such example, a CG force field was developed for biological lipid molecules based on Boltzmann inversion criteria by matching the distributions at the potential level that captures specific biological properties including self-assembly into various morphologies.¹⁷ In another approach, systematic CG models for a variety of biological systems have been developed by introducing multiscale force-matching CG methods.¹⁴ A further method captures correlated motion by matching covariance Hessian matrices.^{15,16} These approaches widen the use of CG methods. Alternatively, CG parameters can be developed based on previously available generalized parameters for the individual components of the system. In an example of this, the Martini CG force field contains a database of CG parameters for the naturally occurring amino acids based solely on their active coefficients in water and oil.³² These, more general CG parameters are more transferable for simulating systems not considered during the force field development process but may fall short of accuracy if the system-specific details have not been incorporated into the parametrization. For example, a recent CG model of cellulose was unable to reproduce a crystalline fibril structure based on MARTINI parameters.²⁶ The present one-site model is based on Boltzmann inversion criteria, an approach that Klein et al. have used extensively for developing CG force fields for copolymers,^{21,22} and was also adopted for a previous three-site model of cellulose.¹³

Distance distributions between the nonbonded units of interest from corresponding atomistic simulations were used as target observables. The goal is to reproduce the atomistic distributions

by carrying out CG simulations, adjusting the underlying potentials. To this end, the radial distribution function $g(r)$ is the target observable from the atomistic simulations. Since there are two different CG units, OR and CE, three separate nonbonded potentials must be determined (OR–OR, OR–CE, and CE–CE); $g(r)$ for these interactions from the atomistic simulations are shown in Figure 3.

Here, we obtain nonbonded CG potentials iteratively using the target observable (the atomistic $g(r)$) as follows

$$V_{new} = V_{old} + k_B T \ln(g_{cg}(r)/g_{aa}(r)) \quad (5)$$

where the updated potential V_{new} between corresponding CG units in each step is obtained by modifying the potential from the previous simulation, V_{old} , k_B is the Boltzmann constant, T the absolute temperature, and the distributions from the atomistic and CG simulations are $g_{aa}(r)$ and $g_{cg}(r)$, respectively. We began with a simple Lennard-Jones 12-6 potential as the starting CG potential and then modified it according to eq 5. In each iteration the CG simulation is carried out using the updated potential, the distribution functions obtained from the updated trajectory are inserted in eq 5, and the potentials updated. Iteration continues until a satisfactory comparison with the target observables is obtained.

The $g(r)$ for crystalline cellulose is complex with multiple discrete distinct peaks and thus contrasts with the simple, relatively smoother functions of typical bulk/liquid systems. Consequently, it is impractical to employ simple Lennard-Jones-type

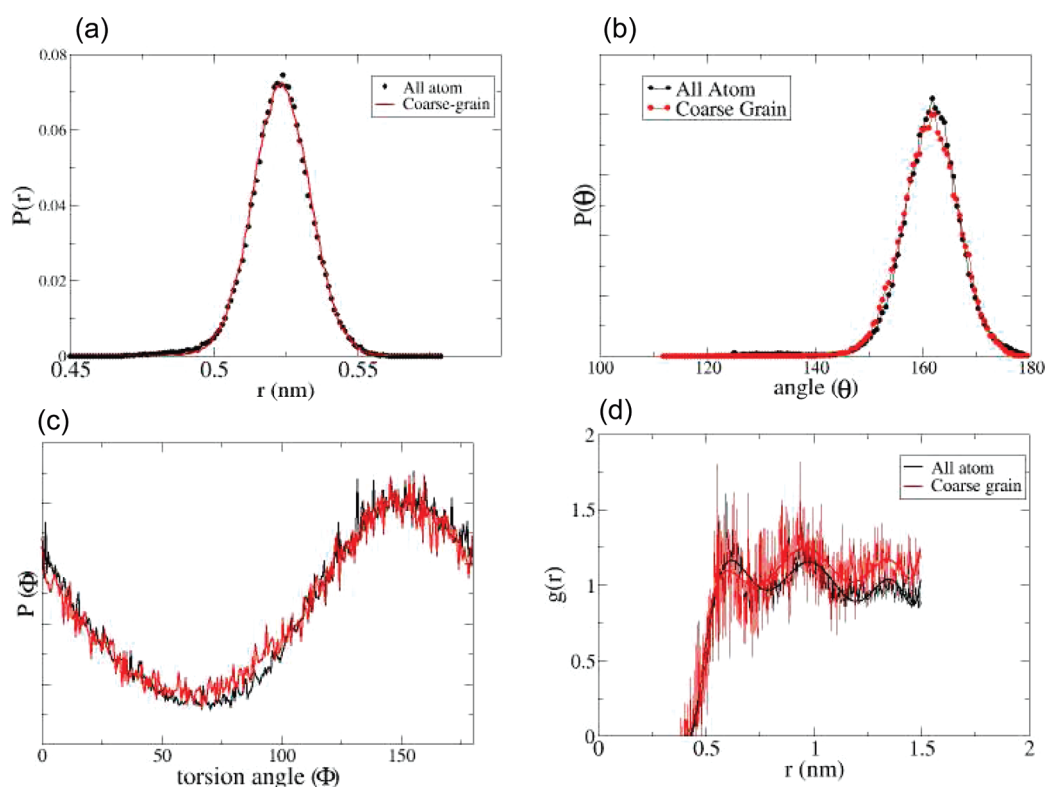


Figure 4. Comparison between atomistic (black) and CG (red) distribution functions of a single cellulose in water: (a) Bond distance, (b) bond angle, and (c) torsional angle, and (d) radial distribution function.

potentials in this case. Using such simplified potentials may also cause instabilities in cellulose crystalline structure, as was observed in previous CG simulation studies of the cellulose fibril.²⁵

Comparison (after five or six iterations) of the target and CG observables for the origin chains indeed showed that this method is unable to capture the fine structure of $g_{aa}(r)$ arising from the crystalline nature of cellulose. Such problems are common when dealing with crystalline materials, as, for example, in the recent study of crystalline fatty acids.³³ As noted in ref 33, eq 5 needs to be modified by introducing a damping factor (δ_{ij}) in order to suppress abrupt changes in the potentials:

$$V_{new} = V_{old} + \delta k_B T \ln[g_{cg}(r)/g_{aa}(r)] \quad (6)$$

where δ is assigned a positive value between 0 and 1. When δ is 1, we recover eq 5 corresponding to a noncrystalline system.

The $g(r)$ obtained including the damping factor was compared with the atomistic $g(r)$ functions, and the iterative procedure repeated with eq 6 until the target and CG distributions were in satisfactory agreement. Final comparison is shown in Figure 3. The CG distribution functions reproduce their atomistic counterparts remarkably well. Importantly, the positions of all the peaks corresponding to the crystal structure are reproduced faithfully by the present CG simulations. Excellent agreement between the atomistic and CG structures demonstrates the ability of the single-bead model to capture the structural features at the monomer level. Hence, this comparison validates the present CG force field for the $I\beta$ crystalline cellulose fibril.

The potentials obtained using eq 6 do not correspond to any simple analytical form, and hence all three nonbonded potentials were tabulated. The Supporting Information gives more details together with the tabulated potentials.

B. CG Model for Noncrystalline Cellulose. The CG force field for the crystalline cellulose fibril was used as a basis from which to develop CG parameters for noncrystalline cellulose. The development procedure used is similar in spirit to that of the crystalline fibril. However, while there is a unique experimental structure corresponding to any given crystalline phase of cellulose, for amorphous cellulose, this is no longer the case. Hence, in order to develop CG parameters for amorphous cellulose, the approach adopted was modified. To this end, two simulation systems representing two extreme scenarios were considered: one for fully crystalline cellulose and the other for fully amorphous. For the fully amorphous systems an atomistic model was constructed consisting of 9 fully hydrated single cellulose chains in TIP3P water, each chain consisting of 10 glucose monomers. An MD simulation of this system was carried out for 20 ns, as was the case for the crystalline system.

Bonded Parameters. A CG parameter set for the amorphous system was mapped from the atomistic system in a similar manner to the crystalline case. Again, the bond distance and the bond and torsion angle potentials were obtained by iterative CG simulations. The obtained parameter set for the single chains differs significantly from that for crystalline cellulose. One major difference is clearly that there is no distinction of origin or center chains, all chains being indistinguishable. Hence, there is a single distribution for each of the bond distances and angles and the torsion angles.

The bond distance and angle distributions are represented with harmonic potentials as before. However, the torsion angle is more flexible than in the crystalline phase and is thus represented using a periodic potential of the following form, corresponding to a periodic potential with single multiplicity.

$$V_{am}(\Phi) = \sum_{\text{dihedrals}} k_{am}(1 + \cos(\Phi - \Phi_0)) \quad (7)$$

The comparisons of the bond distance and angle and the torsional angle distributions obtained from the atomistic and CG simulations are shown in Figure 4a–c, respectively. The final bonded parameter values obtained are listed in Table 1.

Nonbonded Parameters. The complexity of developing nonbonded interaction parameters is greatly reduced in the amorphous case compared to the crystalline fibril due to the presence of a single type of CG bead (named AM). Hence, nonbonded interaction potentials needed to be developed only for AM–AM interactions. Accordingly, the radial distribution function for AM obtained from the atomistic simulations is shown in Figure 4d. As can be seen, the $g_{am}(r)$ in this case does not show discontinuities, in contrast to the crystalline system. The CG distributions were obtained starting with the following Lennard-Jones potential:

$$V_{am}(r_{ij}) = 27/4(\epsilon_{ij})((\sigma/r_{ij})^9 - (\sigma/r_{ij})^6) \quad (8)$$

where ϵ_{ij} and σ_{ij} represent the potential well depth and the contact distance between i and j . The $g_{ij}(r)$ so obtained was used with eq 6 to refine the potential, which in turn was used to carry out new CG simulations. By following the iterative procedure described in Section II, final ϵ and σ values were determined. The final parameters for the fully amorphous system are listed in Table 1. Note that the scaling factor δ (eq 6) was not needed in this case.

III. SIMULATION DETAILS

The CG simulations were performed using stochastic dynamics in order to incorporate solvent effects in an implicit fashion. The dynamics of the individual CG particles was governed by a Langevin equation:

$$m_i(dv_i/dt) = F_i(r_i) - m_i\xi_i v_i + \mathbf{R}_i \quad (9)$$

where the CG particle's mass, velocity, and position are represented by m_i , v_i , and r_i , respectively. F_i is the systematic force on particle i calculated using the interaction potentials described in the previous sections. ξ_i is the friction coefficient, and \mathbf{R}_i is the random force. While longer time step (10–20 fs) was possible for noncrystalline CG fibrils, we were unable to use a time step larger than 1 fs for fully crystalline fibrils because of the discrete nature of the fitted potential. Nevertheless, for the sake of consistency, we have decided to use 1 fs time step for both the crystalline and the noncrystalline fibril simulations. As a result we have chosen a GROMACS default value of 0.2 ps^{-1} for the damping coefficient in all the simulations.

IV. RESULTS AND DISCUSSION

A. Crystalline Cellulose Fibril. Previous CG simulation studies of cellulose fibrils have encountered problems with the stability of the fibril. Hence, as a first test of the present CG model, we examine this aspect. To this end, two different CG simulations were carried out, with different initial conditions. Both of these simulations started with a CG cellulose fibril in the crystalline 1β form. In general, CG simulations would allow larger time steps

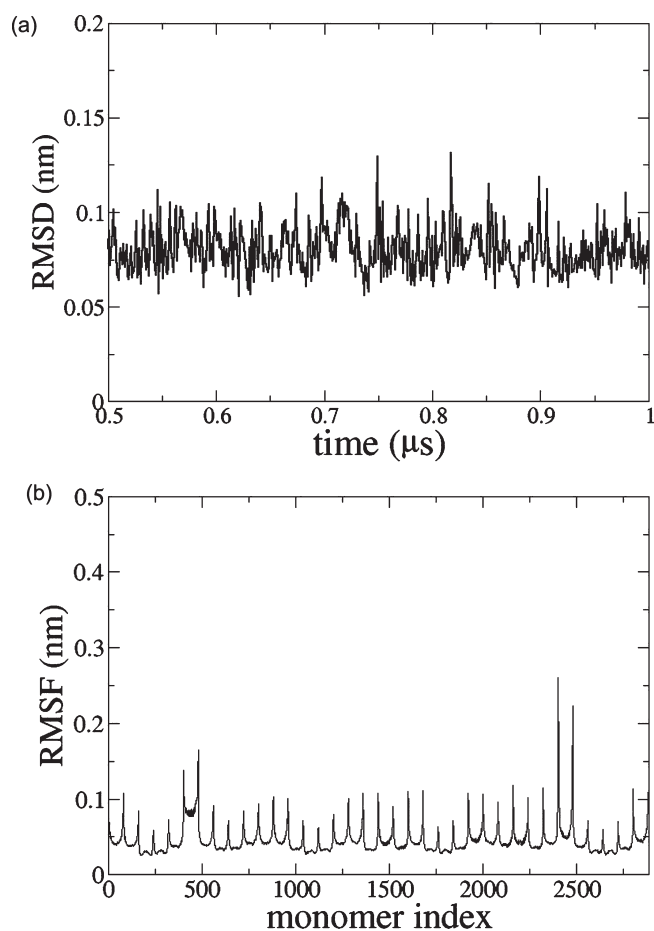


Figure 5. Stability of the CG crystalline cellulose fibril. (a) RMSD is plotted as a function of simulation time. (b) RMSF for each CG monomer within crystalline cellulose fibril.

compared to atomistic simulations. However, in the present study, crystalline nature of the fibril required smaller time steps, as the potentials are relatively rough compared to traditional empirical potentials. After several test runs with different time steps, a time step of 1 fs was chosen, and the production runs were carried out for at least $1 \mu\text{s}$ after the equilibration. The results presented in the following were obtained by averaging over the two separate trajectories.

The stability of the crystalline structure can be assessed by calculating the fluctuations within the fibril (Figure 5a). The root-mean-square displacement (RMSD) is $<0.11 \text{ nm}$ over the entire simulation time ($>1 \mu\text{s}$). This result is in good agreement with a recent atomistic simulation study of crystalline cellulose fibril.¹² The RMSD for the individual (origin and center) sheets shows similar behavior. Considering the fact that each CG site represents a glucose monomer of diameter $\sim 0.52 \text{ nm}$, these fluctuations are relatively small. The stability of the cellulose crystalline structure is confirmed by the root-mean-square fluctuations (RMSF) of individual monomers, shown in Figure 5b. The monomers at the chain ends exhibit relatively larger fluctuations compared to monomers in the middle part of the chain. In other words, the chain ends are relatively flexible in the fibril structure and might disrupt the crystal were the cellulose chains to be too short. Using longer cellulose chains avoids such unwanted effects. Further, the largest fluctuations were observed for the monomers 401–480 and 2401–2480, which belong to corner chains in the

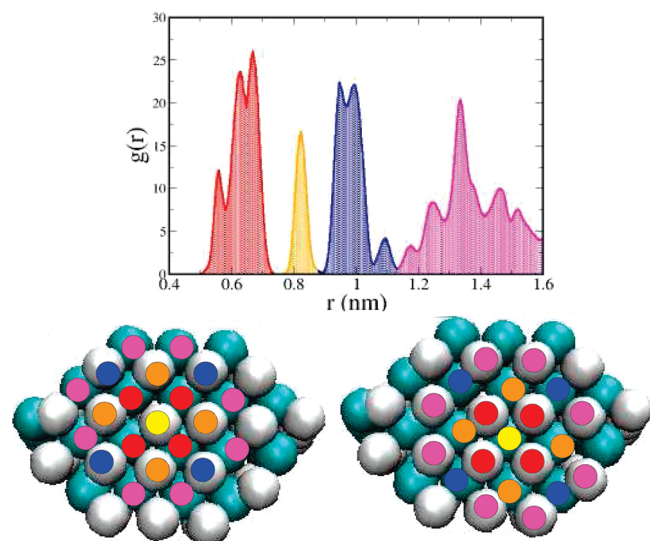


Figure 6. Radial distribution function for a crystalline cellulose fibril. Peaks are color coded. The cellulose fibril is shown in two schematic representations. The fibril structure on the right reveals the nearest neighbors around a selected origin chain monomer, while that on the left shows the same for a center chain monomer.

fibril. This suggests that the corner chains are relatively flexible compared to other chains in the fibril.

In order to examine the structural details we have calculated the pair distribution functions. In Figure 6 the radial distribution function $[g(r)]$ for crystalline fibril is plotted. This system shows rich structural features even beyond 2 nm. Individual peaks were analyzed in detail and are schematically shown in the figure, color coded so as to indicate the responsible CG sites within the fibril. The broader set of peaks around 0.64 nm arises from first neighbor interactions between opposite-type monomers (origin and center or center and origin). The first and second nearest neighbors of the same kind (origin or center monomers) produce two peaks, at 0.81 and 0.98 nm, respectively. Second nearest neighbors of different kinds give rise to the broader peak at 1.38 nm. These structural details are important when comparing crystalline and amorphous cellulose structures, as discussed in detail in the later sections pertaining to amorphous structures.

B. Noncrystalline Cellulose Fibril. As a first check, we applied the fully amorphous CG parameter set to a crystalline fibril. The fibril lost its structure and rapidly collapsed to form a near spherical aggregate (<100 ns). The RMSD for simulations using crystalline and fully amorphous potentials sets is shown in Figure 7. The fully amorphous structure shows very high RMSD values (12 nm) compared to that of crystalline fibril (0.11 nm).

C. Generating Noncrystalline Cellulose Fibril Structures. In this section we describe the study of noncrystalline fibrils, generated by using a combination of fully crystalline and fully amorphous potentials with a coupling parameter as follows

$$V_{cg}(r_{ij}) = \lambda V_{cr}(r_{ij}) + (1 - \lambda)V_{am}(r_{ij}) \quad (10)$$

When $\lambda = 0$ the above potential represents a fully amorphous state, whereas for $\lambda = 1$ we recover crystalline behavior. The above combination rule is applied for both bonded and non-bonded interactions separately.

Fifteen different simulation systems were constructed with a nonlinear spacing of λ values between 0 and 1. Final snapshots

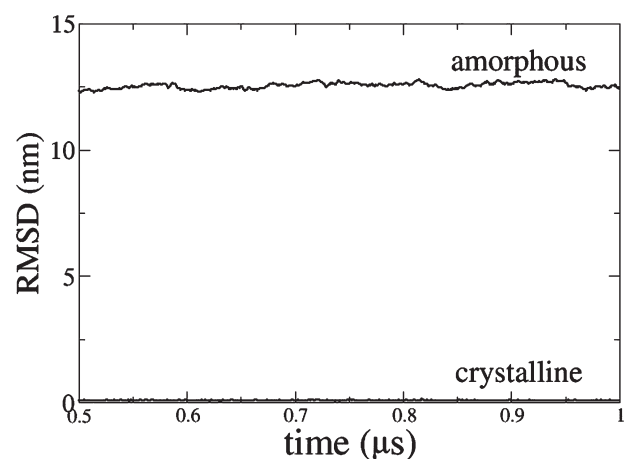


Figure 7. RMSD for the fully amorphous system for the 1 μ s CG simulation. For comparison, RMSD for the fully crystalline (CG) cellulose is also shown.

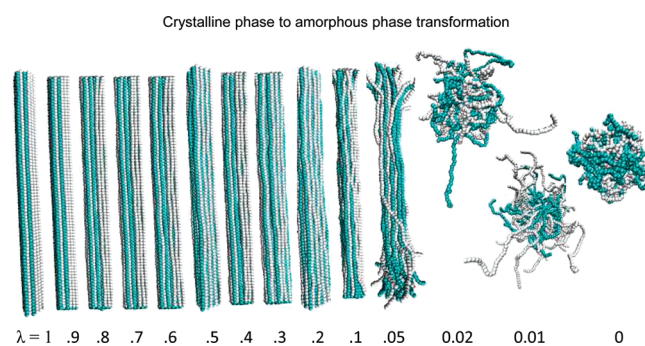


Figure 8. Transition of cellulose fibril from crystalline structure to amorphous (noncrystalline) structure with decreasing coupling parameter (λ) values. Final snapshots obtained from simulations with different λ values are shown; $\lambda = 1$ and 0 correspond to crystalline and fully amorphous phases, respectively.

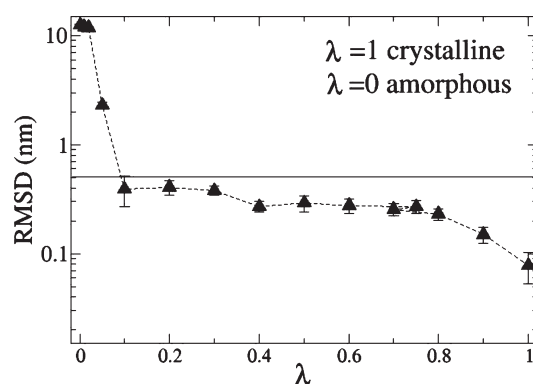


Figure 9. RMSD values for systems with different λ values.

obtained for the selected λ values after 1 μ s simulation are shown in Figure 8. The cellulose deviates from the fully crystalline structure for all values of $\lambda < 1$. However, the transition to a fully amorphous structure accelerates when decreasing λ below 0.1. This is confirmed in Figure 9 in which the average RMSD values for the systems with different λ values are plotted. The RMSD in the fully crystalline phase is near zero. Decreasing λ values increase the RMSD, indicating the emergence of noncrystalline

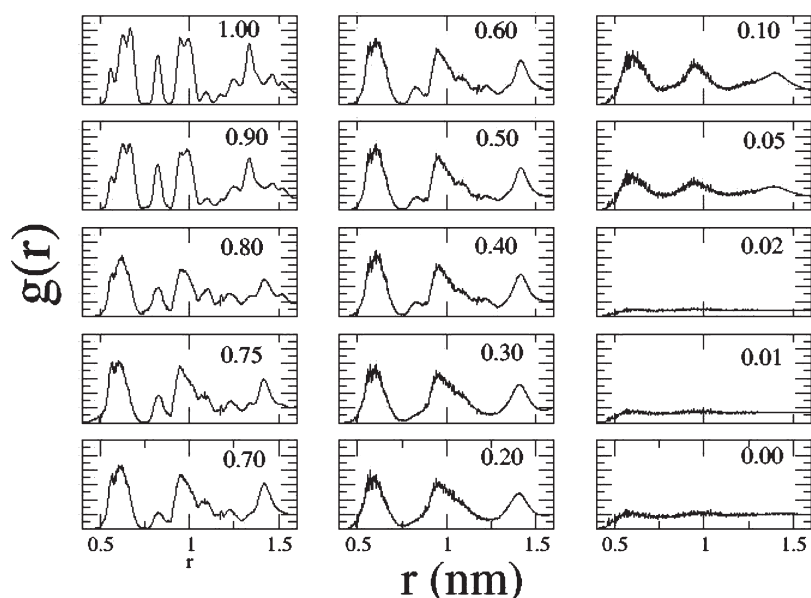


Figure 10. Gradual progression of crystalline to amorphous transition as evident by radial distribution functions. Each figure corresponds to a simulation carried out with the specified λ value.

behavior. However, the RMSD shows a significant crossover between $\lambda = 0.1$ and 0.02 , suggesting a major disruption in the crystalline structure. A similar trend was also observed in RMSF (Supporting Information). For $\lambda < 0.1$, single chain behavior becomes dominant, and the RMSD fluctuations are larger than the size of the CG monomers (>0.52 nm). In fully amorphous structures ($\lambda = 0$) the RMSD is found to be as large as 12 nm. Together, these results provide guidelines for the interaction range over which a possible transition from crystalline to amorphous structure occurs. In the present formalism, fibrils with $\lambda < 0.1$ represent fully noncrystalline structures.

Analyzing fibril structures as a function of λ reveals interesting features. In Figure 10 the pair distribution function $g(r)$ for all the systems with different λ values is plotted. The crystalline structure ($\lambda = 1.0$) gradually disappears with decreasing λ , as evidenced by a gradual disappearance of corresponding peaks in $g(r)$. With decreasing λ , the first peak to show changes is the peak at 0.81 nm, which starts to diminish when λ is decreased by as little as 10%. The peak height then gradually further decreases and ultimately merges with the neighboring peak at 0.98 nm, thereby forming a single broader peak. As explained above, the peak at 0.81 nm corresponds to first nearest neighbors of the same kind, and the peak at 0.98 nm corresponds to second nearest neighbors of different kinds. Combined with the peak position analysis, the merging of these two peaks illustrates how the distinction between the second neighbors of the same or different kind of beads becomes negligible. The structure defining origin and central chains is gradually lost at distances greater than 0.8 nm. Similarly, the peaks from distinct third neighbors merge to a single broader peak around 1.33 nm. Thus the simulations trace the disappearance of long-range structure with decreasing λ . Nevertheless, the presence of peak around 0.63 nm indicates that a short-range structure within the fibril is still present. Further decreasing λ results in the complete loss of crystalline order, as evident from the $g(r)$ at $\lambda < 0.1$. For the λ values 0.01 – 0.0 , the $g(r)$ resembles that of typical dense liquid systems, with the peaks corresponding to neighboring shells of the monomers.

V. CONCLUSIONS

In this paper we have presented efficient and reliable CG models for crystalline and amorphous cellulose fibrils. Each monomer in the cellulose chain is mapped on to a single CG bead. In order to model the corresponding experimental cellulose 1β structure, the cellulose fibril is constructed with distinct chains of origin and center sheets. The target observables were obtained from extensive atomistic simulations, and the optimized CG model obtained is in good agreement with the corresponding atomistic distributions. The robustness of the present CG force field for crystalline cellulose is demonstrated by the stability of cellulose fibrils for several hundreds of nanoseconds without using constraints. To our knowledge, this is the first such unconstrained CG model that can be used to study cellulose fibrils.

Over the course of the CG simulation (>1 μ s), mean bead fluctuations are ~ 0.1 – 0.3 nm, indicating that the fluctuations in the crystalline structure are much smaller than the size of the CG glucose monomer itself (0.52 nm). In natural samples, amorphous and crystalline cellulose coexist.³⁵ Hence, in order to enable a more versatile model of cellulose structure, we extended the crystalline CG model to amorphous cellulose. An individual cellulose chain completely solvated (in water) was considered as an “extreme” case of noncrystalline fully amorphous cellulose, and a corresponding CG parameter set was developed. The use of a coupling parameter λ allows a combinations of potentials for fully crystalline ($\lambda = 1$) and fully amorphous ($\lambda = 0$) systems, thereby producing a series of structures intermediate between crystalline and amorphous states. This λ coupling approach allows the CG study of various partially crystalline and noncrystalline cellulose structures. Decreasing λ by as little as 10% introduces significant noncrystallinity into the cellulose structures, and the simulations exhibit a clear crossover from partially crystalline to fully amorphous behavior around $\lambda < 0.1$.

Previous experimental and simulation studies have shown distinct behavior between the center and the origin planes due to their slightly different hydrogen-bonding patterns.^{12,36} The present results provide noncrystalline models in which the specific

hydrogen bonding is disrupted, leading to the breakdown of short-range order. Analysis of the radial distribution functions shows how as a function of λ the distinction between the center and the origin planes evolves. For $\lambda < 0.3$, the distinction between the two types of planes exists up to the first nearest neighbors, while no distinction was found beyond 0.8 nm, a distance corresponding to second and further neighbors.

In conclusion, the present CG model allows the exploration of both crystalline and amorphous cellulose fibril structures for length- and time-scales beyond the reach of atomistic simulations. A systematic method is presented for generating and representing both crystalline and amorphous cellulose states. Natural cellulose fibrils consist predominately of a mixture of the crystalline phases I α and I β . After biomass treatment, other crystalline forms, such as cellulose II, III, can be formed. Although the present study has focused on the I β cellulose, similar methodology can be easily extended to other crystalline forms. Further, the introduction of a coupling parameter λ provides a unified way of generating cellulose fibril structures with different degrees of crystallinity, thereby enabling the modeling of extensive cellulose fibrils with both crystalline and amorphous characteristics. The mixed CG models could also allow the study of the crystalline-to-amorphous transition of celluloses using an adaptive resolution scheme.³⁷ Future studies will include the importance of explicit solvent incorporation along with the interactions between crystalline and amorphous cellulose structures. Furthermore, increasing the complexity of the CG models to incorporate other biomolecules, such as lignin and hemicellulose, will be important in understanding biomass recalcitrance, a central theme in biomass-based renewable energy. Work in these directions is presently underway.

ASSOCIATED CONTENT

S Supporting Information. Details on tabulated potentials used in this work along with supporting results that were not presented in manuscript are presented in the Supporting Information. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: smithjc@ornl.gov.

ACKNOWLEDGMENT

The authors would like to thank L. Petridis for help and valuable scientific discussions. G.S. thanks R. Schulz for the help with GROMACS software and B. Lindner for helping with structural characterizations and for providing the trajectories of the atomistic crystalline cellulose fibril. This research is sponsored by DOE's Scientific Discovery through Advanced Computing (SciDAC) program through DOE's Office of Advanced Scientific Computing Research (ASCR) and Biological and Environmental Research (BER) under FWP ERKJE84 and performed at Oak Ridge National Laboratory (ORNL). ORNL is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

REFERENCES

(1) Habibi, Y.; Lucia, L. A.; Rojas, O. J. *Chem. Rev.* **2010**, *110*, 3479–3500.

- (2) Metthew, J. F.; Skopec, C. E.; Mason, P. E.; Zuccato, P.; Torget, R. W.; Sugiyama, J.; Himmel, M. E.; Brady, J. W. *Carbohydr. Res.* **2006**, *341*, 138–152.
- (3) Ding, S. Y.; Himmel, M. E. *J. Agri. Food. Chem.* **2006**, *54*, 597–606.
- (4) Demain, A. L.; Newcomb, M.; Wu, J. H. D. *Microbiol. Mol. Biol. Rev.* **2005**, *69*, 124.
- (5) Lynd, L. R.; Cushman, J. H.; Nicholas, R. J.; Wyman, C. E. *Science* **1991**, *251*, 1318.
- (6) Himmel, M. E.; Ding, S.-Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D. *Science* **2007**, *315*, 804–807.
- (7) Vasella, A.; Davics, G. J.; Bohm, M. *Curr. Opin. Chem. Bio.* **2008**, *12*, 539–629.
- (8) Nishiyama, Y.; Langan, P.; Chanzy, H. *J. Am. Chem. Soc.* **2002**, *124*, 9074–9082.
- (9) Nishiyama, Y.; Sugiyama, J.; Chanzy, H.; Langan, P. *J. Am. Chem. Soc.* **2003**, *125*, 14300–14306.
- (10) Zhong, L.; Matthews, J. F.; Crowley, M. F.; Rignall, T.; Talon, C.; Cleary, J. M.; Walker, R. C.; Chukkappalli, G.; McCabe, C.; Nimlos, M. R.; Brooks, C. L., III; Himmel, M. E.; Brady, J. W. *Cellulose* **2008**, *15*, 261–273.
- (11) Zhong, L.; Matthews, J. F.; Hansen, P. I.; Crowley, M. F.; Cleary, J. M.; Walker, R. C.; Nimlos, M. R.; Brooks, C. L., III; Adney, W. S.; Himmel, M. E.; Brady, J. W. *Carbohydr. Res.* **2009**, *344*, 1984–1992.
- (12) Gorss, A. S.; Chu, J.-W. *J. Phys. Chem. B* **2010**, *114*, 13333–13341.
- (13) Nielsen, S.; Lopez, C.; Srinivas, G.; Klein, M. L. *J. Phys.: Condens. Matter* **2004**, *16*, R481.
- (14) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.
- (15) Moritsugu, K.; Smith, J. C. *Biophys. J.* **2008**, *95*, 1639–1648. *Biophys. J.* **2007**, *93*, 3460–3469.
- (16) Voltz, K.; Trylska, J.; Tozzini, V.; Kurkal-Siebert, V.; Langowski, J.; Smith, J. C. *J. Comput. Chem.* **2008**, *29*, 1429–1439.
- (17) Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L. *J. Phys. Chem. B* **2001**, *105*, 4464. Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Moore, P. B.; Klein, M. L. *J. Phys. Chem. B* **2001**, *105*, 9785.
- (18) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. *J. Chem. Phys.* **2008**, *128*, 244115.
- (19) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750–760.
- (20) Queyroy, S.; Neyertz, S.; Brown, D.; Muller-Plathe, F. *Macromolecules* **2004**, *37*, 7338–7350.
- (21) Srinivas, G.; Discher, D. E.; Klein, M. L. *Nat. Mat.* **2004**, *3*, 638–644.
- (22) Srinivas, G.; Pitera, J. W. *Nanolett* **2008**, *8*, 611–618.
- (23) Molinero, V.; Goddard, W. A., III *J. Phys. Chem. B* **2004**, *108*, 1414–1427.
- (24) Liu, P.; Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 11566–11575.
- (25) Bu, L.; Beckham, G. T.; Crowley, C. F.; Chang, C. H.; Matthews, J. F.; Bomble, Y. J.; Adney, W. S.; Himmel, M. E.; Nimlos, M. R. *J. Phys. Chem. B* **2009**, *113*, 10994–11002.
- (26) Wohler, J.; Berglund, L. A. *J. Chem. Theory Comput.* **2011**, *7*, 753–760.
- (27) Schultz, R.; Lindner, B.; Petridis, L.; Smith, J. C. *J. Chem. Theory Comput.* **2009**, *5*, 2798–2808.
- (28) Pingali, S. V.; Urban, V. S.; Heller, W. T.; McGaughey, J.; O'Neill, H.; Foston, M.; Myles, D. A.; Ragauskas, A.; Evans, B. R. *Biomacromolecules* **2010**, *11*, 2329–2335.
- (29) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **2004**, *4*, 187–217.
- (30) Jorgensen, W. L.; Chandrasekar, J.; Madhura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (31) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (32) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; deVries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

- (33) Hadley, K. R.; McCabe, C. J. *Chem. Phys.* **2010**, *132*, 134505.
- (34) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. J. *Chem. Theory Comput.* **2008**, *4*, 435–447.
- (35) Dumitriu, S. *Polysaccharides: Structural Diversity and Functional Versatility*, 2nd ed.; Marcel Dekker: New York, 2005.
- (36) Yui, T.; Nishimura, S.; Akiba, S.; Hayashi, S. *Carbohydr. Res.* **2006**, *341*, 2521–2530.
- (37) Praprotnik, M.; Site, L. D.; Kremer, K. *Annu. Rev. Phys. Chem.* **2008**, *59*, 545–571.

Toward Quantitative Structure–Property Relationships for Charge Transfer Rates of Polycyclic Aromatic Hydrocarbons

Milind Misra,[†] Denis Andrienko,^{*,§,‡} Björn Baumeier,^{§,‡} Jean-Loup Faulon,^{||} and O. Anatole von Lilienfeld^{*,†,§,⊥}

[†]Advanced Device Technologies Department, Sandia National Laboratories, Albuquerque, New Mexico 87185-1322, United States

[‡]Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany

[§]Institute of Pure and Applied Mathematics, University of California Los Angeles, Los Angeles, California 90095-7121, United States

^{||}Institute of Systems & Synthetic Biology, University of Evry, 5 rue Henri Desbrières, 91030, Evry cedex, France

S Supporting Information

ABSTRACT: Quantitative structure–property relationships (QSPRs) have been developed and assessed for predicting the reorganization energy of polycyclic aromatic hydrocarbons (PAHs). Preliminary QSPR models, based on a combination of molecular signature and electronic eigenvalue difference descriptors, have been trained using more than 200 PAHs. Monte Carlo cross-validation systematically improves the performance of the models through progressive reduction of the training set and selection of best performing training subsets. The final biased QSPR model yields correlation coefficients q^2 and r^2 of 0.7 and 0.8, respectively, and an estimated error in predicting reorganization energy of ± 0.014 eV.

I. INTRODUCTION

A key property of organic semiconducting materials is that their conducting properties can be tuned by optimizing their chemical structure.^{1–5} A practical route to do this includes the synthesis of a new compound, optimization of its processing conditions, fabrication of the device, and measurement of its performance (properties). By repeating this procedure, one can formulate structure–processing–property relationships and proceed with the rational design of organic semiconductors.

It is desirable to assist the design by optimizing material properties using computer simulations. First, methods are required that are capable of predicting the property of interest starting from the chemical structure, preferably without fitting parameters. The second step consists of correlating these properties with the corresponding structures for a specified training set of compounds and formulating quantitative structure–property relationships (QSPRs). Finally, improved compounds are identified for a specific property range.

For organic semiconductors, already the first step in this scheme is nontrivial since charge carrier mobility depends on molecular geometry, electronic structure, and global percolation pathways for charge carriers. Without discussing any details, this represents a typical multiscale problem, and attempts to solve it constitute an entire research field.^{6–22} Current experience suggests that it is very difficult to directly evaluate charge carrier mobility as a property of interest for an arbitrary chemical compound, since several assumptions are necessary regarding material morphology, the type of transport, and the model used to describe it. One could, however, ask whether it is possible to find adequate QSPRs that relate chemical structure to charge transport properties, the link between chemical structure and mobility being established first.

In this paper, we construct and assess the quality of several such QSPRs in the context of organic semiconductors. As a test

system, we use polycyclic aromatic hydrocarbons (PAHs). PAHs or, more specifically, discotic liquid crystals have already found application in organic solar cells and field effect transistors.^{2,23,24} A typical chemical structure of a discotic liquid crystal consists of a flat conjugated core with side chains attached to its periphery. Discotics self-assemble into columnar structures with aromatic cores stacked on top of each other. Overlap of the π orbitals of these cores enables charge transport along columns, rendering these materials one-dimensional semiconductors. The efficiency of charge transport can be engineered by either varying the shape and size of the conjugated core or influencing their packing through the modification of side chains.

Due to structural, dynamic, and energetic disorder, charge transport in discotic liquid crystals occurs via charge carrier hopping between the neighboring molecules. The rate of hopping is given by the high-temperature nonadiabatic Marcus theory:^{6,25,26}

$$\omega = \frac{J^2}{\hbar} \sqrt{\frac{\pi}{\lambda k_B T}} \exp \left[-\frac{(\Delta G - \lambda)^2}{4k_B T \lambda} \right] \quad (1)$$

where J is the electronic coupling matrix element between the donor and acceptor molecules, λ is the reorganization energy, ΔG is the free energy difference between the initial and final states, and T is the temperature.

Equation 1 identifies several parameters important for charge transport. The transfer integral J is related to the overlap of electronic orbitals, highest occupied molecular orbital (HOMO) for the hole and lowest molecular orbital (LUMO) for the electron transport. As such, it is very sensitive to the relative position and orientation of neighboring molecules.^{13,17,27}

Received: April 5, 2011

Published: June 02, 2011

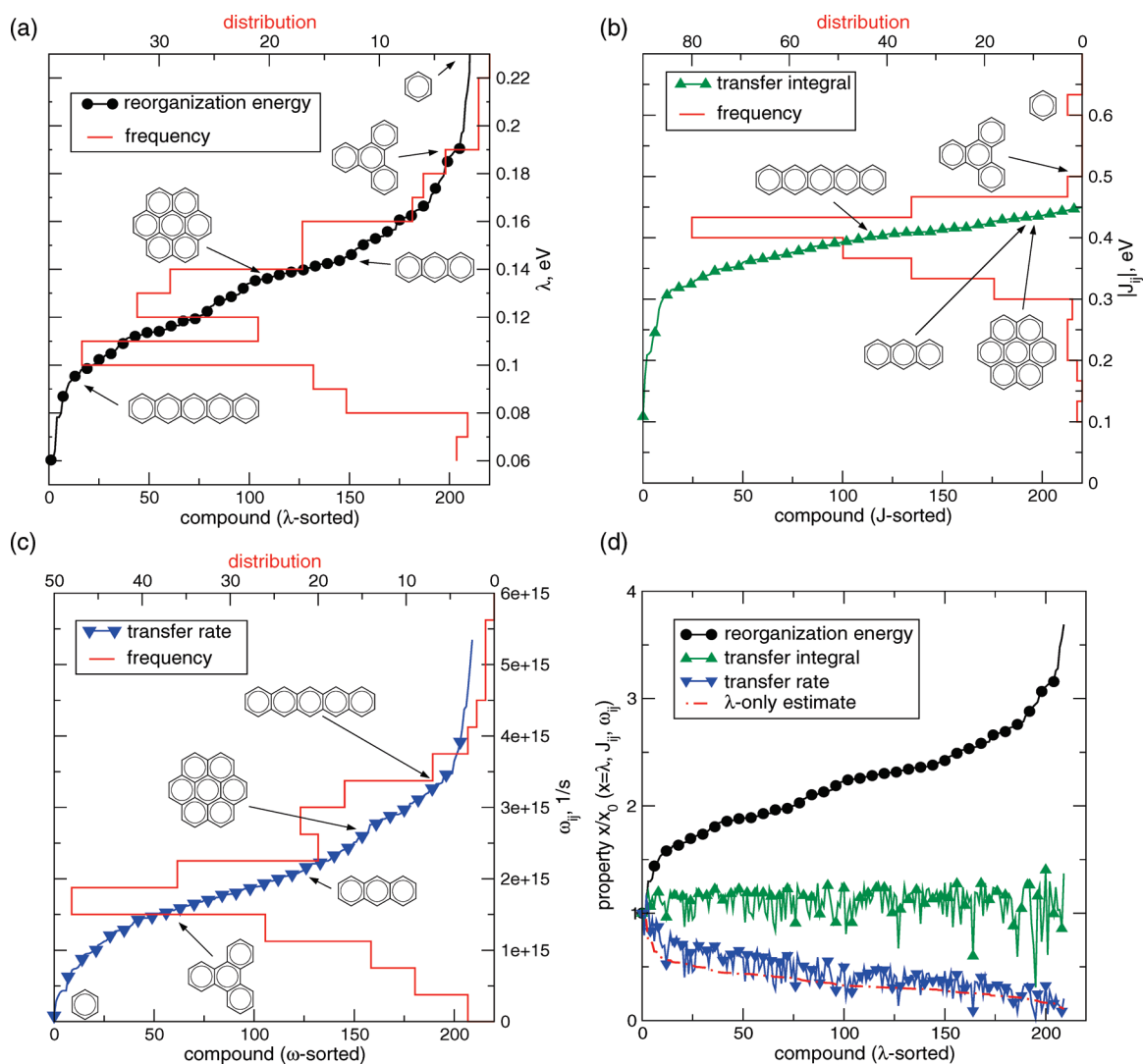


Figure 1. Analysis of reorganization energies, transfer integrals, and transfer rates in the PAH compound data set: Sorted values of (a) λ , (b) J , and (c) ω as functions of compound indices, as well as the respective compound distributions (red lines). For illustrative purposes, several compounds at λ values of 0.3 eV (benzene), 0.19 eV (triphenylene), 0.14 eV (anthracene), 0.13 eV (coronene), and 0.09 eV (pentacene) are shown. Panel d shows a combined plot of the three properties against the λ -sorted compound index. The red dash-dotted line indicates a λ -only prediction of the transfer rates based on eq 1, assuming a constant transfer integral for the entire data set.

In columnar phases of discotics, the maximum of the transfer integral is achieved in a face-to-face molecular arrangement, with the typical intermolecular distance of $d = 3.5 \text{ \AA}$.^{24,28–30} In what follows, we assume such an “ideal” molecular arrangement, since it maximizes charge transport and hence provides an upper bound for the charge mobility which can be reached experimentally. We ignore the distribution in transfer integrals due to thermal fluctuations as well as static defects in morphology. Another parameter, ΔG , is the free energy difference between the states with charge localized on a donor or an acceptor of the charge transfer complex. For an ideal face-to-face arrangement, this contribution vanishes due to equivalence of the initial and final states. Finally, the internal reorganization energy, λ , expresses the strength of electron–phonon coupling and has an exponential impact on the transfer rate, with small λ favoring more efficient charge transport.

For an ideal face-to-face columnar alignment, the mobility of the charge carrier along the column is proportional to the

hopping rate, eq 1, with $\Delta G = 0$:

$$\mu = \frac{\omega d^2}{k_B T} = \frac{J^2 d^2}{\hbar k_B T} \sqrt{\frac{\pi}{k_B T \lambda}} \exp\left[-\frac{\lambda}{4k_B T}\right] \quad (2)$$

where d is the distance between neighboring sites. We can therefore argue that large hopping rates (that is, large transfer integrals, small reorganization energies) favor high charge mobilities. Hence, the potential descriptors shall link the chemical structure of a compound with the hopping rate, or, alternatively, J and λ .

In this study, we develop appropriate structure–mobility QSPRs. To do this, we first present how the PAH compound data set was generated and used to select the parameters dominating the charge transport in columnar phases of discotics. We then present two descriptors and assess their performance within preliminary QSPR models. Finally, a robust QSPR model is developed using Monte Carlo cross-validation for variable training/test set ratios.

II. COMPOUND DATA SET

Even for an “ideal” molecular packing, it is not immediately obvious which of the two physical parameters, J or λ , is more important for charge transfer rate prediction. In order to identify the dominant physical parameter and set up the reference values for QSPRs, we have generated a compound data set of PAHs and analyzed its properties. Starting from benzene, we have appended additional aromatic rings at random available bonds. We have used standard carbon–carbon and carbon–hydrogen bond lengths and angles, checking for atom overlaps as well as aromaticity (Hückel rule) and discarding multiple copies of the same PAH. Thus, a data set of 211 closed shell aromatic PAHs with up to nine benzene rings has been generated.

For hole transport, the reorganization energy can be written as a sum of the relaxation energies in neutral and positively charged states

$$\lambda = E_n^+ - E_n^0 + E_c^0 - E_c^+ \quad (3)$$

where E_g^q is an energy of the compound in charge state q and geometry g . $q = 0$ corresponds to a neutral molecule and $q = +$ to a cation. $g = n$ indicates optimized geometry of a neutral molecule, while $g = c$ corresponds to an optimized cation geometry. Hence, four calculations per compound are necessary, two geometry optimizations for the neutral (E_n^0) and cationic (E_c^+) species and two single point energy calculations for the cationic species in the neutral geometry (E_n^+) and for the neutral species in the cationic geometry (E_c^0).

Reorganization energies were computed using density functional theory (DFT; B3LYP functional,³¹ the 6-311++g(d,p) basis set) using the Gaussian 03 package.³² Figure 1a shows the values of λ in ascending order together with their distribution in the data set. The reorganization energies of 211 compounds are spread from 0.06 to 0.30 eV.

Transfer integrals J were calculated for a cofacial geometry and molecular separation of 3.5 Å using Zerner’s Intermediate Neglect of Differential Overlap method as implemented in the Molecular Orbital Overlap package.³³ Figure 1b shows the resulting values of J in ascending order and their distribution within the data set. The transfer integrals span a range of 0.1 to 0.5 eV, which is relatively large due to the assumed columnar stacking of the molecules. This distribution is sharply peaked around 0.4 eV, indicating that there are only small variations of J within the data set. The corresponding distribution of transfer rates ω is shown in Figure 1c.

Figure 1d combines the three parameters $x = \lambda, J$, and ω plotted as a function of the λ -sorted compound index. All values are shown relative to the value of the compound with index zero (x_0). This representation illustrates that among the three parameters the reorganization energy has the largest relative variance and that the transfer integrals only slightly fluctuate around a constant value. To further support this conclusion, we have included a λ -only estimate of the transfer rates, with a constant transfer integral for the entire data set. The result, shown in Figure 1d, corroborates the assumption that, for an “ideal” packing considered here, the reorganization energy is a dominant factor influencing the transfer rates and thereby the charge carrier mobilities. Henceforth, we will concentrate on developing QSPRs for the reorganization energy λ .

III. DESCRIPTORS

QSPR relies on the definition of descriptors that characterize the chemical structure, for instance, the number of atoms,

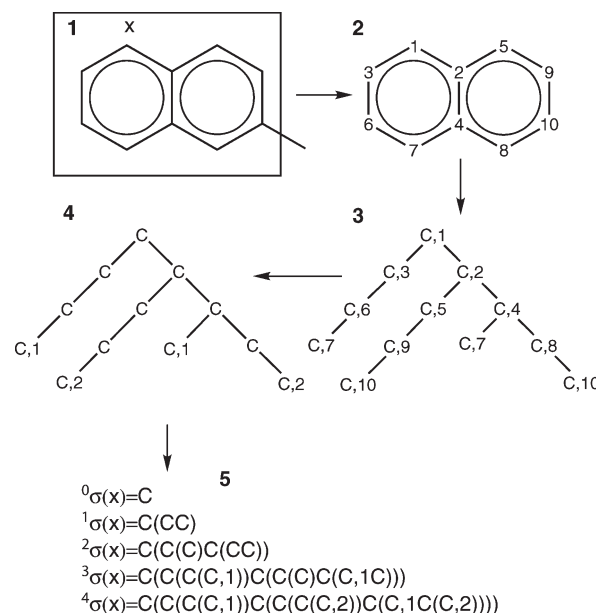


Figure 2. Atomic signatures ${}^h\sigma$ from height $h = 0-4$ for an atom X in 2-methyldecahydronaphthalene. ${}^h\sigma(X)$ is determined as follows: (1) The subgraph containing all atoms at distance 4 from atom X is extracted. (2) This subgraph is canonicalized with atom X having label 1. (3) A tree spanning all edges of the subgraph is constructed. (4) All labels appearing only once are removed, and the remaining labels are renumbered in the order they appear. (5) The atomic signature is determined after reading the tree in a depth-first order, the depth corresponding to height h .

molecular mass, and deviations of the molecular shape from planarity or linearity. Several scalar descriptors that have been investigated but rejected due to their low correlation with the reorganization energy are described in the Supporting Information.

The two descriptors, molecular signature and $\Delta\epsilon$, had the largest correlation with λ . We first discuss their use for the preliminary QSPR models that are based on the full 211-compound PAH data set. We then present the development of “biased” QSPR models after partitioning the data set into a 188-compound training set and a 23-compound test set. Finally, we address the predictive power of the biased models for the test set compounds.

A. Molecular Signature. The molecular signature is a compilation of a set of atomic signatures, $\{\sigma\}$, that occur in a molecule. It was first presented and applied in the context of structure elucidation³⁴ and later defined for acyclic compounds and used in QSPR analyses.³⁵ An atomic signature describes the extended covalent bonding neighborhood of an atom within a molecule up to a certain “height”, h . Figure 2 illustrates how atomic signatures, ${}^h\sigma$, are generated. The molecular signature for a given height is a vector that contains the frequencies of all of the ${}^h\sigma$'s occurring in the molecule. As such, it represents a methodical codification system over an alphabet of atom types.

The MolConverter program from ChemAxon³⁶ was used to convert the xyz-coordinate files of the structures in our PAH data set to corresponding simplified molecular input line entry specification (SMILES) strings. SMILES describe chemical structures and topologies using short textual strings.³⁷ From the SMILES strings, molecular signatures have been determined for individual heights. The correlation between these molecular signatures alone and the reorganization energy is, however, insufficient for making predictions of λ . For example, the correlation

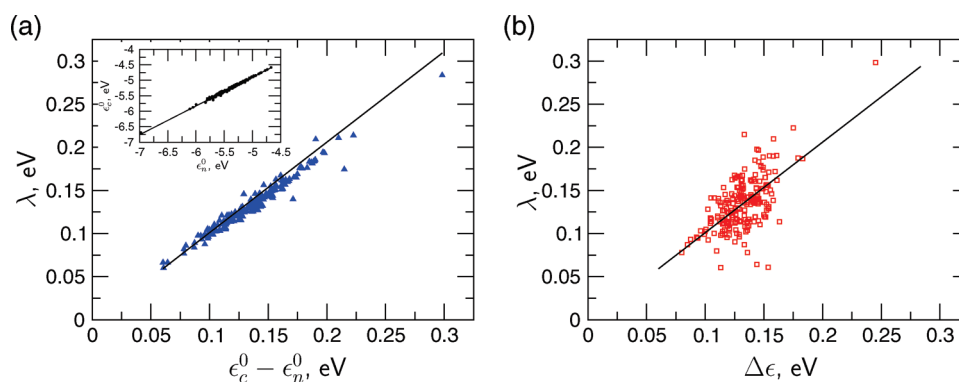


Figure 3. Correlations of calculated reorganization energies λ of the PAH data set with differently predicted values. In panel a, the correlation to $\epsilon_c^0 - \epsilon_n^0$ according to eq 4 results in $\lambda = 1.05 \times (\epsilon_c^0 - \epsilon_n^0) - 0.004$ [eV] with $r^2 = 0.96$ (blue triangles). The inset shows the correlation between the highest occupied electronic eigenvalue of the neutral molecule in cationic geometry, ϵ_c^0 , with the respective value for the neutral geometry ϵ_n^0 , resulting in $\epsilon_c^0 = 0.93 \times \epsilon_n^0 - 0.25$ [eV] and $r^2 = 0.99$. Using this relation to predict λ from ϵ_n^0 only according to $\Delta\epsilon$ in eq 6 yields the correlation shown in panel b with $\lambda = 1.01 \times \Delta\epsilon + 0.001$ [eV], $r^2 = 0.39$ (red squares).

coefficient, r^2 , of molecular signatures with $h \in (0, 1, 2, 3)$ versus λ does not exceed 0.39.

B. HOMO Eigenvalue Difference, $\Delta\epsilon$. As mentioned above, the reorganization energy λ expresses the strength of electron–phonon coupling in the molecule. Thus, a weak correlation of λ and the descriptor based solely on structural features, such as the molecular signature, is not surprising. In order to add electronic properties to descriptors, we note that eq 3 can be rearranged in terms of the difference between vertical excitation energies, linking states of the same geometry but a different number of electrons N_e , yielding

$$\begin{aligned}\lambda &= \delta_n - \delta_c, \\ \delta_n &= E_n^+(N_e - 1) - E_n^0(N_e) \\ \delta_c &= E_c^+(N_e - 1) - E_c^0(N_e)\end{aligned}\quad (4)$$

Here, δ_n is the iso-nuclear change in energy due to removal of an electron from the neutral species in its relaxed geometry, while δ_c is the iso-nuclear change in energy due to addition of an electron to the cationic species in its relaxed geometry.

Based on molecular grand-canonical ensemble DFT,^{38–40} we can Taylor-expand δ_n and δ_c in number of electrons, N_e

$$E(N_e + \Delta N_e) = E(N_e) + \frac{\partial E(N_e)}{\partial N_e} \Delta N_e + O(\Delta N_e^2) \quad (5)$$

For an exact expression for the exchange-correlation potential within density-functional theory, all higher order terms would vanish for $0 \leq \Delta N_e \leq 2$ because the total potential energy of a molecule with fixed external potential changes only linearly as one varies the number of electrons.^{41,42} Since the derivative of the energy with respect to N_e is the eigenvalue of the highest occupied molecular orbital (HOMO),^{43,44} we can combine eqs 4 and 5 and express λ as

$$\begin{aligned}\delta_n &= \frac{\partial E_n^0(N_e)}{\partial N_e} \Delta N_e = \epsilon_n^0(N_e) \Delta N_e \\ \delta_c &= \frac{\partial E_c^0(N_e)}{\partial N_e} \Delta N_e = \epsilon_c^0(N_e) \Delta N_e \\ \lambda &= \epsilon_c^0 - \epsilon_n^0\end{aligned}\quad (6)$$

where $\Delta N_e = -1$ and $\epsilon_n^0(N_e)$ and $\epsilon_c^0(N_e)$ denote the eigenvalues of the highest occupied molecular Kohn–Sham orbitals of the

neutral molecule in the respective optimal neutral and cationic geometries.

The exact form of the exchange-correlation functional is, however, unknown. Moreover, the self-interaction error increases for fractional occupation within widely used functionals.⁴⁵ The difference between electronic eigenvalues of the HOMOs in the neutrally and cationically relaxed geometries, $\epsilon_c^0 - \epsilon_n^0$ yields therefore only an estimate of λ . In our case, we have tested the quality of this approximation for the B3LYP hybrid functional by correlating the λ obtained from the eigenvalues as in eq 6 with the λ obtained from the energies according to eq 3. As shown in Figure 3a, the correlation is very strong with a correlation coefficient r^2 of 0.96. This could be further improved by using functionals that correctly account for fractional occupation numbers.⁴²

The (approximate) determination of the reorganization energy according to eq 6 still requires the optimizations of neutral and cationic geometries, as well as a single-point calculation for the neutral molecule in the cationic geometry. While this is one calculation less than in eq 3, it is inconvenient since ideally one would like to predict λ from ground-state properties of the neutral molecule alone, i.e., without having to calculate ϵ_c^0 . We have therefore probed whether ϵ_c^0 correlates with ϵ_n^0 in the PAH data set. The inset in Figure 3a shows ϵ_c^0 plotted versus the respective ϵ_n^0 . The linear regression yields $\epsilon_c^{\text{pred}} = 0.93\epsilon_n^0 - 0.25$ [eV] with a remarkable correlation of $r^2 = 0.99$. On the basis of these observations, we have used

$$\Delta\epsilon \equiv \epsilon_c^{\text{pred}} - \epsilon_n^0 \quad (7)$$

as an additional scalar descriptor for λ .

Figure 3b shows the correlation of the actual λ from eq 3 with the estimated $\Delta\epsilon$, $\lambda \approx 1.01\Delta\epsilon + 0.001$ [eV]. The regression for this expression, however, yields a rather low correlation coefficient of only $r^2 = 0.39$. Thus, solely an electronic descriptor cannot reliably predict reorganization energies.

IV. QSPR MODELS

From the two preceding sections, it is apparent that when used separately neither the structural molecular signature nor the electronic eigenvalue descriptor $\Delta\epsilon$ are sufficient for reliable quantitative estimates of the reorganization energy in our set of PAHs. Since λ is a measure of the coupling of structural and

Table 1. Preliminary QSPR Models, i–viii, and Corresponding q^2 Values for Multiple Linear Regression (MLR) and Partial Least Squares (PLS), Respectively^a

<i>h</i>	type	# σ	i	ii	iii	iv	v	vi	vii	viii
0–3	MLR	63	0.29	0.62	0.20	0.28	0.29	0.29	0.63	0.64
0–3	PLS	63	0.47	0.47	0.47	0.46	0.47	0.47	0.46	0.45
0–4	PLS	431	0.50	0.50	0.50	0.49	0.50	0.50	0.50	0.26
0–5	PLS	1635	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.31

^aSee section IV.A and the Supporting Information for more details. Here, *h* is the signature height; # σ refers to the number of atomic signatures, i.e., the dimension of the molecular signature vector. These models were generated using the data set of 211 PAHs. Highlighted model ii has been used for the construction of the “biased” QSPR. (i) molecular signatures. (ii) molecular signatures + $\Delta\epsilon$. (iii) molecular signatures + dM (molecular distance). (iv) molecular signatures + dL (molecular linearity). (v) molecular signatures + dP (molecular planarity). (vi) molecular signatures + dH (hydrogen repulsion). (vii) molecular signatures + $dM + dL + dP + dH + \Delta\epsilon$. (viii) vii redundant descriptors removed based on UFS.

electronic degrees of freedom in a molecule, it is natural to attempt a combination of the two descriptors.

A. Preliminary QSPR Models. For the different heights of molecular signatures (see Section IIIA), we have set up different preliminary QSPR models using signatures of heights 0–3 through 0–5 for the PAH compound data set (without outliers). Specifically, leave-one-out cross-validated correlation coefficients (q^2) have been calculated using multiple linear regression (MLR) and partial least-squares (PLS). These coefficients, together with the preliminary models, are listed in Table 1. For the sake of completeness, we also present the results for additional models that are not based on $\Delta\epsilon$ but that combine molecular signature with various other scalar structural descriptors. More technical details can be found in the Supporting Information.

Our results show that while the PLS calculations yield a q^2 of around 0.50, indicating predictability in general for all of the models, they do not suggest a preference for a particular descriptor combination. MLR results, in contrast, indicate a clear preference for the model combination of the height 0–3 molecular signature with $\Delta\epsilon$, which has a q^2 of 0.62 (see model ii in Table 1). For the alternative combinations of molecular signature and various structural scalar descriptors, the corresponding r^2 ranges only from 0.20 to 0.29. An additional model which combines all descriptors considered in this study (see, e.g., model vii in Table 1) does not improve the performance of preliminary QSPR model ii, even when using unsupervised forward selection of the descriptors (see the Supporting Information) to further eliminate redundancy among descriptors (model viii).

Thus, the combination of height 0–3 signatures and $\Delta\epsilon$ in model ii is identified as the optimal starting combination for developing the “biased” QSPR model in the next section. More specifically, the logarithm of λ is estimated by

$$\log_{10} \lambda = c_\lambda \Delta\epsilon + c_0 + \sum_i c_i \sigma_i \quad (8)$$

where *i* runs over all 63 signatures and where c_0 , c_λ , and $\{c_i\}$ are the trained QSPR coefficients of the intercept and the descriptors $\Delta\epsilon$ and signatures, respectively (see the Supporting Information).

B. Biased QSPR Models. We found that the previously identified optimal preliminary QSPR model can further be

developed into a “biased” QSPR model with the help of Monte Carlo cross-validation. To this end, the PAH compound data set was first split into a total 188-compound training set and a test set of 23 compounds, enabling the validation of the biased models. The test set was determined using dissimilarity-based compound selection, as described in the Supporting Information. The resulting test set compounds are shown in Table 2.

Thereafter, out of the total 188-compound training set, subsets with varying percentages *x* were defined, where $x \in (5, 10, 15, \dots, 90, 95)\%$. For each *x*, 10 000 random partitions from among the 188 compounds were generated. All of the random partitions were subjected to training using the preliminary QSPR model ii, i.e., height 0–3 molecular signatures combined with $\Delta\epsilon$ based on PLS.

The models obtained, dubbed M_x^k ($k \in 1, 2, \dots, 10\,000$), were subsequently ranked according to their performance as measured by q^2 . For M_{100} , $q^2 = 0.44$ and $r^2 = 0.53$, which is below the conventional predictive threshold of 0.50. As described in more detail in the Supporting Information, q^2 can be improved by reducing the training subset size *x*, followed by the QSPR model training of 10 000 random partitions for each of these reduced training subsets.
























Figure 4 illustrates the results for varying percentage *x*. The average q^2 , i.e., the average of the cross-validated correlation coefficients over all randomly chosen partitions, declines progressively as the training set size decreases. The standard deviation around that average, however, increases even more, thereby enabling us to identify “biased” models M_x^I , M_x^{II} , M_x^{III} , etc., namely models that yield the respective best, second best, third best, etc. q_x^2 out of all 10 000 models that have been trained for each partition at a particular *x*. This behavior is in line with ref 46.

C. Test Set Results of Biased QSPR Models. Figure 4 demonstrates overfitting, namely, that predicting the reorganization energy based on the biased (best performing) Monte Carlo models will always be the more favorable the smaller the subset is. From the behavior of q^2 versus *x*, one could therefore be tempted to deduce that the optimal model should be based on the smallest training subset. Obviously, q^2 is not a sufficient requirement for the predictability of a model, and only external validation provides a sound assessment of a QSPR model.⁴⁷ Thus, to determine the optimal size of the training subset, we compute the root-mean-square (RMS) deviation of predicted λ from actual λ for the 23 test set molecules using the biased M_x^I , where $x = 5, 10, \dots, 100$. Note that we excluded two outliers from the test set since they had the largest residuals and corresponded to extreme λ values (maximum and minimum) within the entire compound data set.

As shown in Figure 4, as *x* decreases from 100 to 60%, RMS remains roughly constant (~ 16 meV) and starts to strongly increase in oscillatory fashion for subsets smaller than 55%. Since RMS is minimal at $x = 40\%$ (14 meV), we define the corresponding biased model M_{40} as our best QSPR model for predicting λ of PAHs. In contrast, model M_{100} has a higher RMS of 17 meV. Biased QSPR model M_{40} does not only have a lower RMS deviation but also exhibits improved correlation coefficients, $q^2 = 0.70$ and $r^2 = 0.80$. Table 2 lists the residuals for the predictions of λ based on models M_{100} and M_{40} .

In summary, model M_{100} predicts the reorganization energy of more than 75% compounds within a reasonable margin of error (± 20 meV). The biased model M_{40} , however, predicts a larger number (>85%) of test set compounds within the same error margin of ± 20 meV.

Table 2. Actual λ 's [eV] and Residuals [meV] of Predicted λ 's of the Test Set of Compounds for the Two Biased Models, M_{100} and M_{40}

ID	λ	M_{100}	M_{40}	structure	ID	λ	M_{100}	M_{40}	structure	ID	λ	M_{100}	M_{40}	structure
0	0.30	-111	-100		1	0.19	1	8		2	0.22	-43	-27	
7	0.11	4	5		41	0.13	-5	5		43	0.13	-9	-10	
55	0.20	-21	-25		83	0.11	8	13		86	0.14	-10	-15	
107	0.11	11	-11		116	0.10	7	2		124	0.12	-3	6	
129	0.09	-11	-9		143	0.13	6	8		158	0.12	-2	2	
166	0.11	13	13		185	0.13	-13	-4		193	0.10	6	8	
195	0.09	25	18		207	0.06	72	77		209	0.10	31	20	
211	0.16	-34	-29		215	0.15	-3	0						

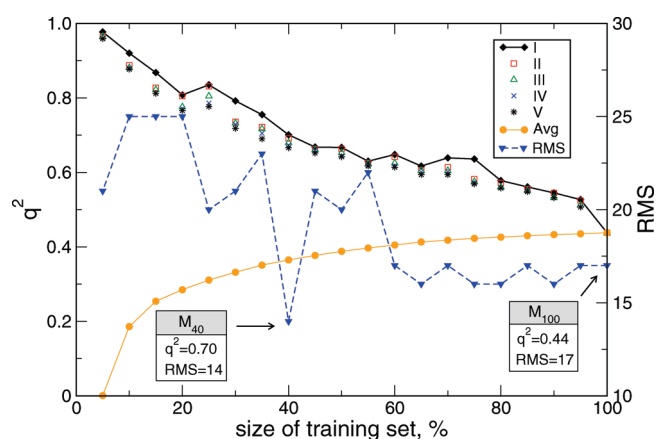


Figure 4. $q^2(M_x^k)$ of the best five partitions $k \in (I, II, III, IV, V)$ (out of 10 000) as a function of training subset size x in steps of 5%. Yellow represents the average q^2 of 10 000 partitions. Root mean square (RMS) deviation [meV] of actual λ from λ predicted by $M_x^k = f$ for test set compounds in Table 2. M_{40} refers to the best partition (I) and is dubbed the biased QSPR model.

V. CONCLUSIONS

On the basis of conceptual density functional theory, we have developed a frontier orbital eigenvalue descriptor $\Delta\varepsilon$ for the empirical prediction of reorganization energies, λ . For a compound data set of over 200 polycyclic aromatic hydrocarbons, we have investigated the performance of various QSPR models aimed at predicting reorganization energies of PAHs based on a combination of a structural and an electronic descriptor, molecular signature, and $\Delta\varepsilon$, respectively. For the entire data set, we find that preliminary QSPR models yield at best a correlation coefficient of $q^2 = 0.5$. Monte Carlo cross-validation with training subsets

permits the definition of a “biased” model with significantly better performance, yielding a q^2 and r^2 of 0.70 and 0.80, respectively, and a root-mean-square deviation of predicted from actual λ of 0.014 eV. Additional scalar structural descriptors, such as average interatomic distance, deviation from linearity, or deviation from planarity yielded only negligible improvement when combined with molecular signature. Furthermore, we have confirmed the basic assumption of selection algorithms based on dissimilarity, which requires that compounds spanning structure/descriptor space also span property/activity space. The main drawback of the proposed descriptor is that it does not account accurately enough for the changes of the molecular geometry upon charging/discharging.

■ ASSOCIATED CONTENT

S Supporting Information. Details of scalar descriptors, molecular signatures, methodology of the QSPR analysis, PLS coefficients for the QSPR models M_{40} and M_{100} , and complete reference Gaussian 03. This information is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: denis.andrienko@mpip-mainz.mpg.de; anatole@alcf.anl.gov.

Present Addresses

[†]Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, United States

■ ACKNOWLEDGMENT

M.M. and O.A.v.L. acknowledge support from SNL Truman Program LDRD project No. 120209. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin

Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. This work was partially supported by the DFG program IRTG 1328, DFG grant SPP 1355, and BMBF grant MESOMERIE.

REFERENCES

- (1) McCulloch, I.; Heeney, M.; Bailey, C.; Genevicius, K.; Macdonald, I.; Shkunov, M.; Sparrowe, D.; Tierney, S.; Wagner, R.; Zhang, W.; Chabinyk, M.; Kline, R.; McGehee, M.; Toney, M. *Nat. Mater.* **2006**, *5*, 328–333.
- (2) Wu, J. S.; Pisula, W.; Müllen, K. *Chem. Rev.* **2007**, *107*, 718–747.
- (3) Feng, X.; Marcon, V.; Pisula, W.; Hansen, M. R.; Kirkpatrick, J.; Grozema, F.; Andrienko, D.; Kremer, K.; Müllen, K. *Nat. Mater.* **2009**, *8*, 421.
- (4) Nelson, J.; Kwiatkowski, J.; Kirkpatrick, J.; Frost, J. *Acc. Chem. Res.* **2009**, *42*, 1768–1778.
- (5) Yan, H.; Chen, Z.; Zheng, Y.; Newman, C.; Quinn, J.; Dotz, F.; Kastler, M.; Facchetti, A. *Nature* **2009**, *457*, 679–U1.
- (6) Bredas, J.; Calbert, J.; da Silva, D.; Cornil, J. *Proc. Natl. Acad. Sci.* **2002**, *99*, 5804–5809.
- (7) Nagata, Y.; Lennartz, C. *J. Chem. Phys.* **2008**, *129*, 034709.
- (8) Kwiatkowski, J. J.; Nelson, J.; Li, H.; Bredas, J. L.; Wenzel, W.; Lennartz, C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1852–1858.
- (9) Cheung, D. L.; Troisi, A. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5941–5952.
- (10) Bredas, J. L.; Beljonne, D.; Coropceanu, V.; Cornil, J. *Chem. Rev.* **2004**, *104*, 4971–5003.
- (11) Coehoorn, R.; Pasveer, W. F.; Bobbert, P. A.; Michels, M. A. J. *Phys. Rev. B* **2005**, *72*, 155206.
- (12) Coropceanu, V.; Cornil, J.; da Silva, D. A.; Olivier, Y.; Silbey, R.; Bredas, J. L. *Chem. Rev.* **2007**, *107*, 2165–2165.
- (13) Kirkpatrick, J.; Marcon, V.; Nelson, J.; Kremer, K.; Andrienko, D. *Phys. Rev. Lett.* **2007**, *98*, 227402.
- (14) Kirkpatrick, J.; Marcon, V.; Kremer, K.; Nelson, J.; Andrienko, D. *J. Chem. Phys.* **2008**, *129*, 094506.
- (15) Marcon, V.; Kirkpatrick, J.; Pisula, W.; Andrienko, D. *Phys. Status Solidi B* **2008**, *245*, 820–824.
- (16) Marcon, V.; Breiby, D.; Pisula, W.; Dahl, J.; Kirkpatrick, J.; Patwardhan, S.; Grozema, F.; Andrienko, D. *J. Am. Chem. Soc.* **2009**, *131*, 11426–11432.
- (17) Olivier, Y.; Muccioli, L.; Lemaure, V.; Geerts, Y.; Zannoni, C.; Cornil, J. *J. Phys. Chem. B* **2009**, *113*, 14102–14111.
- (18) Troisi, A.; Cheung, D. L.; Andrienko, D. *Phys. Rev. Lett.* **2009**, *102*, 116602.
- (19) Vehoff, T.; Baumeier, B.; Troisi, A.; Andrienko, D. *J. Am. Chem. Soc.* **2010**, *132*, 11702–11708.
- (20) Vehoff, T.; Chung, Y. S.; Johnston, K.; Troisi, A.; Yoon, D. Y.; Andrienko, D. *J. Phys. Chem. C* **2010**, *114*, 10592–10597.
- (21) Vehoff, T.; Baumeier, B.; Andrienko, D. *J. Chem. Phys.* **2010**, *133*, 134901.
- (22) Lukyanov, A.; Andrienko, D. *Phys. Rev. B* **2010**, *82*, 193202.
- (23) Schmidt-Mende, L.; Fechtenkötter, A.; Müllen, K.; Moons, E.; Friend, R. H.; MacKenzie, J. D. *Science* **2001**, *293*, 1119–1122.
- (24) Li, J.; Kastler, M.; Pisula, W.; Robertson, J.; Wasserfallen, D.; Grimsdale, A.; Wu, J.; Müllen, K. *Adv. Funct. Mater.* **2007**, *17*, 2528–2533.
- (25) Marcus, R. A. *Rev. Mod. Phys.* **1993**, *65*, 599.
- (26) Hutchison, G. R.; Ratner, M. A.; Marks, T. J. *J. Am. Chem. Soc.* **2005**, *127*, 2339.
- (27) Olivier, Y.; Lemaure, V.; Bredas, J.; Cornil, J. *J. Phys. Chem. A* **2006**, *110*, 6356–6364.
- (28) Pisula, W.; Kastler, M.; Wasserfallen, D.; Pakula, T.; Müllen, K. *J. Am. Chem. Soc.* **2004**, *126*, 8074–8075.
- (29) Pisula, W.; Tomovic, Z.; Simpson, C.; Kastler, M.; Pakula, T.; Müllen, K. *Chem. Mater.* **2005**, *17*, 4296–4303.
- (30) Kastler, M.; Pisula, W.; Wasserfallen, D.; Pakula, T.; Müllen, K. *J. Am. Chem. Soc.* **2005**, *127*, 4286–4296.
- (31) Stevens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1993**, *98*, 11623.
- (32) Frisch, M. J. et al. *Gaussian 03*, revision B.05; Gaussian, Inc.: Wallingford, CT, 2003.
- (33) Kirkpatrick, J. *Int. J. Quantum Chem.* **2008**, *108*, 51.
- (34) Faulon, J. L. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1204–1218.
- (35) Visco, J.; Pophale, R. S.; Rintoul, M. D.; Faulon, J. L. *J. Mol. Graphics Modell.* **2002**, *20*, 429–438.
- (36) <http://www.chemaxon.com> (accessed July 2011).
- (37) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- (38) von Lilienfeld, O. A.; Tuckerman, M. E. *J. Chem. Phys.* **2006**, *125*, 154104.
- (39) Marcon, V.; von Lilienfeld, O. A.; Andrienko, D. *J. Chem. Phys.* **2007**, *127*, 064305.
- (40) von Lilienfeld, O. A. *J. Chem. Phys.* **2009**, *131*, 164102.
- (41) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. *Phys. Rev. Lett.* **1982**, *49*, 1691.
- (42) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *Phys. Rev. Lett.* **2009**, *102*, 066403.
- (43) Janak, J. F. *Phys. Rev. B* **1978**, *18*, 7165.
- (44) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *Phys. Rev. B* **2008**, *77*, 115123.
- (45) Zhang, Y.; Yang, W. *J. Chem. Phys.* **1998**, *109*, 2604.
- (46) Shao, J. *J. Am. Stat. Assoc.* **1993**, *88*, 486–494.
- (47) Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.

Prediction of Charge Mobility in Amorphous Organic Materials through the Application of Hopping Theory

Choongkeun Lee,[†] Robert Waterland,[‡] and Karl Sohlberg^{*,†}

[†]Department of Chemistry, Drexel University, Philadelphia, Pennsylvania 19104-2875, United States

[‡]Central Research and Development, E. I. du Pont de Nemours & Co., Inc., Wilmington, Delaware 19880-0320, United States

ABSTRACT: The application of hopping theory to predict charge (hole) mobility in amorphous organic molecular materials is studied in detail. Application is made to amorphous cells of *N,N'*-diphenyl-*N,N'*-bis-(3-methylphenylene)-1,1'-diphenyl-4,4'-diamine (TPD), 1,1-bis-(4,4'-diethylaminophenyl)-4,4-diphenyl-1,3-butadine (DEPB), *N4,N4'*-di(biphenyl-3-yl)-*N4,N4'*-diphenylbiphenyl-4,4'-diamine (mBPD), *N1,N4*-di(naphthalen-1-yl)-*N1,N4*-diphenylbenzene-1,4-diamine (NNP), and *N,N'*-bis[9,9-dimethyl-2-fluorenyl]-*N,N'*-diphenyl-9,9-dimethylfluorene-2,7-diamine (pFFA). Detailed analysis of the computation of each of the parameters in the equations for hopping rate is presented, including studies of their convergence with respect to various numerical approximations. Based on these convergence studies, the most robust methodology is then applied to investigate the dependence of mobility on such parameters as the monomer reorganization energy, the monomer–monomer coupling, and the material density. The results give insight into what will be required to improve the accuracy of predictions of mobility in amorphous organic materials, and what factors should be controlled to develop materials with higher (or lower) charge (hole) mobility.

1. INTRODUCTION

Organic conducting materials are receiving significant scrutiny for possible application in the development of lightweight, cheap, and flexible electronic devices, such as organic light-emitting diodes (OLED), data storage systems, field-effect transistors (FET), or solar cells.^{1–3} While organic conductors have numerous advantages over established technology, the hole mobility of most organic crystals ($\sim 10^{-2}$ cm²/V s) is much smaller than that of the inorganic materials in common use. Recently, however, organic crystals, such as pentacene and phthalocyanine, have been produced with improved conducting properties, thereby invigorating the field. The mobility of pentacene has been improved from 2×10^{-3} to 1.5 cm²/V s and that of phthalocyanine from 10^{-3} to 0.02 cm²/V s using improved fabrication technology.⁴

Some characteristics of high-performance organic electronic materials are well established. For example, discotic liquid crystal molecules that have one or more strong resonance rings at their center often show good charge mobility. This results from a low reorganization energy and a large electronic splitting.^{5–10} The strong resonance ring tends to make the molecules form columnar phases, leading to a pseudo-one-dimensional charge-transfer pathway. Such a one-dimensional charge-transfer path was observed in our previous work on phthalocyanine, in which we showed that the coupling matrix element is high only for the face to face charge-transfer dimer.¹¹ For this face to face charge transfer, the site energies are nearly identical, producing very low energetic disorder. The large coupling matrix element and the low energetic disorder combine to yield high charge mobility.

Although the mechanism of charge transfer in organic crystals has been studied for several decades, the capability to predict charge mobility in amorphous materials is still very limited. Tse and colleagues have reviewed three general approaches to modeling charge transport.¹² Historically, charge transport was

modeled with macroscopic phenomenological models. The currently most popular approach is to employ the Gaussian disorder model (GDM). The GDM approach, which is most appropriate when charge transport is dominated by energetic disorder in the hopping sites,¹³ was developed by Bässler and colleagues^{14–16} and used extensively by others.^{17,18} The third approach, and the one adopted for the present work, is to apply Marcus hopping theory, which is especially useful when polaron effects are significant.¹³ In amorphous organic materials, charge mobility cannot be predicted by simple application of Marcus–Hush theory, however, because all molecules are inequivalent due spatial disorder, so energetic disorder in the matrix cannot be disregarded.^{19–21} To predict the charge mobility adequately, the site energy must be reliably incorporated.

We seek to identify a reliable and computationally efficient computational method for screening amorphous molecular solids for high hole mobility. Toward this end, in this study we explore the application of hopping theory to calculate charge (hole) mobility in amorphous *N,N'*-diphenyl-*N,N'*-bis-(3-methylphenylene)-1,1'-diphenyl-4,4'-diamine (TPD), 1,1-bis-(4,4'-diethylaminophenyl)-4,4-diphenyl-1,3-butadine (DEPB), *N4,N4'*-di(biphenyl-3-yl)-*N4,N4'*-diphenylbiphenyl-4,4'-diamine (mBPD), *N1,N4*-di(naphthalen-1-yl)-*N1,N4*-diphenylbenzene-1,4-diamine (NNP), and *N,N'*-bis[9,9-dimethyl-2-fluorenyl]-*N,N'*-diphenyl-9,9-dimethylfluorene-2,7-diamine (pFFA) (for which the chemical structures are shown in Figure 1). For comparison, we also calculate charge (hole) mobility in crystalline tetracene and pentacene. First, we describe the theoretical/computational approach. Next, we report detailed studies of the various approximations employed. Finally, numerous factors controlling the mobility are investigated, including the influence of the

Received: May 21, 2011

Published: June 29, 2011

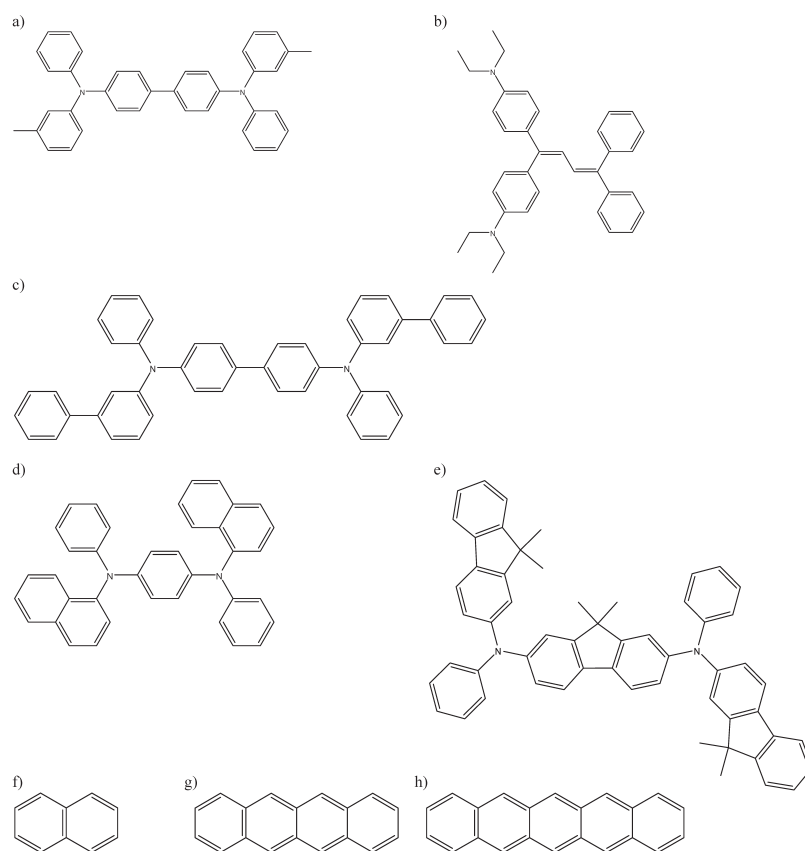


Figure 1. Chemical structures of the monomers of the amorphous materials studied here: (a) TPD, (b) DEPB, (c) mBPD, (d) NNP, (e) pFFA, (f) naphthalene, (g) tetracene, and (h) pentacene.

dielectric properties of the material matrix on the reorganization energy, the dependence of charge mobility on material density, and the method of treating site energy.

2. THEORY AND COMPUTATIONAL METHODS

2.1. Overview. Our approach to computing charge mobility in amorphous molecular solids is based on modeling the following fundamental step in the charge-transport process. If “M” represents the basic monomer unit of the material, (a single molecule) we consider the rate of hopping of a charge between one monomer and an adjacent one:



in the presence of a disordered background distribution of M. The disordered background distribution represents the material bulk.

To model the material bulk, we construct an ensemble of representative unit cells, using a procedure described in detail in Section 2.2. These unit cells are constructed to be sufficiently large so that for a representative monomer near the center of the cell, the nearest-neighbor interactions involve other monomers within the same cell. We refer to these cells as “amorphous cells”, while acknowledging that a material built by translational repetition of such a cell would have periodicity on the length scale of the cell dimension.

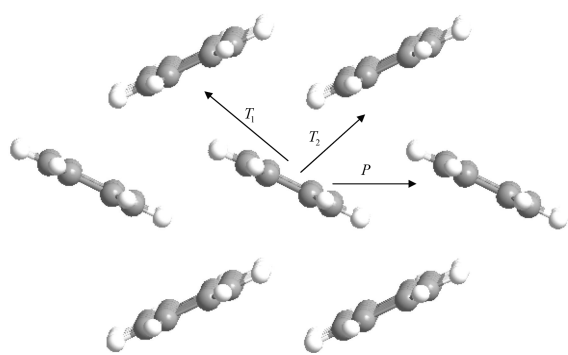
Next, for each cell, hopping theory is applied to a selected monomer within the cell to compute the rate of charge hopping to the adjacent monomers. These hopping rates are then

combined to determine a hole mobility for the cell. The ensemble of cells therefore gives rise to a distribution of mobility values, and the average value of mobility over the entire ensemble is taken as the estimate of mobility in the material. This approach, outlined above and detailed in the balance of this paper, explicitly incorporates energetic disorder of the sites in the material and does so without the introduction of an adjustable “disorder parameter” common to simulation approaches to computing mobility.^{22–24} The use of an explicit cell model also avoids the severe assumption of isotropic monomer–monomer interactions. Instead, each such interaction is treated explicitly.

2.2. Construction of Amorphous Cells. Amorphous cell representations of TPD, DEPB, mBPD, NNP, and pFFA were generated using the amorphous cell construction and forcite modules of Accelrys’ Materials Studio 5.0 commercial code.²⁵ There are no material density measurements for any of the amorphous organic materials studied in this paper. Since the charge mobility depends exponentially on density (vide infra), we first accurately determined material densities using the procedure developed by Rigby.²⁶ In brief, the molecular monomer structures shown in Figure 1a–e were first optimized using molecular mechanics and the COMPASS force field.²⁶ Ten cubic lattice amorphous cells were prepared for each material. Each cell was filled with sufficient copies of the optimized monomer structures to result in approximately 1600 atoms/cell. The starting cubic cell parameter was selected to make the density close to 1 g/cm³, and the cells were energy minimized after generation. Each cell was subjected to four sequential forcite molecular dynamics runs: (i) a constant volume and temperature (NVT)

Table 1. X-ray Crystal Cell Parameters for Tetracene and Pentacene

	space group	<i>a</i> (Å)	<i>b</i> (Å)	<i>c</i> (Å)	α	β	γ
tetracene	$P\bar{1}$	7.98	6.14	13.57	101.3	113.2	87.5
pentacene	$P\bar{1}$	6.28	7.79	14.56	76.4	87.6	84.7

**Figure 2.** Neighbor interactions (P, T1, and T2) in the tetracene crystal.

velocity scaled run of 10 000 steps; (ii) a NVT run of 40 000 steps using the Andersen thermostat; (iii) a constant pressure and temperature (NPT) run of 10 000 steps; and (iv) a final production NPT run of 100 000 steps. Both NPT runs employed the Andersen and Berendsen thermostat and barostat and an applied external pressure of 0.0001 GPa. The time step for all of the dynamics runs was 1 fs. Full atomic coordinate frames were written out every 1000 steps of the production runs and used to compute the mean density and its standard deviation. The computed densities of TPD, DEPB, mBPD, NNP, and pFFA are 1.07, 1.01, 1.10, 1.14, and 1.02 ± 0.01 g/cm³, respectively.

Subsequently these computed densities were used to generate more than 90 different amorphous cells for each material, each with ca. 20 molecules (24 and 15 molecules for NNP and pFFA, respectively) in the unit cell. Each generated cell was energy minimized to remove close contacts generated by the cell packing algorithm. These amorphous cells were generated in a cubic lattice, $a = 26.9129, 26.6633, 26.8210, 26.1847, \text{ and } 26.5138$ Å for TPD, DEPB, mBPD, NNP, and pFFA, respectively.

The crystal structures of tetracene and pentacene were built up using X-ray diffraction data^{27,28} with Material Studio.²⁵ Details of their structures are given in Table 1. In $P\bar{1}$ symmetry, there are 3 nearest-neighbor interactions, T1, T2, and P-type, and 1 long range interaction, L-type.^{19,20} The three nearest neighbor interactions are displayed in Figure 2.

2.3. Mobility. To compute mobility for a given amorphous cell, we apply a generalization of Deng and Goddard's¹⁹ implementation of Marcus–Hush theory.²⁹ In this approximation, mobility (μ) is given in terms of the charge diffusion coefficient D , unit charge e , temperature T , and Boltzmann constant k_B by an expression attributed to Einstein:

$$\mu = \frac{eD}{k_B T} \quad (2)$$

The diffusion coefficient is determined from a weighted average of the hole hopping rates W_{ij} from a representative monomer cation A_i^+ in the material to each of its interacting neutral neighbors A_j . The weight on each term is given by the product of the square of the distance to the neighbor r_{ij} and by the

normalized hopping probability to the neighbor P_{ij} .

$$D = \frac{1}{2n} \sum_{i \neq j} r_{ij}^2 W_{ij} P_{ij} \quad (3)$$

The summation runs over all monomers within a cutoff radius $r_c = 12$ Å of the representative central molecule. (See Convergence Tests Section below.) The charge-hopping rate (without external electric field) is assumed to be given by the general expression:^{19–21}

$$W_{ij} = \frac{V_{ij}^2}{\hbar} \left(\frac{\pi}{\lambda k_B T} \right)^{1/2} \exp \left(- \frac{(\Delta E_{ij} + \lambda)^2}{4\lambda k_B T} \right) \quad (4)$$

Where V_{ij} is the coupling matrix element, λ is the reorganization energy, \hbar is Planck's constant, and ΔE_{ij} is the energy difference between the initial and final states in the hopping process. (This is nominally the energy difference between the reactants and the products in eq 1, which is, in general, not equal to zero because the presence of the disordered background removes the apparent symmetry of the reaction.) Note that throughout this work we make the implicit assumption that the free energy change for the hopping process (ΔG) is well approximated by the total energy change for the hopping process (ΔE_{ij}). This assumption is supported by the fact that the nuclear structures of the initial and final states are essentially the same in the crossing region,³⁰ so entropic considerations to both states are plausibly very similar. To apply eq 4 it is necessary to compute the system-specific parameters λ , $V_{ij}(r_{ij})$, and ΔE_{ij} . We will now present details of the computation of each of these parameters.

2.4. Reorganization Energy, λ . The reorganization energy reflects the geometry relaxation in the two monomers involved in a charge transfer when their electronic state changes. In the case of hole transport, the reorganization energy is determined by four energies, (the Nelson four-point method):^{31,32}

$$\lambda = E_+^* - E_+ + E^* - E \quad (5)$$

where E_+ and E are the optimized energies of the cationic and neutral forms of a single monomer, E_+^* is the energy of the monomer cation at the neutral geometry, and E^* is the energy of the neutral monomer at the cation geometry.^{19–21,31,32} These energies were obtained from standard quantum chemical electronic structure calculations on a single monomer.

Monomer chemical structures of the materials studied in this paper are shown in Figure 1. Initial structures were fully optimized to obtain the reorganization energy. The cationic and neutral forms of each structure were fully optimized to obtain E_+ and E . E_+^* and E^* were obtained using single point energy calculations on the optimized neutral and cationic structures. Optimization was initially performed at the AM1 level³³ and then at the B3LYP/6-31G(d) level of theory.^{34–39} This two-step optimization procedure is primarily to reduce overall computational expense,^{11,40} but it also yields reorganization energies at both the AM1 and DFT levels of theory. These were both tested for use in computing mobility. An unrestricted model was used for the optimized structure in the cationic state. Some spin contamination was observed at the AM1 level, but spin contamination was typically negligible at the DFT level. All electronic structure calculations were carried out with the GAMESS program.⁴¹

2.5. Coupling Matrix Element, $V_{ij}(r_{ij})$. The coupling matrix element is dictated by overlap of molecular orbitals and is strongly dependent upon the relative position and orientation

of the neighboring molecules.^{5,10,11,19–21,42} Several methods are in common use for determining the coupling matrix element. These have been previously compared in detail^{21,30} and will therefore only be reviewed briefly here.

2.5.1. Dimer Splitting. A common approach is the dimer splitting method, where V_{ij} is expressed as

$$V_{ij} = \frac{1}{2} \sqrt{(E_{\text{HOMO}} - E_{\text{HOMO}-1})^2 - (\varepsilon_i - \varepsilon_j)^2} \quad (6)$$

where E_{HOMO} and $E_{\text{HOMO}-1}$ are energies of the highest occupied and the second highest occupied molecular orbitals (HOMO and HOMO-1) of the optimized dimer structure, respectively, and $\varepsilon_i(\varepsilon_j)$ is energy of site $i(j)$. Note that these energies always appear as a difference, the difference in energy between the two molecules that results from their unique positions within the material. This difference arises not only from the differing degree to which the two monomers in an asymmetric dimer polarize each other but also from polarization of the surrounding molecules. This latter intermolecular effect is neglected in the present work. Considerable discourse on site energies may be found in the appendix of refs 20 and 43. We summarize a few essential points here: The approximation of using dimer splitting is simple to apply in an ideal system but encounters problems for complex real systems.^{21,44} In an ideal system, the HOMO and HOMO-1 of the dimer are formed from a linear combination of the HOMOs of the two isolated monomers, but in a complex real system, there are many energetically closely placed orbitals, and it is difficult to identify the pair of split orbitals in the dimer that originate from the monomer HOMOs. Often there is sufficient mixing that no such pair exists. Detailed discussion of this matter may be found in ref 21.

In a crystal, the site energy difference term in eq 6 often vanishes because every molecule is equivalent in the crystal, feels an identical chemical environment, and is therefore polarized identically. When the interacting monomers are not symmetry related, however, each molecule has different chemophysical interactions, inducing a different degree of polarization in its surroundings when charged, and the site energy difference is not zero.^{11,21,42} The coupling matrix element in these asymmetric systems has been investigated by Valeev et al., who showed that the coupling matrix element in a noncofacially stacked dimer should consider the site energy correction because the monomers polarize each other differently.⁴²

One approximate method to estimate the site energy difference in the calculation of the coupling matrix element is to use the HOMO level of each isolated molecule in dimer for its site energy. This approach neglects the influence of the matrix surrounding the dimer, which we will show below is a severe approximation.

2.5.2. Semiempirical Approximation to the Coupling Matrix Element. Another way to estimate the coupling matrix element, which skirts the site energy problem, is to use a semiempirical approach. A few different methods have been introduced.⁴⁵ We tested one of the most common approximations, which assumes

$$V = 1.75S_{ij} \frac{(E_i + E_j)}{2} \quad (7)$$

Here, E_i and E_j are HOMO energies of the two isolated monomers, and S_{ij} is the overlap integral between the orbitals of the isolated monomer in the dimer geometry, $\langle \phi_i | \phi_j \rangle$. In this approximation, the coupling matrix element strongly depends on the relative orientation of the monomers within the dimer. The principal disadvantages of the semiempirical approximation are

that it invokes an empirical parameter (1.75) and neglects the effect of the matrix surrounding the dimer.

2.5.3. Fock Transfer Integral. Within a one-electron product description of the electronic structure, a more rigorous way to determine the coupling matrix element that also avoids the issue of site energy difference is by direct computation of the Fock transfer integral (FTI).^{21,44,46–53} Several direct calculation methods have been investigated.^{21,44,46–53} Troisi and Orlandi used the FTI to calculate the coupling matrix element in DNA and obtained satisfactory results in comparison with experiments.⁴⁴ In their method, the coupling matrix element is written as

$$V = \langle \varphi_{a,\text{HOMO}} | F_{ab} | \varphi_{b,\text{HOMO}} \rangle \quad (8)$$

where F is the Kohn–Sham–Fock matrix for the dimer and φ_{HOMO} is approximated by Löwdin orthogonalization of monomer HOMO orbitals. After orthogonalization through terms of second order, the coupling matrix element is given by

$$\begin{aligned} V &= \langle \varphi_a' | F | \varphi_b' \rangle \\ &= (\langle \varphi_a | \alpha + \langle \varphi_b | \beta \rangle | F | (\alpha | \varphi_b \rangle + \beta | \varphi_a \rangle)) \\ &= \alpha \beta \langle \varphi_a | F | \varphi_a \rangle + \alpha^2 \langle \varphi_a | F | \varphi_b \rangle + \beta^2 \langle \varphi_b | F | \varphi_a \rangle \\ &\quad + \alpha \beta \langle \varphi_b | F | \varphi_b \rangle \end{aligned} \quad (9)$$

where $\alpha = 1 + 3/8S^2$ and $\beta = -1/2S$, and S is the overlap integral of the monomer HOMOs in the dimer geometry.²¹

We calculated the coupling matrix element using all three different techniques: dimer splitting (with and without use of the monomer HOMO energies to approximate the site energies), semiempirical method given by eq 7,⁴⁵ and direct calculation of the FTI as given by eq 9.²¹ For the two acenes, the coupling matrix element was only calculated by dimer splitting and direct calculation of the FTI because the site energy difference vanishes for equivalent sites in the crystalline acenes. Finally, as detailed below, the FTI was selected for use in the present work.

2.6. Site Energies. Even when the coupling matrix element (V_{ij}) is computed by the FTI method, it is necessary to compute the site energy difference. This is because when predicting charge mobility, the site energy correction appears in two places. One case is when the coupling matrix element is calculated by the dimer-splitting method as expressed by eq 6. The other case is for introducing the energetic disorder of the hopping sites into eq 4, i.e., to evaluate the exponential term in eq 4. We tested two approaches for approximating the site energy difference. In the first approach, the HOMO level of each isolated monomer in the dimer geometry is taken as its site energy. The second approach is to approximate the site energy with an empirical potential. This potential is taken to be a sum of three separate potentials: Coulombic, non-Coulombic intermolecular, and intramolecular interaction.^{10,54–57}

$$V_{\text{total}} = V_{\text{Coulomb}} + V_{\text{non-Coulomb}} + V_{\text{intra}} \quad (10)$$

To calculate the Coulomb interactions, the Wolf method⁵⁸ was employed

$$\begin{aligned} V_{\text{Coulomb}} &= \sum_{i>j} \frac{q_i q_j}{4\pi\epsilon\epsilon_0} \left(\frac{\text{erfc}(\alpha r_{ij})}{r_{ij}} - \frac{\text{erfc}(\alpha r_c)}{r_c} \right) \\ &\quad + \left(\frac{\text{erfc}(\alpha r_c)}{r_c} + \frac{2\alpha}{\sqrt{\pi}} \frac{\exp(-\alpha^2 r_c^2)}{r_c} \right) (r_{ij} - r_c) \end{aligned} \quad (11)$$

where ϵ_0 is the permittivity of vacuum, ϵ is the dielectric constant, and q_i and q_j are the MOPAC charge for AM1 level (Mulliken charge for DFT level) at atom i and j , respectively. The dielectric constant is used as 3, which is typical of many electro-optic materials. The cutoff radius, r_c , was 12 Å, and the damping factor, α , was 0.2 \AA^{-1} .

The non-Coulombic intermolecular interactions are given by Grimme's method,⁵⁷ which is expressed as

$$V_{\text{non-Coulomb}} = -\frac{C_6^{ij}}{r_{ij}^6} \left(\frac{1}{1 + \exp\left(-d\left(\frac{r_{ij}}{S_R R_{vdW}} - 1\right)\right)} \right) \quad (12)$$

where r_{ij} is the atom–atom separation, C_6^{ij} is dispersion coefficient, R_{vdW} is equilibrium van der Waals separation, d is damping coefficient, and S_R is scaling factor for R_{vdW} . The geometric mean and simple average combination rules were employed for C_6^{ij} and R_{vdW} , respectively:

$$C_6^{ij} = \sqrt{C_6^i C_6^j}, \quad R_{vdW} = \frac{R_i + R_j}{2} \quad (13)$$

The dispersion coefficients and the van der Waals radii for the different atoms were obtained from Grimme's publication,⁵⁷ and d and S_r were set as 1 and 0.5, respectively.

The intramolecular interactions were employed by the amber force field as

$$V_{\text{intra}} = \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} K_\phi (1 - \cos(\phi - \phi_{eq})) \quad (14)$$

Where r_{eq} , θ_{eq} , and ϕ_{eq} are determined by the optimized structures. We set $K_r = 400 \text{ kcal/mol \AA}^2$, $K_\theta = 70 \text{ kcal/mol radian}^2$, and $K_\phi = 30 \text{ kcal}$.

3. CONVERGENCE TESTS

The charge mobility in molecular materials as approximated by Marcus–Hush theory depends on the reorganization energy and the coupling matrix element. These two parameters can be calculated by using electronic structure methods. In principle, any of the quantum electronic structure methods in the toolkit of computational chemistry can be applied to compute these quantities, but in practice, we require a computationally efficient approach. This is especially true for computing the coupling matrix element because of the large number of such computations involved in predicting hole mobility based on an ensemble of amorphous cells. Two approaches are reported here: First, both the reorganization energy and the coupling matrix elements were computed with the semiempirical AM1 method. Since many fewer calculations of the reorganization energy are required than calculations of the coupling matrix element, it is reasonable to apply more advanced methodology for the computation of reorganization energy. In the second approach, the reorganization energy is computed at the DFT/B3LYP/6-31G(d)^{34–39} level, and the coupling matrix elements are computed at the AM1³³ level. The second approach has proven successful in the past^{20,21} and is termed the “hybrid” method. Using these

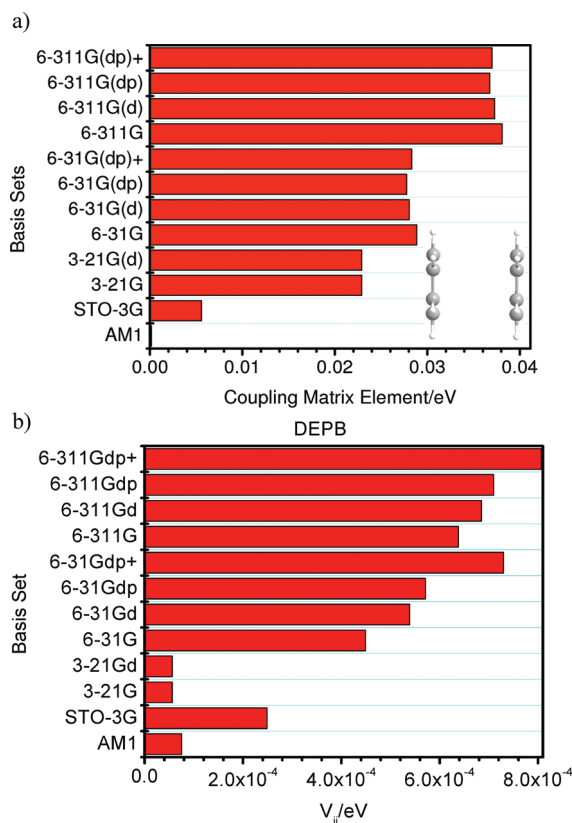


Figure 3. Coupling matrix element dependence on basis set as calculated by the FTI method: (a) parallel naphthalene and (b) DEPBB dimers.

approaches, we performed numerous convergence tests to identify an appropriate set of assumptions.

3.1. Basis Set Selection. To study the influence of basis set change, we calculated the coupling matrix element in the parallel naphthalene dimer and a DEPBB dimer as well as the reorganization energy of mBPD using an assortment of different basis sets.

As shown in Figure 3a, the coupling matrix element (FTI method) in the naphthalene dimer is more strongly influenced by the size of the basis set than by the addition of polarized or diffuse functions. The coupling matrix element appears to be reasonably well converged at the triple- ζ level and is only about 25% too small at the double- ζ level. With a minimal basis set, and at the AM1 level,³³ (which has an implicit minimal basis) the coupling matrix element is significantly underestimated. This underestimation arises from the fact that π – π overlap is significantly underestimated in a minimal basis. The parallel naphthalene dimer therefore represents an extreme case because the interaction is exclusively of the π – π type. The calculated coupling matrix element is influenced by the wave function overlap between the monomers in a dimer, which depends on their relative spatial orientation.⁵⁹ Figure 3b shows the results of a similar convergence test for an asymmetric DEPBB dimer. Note that while AM1 still underestimates the coupling matrix element, in this case it is too small by about a factor of 6, as compared to more than 100 in the case of the parallel naphthalene dimer.

Figure 4 shows the reorganization energy of mBPD as computed with different basis sets. The reorganization energy does not show strong basis set dependence. Consequently, the predicted mobility is not highly sensitive to the basis set used in the

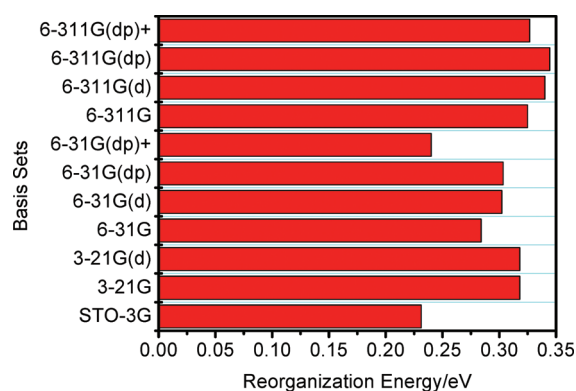


Figure 4. Reorganization energy of mBPD calculated using different basis sets.

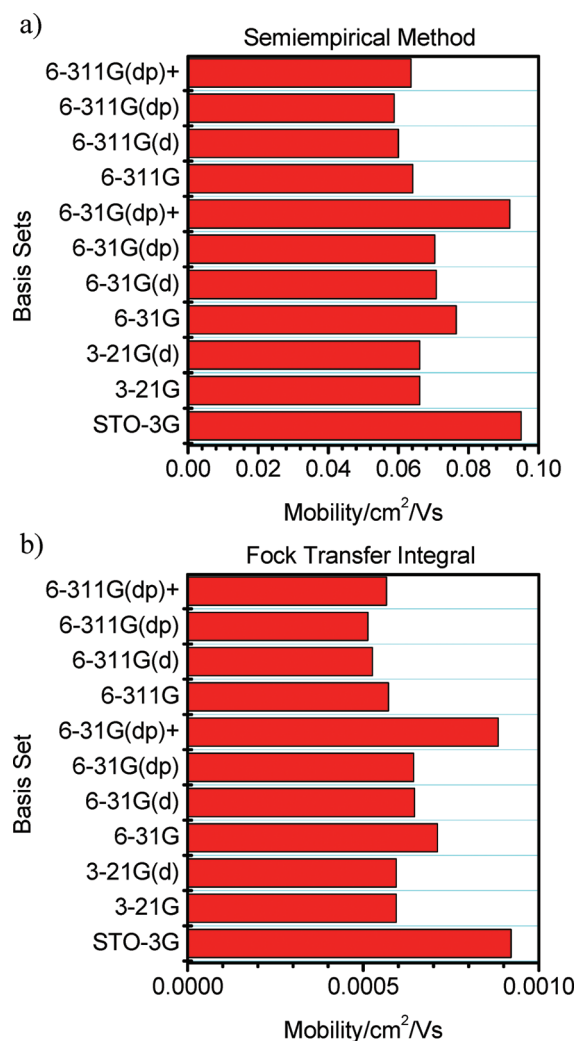


Figure 5. Dependence of predicted charge mobility of mBPD on basis set. The coupling matrix element was computed by (a) semiempirical and (b) FTI methods.

computation of λ . Even when a minimal basis set is used, λ is not grossly inaccurate.

Figure 5 shows predicted charge mobility of mBPD as computed with various basis sets. Tests were performed with the

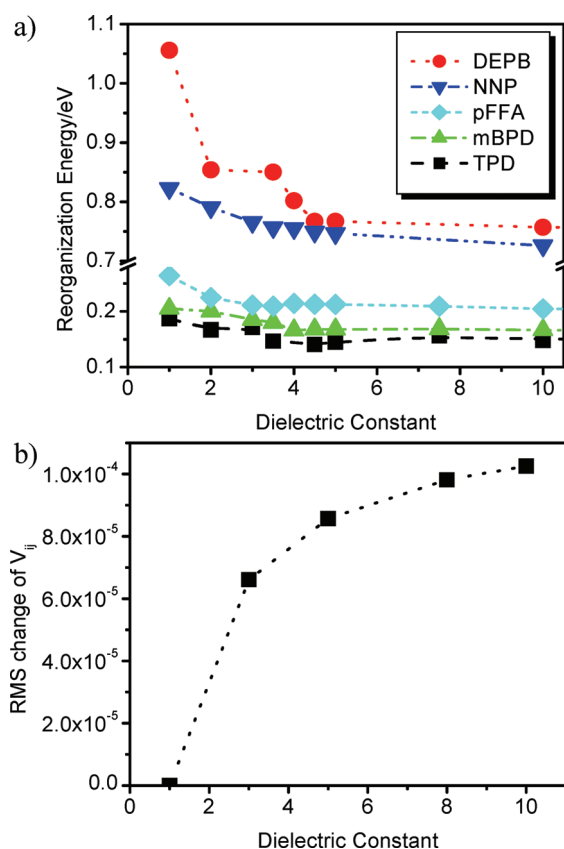


Figure 6. Dependence on dielectric constant for (a) reorganization energy and (b) coupling matrix element.

coupling matrix element calculated by the semiempirical method (Figure 5a) and by the FTI (Figure 5b). The predicted mobility does not change much with change in basis set, except with the 6-31G(dp)+^{36–39} and STO-3G⁶⁰ basis sets. This is similar to the case of the reorganization energy, which similarly shows significant deviations only for these two basis sets, suggesting that the sensitivity of the predicted charge mobility to basis set is dominated by the reorganization energy. Based on the above convergence tests, the 6-31G(d)^{36–39} basis set is selected for all DFT calculations as the best compromise of efficiency and completeness.

3.2. Dielectric Constant. The influence of the dielectric properties of the surrounding matrix on the reorganization energy is shown in Figure 6a. The reorganization energy decreases with increasing dielectric constant. This result demonstrates that the reorganization energy depends on the chemical environment. For this reason, $\epsilon = 3$ was used throughout this study, as this value is representative of many conducting organic materials.⁶¹

No significant dependence of the coupling matrix element on dielectric constant was found. Figure 6b shows the root-mean-square change in the computed coupling matrix element as computed with dielectric constant (ϵ) = 1.0 and as computed with dielectric constant set to the value given on the horizontal axis for an ensemble of 60 mBPD dimers. Note that the change is trivial, being smaller than typical values of the coupling matrix element by a factor of $\times 100$.

The reorganization energy reflects the structural relaxation that accompanies oxidation/reduction. In a high-dielectric medium, the relaxation of the molecule is restricted, so the

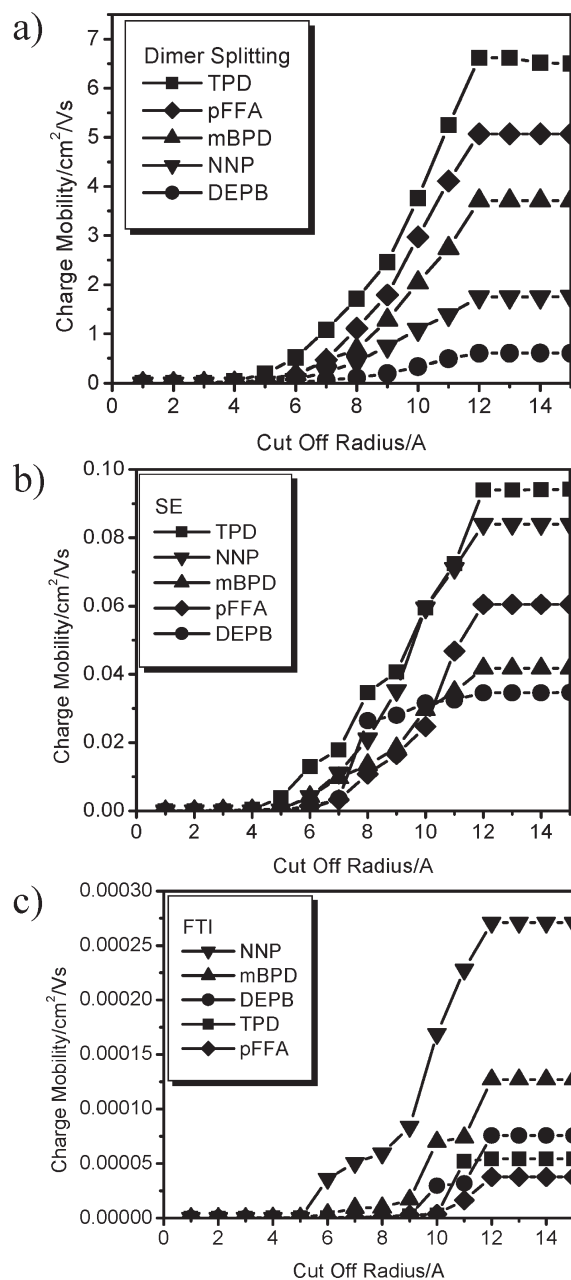


Figure 7. Dependence of predicted charge mobility on cut off radius for the five amorphous materials. B3LYP/6-31G(d) was used to compute reorganization energy and AM1 to compute coupling matrix elements by three different methods: (a) Dimer splitting method without the site energy correction was used, (b) dimer splitting method with HOMO level for the site energy was used, (c) FTI method with sum of empirical potential functions for the site energy was used.

reorganization energy gets smaller. In an amorphous material, each monomer is not equivalent because each feels different chemophysical interactions due to spatial disorder, so the reorganization energy has to include the effect by the local environment. This is the origin of the energetic disorder of the hopping sites.

3.3. Cut-Off Radius Selection. Figure 7 shows the computed mobility in TPD as a function of the cutoff radius used in the weighted average summation of hopping rates given by eq 3. Periodic boundary conditions were employed to find the neighbors

outside of the explicit unit cell. Figure 7a shows the case where V_{ij} is computed by dimer splitting. Figure 7b shows the case where V_{ij} is computed by the semiempirical method, and Figure 7c shows the case where V_{ij} is computed by FTI. In all cases the predicted mobility is converged for $R_c \geq 12$ Å, which was therefore the value used throughout the rest of the investigation.

3.4. Selection of the Method Used for Computing Coupling Matrix Element. The calculated mobility values for the five amorphous materials and two crystalline acenes wherein the coupling matrix element were computed by the dimer splitting method are shown in Figure 8 (denoted “D.S.”). The mobilities of the amorphous materials are greatly overestimated but not for the acenes. We can assume that the origin of this difference lies in the two parameters: reorganization energy and coupling matrix element. In a crystal of high symmetry, the dimer splitting method is often a good way to estimate coupling matrix element, but it is not good for amorphous materials because the site energy different term in eq 6 is not small and cannot be assumed to vanish. Therefore, to predict the charge mobility in amorphous materials adequately, the site energy correction is needed.

One way to approximate the site energy is to use the HOMO level of each monomer as its site energy. The predicted charge mobility for each of the systems studied here using dimer splitting with the HOMO energy approximation for site energy is shown in Figure 8 (denoted “HOMO”). The predicted values improve over the case here where dimer splitting alone is used, but they are still greatly overestimated and show essentially no correlation to experimental values from the literature. It seems dimer splitting, with or without HOMO level approximation to the site energy, does not work well in these amorphous systems.

A method of approximating the coupling matrix element that is one step more advanced than dimer splitting is to use a semiempirical method. Here we tested a method given by eq 7. The coupling matrix element is strongly related to the relative spatial orientation of the monomers in the dimer. This semiempirical method reflects the orientational dependence of coupling matrix element because it uses the overlap integral between the monomers in the dimer geometry. The predicted charge mobilities (Figure 8, denoted “Semi”) are improved over those predicted by D.S., but the values are still significantly overestimated and show weak correlation to experimental values, at best.

Using the FTI is another way to obtain the coupling matrix element that avoids the site energy difference problem.^{21,44,46–53} The predicted charge mobility using the coupling matrix element as computed with FTI method is shown in Figure 8c and d. The predicted values are much improved over those obtained using the semiempirical method of calculating the coupling matrix element. When the site energy difference is neglected, however, (SP = 0) there are still some substantial disagreements with the experimental results. As discussed in Section 2.6, we tested two different methods to approximate the energetic disorder. One method is using the HOMO level of the monomer to approximate its site energy. The other method is to approximate each site’s energy with an empirical potential function (denoted “EPE” in Figure 8). Agreement is better when the site energies are approximated using the empirical potential approach. The predicted charge (hole) mobilities for all of the amorphous materials are within about one order of magnitude of the experimental results. For the acenes, the predicted values are

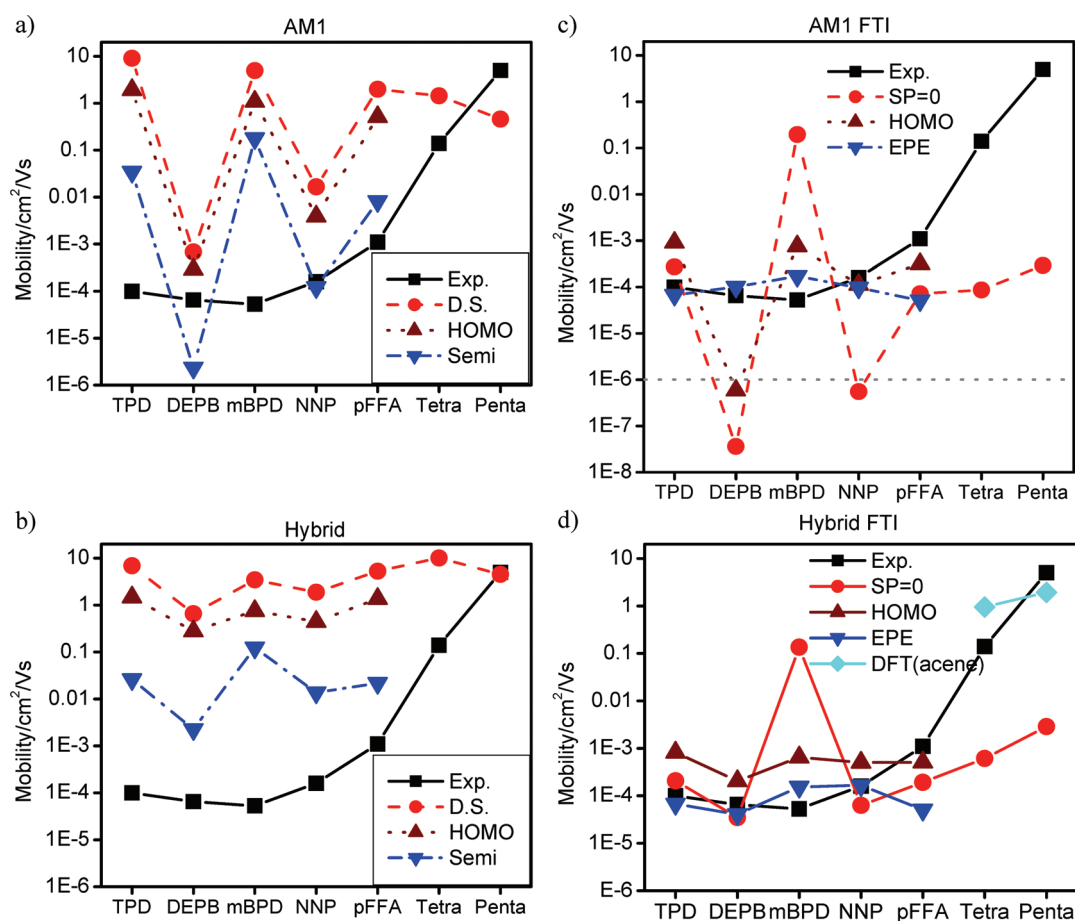


Figure 8. Predicted charge mobility for the five amorphous materials studied here: (a) using AM1 for all electronic structure calculations with the coupling matrix element calculated by three different methods; (b) using B3LYP/6-31G(d) to compute reorganization energy and AM1 to compute coupling matrix elements by three different methods; (c) using AM1 for all electronic structure calculations and three different methods to estimate the site energy difference; and (d) using B3LYP/6-31G(d) to compute reorganization energy, AM1 to compute coupling matrix elements and three different methods to estimate the site energy difference. [The dotted horizontal line depicts the lower limit of the vertical axis used in (a), (b), and (d) for comparison.] The technique marked “EPF” is the most robust approach considered here and is recommended. To obtain the DFT(acene) points, DFT was used for all electronic calculations.

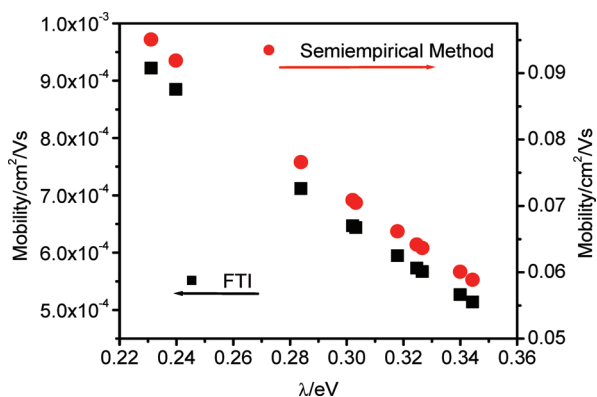


Figure 9. Charge mobility of mBPD as a function of the reorganization energy calculated using different basis sets. Black squares represent calculations where the coupling matrix element was computed with the FTI method and red circles with the semiempirical method.

severely underestimated in comparison to published experimental results. The origin of this error appears to be the small wave function overlap in the parallel dimer in the crystal of the

acenes (see Figure 2). Since π – π overlap is artificially low in small basis set calculations, the coupling matrix element is underestimated. When the coupling matrix elements are calculated at the DFT/B3LYP/6-31G(d)^{34–39} level of theory for the crystalline acenes (this is not computationally practical for the amorphous materials), we again see order of magnitude agreement with experimental results (denoted “DFT” in Figure 8d).

To further compare the semiempirical and the FTI methods, we plot the charge mobility of mBPD as a function of the reorganization energy in Figure 9. The different values of reorganization energy result from the use of different basis sets. The calculated charge mobility linearly decreases with increasing reorganization energy. The results from the semiempirical method and the FTI show the same slope, only the scale is different. This shows that the absolute magnitude of the mobility is strongly related to the reorganization energy, but the same basic physics of the charge transport is captured in both methods of estimating the coupling matrix element. It is reasonable to surmise that the two methods will predict qualitatively similar trends and therefore will be useful for screening materials based on relative mobilities.

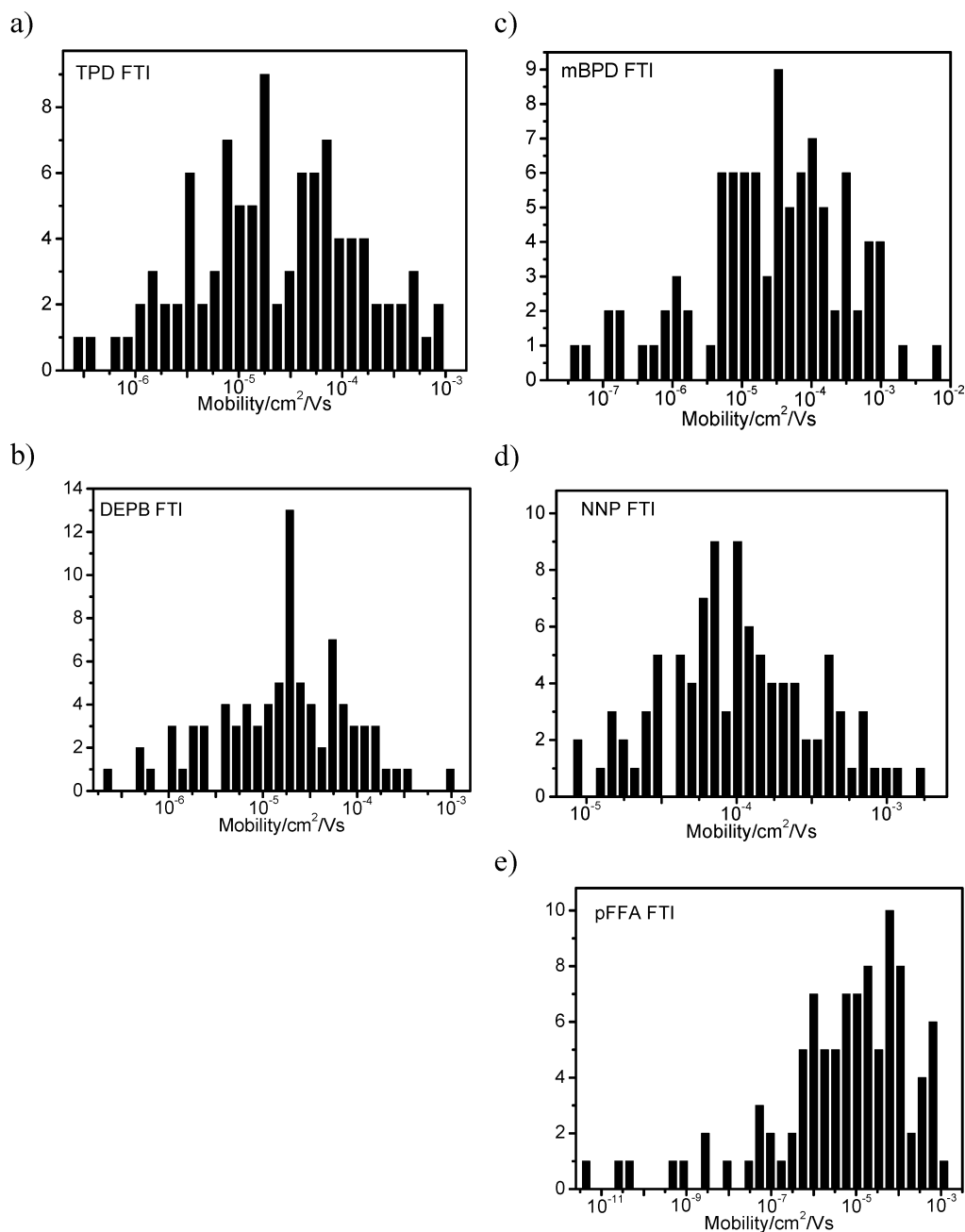


Figure 10. Histograms showing the computed hole mobilities for ensembles of ~ 100 amorphous cells for each of the five systems studied here: (a) TPD, (b) DEPB, (c) mBPD, (d) NNP, and (e) pFFA.

4. DISCUSSION

Histograms based on the predicted mobilities for ~ 100 amorphous cells of each of several different species considered here are presented in Figure 10. Note that the spectrum of values exhibits a clear peak in this log-scale plot. Throughout this work, the average of such an ensemble is reported as the computed mobility value. It is important to note, however, that the range of values is quite broad, typically spread over 2–3 orders of magnitude.

The calculated and (where available) experimental reorganization energy values for the species considered here are given in Table 2. The calculated values at DFT level (B3LYP/6-31G(d))^{34–39} agree well with the experimental values. The experimental values

of the charge mobility are also in Table 2. Two acene molecules show high charge mobility due to low reorganization energy. In amorphous materials, the charge mobility is less strongly related to the reorganization energy.

Figure 11 shows the dependence of charge mobility on material density. To evaluate this dependence, the cells were built up in two different ways. The first was by expansion of cell. In the expanded cell, the individual molecules in the cell have the same relative orientation but have different center to center distances. The second approach was reconstruction of cell, wherein the new cells were developed with different size cubic parameter, so the orientations or the monomers within the cells changed. The results are essentially the same in the two cases.

Table 2. Calculated Reorganization Energy and Corresponding Experimental Data for Several Amorphous Materials as well as Crystalline Tetracene and Pentacene

	TPD	DEPB	mBPD	NNP	pFFA	tetracene	pentacene
λ /eV	AM1	1.1023	0.2706	0.8395	0.3036	0.2684	0.2740
	DFT ¹	0.2740	0.3022	0.3803	0.2189	0.1119	0.09349
	expt.	0.29 ^a				0.1176 ^b	0.0992 ^b
	our work ²	6.7×10^{-5}	4.1×10^{-5}	1.5×10^{-4}	1.7×10^{-4}	5.1×10^{-5}	1.9
mobility, cm ² /V s	expt.	$\sim 10^{-4,c}$	$5-8 \times 10^{-5,c}$	$5.3 \times 10^{-5,d}$	$1.6 \times 10^{-4,d}$	$1.1 \times 10^{-3,e}$	3, ^h
						0.4 ^g	5~7 ⁱ

¹ DFT values were calculated at B3LYP/6-31G(d) theory level using GAMESS.^{34–39,41} Hybrid method was adopted for amorphous material and DFT for acenes. V_{ij} was calculated by FTI method. ^a Refs 62 and 63 ^b Refs 64 and 65 ^c Ref 66. ^d Ref 67. ^e Ref 68. ^f Ref 69. ^g Ref 70. ^h Ref 71. ⁱ Ref 72.

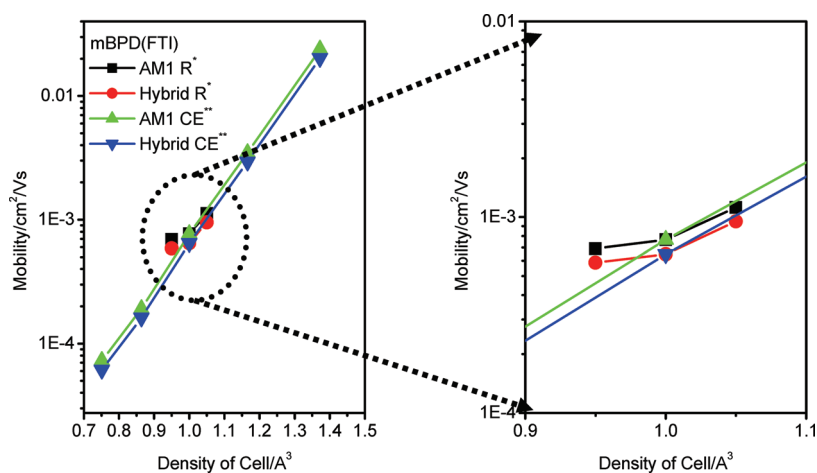


Figure 11. Predicted charge mobility of mBPD as a function of cell density; * indicates complete reconstruction of the cell and ** indicates simple cell expansion.

The charge mobility is exponentially increased with cell density, because the coupling matrix element strongly depends on wave function overlap and decreases exponentially with increasing center to center distance. Reliable estimation of material density is therefore critical for the accurate prediction of hole mobility in amorphous organic materials.

According to eqs 1–3, the charge-hopping rate depends exponentially on the reorganization energy and quadratically on the coupling matrix element. In consulting the above results, it is clear that for the amorphous materials, the charge mobility depends linearly on the reorganization energy and exponentially on the coupling matrix element. Since these two parameters, the reorganization energy and the coupling matrix element, influence the charge mobility, to improve the charge mobility of a material, those two parameters need to be controlled. The reorganization energy cannot be changed without altering the molecule. In our previous work¹¹ on phthalocyanine, we showed that the coupling matrix dominates charge mobility. In a strong conjugated system, like phthalocyanine and many other discotic molecules, the charge mobility depends more strongly on the coupling matrix element than on the reorganization energy.⁷ This arises because the system is not changed much during the redox process because of the presence of the strong resonance structure, so the reorganization is small. In an amorphous material, therefore, changing the orientation (changing the coupling matrix element) is a better way to control the charge mobility.

To study the relation between the predicted charge mobility and the material density, we plotted the predicted charge mobility as

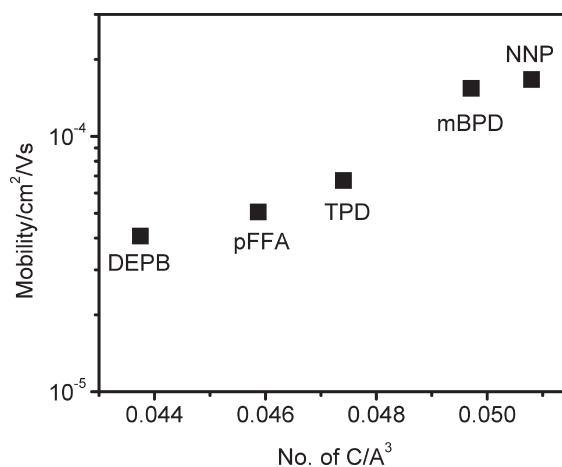


Figure 12. Predicted charge mobility of the amorphous materials as a function of number of carbon atoms in a unit volume. The coupling matrix element was calculated by the FTI method, and the site energy was estimated with an empirical potential (eq 10).

a function of the number of carbons in unit volume, as shown in Figure 12. There is a strong relation between the two parameters. The predicted mobility increases exponentially with the carbon density. This result is similar to the relation of the mobility and the cell density of mBPD shown in Figure 11. It again follows that the predicted mobility depends more strongly on the coupling

matrix than on the reorganization energy in the amorphous molecules. This result shows that to develop new material with high charge mobility, a potentially fruitful approach is to increase the density of the material.

5. CONCLUSIONS

The application of a generalization of Deng and Goddard's implementation of Marcus hopping theory to the estimation of hole mobility in amorphous organic materials has been presented here in detail. The approach is based on determining an average mobility based on an ensemble of amorphous cells representative of the material. In the most successful approach, the reorganization energy is calculated at the B3LYP/6-31G(d)^{34–39} level and includes an adjustment of the dielectric constant to match that of the material. The coupling matrix elements are calculated at the AM1 level by the FTI method based on orthogonalized monomer HOMOs. The AM1 method is used for computational efficiency, since thousands of such calculations are required. The hopping site energies are approximated with an empirical potential energy function that includes Coluombic, non-Coulombic, and intramolecular interaction terms. All adjacent monomers within 12 Å of the representative central molecule are included in the calculation.

The results show that unlike crystalline organic molecular solids, charge (hole) mobility is more highly dependent on the coupling matrix elements than on the monomer reorganization energy in these amorphous organic materials that have similar reorganization energy. Furthermore, the mobility is highly sensitive to material density, showing that a reliable estimate or mobility requires accurate prediction of the material density.

It is likely that the use of the AM1 method for the computation of the coupling matrix elements underestimated their magnitude quite significantly. It would be desirable to apply more advanced electronic structure methodology in their computation, but this is impractical due to the sheer number of such calculations that are required to apply hopping theory to an ensemble of amorphous cells. This is probably the most pressing issue that needs to be addressed to improve the reliability of predictions of charge (hole) mobility in these materials. An electronic structure method that provides the reliability of a split-valence basis DFT calculation at the cost of AM1 is needed. Nevertheless, the approach outlined here appears to provide order of magnitude estimates of charge (hole) mobility in amorphous organic materials.

AUTHOR INFORMATION

Corresponding Author

*E-mail: sohlbergk@drexel.edu.

ACKNOWLEDGMENT

This research was supported by National Science Foundation grant CHE0449595 and E. I. du Pont de Nemours & Co., Inc. Thanks also to Dr. K. Dobbs for his sustained encouragement of this research.

REFERENCES

- (1) Murphy, A. R.; Fréchet, J. M. J. *Chem. Rev.* **2007**, *107*, 1066.
- (2) Wang, L.; Nan, F.; Yang, X.; Peng, Q.; Li, Q.; Shuai, Z. *Chem. Soc. Rev.* **2010**, *39*, 423.
- (3) Kumar, A.; Liao, H.-H.; Yang, Y. *Org. Electron.* **2009**, *10*, 1615.

- (4) Dimitrakopoulos, C. D.; Mascaro, D. J. *IBM J. Res. Dev.* **2001**, *45*, 11.
- (5) Tant, J.; Geerts, Y. H.; Lehmann, M.; Cupere, V. D.; Zucchi, G.; Laursen, B. W.; Bjørnholm, T.; Lemaure, V.; Marcq, V.; Burquel, A.; Hennebicq, E.; Gardebien, F.; Viville, P.; Beljonne, D.; Lazzaroni, R.; Cornil, J. *J. Phys. Chem. B* **2005**, *109*, 20315.
- (6) Laschat, S.; Baro, A.; Steinke, N.; Giesselmann, F.; Hägele, C.; Scalia, G.; Judele, R.; Kapatsina, E.; Sauer, S.; Schreivogel, A.; Tosoni, M. *Angew. Chem., Int. Ed.* **2007**, *46*, 4832.
- (7) Andrenko, D.; Kirkpatrick, J.; Marcon, V.; Nelson, J.; Kremer, K. *Phys. Status Solidi B* **2008**, *245*, 830.
- (8) Lemaure, V.; Filho, D. A.; Da, S.; Corpceanu, V.; Lehmann, M.; Geerts, Y.; Pirijs, J.; Debije, M. G.; van de Craats, A. M.; Senthilkumar, K.; Siebbeles, A.; Warman, L. D.; Brédas, J. M.; Cormin, J.-L. *J. Am. Chem. Soc.* **2004**, *126*, 3271.
- (9) Crispin, X.; Cornil, J.; Friedlein, R.; Okudaira, K. K.; Lemaure, V.; Crispin, A.; Kestemont, G.; Lehmann, M.; Fahlman, M.; Lazzaroni, R.; Geerts, Y.; Windin, G.; Ueno, N.; Brédas, J.-L.; Salaneck, W. R. *J. Am. Chem. Soc.* **2004**, *126*, 11889.
- (10) Andrienko, D.; Marcon, V.; Kremer, K. *J. Chem. Phys.* **2006**, *125*, 124902.
- (11) Lee, C.; Sohlberg, K. *Chem. Phys.* **2010**, *367*, 7.
- (12) Tse, S. C.; Cheung, C. H.; Su, S. K. Charge transport and injection in amorphous organic semiconductors. In *Organic Electronics: Materials, Processing, Devices and Applications*; So, F., Ed.; CRC Press: Boca Raton, FL, 2010; pp 61–109.
- (13) Arkhipov, V. I.; Fishchuk, I. I.; Kadashchuk, A.; Bäessler, H. Charge transport in disordered organic semiconductors. In *Photophysics of Molecular Materials: From single molecules to single crystals*; Lanzani, G., Ed.; Wiley-CVH: Weinheim, Germany, 2006; pp 261–366.
- (14) Fishchuk, I. I.; Kadashchuk, A.; Bäessler, H.; Nešpůrek, S. *Phys. Rev. B* **2003**, *67*, 224303.
- (15) Bäessler, H.; Schweitzer, B. *Acc. Chem. Res.* **1999**, *32*, 173.
- (16) Bäessler, H. *Phys. Status Solidi* **1993**, *175*, 15.
- (17) Tsung, K. K.; So, S. K. *Appl. Phys. Lett.* **2008**, *92*, 103315.
- (18) Cheung, C. H.; Tsung, K. K.; Kwok, K. C.; So, S. K. *Appl. Phys. Lett.* **2008**, *93*, 083307.
- (19) Deng, W.-Q.; Goddard, W. A. *J. Phys. Chem. B* **2004**, *108*, 8614.
- (20) Rossi, M.; Sohlberg, K. *J. Phys. Chem. C* **2009**, *113*, 6821.
- (21) Rossi, M.; Sohlberg, K. *J. Phys. Chem. C* **2010**, *114*, 12173.
- (22) Schönherr, G.; Bäessler, H. *Philos. Mag. B* **1981**, *44*, 47.
- (23) Novikov, S. V.; Dunlam, D. H.; Kenkre, V. M.; Pattis, P. E.; Vannikov, A. V. *Phys. Rev. Lett.* **1998**, *81*, 4472.
- (24) Nešpůrek, S.; Sworakovsky, J. *Thin Solid Films* **2001**, *393*, 168.
- (25) *Material Studio*; Accelrys Inc.: San Diego, CA.
- (26) Rigby, D. *Fluid Phase Equilib.* **2004**, *217*, 77.
- (27) Robertson, J. M.; Sinclair, V. C.; Trotter, J. *Acta Crystallogr.* **1961**, *14*, 697.
- (28) Campbell, R. B.; Robertson, J. M.; Trotter, J. *Acta Crystallogr.* **1961**, *14*, 705.
- (29) Marcus, R. A. *J. Chem. Phys.* **1955**, *24*, 966.
- (30) Marcus, R. A.; Sutin, N. *Biochim. Biophys. Acta* **1985**, *115*, 265.
- (31) Wu, Q.; Voorhis, T. V. *J. Phys. Chem. A* **2006**, *110*, 9212.
- (32) Nelsen, S. F.; Blackstock, S. C.; Kim, Y. *J. Am. Chem. Soc.* **1987**, *109*, 677.
- (33) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (34) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.
- (35) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (36) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724.
- (37) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.
- (38) Hariharan, P. C.; Pople, J. A. *Mol. Phys.* **1974**, *27*, 209.
- (39) Gordon, M. S. *Chem. Phys. Lett.* **1980**, *76*, 163.
- (40) Lee, C.; Park, S. -K.; Min, K. -C.; Kim, Y.; Lee, N. -S. *Bull. Korean Chem. Soc.* **2008**, *29*, 1951.
- (41) 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.
- (42) Valeev, E. F.; Coropceanu, V.; da S. Filho, D. A.; Salman, S.; Brédas, J.-L. *J. Am. Chem. Soc.* **2006**, *128*, 9882.
- (43) Norton, J. E.; Brédas, J.-L. *J. Am. Chem. Soc.* **2008**, *130*, 12377.
- (44) Troisi, A.; Orlandi, G. *Chem. Phys. Lett.* **2001**, *344*, 509.
- (45) Hare, C. R.; Sleight, T. P.; Cooper, W.; Clarke, G. A. *Inorg. Chem.* **1968**, *7*, 669.
- (46) Kieninger, M.; Suhai, S. *J. Comput. Chem.* **1996**, *17*, 1508.
- (47) Fujita, T.; Nakai, H.; Nakatsuji, H. *J. Chem. Phys.* **1996**, *104*, 2410.
- (48) Yin, S.; Yi, Y.; Li, Q.; Yu, G.; Liu, Y.; Shuai, Z. *J. Phys. Chem. A* **2006**, *110*, 7138.
- (49) Orlandi, G.; Troisi, A.; Zerbetto, F. *J. Am. Chem. Soc.* **1999**, *121*, 5392.
- (50) Wu, Q.; Voorhis, T. V. *J. Phys. Chem. A* **2006**, *110*, 9212.
- (51) Wu, Q.; Voorhis, T. V. *J. Chem. Phys.* **2006**, *125*, 164105.
- (52) Yang, X.; Li, Q.; Shuai, Z. *Nanotechnology* **2007**, *18*, 424029.
- (53) Yang, X.; Wang, L.; Wang, C.; Long, W.; Shuai, Z. *Chem. Mater.* **2008**, *20*, 3205.
- (54) Williams, D. E. *J. Comput. Chem.* **2001**, *22*, 1154.
- (55) Nagata, Y.; Lennartz, C. *J. Chem. Phys.* **2008**, *129*, 034709.
- (56) Foster, M. E.; Sohlberg, K. *J. Chem. Theory Comput.* **2010**, *6*, 2153.
- (57) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
- (58) Wolf, D.; Keblinski, P.; Phillpot, S. R.; Eggebrecht, J. *J. Chem. Phys.* **1999**, *110*, 8254.
- (59) Brédas, J.-L.; Calbert, J. P.; da Silba Filho, D. A.; Cornil, J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 5804.
- (60) Slater, J. C. *Phys. Rev.* **1930**, *36*, 57.
- (61) Schildkraut, J. S.; Buettner, A. V. *J. Appl. Phys.* **1991**, *72*, 1888.
- (62) Lin, B. C.; Cheng, C. P.; Lao, Z. P. M. *J. Phys. Chem. A* **2003**, *107*, 5241.
- (63) Malagoli, M.; Brédas, J.-L. *Chem. Phys. Lett.* **2000**, *327*, 13.
- (64) Podzov, V.; Sysoev, S. E.; Loginova, E.; Pudalov, V. M.; Gershenson, M. E. *Appl. Phys. Lett.* **2003**, *83*, 3504.
- (65) Podzov, V.; Pudalov, V. M.; Gershenson, E. *Appl. Phys. Lett.* **2003**, *82*, 1739.
- (66) Mori, T.; Sugimura, E.; Mizutani, T. *J. Phys. D* **1993**, *26*, 452.
- (67) Shirota, Y.; Kageyama, H. *Chem. Rev.* **2007**, *107*, 953.
- (68) Okumoto, K.; Shirota, Y. *Mater. Sci. Eng. B* **2001**, *85*, 135.
- (69) Butko, V. Y.; Chi, X.; Ramirez, A. P. *Solid State Commun.* **2003**, *128*, 431.
- (70) de Boer, R. W.; Klapwijk, T. M.; Morpurgo, A. F. *Appl. Phys. Lett.* **2003**, *83*, 4345.
- (71) Klauk, H.; Halik, M.; Zschieschang, U.; Schmid, G.; Radlik, W.; Weber, W. *J. Appl. Phys.* **2002**, *92*, 5259.
- (72) Kelley, T. W.; Muyres, D. V.; Baude, P. F.; Smith, T. P.; Jones, T. D. *Mater. Res. Soc. Symp. Proc.* **2003**, *771*, L651.

Improved Prediction of Properties of π -Conjugated Oligomers with Range-Separated Hybrid Density Functionals

Ulrike Salzner* and Aykut Aydin

Department of Chemistry, Bilkent University, 06800 Bilkent, Ankara, Turkey

ABSTRACT: Range-separated hybrid functionals along with global hybrids and pure density functionals have been employed to calculate geometries, ionization energies (IP)s, electron affinities (EA)s, and excitation energies of neutral and oxidized polyenes, thiophene, and furan oligomers. Long-range correction with 100% HF exchange solves the problem of density functional theory with incorrect chain length dependence of IPs and energy gaps. There is a possibility of overcorrection, if the short-range part of the functional with no or low HF exchange is too small. The wB97XD functional with 22% of HF exchange in the short-range and a range-separation parameter of 0.2 seems to be just right for conjugated systems at all chain lengths. The wB97XD functional additionally produces negative orbital energies in very good agreement with IPs and EAs. With correct orbital energies, band gaps correspond to transport gaps (E_t) and not to optical gaps (E_g). E_t is much larger than E_g in the gas phase, but the difference is significantly smaller in the solid state. The accuracy of the negative orbital energies is good down to about 30 eV so that valence and innervalence PE spectra can be modeled. wB97XD is therefore suitable for calculating band structures of conjugated polymers employing orbital energies.

INTRODUCTION

Crucial properties for theoretical research on conducting organic polymers (COPs) are ionization potentials (IP)s, electron affinities (EA)s, and excitation energies of neutral and charged species. These properties relate to experimental observables and determine ease of doping, stability in neutral and doped states, optical band gap (E_g), and conductivity. Instead of calculating states, band structure calculations are often used, which means that orbital energies and orbital energy differences are used as approximations for energies and densities of states. Therefore, an adequate theoretical level should correctly predict IPs and EAs, and both should match the corresponding negative orbital energies. To evaluate the density of states, lower and higher lying orbital energies have to match states, too. Because all of the properties depend on geometries, i.e., bond length alternation (BLA) in neutral systems and defect sizes in ions, geometries have to be predicted accurately.

In recent years, DFT has gradually replaced semiempirical methods,^{1–4} Hartree–Fock (HF) theory,^{5–7} and perturbation theory^{8,9} in research on COPs because it includes electronic interactions self-consistently,¹⁰ accounts for electron correlation,¹¹ and avoids spin-contamination in open-shell systems.^{12,13} Despite many successes of DFT, a couple of vexing problems remain, however. Foremost, there is the incorrect chain length dependence of IPs¹⁴ and excitation energies¹⁵ of neutral systems, which leads to underestimation of polymer properties no matter whether periodic boundary conditions or extrapolation methods are applied. Another important issue is the mismatch between the negative energy of the highest occupied molecular orbital ($-\epsilon_{\text{HOMO}}$) and IP_{ΔSCF} ($E_{\text{cat}} - E_{\text{neutral}}$) as well as the experimental IP. $-\epsilon_{\text{HOMO}}$ should be identical with the IP_{ΔSCF} with the exact exchange-correlation functional,^{16,17} but it differs by a couple electronvolts with most approximate functionals.^{18–21} Finally, there is controversy about geometries, as pure DFT underestimates bond length

alternation,^{22,23} predicts larger defect sizes than semiempirical methods,²⁴ and does not confirm the existence of bound bipolarons in doped systems.^{25–27}

Some of the problems of DFT with extended π systems can be ameliorated by the inclusion of HF exchange.^{18,23,28} It was shown, for instance, that bond length alternation increases to reasonable values with hybrid functionals,²³ that defect sizes predicted with hybrid functionals agree with those at the MP2 level,¹² that HOMO–LUMO gaps ($E_{\text{H–L}}$) can be adjusted with HF exchange to reproduce E_g ,^{23,28} and that orbital energy differences agree better with differences between states.²⁹ Nonetheless, no global hybrid functional is able to predict the correct chain length dependence of IPs,¹⁴ E_g 's,¹⁵ or $E_{\text{H–L}}$'s.³⁰ These errors are a consequence of the self-interaction error (SIE) that arises in DFT because exchange is approximated.

The SIE is now addressed with range-separated density functionals^{31,32} that eliminate the SIE by using 100% HF exchange for the long-range part of the exchange-correlation functional.^{32–43} Therefore, SIE-free functionals with the correct long-range behavior of the exchange-correlation potential hold promise for improving the accuracy of theoretical results on extended π systems. It is the goal of this investigation to test a selection of new density functionals with respect to IPs, EAs, fundamental gaps (E_t) and E_g 's, bipolaron binding energies, and excitation energies of doped π systems. Guided by the availability of experimental data, polyenes, thiophene, and furan oligomers were chosen as representative systems.

METHODS

Structures of oligomers were optimized with a range of density functionals, as summarized below. If not stated otherwise, the

Received: May 21, 2011

Published: July 06, 2011

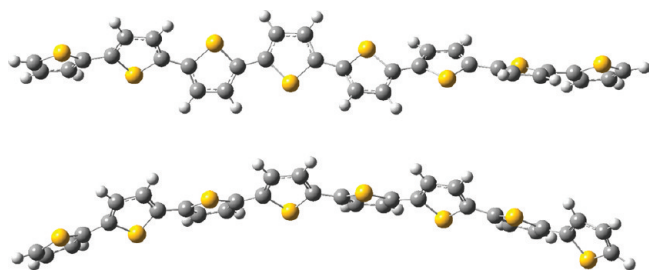


Figure 1. 8T with same sign dihedral angles (top) and alternating sign dihedral angles (bottom).

6-311G* basis set was employed. Thiophene and furan oligomers of different size are designated nT and nF , respectively, n being the number of rings. Polyenes⁴⁴ and furanes⁴⁵ are planar. Because bithiophene is nonplanar in the gas phase,⁴⁶ thiophene oligomers were allowed to deviate from planarity. There are two main structural alternatives: one with all interring dihedral angles having the same sign and one with alternating signs (see Figure 1). The latter are slightly lower in energy and have about 0.1 eV smaller excitation energies. This conformation leads to bending and may not be adopted in a crystal. The lower energy forms were used in the following because they are the lower energy structures in the gas phase and probably in solution. In all cases, planarization energies and influences on other properties due to bending are relatively small. Cation energies were obtained on the structures of the neutral forms to evaluate vertical IPs and with optimized geometries to evaluate UV spectra upon doping. Vertical IPs were calculated with the Δ SCF method as the difference between ground state energies of cation and neutral species and are abbreviated as $IP_{\Delta SCF}$.

Upon structure optimization, cations and anions become planar and straight with pure DFT and with global hybrids. With a range-separated hybrid functional, planarization occurs over five rings, but the chain ends remain slightly twisted, as shown in Figure 2 for $19T^+$ at the wB97XD/6-31G* level (description, see below). Excitation energies are calculated with time-dependent density functional theory. Most calculations were performed with Gaussian 09.⁴⁷ Calculations with BNL were done with QChem.⁴⁸

The density functionals are grouped into pure local density functionals, gradient corrected (nonlocal) functionals (GGA), meta-GGAs that depend explicitly on the kinetic energy density,^{42,49–51} global hybrid functionals that use a constant amount of exact (HF) exchange,⁴² and range-separated functionals³² that use different amounts of DFT and HF exchange for short and long ranges. For comparison, HF theory was included as well. The following density functionals were used:

Local Functionals. SVWN: Slater exchange⁵² and correlation functional of Vosko, Wilk, and Nussair,⁵³ identical to local spin density approximation (LSDA). SVWN5: same as SVWN with correlation functional V from ref 53.

Gradient Corrected (GGA) Functionals. BLYP: gradient corrected functional with Becke exchange⁵⁴ and correlation functional of Lee, Yang, and Parr.⁵⁵ BP86: gradient corrected functional with Becke exchange⁵⁴ and correlation functional of Perdew.⁵⁶ PBEPBE: Perdew, Burke, and Ernzerhof^{57,58} exchange and Perdew, Burke, and Ernzerhof correlation functionals.^{57,58} OLYP: OPTX modification by Handy⁵⁹ of Becke exchange⁵⁴ and Lee, Yang, and Parr correlation functional.⁵⁵

Meta-GGA (Containing Explicit Dependence on the Kinetic Energy Density).^{42,49–51} TPSS: Exchange and correlation functional of Tao, Perdew, Staroverov, and Scuseria.⁵⁰

Global GGA Hybrids (Constant Amount of HF Exchange). B3LYP: Becke's three-parameter hybrid functional with 20% HF exchange⁶⁰ and correlation functional of Lee, Yang, and Parr.⁵⁵ B3P86: Becke's three-parameter hybrid functional with 20% HF exchange⁶⁰ with the correlation functional of Perdew.⁵⁶ B3P86–30%: same as B3P86, but HF exchange is increased to 30%. PBE1PBE: also known as PBE0, hybrid version (25% exchange and 75% correlation weighting) by Adamo and Barone⁶¹ of Perdew, Burke, and Ernzerhof^{57,58} exchange and Perdew, Burke, and Ernzerhof correlation functionals.^{57,58}

Global meta-GGA Hybrids (Constant Amount of HF Exchange). M06: meta hybrid functional with 27% HF exchange. M06-HF: 100% HF exchange.

Range Separated Hybrid Functionals without Short-Range HF Exchange. All local, GGA, and meta-GGA functionals listed above were also used in their long-range corrected version. For the long-range correction (LC) Hirao et al.'s range separation scheme³² without short-range HF exchange and with a range-separation parameter $\gamma = 0.4$ was employed. LC-wPBE: long-range corrected^{35,42,49,51} Perdew, Burke, and Ernzerhof^{57,58} functional. wB97: range-separated version⁶² of Becke's 97 functional $\gamma = 0.4$.⁶³

Range Separated Hybrid Functionals with Short-Range HF Exchange. CAM-B3LYP: Coulomb attenuated version of B3LYP by Yanai et al.,⁶⁴ includes 19% HF exchange at short range and 65% HF exchange at long range, $\gamma = 0.33$. wB97X: same as wB97 but containing 16% short-range HF exchange and $\gamma = 0.3$.⁶² wB97XD: same as above with additional dispersion correction, containing 22% short-range HF exchange and $\gamma = 0.2$.⁶⁵

Range Separated Hybrid Functional with γ Tuning. BNL: Baer–Neuhauser–Livshits^{37,43} using LDA exchange, the LYP⁵⁵ correlation functional subtracting a small part of the Savin exchange.⁴³ γ Tuning means that the range-separation parameter is adjusted to reproduce $IP_{\Delta SCF} = -\epsilon_{HOMO}$. Decreasing the value of γ increases the extension of the short-range part. $\gamma = 0$ reduces range-separated functionals to pure DFT or global hybrid functionals with constant HF exchange equal to that of the basic functional.

RESULTS

Degree of Planarity of Thiophene Oligomers. $6T$ ⁶⁶ and $8T$ ⁶⁷ have been crystallized, and their structural parameters have been determined. Although $2T$ is nonplanar in the gas phase with an interring angle of 148° ,⁴⁶ $6T$ and $8T$ are planar in the crystal. We are not aware of gas-phase structures for $6T$ and $8T$. Theoretical results indicate that the interring dihedral angle in $6T$ is larger than in $2T$ but that $6T$ is slightly nonplanar in the gas phase. For instance, the interring angles are 138.6° ($2T$) and 145.7° ($6T$) at the MP2/6-311G* level; the energy for planarization of $6T$ is 1.4 kcal/mol. The agreement with the experiment for $2T$ is best with LC-BLYP, 147.0° ; CAM-B3LYP, 147.0° ; and B3P86–30%/6-311G*, 149.9° . Planarity increases more from $2T$ to $6T$ with global hybrids than with range-separated functionals to 150.6° at LC-BLYP, 152.2° at CAM-B3LYP, and 156.9° at B3P86–30%. In general, more HF exchange leads to stronger nonplanarity.

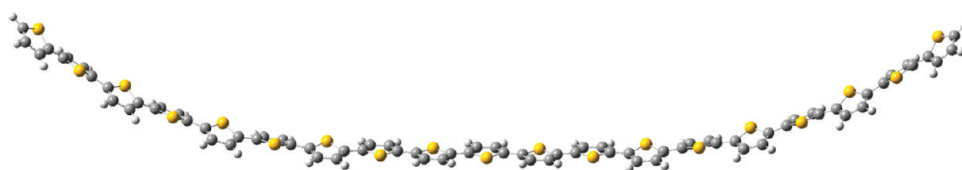


Figure 2. 19T⁺ optimized at wB97XD/6-31G*.

Table 1. Theoretical Bond Lengths and Dihedral Angles between the Innermost Rings for 6T and 6F

method	outer ring double	ring single	inner ring double	inter ring	BLA	dihedral
6T						
exptl ⁶⁶	1.370/1.376	1.400/1.403	1.367/1.380	1.444	0.070	180.0
LSDA/6-311G*	1.378	1.397	1.378	1.421	0.043	163.4
LC-SVWN/6-311G*	1.352	1.411	1.352	1.451	0.099	149.0
LC-SVWN5/6-311G*	1.353	1.413	1.353	1.453	0.100	148.6
BLYP/6-311G*	1.392	1.416	1.393	1.443	0.050	163.1
LC-BLYP/6-311G*	1.353	1.413	1.353	1.450	0.097	150.6
B3LYP/6-311G*	1.378	1.413	1.378	1.443	0.065	158.5
CAM-B3LYP/6311G*	1.366	1.416	1.366	1.450	0.084	152.2
BNL-0.182/6-311G*	1.400	1.440	1.400	1.473	0.073	149.9
LC-BP86/6311G*	1.354	1.409	1.354	1.446	0.092	151.2
B3P86/6-311G*	1.376	1.408	1.376	1.438	0.062	159.2
B3P86-30/6-311G*	1.371	1.409	1.371	1.441	0.070	156.9
PBE1PBE/6-311G*	1.375	1.409	1.375	1.440	0.065	158.3
LC-PBEPBE/6-311G*	1.354	1.408	1.354	1.446	0.092	150.7
LC-wPBE/6-311G*	1.362	1.416	1.362	1.456	0.094	149.1
M06/6-311G*	1.371	1.408	1.371	1.439	0.068	158.4
M06-HF/6-311G*	1.364	1.426	1.365	1.462	0.098	146.6
wB97/6-311G*	1.368	1.424	1.367	1.463	0.095	150.9
wB97x/6-311G*	1.366	1.420	1.366	1.458	0.091	150.7
wB97xd/6-311G*	1.368	1.417	1.368	1.453	0.084	150.2
wB97xd/6-31G*	1.371	1.418	1.371	1.452	0.081	154.4
OLYP/6-311G*	1.393	1.410	1.390	1.442	0.053	160.8
LC-OLYP/6-311G*	1.352	1.410	1.352	1.448	0.096	150.2
TPSSTPSS/6-311G*	1.389	1.411	1.389	1.438	0.049	169.3
LC-TPSSTPSS/6-311G*	1.353	1.409	1.353	1.447	0.094	150.2
HF/6-311G*	1.351	1.429	1.351	1.462	0.112	148.3
MP2/6-311G*	1.393	1.409	1.393	1.447	0.054	145.7
MP2/6-311G*	1.391	1.405	1.391	1.439	0.049	180.0
6F						
exptl ⁴⁵	1.362	1.408	1.358	1.432	0.074	180
LSDA/6-311G*	1.372	1.405	1.372	1.411	0.039	180
BLYP/6-311G*	1.386	1.424	1.387	1.431	0.044	180
LC-BLYP/6-311G*	1.347	1.420	1.347	1.436	0.089	180
B3LYP/6-311G*	1.372	1.420	1.372	1.429	0.057	180
CAM-B3LYP/6311G*	1.361	1.422	1.361	1.432	0.071	180
B3P86/6-311G*	1.370	1.415	1.370	1.425	0.055	180
B3P86-30/6-311G*	1.365	1.416	1.365	1.427	0.062	180
wB97xd/6-311G*	1.362	1.423	1.362	1.437	0.075	180

Bond Lengths of Thiophene and Furan Oligomers. The double bond lengths in the inner rings of 6T single crystals vary from 1.367 to 1.380 Å.⁶⁶ The ring single bond lengths of the two inner rings are 1.400 Å and 1.403 Å, and the central inter-ring single bond length is 1.444 Å. The BLA is therefore ~0.07 Å.

Bond lengths in 8T⁶⁷ are as follows: double bonds of inner rings, 1.38 Å; inner ring single bonds, 1.41 Å; and central inter ring single bond, 1.43 Å. BLA in 8T is therefore 0.05 Å. Because there is little difference between gas-phase structures of 6T and 8T at the B3LYP/6-311G* level, part of the differences between

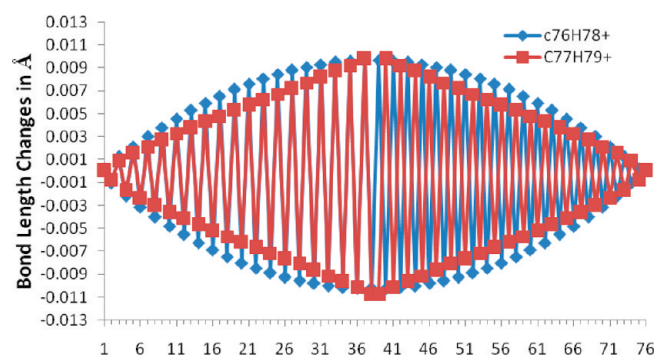


Figure 3. Differences in C–C bond lengths between $C_{76}H_{78}$ and $C_{76}H_{78}^+$ and between $C_{76}H_{78}$ and $C_{77}H_{79}^+$ at BLYP/6-31G*.

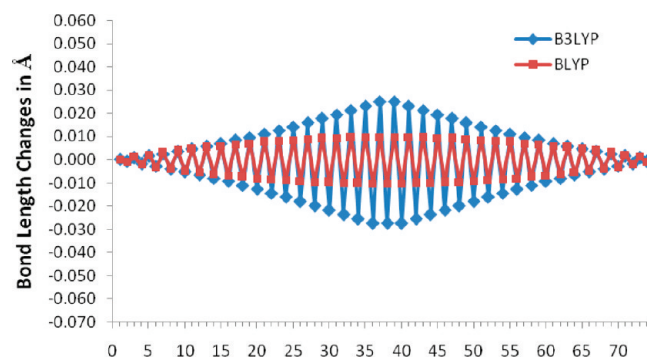


Figure 4. Differences in C–C bond lengths between $C_{76}H_{78}$ and $C_{75}H_{77}^+$ at BLYP/6-31G* and B3LYP/6-31G*.

experimental data of 6T and 8T might be due to experimental uncertainties. 6F has slightly shorter inner-ring double bonds (1.358/1.362 Å) and a shorter inter-ring single bond (1.432 Å) than 6T. The central inter-ring single bond in 6F is 1.408-Å-long.⁴⁵ BLA is thus 0.074 Å, slightly larger than in 6T. In Table 1, bond lengths of fully optimized nonplanar structures of 6T and of planar 6F at various levels of theory are summarized. Planarization has a little effect on the bond lengths in 6T.

The general trends are as follows: pure DFT (LSDA) tends to overestimate double bond lengths and to underestimate single bond lengths. BLA is therefore too small. Gradient correction (BLYP) predicts single and double bonds to be too long. Hybrid functionals (especially B3P86–30% and M06) are very accurate. Long-range correction shortens double bonds and comes sometimes close to values obtained with HF theory. Single bond lengths are overestimated but not as much as with HF theory. As a result, BLA is too large with all range-separated functionals. CAM-B3LYP with only 65% of long-range HF exchange and $\gamma = 0.33$ and wB97xd with a rather large short-range part ($\gamma = 0.2$) are doing better than the LC-functionals with 100% HF exchange and $\gamma = 0.4$. CAM-B3LYP and wB97XD perform better for 6F than for 6T. BNL predicts all bonds to be too long, but BLA is accurate. HF theory overestimates BLA because it overestimates single bond lengths and underestimates double bond lengths. Correlation at the MP2 levels shortens the double bonds correctly but lengthens the single bonds too much. The closest agreement between theory and experiment for 6T is reached with global hybrid functionals. For 6F, B3P86–30%, CAM-B3LYP, and wB97XD give the best results.

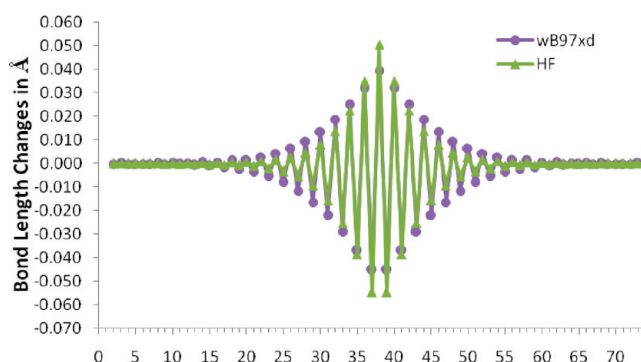


Figure 5. Differences in C–C bond lengths between $C_{76}H_{78}$ and $C_{75}H_{77}^+$ at wB97XD/6-31G* and HF/6-31G*.

Defect Sizes. Defect sizes of $C_{41}H_{43}^+$, $C_{101}H_{103}^+$, and $19T^+$ were previously investigated with pure DFT and with the B3P86–30% global hybrid functional using Stevens–Bach–Krauss pseudopotentials (SBK).^{12,13,24} Here, we are repeating some of the calculations with range-separated hybrids, i.e., LC-wPBE and wB97XD, and with the “BLYP series”, i.e., BLYP, B3LYP, CAM-B3LYP, and LC-BLYP. Because of the size of the systems, 6-31G* and SBK basis sets were used. Defect sizes decrease by about six bonds (or about one ring) with 6-31G* compared to SBK in $19T^+$ and $C_{77}H_{79}^+$ with the wB97XD functional. Differences in bond lengths between 6-311G* and 6-31G* basis sets can be seen in Table 1 for wB97XD. 6-31G* lengthens the double bonds and slightly shortens the single bonds compared to 6-311G*.

Test calculations were done on $19T^+$ and $25T^+$ and on $C_{75}H_{77}^+$ and $C_{101}H_{103}^+$ to determine the required chain lengths for converged defect size. With LC functionals, defect sizes are converged for $19T^+$ and $C_{75}H_{77}^+$. With global hybrids, defects are delocalized and do not converge with increasing chain length. Figure 3 shows differences in C–C bond lengths between neutral $C_{76}H_{78}$ and $C_{76}H_{78}^+$ and $C_{77}H_{79}^+$ cations with BLYP. The expectation value of the spin operator ($\langle S^2 \rangle$) for $C_{76}H_{78}^+$ is 0.77. Figure 3 shows that odd- and even-numbered systems have similar defect extensions and shapes when the wave function of the open-shell system is not spin-contaminated. As soon as HF exchange is included, unreasonably high $\langle S^2 \rangle$ values are obtained for long polyene cations, while global hybrids produce good $\langle S^2 \rangle$ values for thiophene oligomer cations.^{12,13} Spin contamination is avoided with polyenes by employing odd-numbered systems and investigating charged defects with closed-shell calculations.

The effect of HF exchange on defect size is established in Figures 4 (BLYP and B3LYP) and 5 (wB97XD and HF) for $C_{75}H_{77}^+$ (bond lengths are compared to those of $C_{76}H_{78}$, from which the innermost bond was deleted to achieve the same chain length and position of double bonds as in $C_{75}H_{77}^+$). Increasing HF exchange leads to defect localization. While defects are delocalized at BLYP and B3LYP, they spread over about 48 bonds (SBK) and 42 bonds (6-31G*) with the wB97XD functional and over about 30 bonds at the HF/6-31G* level. The criterion for determining defect size is that bond lengths in the cation differ by more than 0.002 Å from those of the neutral molecule.

Oligothiophene cations have one unpaired electron and have to be treated in the open-shell formalism. The $\langle S^2 \rangle$ values for $19T^+$ are BLYP, 0.75; B3LYP, 0.76; CAM-B3LYP, 0.88; wB97XD, 0.90; LC-wPBE, 1.09; and LC-BLYP, 1.11. The values

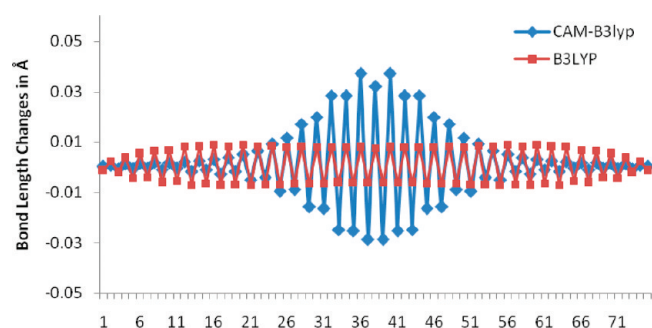


Figure 6. Differences in bond length between 19T and 19T⁺ at B3LYP/SBK (red squares) and CAM-B3LYP/SBK (blue diamonds).

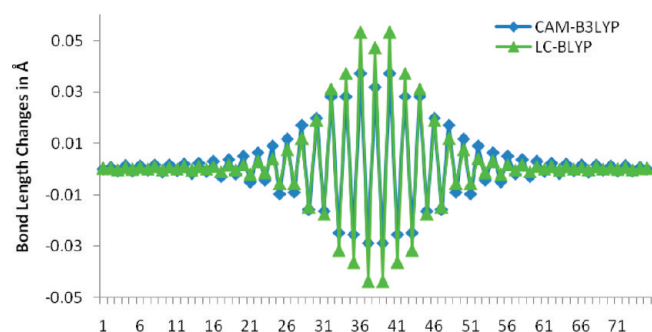


Figure 7. Differences in bond length between 19T and 19T⁺ at CAM-B3LYP/SBK (blue diamonds) and LC-BLYP/SBK (green triangles).

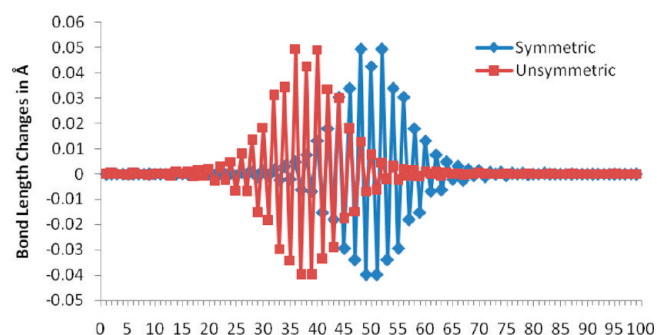


Figure 8. Localized defects at LC-wPBE/SBK in different positions on 25T⁺.

do not increase from 19T⁺ to 25T⁺. As shown in Figure 2, cations are nonplanar with range-separated hybrid functionals. The planarization energy is very small, however, e.g., 2.24 kcal/mol for 19T⁺ at wB97XD/6-31G*. Because the defect sizes are virtually the same in planar and nonplanar forms at wB97XD/6-31G*, planar cations were used in the following. In Figure 6, bond length changes upon ionization in 19T⁺ are shown at the B3LYP/SBK and CAM-B3LYP/SBK levels and in Figure 7 at CAM-B3LYP/SBK and LC-BLYP/SBK. The inclusion of large amounts of long-range HF exchange changes the defect shape and extension from totally delocalized at B3LYP with a tendency to split into two separate defects to a localized defect with a maximum in the middle of the chain. LC correction to BLYP localizes the defect more than CAM-B3LYP. LC-wPBE results (not shown) are almost identical to those with LC-BLYP. With the LC correction and with wB97XD, defects spread over about 9 or 10 thiophene rings. With

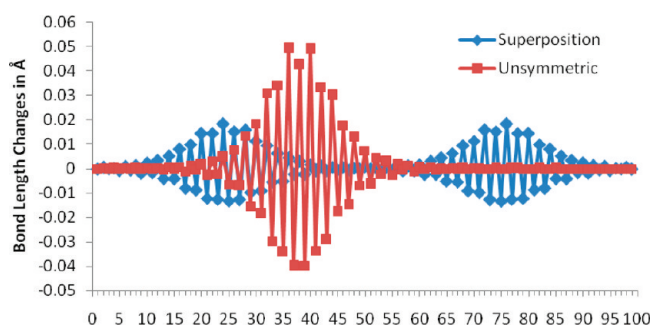


Figure 9. Two defect types on 25T⁺ at the LC-wPBE/SBK* level.

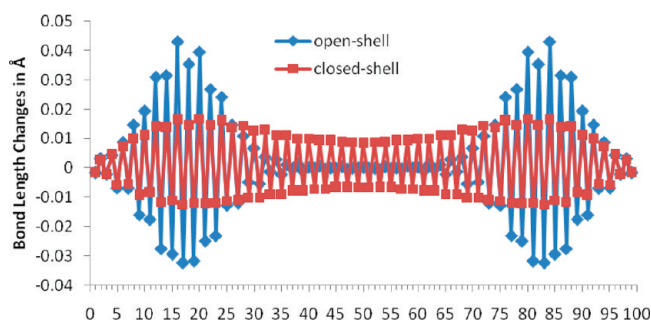


Figure 10. Comparison of defects on 25T²⁺ with closed- and open-shell calculations at the wB97XD/SBK level of theory.

CAM-B3LYP, the defects spread over about 15 rings. The defect size in 19F⁺ is 10 rings at wB97XD/6-31G*.

As demonstrated in Figure 8, it is possible to move the defect from the middle of the chain closer to one end. This move changes neither the energy of the cation nor the defect size and shape. Therefore, defects are converged with respect to chain length with the LC-corrected functionals.

It is also possible to obtain cations that exhibit two separated defects, as seen in Figure 9 at LC-wPBE/SBK. A cation with two half charges at either end of the chain is reminiscent of the DFT problem with the unphysical dissociation of symmetrical radical cations into two fragments with half charges. Such a state might be a superposition of two nonsymmetrical cations and a valid solution of the Schrödinger equation, as discussed by Vydrov and Scuseria⁴⁹ for symmetrical radical cations. The problem with pure DFT is that it underestimates the energy of such solutions and predicts them to be global minima, whereas they should be energetically degenerate with the radical/cation pair. HF, in contrast, overestimates the energy of the linear combination. At LC-wPBE, the split charge state lies 12 kcal/mol above the single defect cation, suggesting that the problem is overcorrected (too close to the HF result).

Defects in Dications. Geometries of 25T dications were optimized with LC-wPBE and wB97XD functionals. Figure 10 shows the defects obtained with closed-shell and open-shell ($\langle S^2 \rangle = 1.44$) calculations at wB97XD/SBK. It is clearly visible that even with closed-shell calculations there is a tendency for splitting the defect and placing the two charges at opposite ends of the molecule. Open-shell calculations lead to a complete separation of the two charges with an undistorted chain segment in the middle. Figure 11 shows that LC-wPBE ($\langle S^2 \rangle = 1.98$) leads to defects virtually identical to those of wB97XD/SBK. With both methods, the defect size is nine rings.

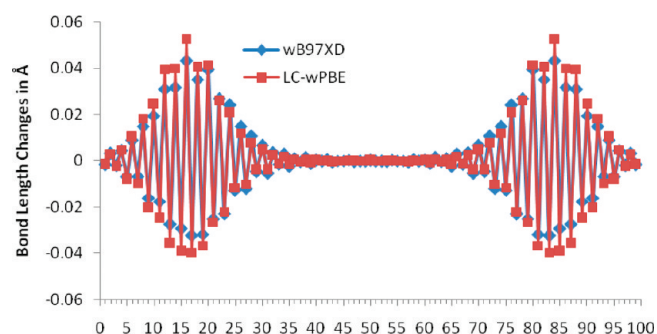


Figure 11. Comparison of defects on $25T^{2+}$ with open-shell calculations at the wB97XD/SBK and LC-wPBE/SBK levels of theory.

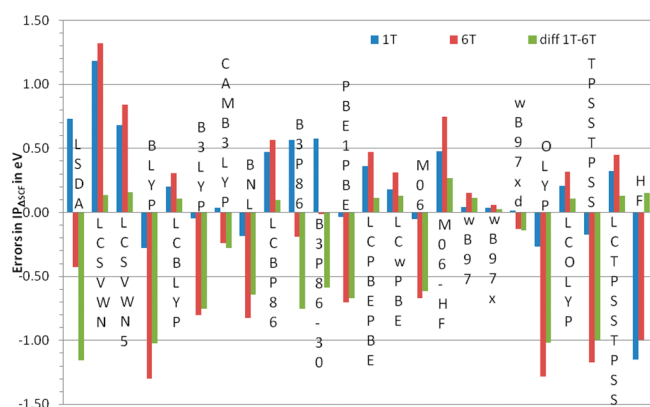


Figure 12. Errors in $IP_{\Delta SCF}$ of 1T and 6T at various levels of theory.

Ionization Energies. Gas phase ionization energies of neutral 1T through 5T were measured by Jones et al.,⁶⁸ for 3T through 8T by da Silva Filho et al.,⁶⁹ and for 1F through 4F by Distefano et al.⁷⁰ It was established recently¹⁴ that $IP_{\Delta SCF}$'s of 1T, 1F, and pyrrole are predicted quantitatively at the B3LYP/6-311G* level and that the higher IPs can be obtained by adding the TDB3LYP excitation energies of the cation to $IP_{\Delta SCF}$. For longer oligomers, the accuracy declines because B3LYP, like any other global hybrid functional, increasingly underestimates $IP_{\Delta SCF}$ with growing chain length. To find out whether range-separated hybrid functionals improve the chain length dependence, vertical $IP_{\Delta SCF}$'s were calculated for 1T and 6T. In Figure 12, the errors compared to the experimental IPs of 8.87 eV (1T)⁶⁸ and 6.98 eV (6T)⁶⁹ are plotted. 1T entries include a zero point energy (ZPE) correction of 0.059 eV. For 6T, the ZPE correction is negligibly small. The third (green) bar shows the difference in errors between the $IP_{\Delta SCF}$'s of 1T and 6T. A short green bar indicates therefore correct chain length dependence. Figure 12 reveals that all range-separated functionals with the exception of BNL have similar errors for 1T and 6T and therefore the correct chain length dependence. The absolute $IP_{\Delta SCF}$'s, however, are not necessarily correct. The functionals that predict $IP_{\Delta SCF}$'s of 1T and 6T within about 0.3 eV of the experiment are LC-BLYP, CAM-B3LYP (with some problems with chain lengths dependence), LC-wPBE, the wB97 series, and LC-OLYP.

Figure 13 depicts the chain length dependence of B3LYP, BNL, CAM-B3LYP, wB97xd, and LC-BLYP $IP_{\Delta SCF}$'s compared to experimental results. B3LYP $IP_{\Delta SCF}$'s fall off too fast, and BNL follows this trend. The reason is probably that with BNL, γ tuning by matching $IP_{\Delta SCF}$ and $-\epsilon_{HOMO}$ energy requires smaller

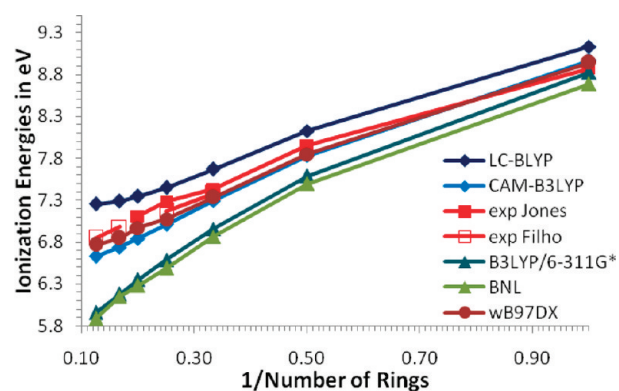


Figure 13. Chain length dependence of $IP_{\Delta SCF}$ of thiophene oligomers.

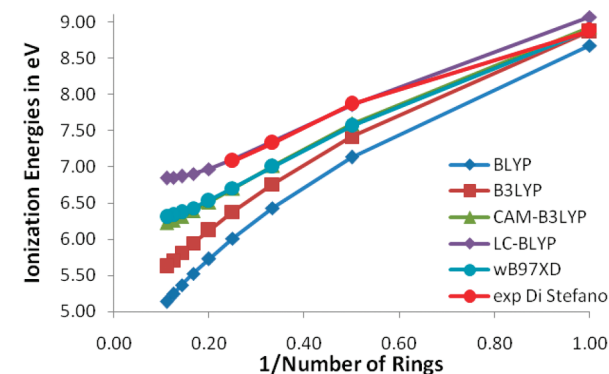


Figure 14. Chain length dependence of $IP_{\Delta SCF}$ of furan oligomers.

values with increasing chain length (for 1T, $\gamma = 0.352$; and for 8T, $\gamma = 0.168$). Thus, the short-range part increases for longer oligomers, and results get closer to those with global hybrids. CAM-B3LYP and wB97xd with constant γ are very close to experimental results and have the correct chain length dependence. LC-BLYP predicts values slightly above experimental values and a leveling off at a relatively short chain length. Such a convergence is not seen in the experimental data. $IP_{\Delta SCF}$'s extrapolated to infinite chain length with second degree polynomial fits are as follows exptl, 6.45 eV; CAM-B3LYP, 6.22 eV; wB97xd, 6.48 eV; LC-BLYP, 7.08 eV.

For furan oligomers IPs are available only up to a chain length of 4 rings.⁷⁰ Figure 14 compares the $IP_{\Delta SCF}$ of 1–9F with several functionals to these experimental values. The first IP of the monomer is predicted quantitatively with B3LYP, CAM-B3LYP, and wB97XD. However, in contrast to results for thiophene oligomers, CAM-B3LYP and wB97XD slightly overestimate the decrease with increasing chain length and underestimate the $IP_{\Delta SCF}$ of long oligomers. LC-BLYP slightly overestimates the IP of the monomer but agrees quantitatively with experimental results for the longer oligomers. LC-BLYP again predicts the onset of convergence, while B3LYP and BLYP predict accelerated decrease of $IP_{\Delta SCF}$'s with increasing chain length. Experimental IPs⁷⁰ extrapolate to about 6.4 eV, LC-BLYP to 6.66 eV, wB97XD to 5.98 eV, and CAM-B3LYP to 5.82 eV.

Excitation Energies and Optical Band Gaps. UV spectra become simpler as oligomers get larger. Because we are interested in long oligomers and polymers, monomers are not representative and were not considered. In Figure 15, errors in

the first excitation energies of 2T and 6T are plotted. The experimental E_g 's are 4.09 eV^{71–73} for 2T and 2.84^{71,72} or 2.92⁷⁴ eV for 6T. Solvent effects were shown theoretically to be small in ref 13. To calculate the errors, we used 4.19 eV as suggested by Andrzejak and Witek⁷³ as the vertical excitation energy for 2T and 2.92 eV for 6T.

For 2T, the HOMO–LUMO (1–1') transition is the leading electron configuration of the first excited state with a CI

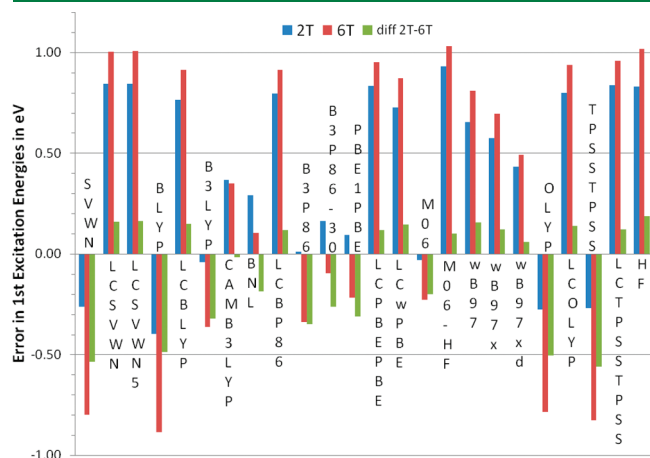


Figure 15. Errors in first excitation energies of neutral 2T and 6T at various levels of theory.

coefficient between 0.68 and 0.70 with all density functionals and at the HF level. There is no other contribution with a CI coefficient above 0.1. Double excitations, which are not included in TDDFT, are of little importance in these systems according to CASSCF calculations.^{75–77} The oscillator strength of the first excitation energy is between 0.38 and 0.40 at all levels of theory employed.

For the first excited state of 6T, pure DFT and global hybrids predict the 1–1' transition to be the only electronic configuration with a CI coefficient above 0.1; CI coefficients range from 0.69 to 0.71. At the HF level, the 1–1' transition has a CI coefficient of 0.58, and there are contributions arising from HOMO–1 to LUMO+1 (2–2') and HOMO–2 to LUMO+2 (3–3') with CI coefficients of 0.27 and 0.16. Range-separated functionals produce values very close to those of HF theory. Estimates of the oscillator strength range from 1.58 to 1.94 (Table 2).

Pure DFT underestimates excitation energies, and errors get larger with increasing chain length. Hybrid functionals are quite accurate for short oligomers, but the chain length dependence is wrong. Long-range correction leads to proper chain length dependence, but excitation energies are too high. Three functionals are better than the rest: B3P86–30% with the smallest absolute errors for 2T and 6T of +0.16 and –0.10 eV but the wrong chain length dependence because of the opposite signs of the errors, CAM-B3LYP with errors of 0.37 and 0.35 eV, and wB97xd with errors 0.43 and 0.49 eV.

Table 2. Spectroscopic Data (Energy (E) and Oscillator Strength (f), for 6T and 6T⁺, and $\langle S^2 \rangle$ for 6T⁺; energies in eV, basis set: 6-311G*)

exptl	6T			6T ⁺			6T ⁺ 1st excited state			6T ⁺ 2nd excited state		
	2.85, ⁷² 2.92 ⁷⁴		$\langle S^2 \rangle$	0.84 ^{74, 79}		$\langle S^2 \rangle$	1.59 ^{74, 79}		$\langle S^2 \rangle$			
	E	f		E	f		E	f				
LSDA	2.121	1.58	0.75	0.914	0.34	0.75	1.746	1.35	0.76			
LC-SVWN	3.924	1.90	0.95	1.222	1.17	0.95	2.120	1.11	1.31			
LC-SVWNS	3.929	1.90	0.96	1.231	1.19	0.97	2.082	1.07	1.34			
BLYP	2.036	1.62	0.75	0.864	0.25	0.75	1.646	1.75	0.76			
LC-BLYP	3.835	1.88	0.98	1.231	1.23	0.99	2.018	1.01	1.38			
B3LYP	2.557	1.86	0.77	0.942	0.45	0.76	1.786	1.78	0.88			
CAM-B3LYP	3.271	1.94	0.85	1.042	0.82	0.82	1.904	1.49	1.13			
BNL-0.182	3.025	1.86	0.78	0.908	0.94		2.001	1.47				
LC-BP86	3.835	1.92	0.98	1.218	1.25	0.99	2.063	1.00	1.42			
BP86	2.079	1.61	0.75	0.880	0.29	0.75	1.678	1.68	0.76			
B3P86	2.833	1.88	0.77	0.953	0.48	0.76	1.828	1.74	0.89			
B3P86–30	2.822	1.95	0.80	0.987	0.57	0.77	1.824	1.73	1.01			
PBE1PBE	2.704	1.92	0.78	0.969	0.54	0.77	1.835	1.72	0.96			
LC-PBEPBE	3.871	1.94	0.98	1.220	1.27	0.99	2.072	0.98	1.44			
LC-wPBE	3.794	1.92	0.98	1.188	1.21	0.98	2.005	1.01	1.43			
M06	2.691	1.85	0.78	0.934	0.52	0.77	1.801	1.67	0.94			
wB97	3.729	1.91	0.94	1.159	1.13	0.94	2.063	1.11	1.33			
wB97x	3.617	1.93	0.91	1.116	1.03	0.89	2.036	1.24	1.24			
wB97xd	3.411	1.96	0.86	1.048	0.87	0.83	1.998	1.45	1.12			
OLYP	2.136	1.62	0.75	0.893	0.31	0.75	1.707	1.71	0.76			
LC-OLYP	3.858	1.90	0.98	1.231	1.24	0.99	2.042	1.01	1.39			
TPSSTPSS	2.093	1.71	0.75	0.898	0.29	0.76	1.703	1.80	0.78			
LC-TPSSTPSS	3.880	1.93	1.01	1.237	1.33	1.02	2.018	0.92	1.48			
HF	3.938	1.87	3.03	1.033	1.86	3.85	5.440	0.13	3.54			

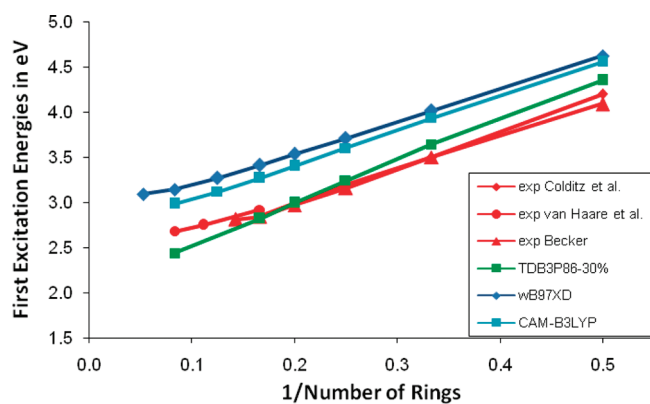


Figure 16. First excitation energies of neutral thiophene oligomers with different density functionals compared to experimental results.^{71,72,74}

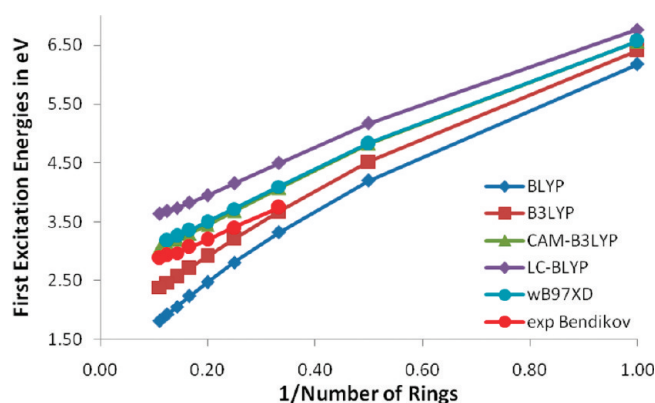


Figure 17. First excitation energies of neutral furan oligomers with different density functionals compared to experimental results.⁴⁵

In Figure 16, excitation energies of 2T–12T with B3P86–30%, CAM-B3LYP, and wB97XD are plotted versus inverse chain length. The extrapolated excitation energies (linear and polynomial fits give almost the same results) are 2.43 eV with the solution data of Colditz et al.,⁷² 2.47 eV according to the solution data of van Haare et al.,⁷⁴ 2.85 eV at wB97XD, 2.67 with CAM-B3LYP, and 2.00 eV at B3P86–30%.

The electronic configurations of excited states of furan oligomers are very similar to those described above for 2T and 6T. The excitation energies are less underestimated with global hybrids, and the errors with CAM-B3LYP and wB97XD are smaller than for thiophene oligomers (Figure 17). Extrapolation with second degree polynomial fits predicts the following ϵ_{\max} values for polyfuran: exptl,⁴⁵ 2.53 eV; CAM-B3LYP, 2.55 eV; and wB97XD, 2.65 eV. LC-BLYP overestimates first excitation energies of neutral oligomers by 0.75 eV and extrapolates to 3.22 eV. BLYP and B3LYP predict polymer values of 0.92 and 1.67 eV.

Excitation Energies of Cations. The 2T cation was produced in a matrix at 77 K. It has a strong and a very weak absorption in the UV spectrum.⁷⁸ 2T⁺ and 3T⁺ cannot be produced in solution. 3T seems to dimerize upon oxidation as the recorded absorption spectrum is identical to that of 6T⁺.⁷⁰ Starting with 4T, cations were observed in solution.^{74,79} Upon oxidation of 4T to 9T,^{74,79} the absorption of the neutral form disappears, and the two new features appear at lower energy. Both bands decrease in energy with increasing chain length. For 12T, the absorption of the

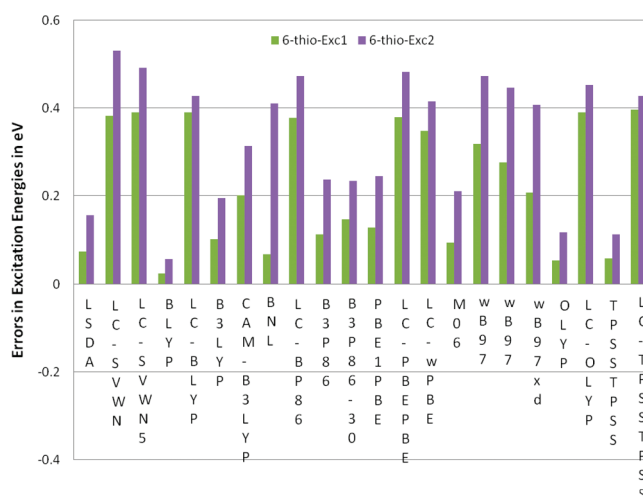


Figure 18. Errors and first and second excitation energies of 6T⁺.

neutral form decreases to half its original intensity upon oxidation, while two new features appear.⁷⁴

Theoretical analysis of the 2T cation spectrum shows that there are four low energy bands, two of them with negligible oscillator strengths.^{13,78} Double excitations contribute between 4 and 7% in the ground and excited states.⁸⁰ Because the lowest energy transition of 2T is not dominated by the 2–1 transition, 2T is not representative for longer oligothiophenes. For 3T to 9T, global hybrid functionals predict two bands for cations, in agreement with experimental results.¹³ The lower energy transition of oxidized thiophene oligomers is due to an electronic transition from HOMO to SOMO (2–1 transition) with a CI coefficient between 0.73 and 0.94. The smaller coefficients are caused by deexcitations, and there is no correlation between CI coefficients and HF exchange. The second excited state is dominated by a transition from SOMO to LUMO (1–1' transition) with a CI coefficient between 0.66 and 0.88. Again, the size of the CI coefficients appears to be random with respect to the functionals, and density functionals that predict small CI coefficients for the 1–2 transition may predict large ones for the 1–1' transition. With all functionals, there are small contributions of the 1–1' transition to the 1–2 state and vice versa. Between 8T and 12T, the 1–2 and 1–1' transitions do not mix anymore, but the 1–1' transition starts to interact with 2–2' and 3–1' electronic configurations and splits into two peaks.¹³ In agreement with experimental results, the third band of 12T⁺ occurs at the same position as the absorption of neutral 12T at the same level of theory.

Excitation energies of 6T and 6T⁺ are summarized in Table 2. Pure DFT and global hybrids produce $\langle S^2 \rangle$ values that are close to the correct value of 0.75 for the ground state of 6T⁺. HF predicts a value of 3.02, and long-range corrected functionals predict values between 0.85 and 1.0. Normally, a deviation of 10% is considered to be acceptable. CAM-B3LYP and wB97XD with values of 0.85 and 0.86 are therefore borderline. Values of around 1 were shown to produce useless spectra for polyene cations.¹² $\langle S^2 \rangle$ values of the first excited state are only slightly above those of the ground state, but spin contamination of the second excited state is problematic. Only pure DFT is adequate in terms of $\langle S^2 \rangle$ for the second excited state; global hybrids are borderline. Overall, there is a pretty good correlation between $\langle S^2 \rangle$ values and excitation energies, with larger spin contamination leading to higher excitation energy.

Table 3. Ground state and Excited State $\langle S^2 \rangle$ Values and Excited State Energies and Oscillator Strengths in the Absorption Spectrum of $12T^+$ with Different Density Functionals

state	0		1		2			3		
exptl74	0.62		1.45		2.68					
	$\langle S^2 \rangle$	E	f	$\langle S^2 \rangle$	E	f	$\langle S^2 \rangle$	E	f	$\langle S^2 \rangle$
BLYP	0.75	0.48	1.02	0.75	1.11	1.70	0.83	1.58 1.60	1.06 0.61	0.81 0.86
B3LYP	0.77	0.48	1.53	0.76	1.35	1.57	1.44	1.96 2.22	1.06 0.74	1.94 0.96
B3P86–30%	0.78	0.42	1.59		1.67	0.81		2.50 2.52	0.48 1.68	
CAM-B3LYP	0.89	0.55	2.06	0.87	1.53 1.80	1.25 0.58	1.54 1.77	2.87 3.23	0.74 0.57	1.26 0.96
LC-BLYP	1.10	0.90	2.29	1.20	1.55 1.90	0.48 0.75	2.68 1.79	3.41 3.50 3.83	0.83 0.32 0.29	1.32 1.43 1.38
wB97XD	0.90	0.62	1.98	0.90	1.60 1.75	0.71 1.23	2.27 1.17	2.99 3.04 3.70	0.66 0.39 0.46	1.03 1.01 0.94
LC-wPBE	1.09	0.84	2.27	1.18	1.51 1.87	0.43 0.81	2.70 1.81	3.30 3.40 3.70	0.58 0.58 0.29	1.34 1.32 1.39

Figure 18 shows errors in excitation energies compared to experimental values. The best agreement with experimental results is achieved with BLYP. OLYP and TPSSTPSS are also quite accurate. The more HF-exchange is used, the more the results deviate from experimental results. Long-range corrected functionals perform worse than global hybrids. Among the long-range corrected functionals, CAM-B3LYP is the best, followed by wB97XD. This is in line with the spin contamination. With all functionals, the error in the second excitation energy is larger than in the first.

The influence of HF exchange on cation spectra of $12T^+$ is summarized in Table 3 and illustrated in Figure 19. For these calculations, the smaller 6-31G* basis set was used. (For $6T^+$, the differences in results with 6-311G* and 6-31G* basis sets are about 0.1 eV.) Spin contamination does not increase from $6T^+$ to $12T^+$ for the ground state (GS) and first excited states. For excited states 2 and 3, only pure DFT produces acceptable $\langle S^2 \rangle$ values.

The effect of HF exchange is spreading the peaks over a wider range and splitting them into several contributions. Oscillator strengths of the first peak increase while the oscillator strengths of the higher energy peaks decrease with increasing HF exchange. The vertical lines in Figure 19 at the experimental peak positions show that BLYP underestimates the peak separations and that LC-BLYP overestimates them. The agreement of peak positions with experimental results is best with CAM-B3LYP and B3P86–30%, which seems to produce oscillator strengths that match the experimental peak heights better. The performance of pure DFT is not as good as for $6T^+$. The close lying peaks are most likely a result of the too close lying orbital energies with pure DFT.

Orbital Energies. With the exact exchange-correlation functional, $-\varepsilon_{\text{HOMO}}$ should match $\text{IP}_{\Delta\text{SCF}}$. The BNL functional produces orbital energies that fulfill this requirement,²⁰ but the BNL functional suffers from incorrect chain length dependence of $\text{IP}_{\Delta\text{SCF}}$ for extended π systems. Because γ is tuned to equalize $-\varepsilon_{\text{HOMO}}$ with $\text{IP}_{\Delta\text{SCF}}$, $-\varepsilon_{\text{HOMO}}$ gets too small as well. It is not possible to find a value for γ that simultaneously produces a correct $\text{IP}_{\Delta\text{SCF}}$ and a match between $\text{IP}_{\Delta\text{SCF}}$ and $-\varepsilon_{\text{HOMO}}$ for long oligomers.^{20,81} LC functionals with fixed γ were also shown to improve orbital energies of small molecules because they achieve cancellation of the Coulomb self-repulsion through the exchange self-interaction term,⁸² as it is known from HF theory.

Differences between $-\varepsilon_{\text{HOMO}}$ and experimental IPs are plotted in Figure 20, differences between $-\varepsilon_{\text{HOMO}}$ and $\text{IP}_{\Delta\text{SCF}}$

in Figure 21. Figure 20 reveals that all pure DFT and global hybrid functionals grossly underestimate $-\varepsilon_{\text{HOMO}}$ IPs. Range-separated functionals lower the orbital energies, and most of them overestimate IPs. Excellent agreement with experimental results is achieved with HF and the wB97 series, wB97XD producing an almost perfect match. Results with CAM-B3LYP and BNL are acceptable.

Figure 21 shows whether $-\varepsilon_{\text{HOMO}}$ IPs are consistent with $\text{IP}_{\Delta\text{SCF}}$'s at a given level of theory, i.e., whether a theoretical level leads to internally consistent results without reference to experimental results. BNL is internally consistent by definition. HF, which yields very good $-\varepsilon_{\text{HOMO}}$ IPs, is internally inconsistent because the $\text{IP}_{\Delta\text{SCF}}$'s are too low. Long-range corrected functionals suffer from overcorrection of the orbital energies and predict larger $-\varepsilon_{\text{HOMO}}$ IPs than $\text{IP}_{\Delta\text{SCF}}$'s. The best results in terms of accuracy compared to experimental results, internal consistency, and chain length dependence are achieved with wB97XD. CAM-B3LYP is acceptable.

The proof that $-\varepsilon_{\text{HOMO}}$ is equal to the first IP with the unknown exact exchange-correlation functional cannot be easily extended to other orbital energies.^{83,84} Therefore, controversy exists about lower lying orbitals and about unoccupied orbitals. The energies of the relevant states can be calculated by adding excitation energies of cations that create holes in the appropriate orbitals to $\text{IP}_{\Delta\text{SCF}}$. The idea to model photoelectron spectra in this way is described in refs 14 and 20.

Figure 22 compares higher ionic states with negative energies of lower lying orbitals for 1T at the wB97XD/6-311G* level. In the valence region, there is a one to one correspondence between negative orbital energies and the corresponding states. The states arise from electronic transitions of β electrons from lower lying orbitals into the SOMO. The contributions (CI coefficients) of these transitions are above 0.98 in the energy range from the first IP down to 15 eV. This means that these states are single-configurational and that the ionization is a single electron process. States with higher ionization energies than 16 eV are multiconfigurational. Plotted in Figure 22 are all states that have a contribution from an electron configuration that has a hole in one of the β orbitals. The lengths of the peaks are the squared CI coefficients of these electronic configurations. It can be seen that the states with holes in β orbitals cluster around the negative energy of the corresponding orbitals. For the whole energy range from 8 to 28 eV, there is an excellent correspondence between states and orbital energies. Figure 23 shows that the match is even better with the BNL functional.

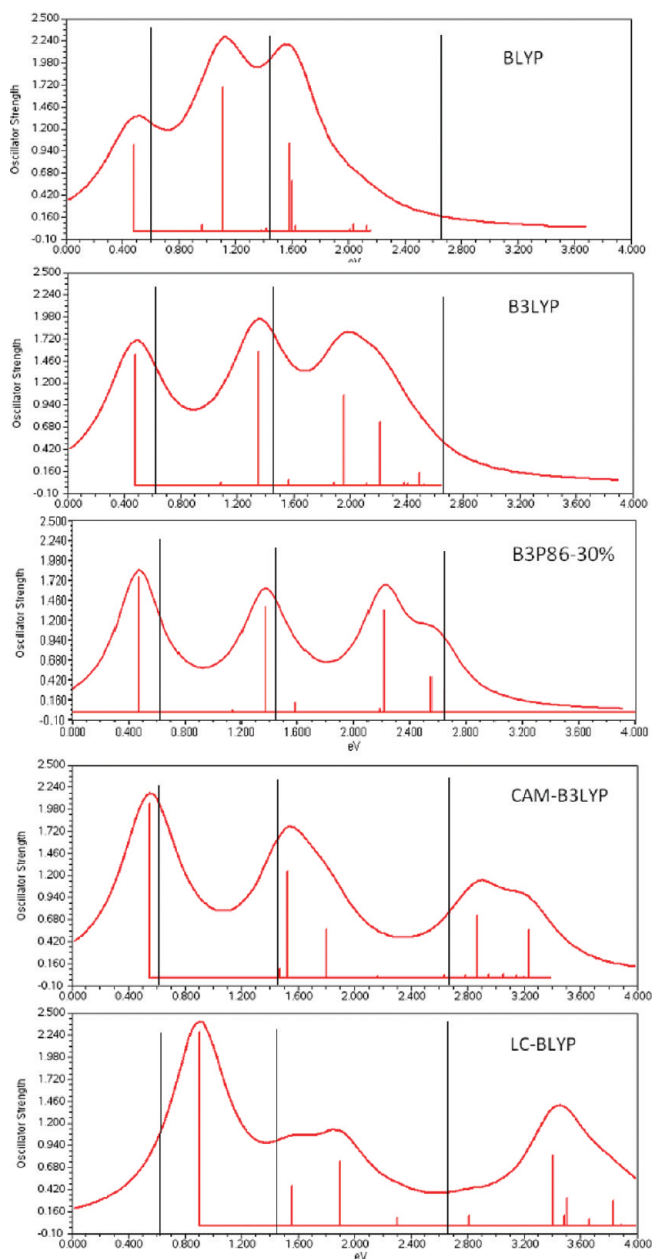


Figure 19. Spectra of $12T^+$ with different density functionals and the 6-31G* basis set. Experimental peak positions are shown as black vertical lines.

Figure 24 compares the wB97XD/6-311G* orbital energies with experimental peak positions. The excellent match of negative orbital energies with photoelectron peaks (PE) shows that the wB97XD functional is capable (like BNL) of predicting reliable orbital energies for small systems without any adjustments.

What is left to show is that the results are not deteriorating with increasing chain lengths, a problem that cannot be solved with BNL. Figure 25 compares negative orbital energies and states to PE peaks of 4T. The resolution of the experimental 4T spectrum is much less than that of 1T, but the overall agreement between peak position and negative orbital energies holds over the whole range of available PE peaks.

The only thiophene oligomer for which experimental EAs seem to exist is 2T. 2T has a positive EA of 0.049 ± 0.005 eV

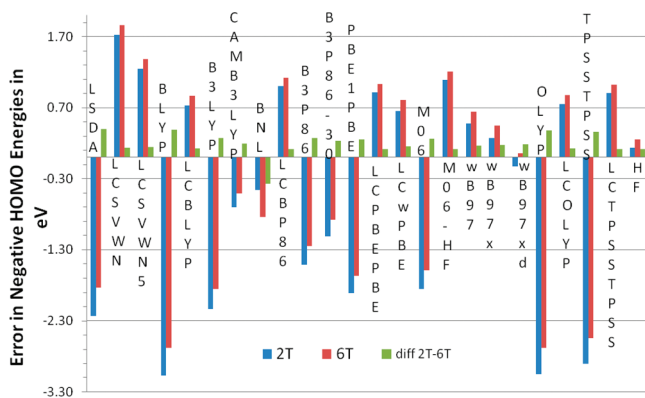


Figure 20. Difference between negative HOMO energies and experimental IPs of 2T and 6T.

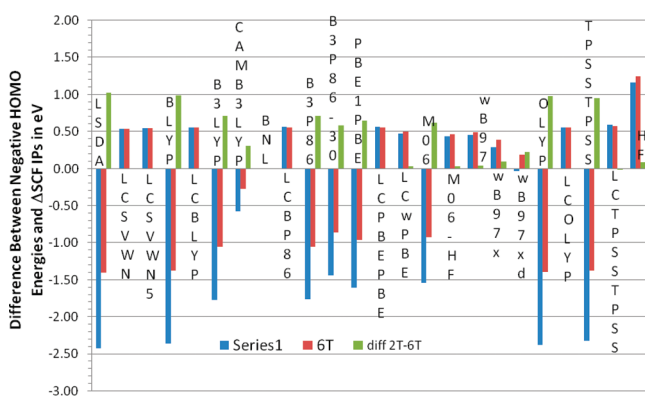


Figure 21. Difference between negative HOMO energies and Δ SCF IPs of 2T and 6T with different density functionals.

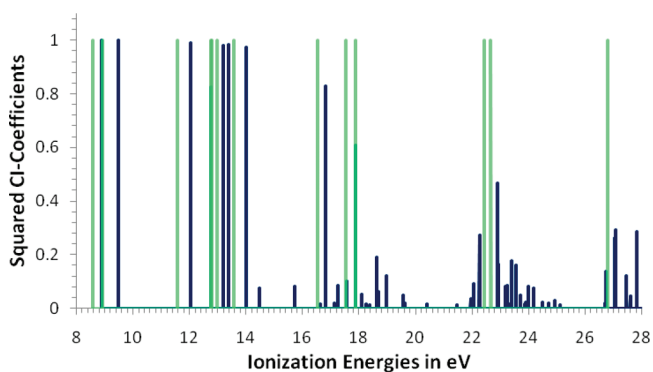


Figure 22. Ground and excited states of $1T^+$ (blue lines) compared to negative orbital energies (green lines) at wB97XD/6-311G*.

according to photoelectron spectroscopy (PES) of the anion.⁸⁶ Since the electron is removed from a stable anion, the EA from the PES experiment is adiabatic. Electron transmission spectroscopy (ETS) produces an attachment energy at 0.2 eV.⁶⁸ Attachment energies are approximate negative vertical EAs.⁶⁸ Thus, $2T^-$ at the geometry of the neutral form is unbound according to ETS.

To evaluate the contribution of the zero point energy (ZPE), frequency calculations were carried out on 2T and $2T^-$ at the wB97XD/6-311+G* level. The ZPE correction for the EA of 2T is

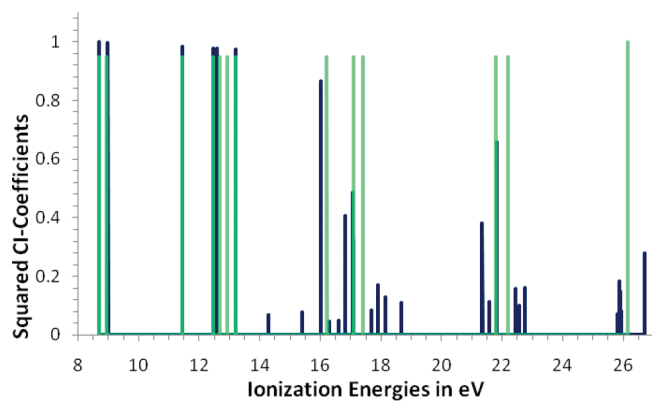


Figure 23. Ground and excited states of $1T^+$ (blue lines) compared to negative orbital energies (green lines) at BNL/6-311G*.

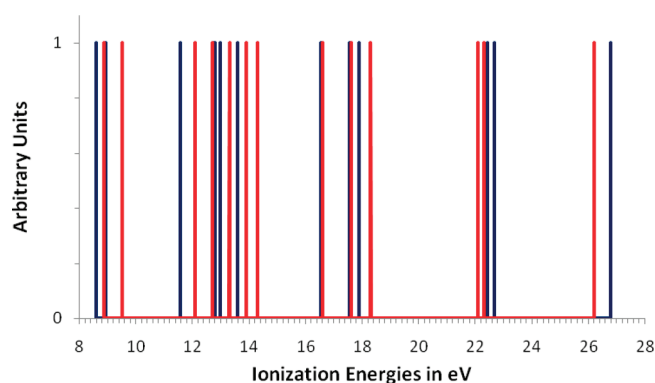


Figure 24. Negative orbital energies (blue lines) at wB97XD/6-311G* and experimental⁸⁵ PE peaks (red lines) of $1T$.

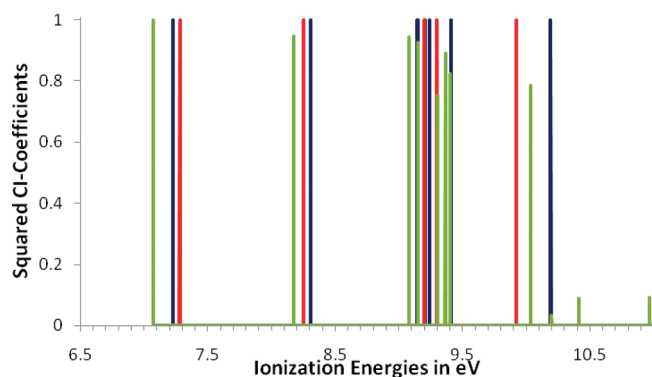


Figure 25. Negative orbital energies (blue lines), states (green lines) at wB97XD/6-311G* and experimental⁸⁵ PE peaks (red lines) of $4T$.

0.135 eV. As this value is larger than the experimental adiabatic EA, $2T^-$ is unbound in its equilibrium structure before ZPE correction. Orbital energies of neutral species correspond to vertical EAs, and a negative EA of $2T$ indicates that LUMO energy, if correct, must be positive. There seem to be no PES experiments on longer oligomers, and ETS data of longer oligomers do not simply correlate with orbital energies from semiempirical calculations.⁶⁸ Therefore, only the vertical EA of $2T$ is compared with experimental values, and $6T$ is checked for internal consistency.

Because vertical and adiabatic EAs of $2T$ are negative before ZPE correction, the electron would simply be removed from the

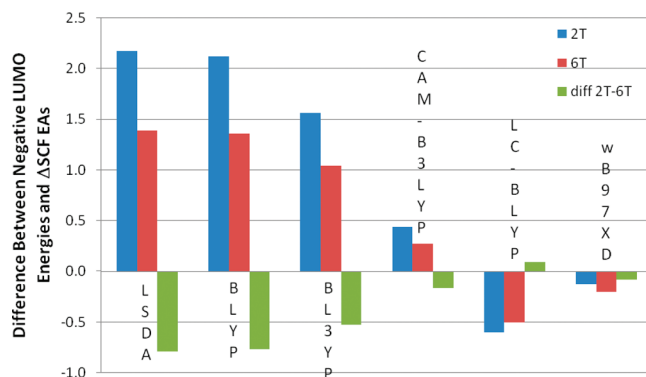


Figure 26. Difference between negative LUMO energies and Δ SCF EAs of $2T$ and $6T$ with different density functionals.

$2T$ molecule with an infinite basis set. With finite basis sets, the electron is forced to stay on the anion, and an attempt can be made to estimate the energy difference between $2T$ and $2T^-$. At the CCSD/aug-CCPVTZ level of theory, the vertical EA of $2T$ is -0.428 eV with ZPE correction, which is reasonably close to the ETS result of -0.2 eV. We were unable to include approximate triples with the aug-CCPVTZ basis set, but with the 6-311G* basis set, approximate triples contribute only 0.019 eV. Density functional theory with the exception of LSDA predicts negative EAs as well. The values of the vertical EAs including the ZPE are LSDA, 0.476 eV; BLYP, -0.224 eV; B3LYP, -0.203 eV; CAM-BLYP, -0.367 eV; LC-BLYP, -0.482 eV; wB97XD, -0.444 eV with the 6-311G* basis set. Inclusion of diffuse functions with 6-311+G* and aug-CCPVTZ basis sets increases the EA by 0.2–0.3 eV.

Figure 26 shows the difference between vertical Δ SCF EAs and LUMO energies of $2T$ and $6T$ with six different density functionals. All levels of theory predict $6T^-$ to be stable. LSDA and BLYP have LUMO orbitals that are lying too low and predict therefore EAs that are significantly larger than the Δ SCF EAs. With global hybrids, i.e., B3LYP, the LUMO is pushed up and the difference between negative LUMO energies and EAs decreases. Long-range correction fixes this problem to a large extent, but the agreement differs from functional to functional. CAM-B3LYP does not correct enough, and LC-BLYP overcorrects. The differences between orbital energies and EAs are larger for $2T$ than for $6T$. wB97XD leads to a near perfect match for $2T$ and $6T$ between negative LUMO energy and EA. These findings are very similar to those for IPs.

DISCUSSION

Performance of Range-Separated Functionals. Taking all results together and setting the maximal allowable error for any property of any system to 0.5 eV, only the wB97XD functional emerges as sufficient. The largest errors with wB97XD are for excited states of cations. Second best is the performance of CAM-B3LYP, which is superior in excitation energies of cations but exceeds errors of 0.5 eV for orbital energies. M06 and TPSSSTPSS do not have any obvious advantages for conjugated π systems. All range-separated functionals are capable of correcting the incorrect chain length dependence of IPs, E_g 's, and orbital energies that make problems with pure DFT and global hybrids. Here, CAM-B3LYP, which contains only 65% long-range HF exchange and has therefore an incorrect long-range limit, tends to be

slightly inferior. Bond length data, IPs and E_g 's show, however, that the LC correction tends to overcorrect and approaches HF results too closely. Therefore, BLA and nonplanarity are somewhat overestimated, and defect sizes may be underestimated with the LC functionals which have no short-range HF exchange and as a result an earlier onset of 100% HF-exchange (larger γ). Thus, range-separated functionals that include a small amount of short-range HF exchange perform decidedly better for conjugated π systems than those without.

A comparison of thiophene and furan oligomers indicates that there are differences between systems. Properties of furan oligomers require more HF exchange and are better predicted with LC functionals than those of thiophenes, although a tendency to overshoot is also predicted here. For furans, CAM-B3LYP might not be enough, and wB97XD appears to be a good compromise. The larger error for the thiophene compared to furan oligomers with LC functionals might be due in part to the overestimated nonplanarity.

Caution has to be applied when cations or anions are to be investigated. Range-separated hybrid functionals are more likely to produce spin-contaminated wave functions than global hybrids. This causes inaccuracies in the higher energy peaks in spectra of doped species. Here, LC functionals overshoot clearly and CAM-B3LYP and wB97XD are more suitable. In particular, because of the excellent match between its negative orbital energies and states, wB97XD is the best choice overall.

The difference between the LC-corrected GGA functionals and CAM-B3LYP, wB97X, and wB97XD is that the latter contain short-range HF-exchange and a larger short range (smaller γ) than the LC functionals. In this respect, the wB97 functional resembles the LC functionals, but it produces generally smaller errors than any of them. Therefore, the superior performance of wB97X and wB97XD seems to be due to a suitable parametrization of the basic functional plus a proper amount of short-range HF exchange in combination with an adequate size of the short range. It is especially encouraging that the functional performs extremely well on properties like excitation energies and orbital energies, which were not included in the training set for the parametrization.

Defect Sizes and Shapes. With range-separated hybrid functionals, singly charged defects shrink and converge to a size of about 9 or 10 rings or 36–40 carbon atoms for polyenes, thiophenes, and furans. As mentioned above, defect sizes approach those at the HF level (30 atoms) and are probably underestimated. In dications, two clearly separated defects at the chain ends form that extend also over about 9– or 10 rings. The chain segment between the two defects is completely undistorted. Range-separated hybrid functionals were shown to correct the error of pure DFT and global hybrids with the wrong dissociation limit of symmetrical radical cations.⁴⁹ Therefore, a crucial error of DFT is eliminated. This error might have influenced results with global hybrids regarding the nonexistence of bound bipolarons. The pronounced separation of two positive charges into two defects located at the ends of the chains confirms therefore that there is no bipolaron binding energy in conducting organic polymers.

Orbital Energies, Transport Gap, and Optical Band Gap. In principle, there are three band gaps: E_g , which is the first peak in the UV spectrum (either the onset of absorption or ϵ_{\max}); the electrochemical band gap E_{el} , which is the difference between oxidation and reduction potential; and E_t , which is the difference between IP and EA. Although the three band gaps are obviously

Table 4. IPs, EAs, and Energy Gaps of Isolated 6T and a Cluster of Five 6T Chains in eV with the wB97XD Functional

property	6T			5 × 6T	diff _{solid-gas}
	6-311G*	6-31G*	6-31G* pl	6-31G*	6-31G*
IP _{ΔSCF} (vert)	6.85	6.61	6.44	5.73	−0.88
−ε _{HOMO}	7.04	6.80	6.64	6.01	−0.79
IP _{ΔSCF} (ad)	6.44	6.23	6.19	5.24	−0.99
EA _{ΔSCF} (vert)	0.61	0.51	0.80	0.92	0.41
−ε _{LUMO}	0.41	0.31	0.50	0.60	0.29
EA _{ΔSCF} (ad)	1.03	0.88	0.94	1.37	0.49
E_t (vert)	6.24	6.10	5.64	4.81	−1.29
E_{H-L}	6.63	6.49	6.14	5.41	−1.08
E_t (ad)	5.41	5.35	5.25	3.87	−1.48
E_g	3.41	3.30	3.03	3.55	0.25

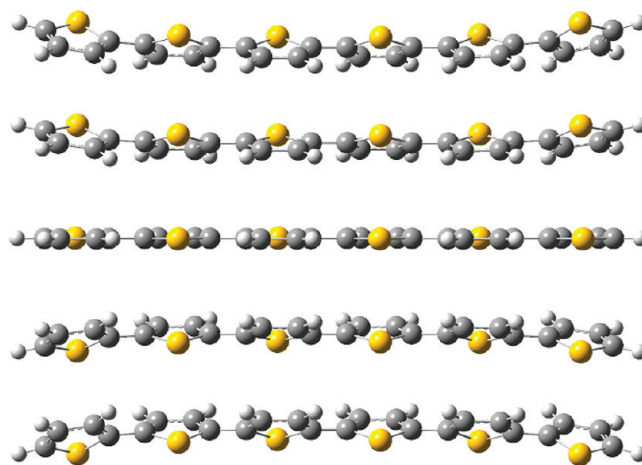


Figure 27. Five molecules of 6T stacked parallel; neutral form optimized at wB97XD/6-31G*.

correlated, the exact relationship between them is rather complicated.^{87,88} In experiments on thin films, E_g and E_{el} are often quite close, differing by several tenths of an electronvolt. From a conceptual point of view, E_{el} and E_t might be expected to be closest because both correspond to the difference between independent electron removal and electron attachment energies, while electron and hole are not completely separated in the UV experiment. The relationship between E_t and E_g has been investigated with PES, inverse PES (IPES), and UV spectroscopy of thin films.⁸⁸ The directly measured difference between IP and EA of 6T is 4.2 eV. E_g is 3.0 eV. After applying a range of corrections for polarization, relaxation, peak centers, difference between surface and bulk, and so on, the difference between E_t and E_g was estimated to be 0.4 eV.⁸⁸

In gas-phase calculations with global hybrid functionals, there is a relatively close match of E_g and E_{H-L} .^{18,28,30} Because accurate negative orbital energies formally match IPs and EAs, E_{H-L} should match E_t rather than E_g .^{87,88} Given the small experimental difference between the two and the considerable uncertainty in the DFT orbital energies, the agreement was usually found to be sufficient. Having now reliable orbital energies at the wB97XD level allows us to address this issue more thoroughly. Vertical and adiabatic IPs and EAs of 6T are collected in Table 4. The E_t of 6T

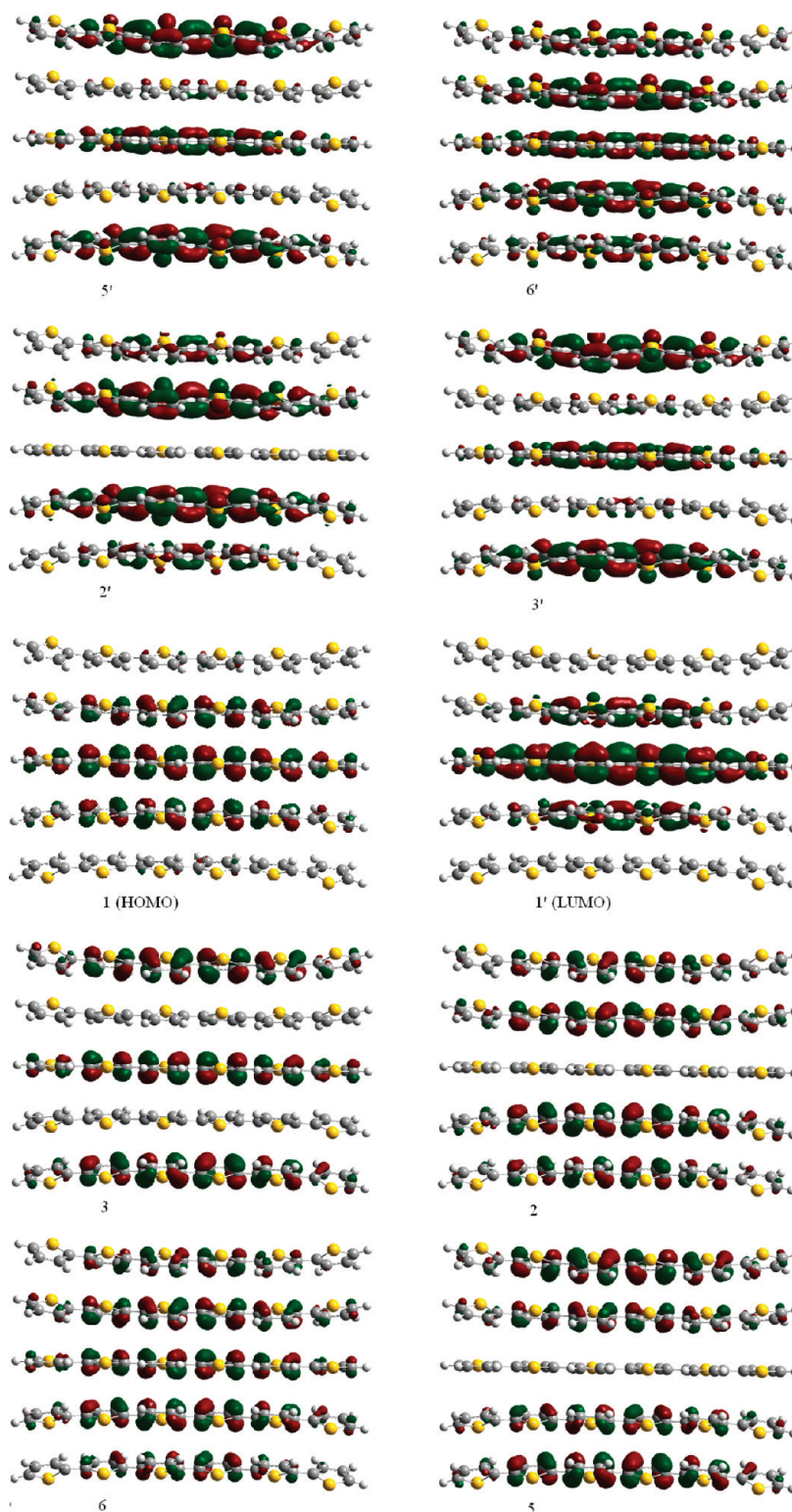


Figure 28. Molecular orbitals of a π -stacked cluster of five 6T molecules.

in the gas phase amounts to 6.24 eV. This is matched fairly well by the HOMO–LUMO gap (E_{H-L}) of 6.63 eV. E_g calculated with TDwB97XD/6-311G* is 3.41 eV. Thus, the difference between E_t and E_g is 2.83 eV for 6T in the gas phase. For comparison, B3LYP yields $E_t = 6.18 \text{ eV} - 1.23 \text{ eV} = 4.95 \text{ eV}$. This value is too small because of the incorrect chain length

dependence of $IP_{\Delta SCF}$ and $EA_{\Delta SCF}$. $E_{H-L} = 5.12 \text{ eV} - 2.27 \text{ eV} = 2.85 \text{ eV}$, which is even smaller because of the incorrect orbital energies with global hybrids. $E_g = 2.56 \text{ eV}$, which is underestimated (by 0.36 eV) because of incorrect chain length dependence. The good match between E_{H-L} and E_g at B3LYP is thus a coincidence.

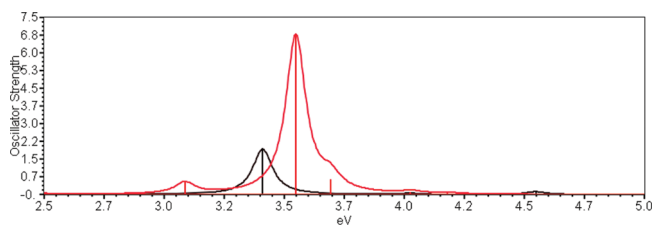


Figure 29. UV spectra of 6T (black) and a cluster of five 6T molecules (red).

Because the wB97XD gas phase values were shown to be reliable throughout this paper, the large discrepancy between the theoretical (2.83 eV) and experimental (0.4 eV) differences between E_t and E_g can only be caused by solid state effects. In order to investigate the influence of neighboring chains, we optimized a cluster of five 6T chains (Figure 27) in neutral, positively, and negatively charged states at the wB97XD/6-31G* level and calculated IPs and EAs. The results are included in Table 4. Although a small π -stacked cluster is a very crude model for a herringbone structured crystal of 6T, we believe that the model catches the essentials at least qualitatively. Since wB97XD includes dispersion, the geometry optimization yielded a reasonable distance between the chains of 4.5 Å. The spacing between chains in the cation is 3.4 Å; in the anion it is 3.6 Å. In agreement with the experimental crystal structures of 6T and 8T, the inner 6T is planar.

Vertical IP, adiabatic IP, and $-\epsilon_{\text{HOMO}}$ decrease from the gas phase to cluster by between 0.8 and 1.0 eV. The electron affinity increases by 0.3–0.5 eV. Thus, the transport gap is reduced by 1.1–1.5 eV in the cluster. About 0.5 eV of that is due to planarization. The first excitation energy of the cluster is 0.25 eV higher than that of 6T. Theory therefore predicts that the difference between $E_t(\text{vert})$ and E_g shrinks to 1.26 eV in the cluster. For comparison with experimental results, we need to consider furthermore that the solid state IPs and EAs are adiabatic.⁸⁸ For the cluster, we should therefore compare $E_t(\text{ad})$ to E_g . This difference is 0.32 eV, in very good agreement with the experimental value of 0.4 eV.⁸⁸ $E_{\text{H-L}}$, albeit 1.08 eV smaller in the cluster than in the gas phase, overestimates E_g of the cluster by 1.86 eV mainly because it approximates $E_t(\text{vert})$ rather than $E_t(\text{ad})$.

The opposite effect of neighboring chains on E_t and E_g is caused on the one hand by the fact that HOMO and LUMO delocalize to some extent over the whole structure (Figure 28), which lowers the IP and increases the EA. On the other hand, E_g increases because the HOMO–LUMO transition is symmetry-forbidden in the cluster. The lowest allowed state involves transitions between higher and lower lying orbitals, namely, 6–1', 5–2', 3–3', 2–5', and 1–6'. The spectra of 6T and the cluster are shown in Figure 29. Both spectra have one main feature at similar energies, but the cluster peak has a much larger oscillator strength. These findings rationalize the puzzling observation that there is a large solid state polarization energy for IPs of organic solids (~ 1.8 eV)⁸⁹ but very little difference between UV absorption energies of organic molecules in the gas phase, in solution, and in the solid state. Our results furthermore predict that the solid state polarization energy is larger for IPs than for EAs, in contrast to earlier assumptions.⁸⁸

CONCLUSIONS

Range-separated functionals solve the DFT problem of incorrect chain length dependence of IPs, excitation energies, and

orbital energies of conjugated systems. Using 100% of long-range HF exchange can overcorrect, however, as is obvious from the results on BLA, twist angles, and orbital energies. Of all range-separated functionals tested, wB97X, wB97XD, and CAM-B3LYP show the best overall performance, especially when excited states of cations are included. The reason seems to be that these functionals include short-range HF exchange and have a larger short range. This pushes the onset of 100% HF exchange further out and leads to less overshooting.

Because range-separated hybrid functionals correct the erroneous dissociation of symmetrical radical cations, a major error of approximate DFT that might have had an impact on the nature of the charge carriers in conducting polymers has been removed. The fact that range-separated functionals not only confirm but enhance the splitting of two charges on one chain into two separated defects provides further support to our earlier claim¹³ that there are no bound bipolarons in conducting polymers.

Among all functionals, wB97XD appears most promising for calculations on conducting polymers. In particular, the excellent match of negative orbital energies (with binding energies of up to 30 eV) and IPs and EAs makes it useful for band structure calculations. Only the higher excited states of charged systems are somewhat problematic.

The good accuracy of wB97XD orbital energies allows assessment of the relationship between $E_{\text{H-L}}$, E_g , and E_t in the gas phase and in a cluster as a model for the solid state. $E_{\text{H-L}}$ approximates $E_t(\text{vert})$ with errors of 0.4 and 0.6 eV in the gas phase and in the cluster, respectively. $E_t(\text{vert})$ exceeds E_g by 2.8 eV in the gas phase and by 1.26 eV in the cluster. As IPs and EAs of thin films are adiabatic, for the cluster, E_g has to be compared to $E_t(\text{ad})$, and the difference is only 0.32 eV. Nonetheless, even if E_g is dominated by a HOMO–LUMO transition, which may or may not be the case, it is not equal to $E_{\text{H-L}}$. Therefore, correct orbital energy differences do not predict optical band gaps, and the “success” of DFT orbital energies with global hybrids to do so is a result of error cancellation between incorrect orbital energies, wrong chain length dependence, and the neglect of solid state polarization effects.

AUTHOR INFORMATION

Corresponding Author

*Tel.: +90 312 290 2122. Fax: +90 312 266 4068. E-mail: salzner@fen.bilkent.edu.tr.

ACKNOWLEDGMENT

This work was supported by TUBITAK, Grant TBAG 109T426.

REFERENCES

- (1) Su, W. P.; Schrieffer, J. R.; Heeger, A. J. *Phys. Rev. B* **1980**, *22*, 2099–2111.
- (2) Bertho, D.; Jouanin, C. *Phys. Rev. B* **1987**, *35*, 626–633.
- (3) Brédas, J.-L. In *Handbook of Conducting Polymers*; Skotheim, T. A., Ed.; Dekker: New York, 1986; pp 859–913.
- (4) Brédas, J. L.; Street, G. B. *Acc. Chem. Res.* **1985**, *18*, 309–315.
- (5) Wendel, H.; Otto, P.; Seel, M.; Ladik, J. *Solid State Commun.* **1985**, *54*, 551–554.
- (6) Otto, P.; Bakhshi, A. K.; Ladik, J. *THEOCHEM* **1986**, *135*, 209–223.
- (7) Otto, P.; Ladik, J. *Synth. Met.* **1990**, *36*, 327–335.
- (8) Villar, H.; Otto, P.; Dupuis, M.; Ladik, J. *Synth. Met.* **1993**, *59*, 97–110.

- (9) Villar, H. O.; Otto, P.; Dupuis, M.; Ladik, J. *Synth. Met.* **1993**, *59*, 97–110.
- (10) Brocks, G. *J. Phys. Chem.* **1996**, *100*, 17327–17333.
- (11) Vogl, P.; Campbell, C. K. *Phys. Rev. Lett.* **1989**, *62*, 2012–2015.
- (12) Salzner, U. *J. Chem. Theory Comput.* **2007**, *3*, 219–231.
- (13) Salzner, U. *J. Chem. Theory Comput.* **2007**, *3*, 1143–1157.
- (14) Salzner, U. *J. Phys. Chem. A* **2010**, *114*, 10997–11007.
- (15) Cai, Z.-L.; Sendt, K.; Reimers, J. R. *J. Chem. Phys.* **2002**, *117*, 5543–5549.
- (16) Williams, A. R.; von Barth, U. In *Theory of the Inhomogeneous Electron Gas*; Lundqvist, S., March, N. H., Eds.; Plenum Press: London, 1983.
- (17) Perdew, J. P.; Levy, M. *Phys. Rev. B* **1997**, *56*, 16021.
- (18) Salzner, U.; Lagowski, J. B.; Pickup, P. G.; Poirier, R. A. *J. Comput. Chem.* **1997**, *18*, 1943–1953.
- (19) Chong, D. P.; Gritsenko, O. V.; Baerends, E. J. *J. Chem. Phys.* **2002**, *116*, 1760–1772.
- (20) Salzner, U.; Baer, R. *J. Chem. Phys.* **2009**, *131*, 231101–4.
- (21) Stowasser, R.; Hoffmann, R. *J. Am. Chem. Soc.* **1999**, *121*, 3414–3420.
- (22) Jacquemin, D.; Adamo, C. *J. Chem. Theory Comput.* **2011**, *7*, 369–376.
- (23) Choi, C. H.; Kertesz, M.; Karpfen, A. *J. Chem. Phys.* **1997**, *107*, 6712–6721.
- (24) Champagne, B.; Spassova, M. *Phys. Chem. Chem. Phys.* **2004**, *6*, 3167–3174.
- (25) Salzner, U. *J. Phys. Chem. A* **2008**, *112*, 5458–5466.
- (26) Zade, S. S.; Bendikov, M. *J. Phys. Chem. B* **2006**, *110*, 15839–15846.
- (27) Zamoshchik, N.; Salzner, U.; Bendikov, M. *J. Phys. Chem. C* **2008**, *112*, 8408–8418.
- (28) Salzner, U.; Pickup, P. G.; Poirier, R. A.; Lagowski, J. B. *J. Phys. Chem. A* **1998**, *102*, 2572–2578.
- (29) Körzdörfer, T.; Kümmel, S.; Marom, N.; Kronik, L. *Phys. Rev. B* **2009**, *79*, 201205–4.
- (30) Zade, S. S.; Zamoshchik, N.; Bendikov, M. *Acc. Chem. Res.* **2011**, *44*, 14–24.
- (31) Savin, A. In *Recent Advances in Density Functional Methods Part I*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; pp 129.
- (32) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.
- (33) Kamiya, M.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2002**, *117*, 6010–6015.
- (34) Jaramillo, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 1068–1073.
- (35) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425–8433.
- (36) Toulouse, J.; Colonna, F.; Savin, A. *Phys. Rev. A* **2004**, *70*, 062505.
- (37) Baer, R.; Neuhauser, D. *Phys. Rev. Lett.* **2005**, *94*, 43002.
- (38) Kamiya, M.; Sekino, H.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2005**, *122*, 234111–10.
- (39) Sato, T.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2005**, *123*, 104307–10.
- (40) Sato, T.; Tsuneda, T.; Hirao, K. *Mol. Phys.* **2005**, *103*, 1151–1164.
- (41) Sekino, H.; Maeda, Y.; Kamiya, M. *Mol. Phys.* **2005**, *103*, 2183–2189.
- (42) Vydrov, O.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 074106–1–9.
- (43) Livshits, E.; Baer, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932–2941.
- (44) Salzner, U. *J. Phys. Chem. B* **2003**, *107*, 1129–1134.
- (45) Gidron, O.; Diskin-Posner, Y.; Bendikov, M. *J. Am. Chem. Soc.* **2010**, *132*, 2148–2150.
- (46) Samdal, S.; Samuelsen, E. J.; Volden, H. V. *Synth. Met.* **1993**, *59*, 259.
- (47) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonneberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT, 2009.
- (48) Shao, Y.; et al. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.
- (49) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109–9.
- (50) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (51) Vydrov, O. A.; Scuseria, G. E.; Perdew, J. P. *J. Chem. Phys.* **2007**, *126*, 154109–9.
- (52) Slater, J. C. In *Quantum Theory of Molecular and Solids*; McGraw-Hill: New York, 1974; Vol. 4.
- (53) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (54) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (55) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (56) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- (57) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (58) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.
- (59) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403–412.
- (60) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (61) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (62) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106–15.
- (63) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554–8560.
- (64) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (65) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.
- (66) Horowitz, G.; Bachet, B.; Yassar, A.; Lang, P.; Demanze, F.; Fave, J.-L.; Garnier, F. *Chem. Mater.* **1995**, *7*, 1337–1341.
- (67) Fichou, D.; Bachet, B.; Demanze, F.; Billy, I.; Horowitz, G.; Garnier, F. *Adv. Mater.* **1996**, *8*, 500–504.
- (68) Jones, D.; Guerra, M.; Favaretto, L.; Modelli, A.; Fabrizio, M.; Distefano, G. *J. Phys. Chem.* **1990**, *94*, 5761–5766.
- (69) da Silva Filho, D. A.; Coropceanu, V.; Fichou, D.; Gruhn, N. E.; Bill, T. G.; Gierschner, J.; Cornil, J.; Bredas, J.-L. *Philos. Trans. R. Soc. London, Ser. A* **2007**, *365*, 1435–1452.
- (70) Distefano, G.; Jones, D.; Guerra, M.; Favaretto, L.; Modelli, A.; Mengoli, G. *J. Phys. Chem.* **1991**, *95*, 9746–9753.
- (71) Becker, R. S.; Seixas de Melo, J.; MaÅšanita, A. n. L.; Elisei, F. *J. Phys. Chem.* **1996**, *100*, 18683–18695.
- (72) Colditz, R.; Grebner, D.; Helbig, M.; Rentsch, S. *Chem. Phys.* **1995**, *201*, 309–320.
- (73) Andrzejak, M.; Witek, H. *Theor. Chem. Acc.* **2011**, *129*, 161–172.
- (74) van Haare, J. A. E. H.; Havinga, E. E.; van Dongen, J. L. J.; Janssen, R. A. J.; Cornil, J.; Brédas, J. L. *Chem.—Eur. J.* **1998**, *4*, 1509–1522.
- (75) Rubio, M.; Merchan, M.; Orti, E. *Chem. Phys. Chem.* **2005**, *6*, 1357–1368.
- (76) Rubio, M.; Merchan, M.; Orti, E.; Roos, B. *J. Chem. Phys.* **1995**, *102*, 3580–3586.
- (77) Rubio, M.; Merchán, M.; Orti, E.; Roos, B. O. *Chem. Phys. Lett.* **1996**, *248*, 321–328.
- (78) Keszthelyi, T.; Grage, M. M.-L.; Offersgard, J. F.; Wilbrandt, R.; Svendsen, C.; Sonnich Mortensen, O.; Pedersen, J. K.; Jensen, H. J. A. *J. Phys. Chem. A* **2000**, *104*, 2808–2823.

- (79) Fichou, D.; Horowitz, G.; Garnier, F. *Synth. Met.* **1990**, *39*, 125–131.
- (80) Rubio, M.; Ortí, E.; Pou-Amerigo, R.; Merchán, M. *J. Phys. Chem. A* **2001**, *105*, 9788–9794.
- (81) Baer, R.; Livshits, E.; Salzner, U. *Annu. Rev. Phys. Chem.* **2010**, *61*, 85–109.
- (82) Tsuneda, T.; Song, J.-W.; Suzuki, S.; Hirao, K. *J. Chem. Phys.* **2010**, *133*, 174101–9.
- (83) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. *Phys. Rev. Lett.* **1982**, *49*, 1691–1694.
- (84) Almbladh, C. O.; Von Barth, U. *Phys. Rev. B* **1985**, *31*, 3231–3244.
- (85) Holland, D. M. P.; Karlsson, L.; von Niessen, W. *J. Electron Spectrosc. Relat. Phenom.* **2001**, *113*, 221–239.
- (86) Schiedt, J.; Knott, W. J.; Le Barbu, K.; Schlag, E. W.; Weinkauff, R. *J. Chem. Phys.* **2000**, *113*, 9470–9478.
- (87) Djurovich, P. I.; Mayo, E. I.; Forrest, S. R.; Thompson, M. E. *Org. Electronics* **2009**, *10*, 515–520.
- (88) Hill, I. G.; Kahn, A.; Soos, Z. G.; Pascal, J. R. A. *Chem. Phys. Lett.* **2000**, *327*, 181–188.
- (89) Sato, N.; Seki, K.; Inokuchi, H. *J. Chem. Soc., Faraday II* **1981**, *77*, 1621–1633.

Distinct Morphologies for Amyloid Beta Protein Monomer: $A\beta_{1-40}$, $A\beta_{1-42}$, and $A\beta_{1-40}$ (D23N)

Sébastien Côté,[†] Philippe Derreumaux,[‡] and Normand Mousseau^{*,†}

[†]Département de Physique and Groupe de Recherche Sur Les Protéines Membranaires (GEPROM), Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal (Québec), Canada

[‡]Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Institut de Biologie Physico Chimique, Institut Universitaire de France, Université Paris Diderot - Paris 7, 13 rue Pierre et Marie Curie, 75005 Paris, France

ABSTRACT: Numerous experimental studies indicate that amyloid beta protein ($A\beta$) oligomers as small as dimers trigger Alzheimer's disease. Precise solution conformation of $A\beta$ monomer is missing since it is highly dynamic and aggregation prone. Such a knowledge is however crucial to design drugs inhibiting oligomers and fibril formation. Here, we determine the equilibrium structures of the $A\beta_{1-40}$, $A\beta_{1-42}$, and $A\beta_{1-40}$ (D23N) monomers using an accurate coarse-grained force field coupled to Hamiltonian-temperature replica exchange molecular dynamics simulations. We observe that even if these three alloforms are mostly disordered at the monomeric level, in agreement with experiments and previous simulations on $A\beta_{1-40}$ and $A\beta_{1-42}$, striking morphological differences exist. For instance, $A\beta_{1-42}$ and $A\beta_{1-40}$ (D23N) have higher β -strand propensities at the C-terminal, residues 30–42, than $A\beta_{1-40}$. The D23N mutation enhances the conformational freedom of the residues 22–29 and the propensity for turns and β -strands in the other regions. It also changes the network of contacts; the N-terminal (residues 1–16) becoming more independent from the rest of the protein, leading to a less compact morphology than the wild-type sequence. These structural properties could explain in part why the kinetics and the final amyloid products vary so extensively between the $A\beta_{1-40}$ and the $A\beta_{1-40}$ (D23N) peptides.

1. INTRODUCTION

Alzheimer's disease (AD) is characterized by the presence of extracellular neuritic plaques and intracellular neurofibrillary tangles in the brain.¹ Senile plaques are made of the amyloid beta ($A\beta$) protein. This protein is naturally produced through the cleavage of the amyloid precursor glycoprotein (APP) by the β - and γ -secretases. Many alloforms with amino acid lengths varying between 39 and 43 are produced. Of these, $A\beta_{40}$ is the most abundant, and $A\beta_{42}$ is the most toxic and aggregation prone.²

Experimental studies using circular dichroism (CD) spectroscopy and electron microscopy³ indicate that $A\beta$ peptides exhibit a transition from random coil to β -sheet during fibrillation. Other experiments reveal that the final amyloid product, the nucleation and elongation rates⁴ are distinct for different alloforms and are modulated by experimental conditions.⁵ The oligomerization pathways also vary for different alloforms^{6,7} and $A\beta$ oligomers, rather than the amyloid fibrils, are the principal neurotoxic agents^{2,8} interacting with receptors,⁹ metal ions,¹⁰ cell membranes,¹¹ and synapses.¹² Despite these significant advances in $A\beta$ amyloid fibril assembly, precise aqueous solution conformations of $A\beta$ monomer are missing. We know from solution NMR that it is mostly coil turn with little β -strand content in water solution,^{13,14} that $A\beta_{42}$ is more structured at the C-terminal than $A\beta_{40}$,¹⁵ and that oxidation of Met35 causes important changes in the monomer structure.¹⁶ Also, hydrogen/deuterium exchange experiments showed that the monomer is completely exposed to the solvent, revealing that it is highly fluctuating.¹⁷ Finally, limited proteolysis/mass spectrometry on $A\beta_{40}$ monomer suggests many turns, such as between Val24 and Lys28, which was observed by solution NMR on the $A\beta_{21-30}$ fragment.¹⁸

Yet detailed conformational knowledge of $A\beta$ monomer is of utmost importance for two reasons. First, the monomer exists in equilibrium with oligomers during polymerization,⁶ interacts with the cell membrane,¹¹ binds to fibril,¹⁹ and may even be neuroprotective.²⁰ Second, since $A\beta$ oligomers as small as dimers have been implicated in AD,²¹ it is important to block oligomerization at the monomer level. To complement biophysical studies, computer simulations are often used. Several molecular dynamics (MD) simulations have already been reported on $A\beta_{40}$ ²² and $A\beta_{42}$ ²³ folding in aqueous solution. Other studies examined the role of ion binding,²⁴ oxidation of Met35,²⁵ and salt bridge Asp23-Lys28²⁶ on the folding of $A\beta$ monomers. These simulations revealed many interesting features, but the generated conformational ensemble is affected by the starting structure used and by the short time scale explored. Moreover, MD simulations easily get trapped in local minima biasing the results when the energy landscape is rugged, as is the case with $A\beta$.

Replica exchange molecular dynamics (REMD) simulation, which enhances conformational sampling,²⁷ has also been used. For the monomer of $A\beta_{40}$ and $A\beta_{42}$, it was coupled to different all-atom force fields with explicit²⁸ and implicit solvent models.²⁹ Each replica was however simulated for 60–100 ns,^{28,29} which may not be enough to sample the complete morphological ensemble. This convergence problem certainly holds for the simulation of $A\beta_{1-39}$.³⁰ A recent extensive all-atom simulation with implicit solvent on the truncated $A\beta_{10-40}$ monomer reached convergence.³¹ However, the predicted percentage of α -helix

Received: September 21, 2010

Published: June 14, 2011

content (~38%) is significantly overestimated compared to the CD analysis (between 9 and 12%^{3,32}) and the weak α -helix signal detected by NMR.¹³ In addition, the N-terminal segment must not be discarded as it has been recently recognized that mutations at positions 6 and 7 alter monomer folding and oligomerization.³³ Finally, two long all-atom simulations based on different force fields and sampling approaches were recently performed on the $A\beta_{42}$ monomer. They reported, however, very different conformational ensembles: with high β -sheet content and forming, for instance, a four-stranded antiparallel β -sheet using Monte Carlo simulated annealing and an implicit solvent³⁴ or an ensemble of predominantly random coil conformations from a 225 ns REMD simulation using AMBER force field with explicit solvent.³⁵

These divergent computational conclusions on the nature of the conformations that characterize the whole ensemble of $A\beta_{40}$ and $A\beta_{42}$ monomer motivated us to re-explore their foldings using a different approach. Here, we use the coarse-grained protein force field (OPEP), which has been tested on widely different systems,³⁶ and we couple it to a hybrid of Hamiltonian and temperature replica exchange molecular dynamics (HT-REMD),³⁷ which allows a more efficient sampling of the conformational space than standard temperature REMD. Along the wild-type $A\beta_{40}$ and $A\beta_{42}$ peptides, we also examined the $A\beta_{40}$ -(D23N) variant implicated in early onset AD.³⁸ To our knowledge, there is no experimental and computational study on the $A\beta_{40}$ -(D23N) monomer. Yet, the mutation D23N has strong consequences on polymerization and fibril product. While the kinetics of $A\beta_{40}$ and $A\beta_{42}$ polymerization shows a lag phase, with $A\beta_{42}$ peptide forming fibrils at a much faster rate, the $A\beta_{40}$ -(D23N) peptide does not display any lag phase.³⁹ In addition, $A\beta_{40}$ and $A\beta_{42}$ form in-register parallel β -sheet fibrils,^{40–42} whereas $A\beta_{40}$ -(D23N) predominant fibrils form antiparallel β -sheets.³⁹ Such fibrillar morphologies are usually seen only for short sequences making therefore $A\beta_{40}$ -(D23N) a very special alloform.

2. METHODOLOGY

We use the implicit coarse-grained potential OPEP 3.2 parameter set³⁶ coupled with HT-REMD³⁷ to describe the morphologies visited by $A\beta_{1-40}$, $A\beta_{1-42}$, and $A\beta_{1-40}$ -(D23N) alloforms using the same protocol. $A\beta_{40}$ has the following amino acid sequence: DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV. For $A\beta_{42}$, two more residues, Ile41 and Ala42, are present at the C-terminal end, and for $A\beta_{40}$ -(D23N), the positively charged residue Asp23 is mutated into the neutral Asn23. We use 23 temperatures following an exponential distribution: 200.0, 229.7, 239.9, 259.4, 261.5, 273.1, 285.1, 297.7, 310.8, 324.6, 338.9, 353.9, 369.5, 385.8, 402.8, 420.6, 439.2, 458.6, 478.6, 500.0, 502.0, 504.0, and 505.0 K. At the highest temperature, we use a potential reduction scale composed of 5 steps: 0.8, 0.7, 0.6, 0.4, and 0.2 that fractionally reduces nonbonded attractive forces. Exchanges between neighboring replicas are attempted every 7.5 ps resulting in an exchange rate of 50–60%. This relatively high exchange rate is due to the use of a coarse-grained protein model coupled with an implicit solvent, which decreases the effective number of degrees of freedom, and to the low secondary structure probability of the peptides, as discussed in Section 3. In the absence of clear structural differences, configurational energies between nearby temperatures tend to be closer.

Bond lengths are constrained with the RATTLE algorithm.⁴³ The simulations are thermalized using Berendsen's thermostat with a coupling constant of 0.1 ps⁴⁴ and have an integration time

step of 1.5 fs. Simulations are started from a random extended conformation with an end-to-end distance of 36.9, 37.5, and 36.9 Å for $A\beta_{1-40}$, $A\beta_{1-42}$, and $A\beta_{1-40}$ -(D23N), respectively. Each monomer is simulated in a sphere of 60 Å radius with reflecting boundary conditions. The size chosen for the sphere minimizes boundary conditions bias. Each replica of $A\beta_{1-40}$ is simulated for 700 ns (giving 19.6 μ s total time), each replica of $A\beta_{1-42}$ for 900 ns (giving 25.2 μ s total time), and each replica of $A\beta_{1-40}$ -(D23N) for 900 ns (giving 25.2 μ s total time). These extensive simulations are necessary to reach convergence as discussed below.

Potential. We choose the implicit solvent coarse-grained potential optimized potential for efficient structure prediction (OPEP) 3.2 because it captures the main interactions during protein folding without costing much computational time, such as an all-atom explicit solvent potential.³⁶ This potential has been shown to recover the native structure of a variety of peptides with widely different secondary and tertiary structures as accurately as all-atom potentials using temperature REMD simulations.⁴⁵ OPEP has also been applied to short and long amyloid sequences giving results with strong similarities with experiments.^{37,46–51} Coupled to a greedy algorithm and a structural alphabet, OPEP was able to locate, using a benchmark of 25 peptides with 9–23 amino acids, lowest energy conformations differing by 2.6 Å C α root-mean-square deviation (rmsd) from the full NMR structures.^{52,53} Briefly, this model approximates each amino acid by 6 beads: N, H_N, C ω , SC, C, and O, where the side chain (SC) is represented by a unique bead with glycines having a H instead of SC. The interaction parameters are finely tuned against protein structures and thermodynamics and include bond lengths and angles, improper torsions, dihedral angles, van der Waals interactions, and two- and four-body cooperative hydrogen bonds.

HT-REMD. In order to determine the conformations of $A\beta$, we use an hybrid of Hamiltonian and temperature replica exchange molecular dynamics (HT-REMD). T-REMD is widely used to simulate protein aggregation with a variety of implicit and explicit solvents.²⁷ Nonetheless, for some proteins, it is observed that T-REMD alone is not sufficient to completely unfold the protein and escape from strong local minima. HT-REMD enhances sampling by reducing nonbonded attractive forces at the highest temperature.³⁷ This allows $A\beta$ to unfold completely into an extended chain.

As T-REMD,^{27,45} HT-REMD has a distribution of temperatures allowing exchanges between neighboring replicas i and j according to the Metropolis criterion that preserves thermodynamic ensembles:

$$\text{Prob}(i \leftrightarrow j) = \min \left\{ 1, \exp \left[\frac{E_i - E_j}{k_B(T_i - T_j)} \right] \right\} \quad (1)$$

where $\text{Prob}(i \leftrightarrow j)$ is the probability of exchange between replicas i and j , E_i and E_j are the energies of replicas i and j , respectively, and T is the temperature.

As H-REMD,⁵⁴ HT-REMD possesses an energy scale at the highest temperature in which nonbonded attractive forces are fractionally reduced. Exchanges between neighboring replicas i and j are governed by

$$\text{Prob}(i \leftrightarrow j) = \min \left\{ 1, \exp \left[-\frac{H_i(X') - H_i(X) + H_j(X) - H_j(X')}{k_B T_{\max}} \right] \right\} \quad (2)$$

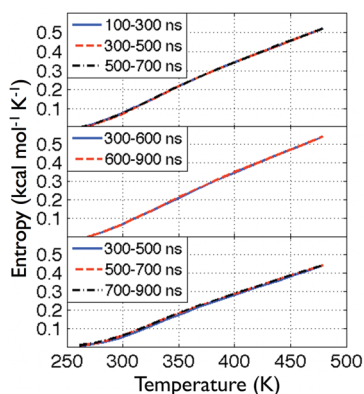


Figure 1. Entropy as a function of temperature $S(T)$ averaged over different time intervals. $A\beta_{40}$, $A\beta_{42}$, and $A\beta_{40}$ (D23N) are shown in the top, middle, and bottom panels, respectively.

where H denotes the Hamiltonian (energy) and X and X' are the configurations of replicas i and j , respectively. By decreasing the nonbonded attractions between atoms at the highest temperature, HT-REMD helps the replicas escape from deep local minima increasing the conformational space sampling.

Analysis. We analyze the secondary and tertiary structures at 325 K as justified in Section 3.2 using the equilibrated time interval of each alloform. The secondary structure is analyzed using STRIDE,⁵⁵ and the tertiary structure is analyzed using contacts between side chains and clustering. Contacts are considered when the distance between two side chain beads is smaller than the sum of their van der Waals radii plus 0.5 Å. For clustering, we first calculate the rmsd between all structures for each alloform. We then find the biggest cluster, remove from the pool of structures all those contained in this cluster, and repeat iteratively until no structure is left.⁵⁶ We select a small threshold of 2 Å C_{α} rmsd to increase the discrimination between clusters.

The weighted histogram analysis method⁵⁷ is used to calculate the free energy, and the entropy is calculated using $S = (E - F)/T$. J-coupling constants between H_N and H_{α} , ${}^3J_{H_N, H_{\alpha}}$, are calculated using the Karplus equation⁵⁸ with three different sets of coefficients.^{35,59,60} Error bars for the J-coupling constants and the secondary structure propensities show the interval of confidence on the mean value given by the bootstrap statistical analysis method.⁶¹

Convergence. Convergence in our simulations is assessed by three criteria. First, we check that, at equilibrium and based on the ergodic principle, the entropy as a function of temperature, $S(T)$, is time independent when averaged over sufficiently long periods as is seen in Figure 1. $A\beta_{40}$ is the most rapidly converged simulation, and $S(T)$ converges within 200 ns time intervals after only 100 ns of equilibration. As $A\beta_{42}$ converges more slowly, the equilibration takes 300 ns, and the entropy becomes time independent when averaged over 300 ns time windows. $A\beta_{40}$ (D23N) requires an equilibration time of 500 ns, and $S(T)$ remains constant within 200 ns time windows. Note that the differences in the time windows to reach equilibrium and obtain converged entropies starting from similar conformations for the three alloforms cannot be related to differences in aggregation properties. They only indicate that the configuration space of $A\beta_{42}$ is more complex than those of $A\beta_{40}$ and $A\beta_{40}$ (D23N). We further note that the entropy of the three alloforms has a similar slope, which is related to their weak secondary structure signals

and their overall behavior as random coil polymers. Second, to further confirm convergence, we look at the variation of secondary structure using the time windows mentioned above for each alloform. The same trend is observed in each time window, and the probabilities of secondary structures per residue vary on average by only 4–5% implying convergence (data not shown). Third, as a final check, we cluster all structures in each time window for each alloform and find that the resulting distribution of clusters is very similar between different time windows. These results confirm that, with our simulation protocol, each alloform has fully converged. As a result, for analysis, we use the equilibrated time interval of each alloform: from 100 to 700 ns for $A\beta_{40}$, 300 to 900 ns for $A\beta_{42}$, and 500 to 900 ns for $A\beta_{40}$ (D23N), totaling 80 000, 80 000, and 53 333 structures, respectively.

Naming Convention. To facilitate analysis, the sequence is often split into four regions: the N-terminal (residues 1–16), the central hydrophobic core or CHC (residues 17–21), the loop region (residues 22–29), and the C-terminal (residues 30–42). The N-terminal is mostly hydrophilic, the CHC corresponds to¹⁷ Leu-Val-Phe-Phe-Ala (LVFFA),²¹ the loop region refers to the residues forming a loop in the fibrillar morphologies, and the C-terminus is mostly hydrophobic.

3. RESULTS

3.1. J-Couplings. Before comparing the properties of the three alloforms, it is important to select the appropriate simulation temperature for comparison with experiments. It is known that even all-atom simulations with explicit solvent generally show a shift between the predicted and experimentally observed melting temperatures, indicating that the comparison between computation and experiments cannot be made directly with the temperature used in the experiments.⁶² This feature has also been observed with the OPEP force field.⁵¹ To find the simulation temperature corresponding best to experiments, we calculate the J-coupling constants, ${}^3J_{H_N, H_{\alpha}}$, of $A\beta_{40}$ and $A\beta_{42}$ at 300 and 325 K (the highest temperature before transition to random coil) and compare those with experimental J-coupling constants measured at 300 K.²⁸ We also examine the use of three sets of parameters for the Karplus equation (namely the Vuister's,⁵⁹ Sgourakis',³⁵ and Schmidt's⁶⁰ parameter sets) in calculating the J-coupling constants from the simulations.

Figure 2 superposes the J-coupling constants calculated for both alloforms at 300 and 325 K using the Vuister's parameter set⁵⁹ on the experimental values.²⁸ In the inset, the Vuister's parameter set is compared to that proposed recently by Sgourakis et al.³⁵ for the simulations at 325 K. Qualitatively, both temperature data sets follow the same trend for $A\beta_{40}$ and $A\beta_{42}$. We note, however, for the Vuister's set, an overall J-coupling shift in the N-terminal region for $A\beta_{40}$ compared to experiment, a shift that is considerably reduced with Sgourakis' set. Quantitatively (Table 1), the Pearson correlation coefficient (PCC) remains nearly the same for $A\beta_{40}$ independently of the simulation temperature and the parameter set used but is significantly improved for $A\beta_{42}$ at 325 K, indicating that this latter temperature is most relevant for comparison with experiment.

Using the same residues and experimental J-coupling constant values²⁸ as benchmark, our results can be compared to the PCC obtained by other computational studies. For $A\beta_{42}$, our PCC value of 0.43 agrees with the values derived from all-atom REMD simulations in explicit solvent using OPLS (PCC of 0.43 using 60 ns per replica) and AMBER99SB (PCC of 0.4–0.5 using

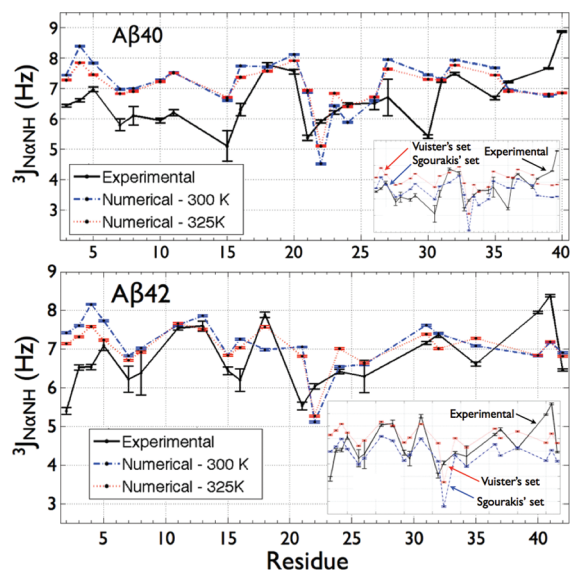


Figure 2. J-coupling constants $^3J_{\text{NH}}$ for $A\beta_{40}$ and $A\beta_{42}$ at 300 and 325 K. The J-coupling constants determined experimentally²⁸ for $A\beta_{40}$ and $A\beta_{42}$ are compared to the predicted J-coupling constants obtained in our simulations. The inset in each figure compares experiment²⁸ (black curve) and the numerically predicted J-coupling constants at 325 K using Sgourakis³⁵ and Vuister's⁵⁹ (blue and red curves, respectively) parameter sets in Karplus equation.⁵⁸ Glycine residues are not included because their experimental values are ambiguous.

Table 1. PCC between Experimental and Calculated J-Coupling Constants^a

parameter set	Vuister et al. ⁵⁹		Sgourakis et al. ³⁵		Schmidt et al. ⁶⁰	
	all	part	all	part	all	part
$A\beta_{40}$ - 325K	0.28	0.45	0.29	0.57	0.29	0.57
$A\beta_{40}$ - 300K	0.26	0.40	0.27	0.53	0.27	0.50
$A\beta_{42}$ - 325K	0.43	0.65	0.42	0.72	0.43	0.71
$A\beta_{42}$ - 300K	0.27	0.52	0.29	0.39	0.27	0.44

^a Glycine residues are not included because their experimental values are ambiguous. The first and the second columns for each parameter set contain, respectively, the PCC obtained when using all the data points and when neglecting the four most problematic residues. For $A\beta_{40}$, we neglected residues 15, 21, 30, and 40 when using Vuister et al.,⁵⁹ residues 22, 30, 39, and 40 when using Sgourakis et al.,³⁵ and residues 22, 30, 39, and 40 when using Schmidt et al.⁶⁰ parameter sets. For $A\beta_{42}$, we neglected residues 2, 21, 40, and 41 when using Vuister et al., residues 2, 22, 40, and 41 when using Sgourakis et al., and residues 2, 22, 40, and 41 when using Schmidt et al. parameter sets.

225 ns per replica) force fields.^{28,35} For $A\beta_{40}$, our PCC value of 0.29 is lower than that obtained by all-atom REMD using OPLS (PCC of 0.66) but is similar to that obtained using all-atom REMD with the AMBER96 force field (PCC of 0.27).²⁸ Our moderate correlation for $A\beta_{40}$ is due a shift of more than ~ 1 Hz for most of the N-terminal residues and the four outliers at positions Glu22, Ala30, Val39, and Val40.

Though there is no physical rationale for eliminating outliers, to compare with Mitternacht et al.,³⁴ we follow their procedure and eliminate the four most significant outliers out of the 24 and 21 experimental data points for $A\beta_{40}$ and $A\beta_{42}$, respectively, and compute the PCC using the Schmidt's parameters set.⁶⁰ Doing

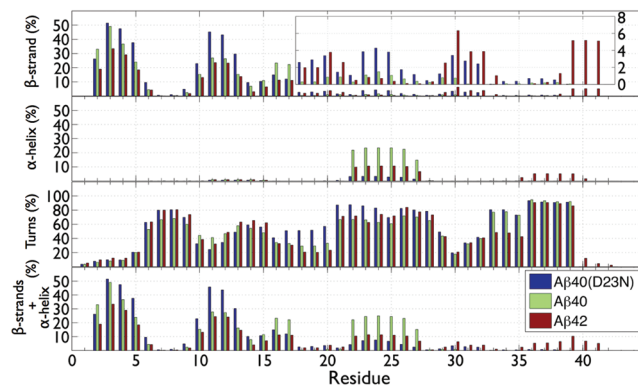


Figure 3. Per residue secondary structure propensities. From top to bottom: β -strand, α -helix, turn, and α -helix plus β -strand propensities in percentage (%). $A\beta_{40}$ is shown in green, $A\beta_{42}$ in red, and $A\beta_{40}(\text{D23N})$ in blue.

so, our PCCs increase from 0.29 to 0.57 for $A\beta_{40}$ and from 0.43 to 0.71 for $A\beta_{42}$, showing that only a few outlying points strongly impact the PCC (Table 1). Using the same approach, our PCC for $A\beta_{42}$, 0.71 is comparable to the PCC of 0.86 determined by Mitternacht et al.³⁴

3.2. $A\beta_{40}$ Properties. Secondary Structure. Averaged over all structures and residues, $A\beta_{40}$ has $7.7 \pm 0.1\%$ β -strand, $3.4 \pm 0.1\%$ α -helix, $50.2 \pm 0.1\%$ turn, and $38.8 \pm 0.1\%$ random coil. Compared to a recent CD study on $A\beta_{40}$ (8% α -helix and 24% β -strand),³² our ensemble shows a comparable propensity for α -helices but underestimates the percentage of β -strands. This experiment, however, may overestimate the percentage of β -strands because NMR shows a very weak signal,^{13,16} and an earlier CD experiment using a more stringent sample separation protocol reports only $\sim 12\%$ of β -strands.³

The secondary structure propensities for each residue in $A\beta_{40}$ are shown in Figure 3. Three β -strands are present from Ala2 to Arg5 with a propensity of 20–50%, from Glu10 to His13 and Lys16 to Leu17 with a propensity of 15–30%, and from Phe18 to Ala30 with a small propensity ($<1\%$). The first two β -strands are stabilized by turns at His6-Gly9 and His13-Gln15. The β -strands 2 and 3 are found in competition with 2 α -helices spanning Glu11-Gln15 with a small probability ($<1\%$) and Glu22-Asn27 with a probability of up to 25%, as depicted in Figure 3. Finally, our analysis also shows a high turn signal ($>60\%$) between residues 21 and 28 and between residues 33 and 39.

Tertiary Structure. Several contacts play a dominant role in the tertiary structure of $A\beta_{40}$, as shown in Figure 4. The C-terminal interacts with the CHC forming hydrophobic contacts between Leu34-Phe19, Val36-Phe19, Met35-Phe19, and Met35-Phe20 with high probabilities and between Ile31-Ile32 and Val18 with slightly lower probabilities. We also observe hydrophobic contacts localized in the CHC between Leu17-Phe19 and Val18-Phe20, indicating a mostly collapsed core. We also find that the electrostatic Asp23-Lys28 contact, present in about 65% of the morphologies, stabilizes a turn between Ala21 and Lys28.

Column 1 of Figure 5 shows the center of the five dominant $A\beta_{40}$ clusters with their respective populations. These five clusters contain 34.3% of all generated conformations. The center of cluster 1 is fully random coil. The centers of clusters 2–5 display two β -strands at the N-terminal residues 2–4 and 10–12, leading to a well-formed or a very flexible β -hairpin for the clusters 4–5 and 2–3, respectively. While the cluster centers do

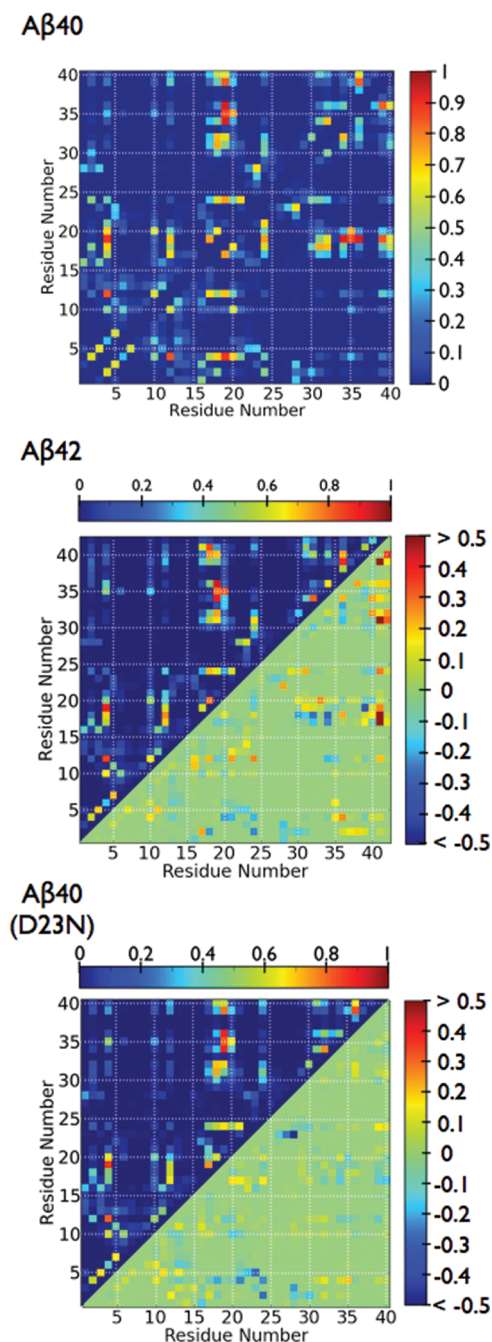


Figure 4. Contact maps between side chains. The contact maps of $A\beta_{40}$, $A\beta_{42}$ and $A\beta_{40}$ (D23N) are shown in the top, middle, and bottom panels, respectively. The upper left corner of each panel depicts the propensity of contact between side chains in each alloform. The lower right corner for $A\beta_{42}$ and $A\beta_{40}$ (D23N) displays the contact differences between each alloform and $A\beta_{40}$. When there is more contacts in $A\beta_{40}$, the propensity is negative (blueish), and the opposite yields a positive propensity (reddish).

not exhibit secondary structure in the CHC or the C-terminal, the loop region is structurally near a α -helical conformation, and we find a α -helix at residues 22–26 with $\sim 45\%$ probability when averaging over all conformations belonging to the top three clusters. All five clusters display contacts between CHC and the C-terminal. Clusters 1–3 also show contacts between residues 1–15 and 21–38, while these contacts are absent in clusters 4 and 5.

3.3. $A\beta_{42}$ Compared to $A\beta_{40}$. *Secondary Structure.* Averaged over all structures and residues, $A\beta_{42}$ shows $2.1 \pm 0.1\%$ α -helix, $6.1 \pm 0.1\%$ β -strand, $48.3 \pm 0.1\%$ turn, and $43.4 \pm 0.1\%$ random coil contents, whereas $A\beta_{40}$ exhibits $3.4 \pm 0.1\%$ α -helix, $7.7 \pm 0.1\%$ β -strand, $50.2 \pm 0.1\%$ turn, and $38.8 \pm 0.1\%$ random coil. Thus, these two sequences have the same overall secondary structure composition. It is in the propensity per residue that differences arise.

The secondary structure propensity for each residue is shown in Figure 3. As for $A\beta_{40}$, there are β -strands at the N-terminal at positions Ala2-Arg5, Glu10-His13, and Lys16-Val17, but their propensities are smaller in $A\beta_{42}$. Noticeably, $A\beta_{42}$ is more prone than $A\beta_{40}$ to form β -strands at the CHC and at positions Ala30-Ile32 and Val39-Ile41, as depicted by the inset in Figure 3. Thus, there is a small random coil to β -strand transition at the CHC and the C-terminal in going from $A\beta_{40}$ to $A\beta_{42}$. Overall, 1.4% of $A\beta_{40}$ and 8.6% of $A\beta_{42}$ structures display β -strands at the CHC or C-terminal. As for $A\beta_{40}$, there are in $A\beta_{42}$ two α -helices at positions Glu11-Gln15 and Glu22-Asn27 (Figure 3). The α -helix at Glu22-Asn27 is however reduced in going from $A\beta_{40}$ ($\sim 22\%$) to $A\beta_{42}$ ($\sim 10\%$). In addition, we find in $A\beta_{42}$ a third helix at positions Met35-Val40 that is absent in $A\beta_{40}$ (Figure 3). On the other hand, both alloforms share a very similar turn profile.

Tertiary Structure. Many hydrophobic contacts contribute to the $A\beta_{42}$ monomer morphologies. Most of them are also present in $A\beta_{40}$, as depicted in Figure 4 that shows the differences in the contact propensity between the two alloforms. These include a number of contacts between the CHC and the C-terminal, particularly between Met35 and the two residues Phe19 and Phe20. The hydrophobic residue Ile41 increases the overall contact occurrence between the CHC and the C-terminal by interacting predominantly with Leu17 and Val18. It also increases contacts inside the C-terminal itself. Finally, the addition of Ile41 and Ala42 causes a small shift of the interactions between the N-terminal and the region 15–35 toward the C-terminal or the N-terminal. This shift may allow the conformational freedom necessary for the CHC to form β -strands and for the region 22–29 to form more easily the specific loop conformation required for fibrillation. Electrostatic interactions also play a role in $A\beta_{42}$ structures. $A\beta_{42}$ and $A\beta_{40}$ have approximately the same propensity of forming a contact between Asp23 and Lys28 ($\sim 60\%$) with a similar contact distance distribution (data not shown).

While the probability of contact occurrence differs between $A\beta_{40}$ and $A\beta_{42}$, both alloforms have a very similar conformational ensemble, and only 16% of all $A\beta_{42}$ conformations are unique, i.e., show a C_{α} -rmsd greater than 2 Å from any observed $A\beta_{40}$ structures by using residues 1–40. The centers of the five dominant $A\beta_{42}$ clusters are shown in column 2 of Figure 5. These five clusters include 50.1% of all conformations. We observe that some of these clusters are structurally similar to the main clusters of $A\beta_{40}$. For instance, clusters 2 and 4 of $A\beta_{42}$ are respectively similar to clusters 2 and 5 of $A\beta_{40}$. The third cluster is similar to the cluster 6 of $A\beta_{40}$ that represents 1.2% of its structural ensemble. On the other hand, the first cluster of $A\beta_{42}$ is only similar to marginal clusters of $A\beta_{40}$ ($<1\%$), and its fifth cluster is not present in $A\beta_{40}$.

To identify the structural differences between $A\beta_{42}$ and $A\beta_{40}$, we cluster all unique structures of $A\beta_{42}$. The first five unique clusters of $A\beta_{42}$ are shown in column 4 of Figure 5 and represent 6% of all generated conformations. We observe that these $A\beta_{42}$ unique morphologies display β -strands at the C-terminal and CHC. In particular, the first, third, and fifth unique clusters show

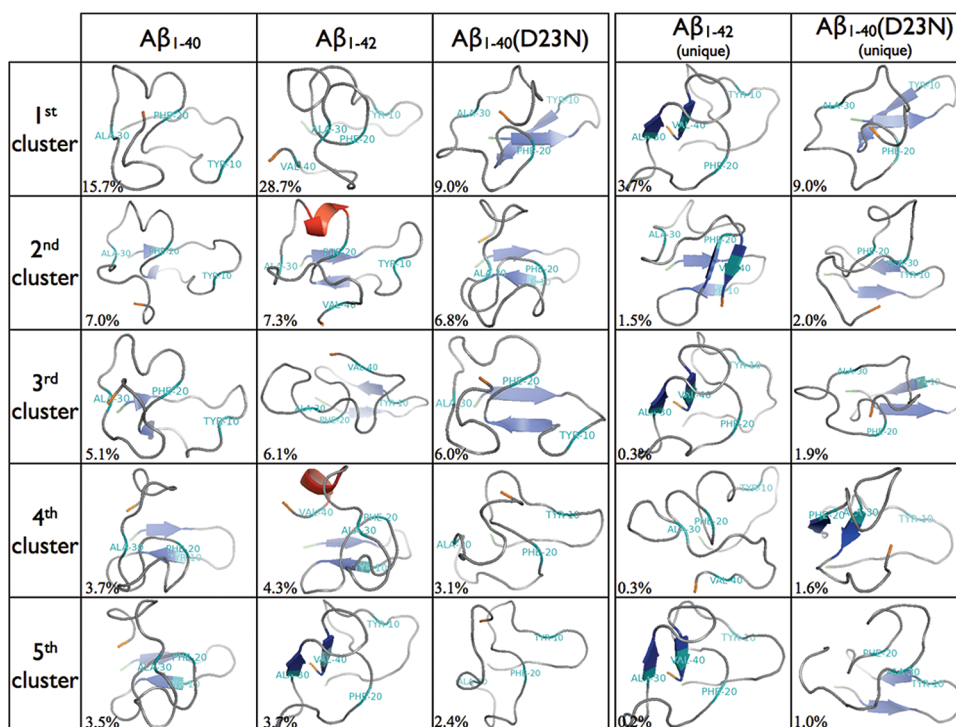


Figure 5. Clusters centers. Centers of the five dominant clusters for $A\beta_{40}$ (first column), $A\beta_{42}$ (second column), and $A\beta_{40}(D23N)$ (third column). The fourth and fifth columns contain, respectively, the centers of the five dominant clusters of $A\beta_{42}$ and $A\beta_{40}(D23N)$ unique morphologies. The percentages shown are with respect to the total number of structures analyzed. The secondary structure was assigned using STRIDE.⁵⁵ The N- and C-termini are shown in green and orange, respectively, for all alloforms, Tyr10, Phe20, Ala30, and Val40, and for $A\beta_{42}$ only, are shown in teal.

a β -hairpin motif with strands spanning the C-terminal at residues 30–32 and 39–41. Structurally, these three clusters are near each other differing only at the N-terminal. Another interesting motif is seen in the second unique cluster, where we note that the extremity of the C-terminal may also interact with the CHC forming a β -sheet between residues 18–20 and 39–41.

3.4. $A\beta_{40}(D23N)$ Compared to $A\beta_{40}$. Secondary Structure. Averaged over all structures and residues, $A\beta_{40}(D23N)$ is composed of $10.1 \pm 0.1\%$ β -strand vs $7.7 \pm 0.1\%$ in $A\beta_{40}$, $0.5 \pm 0.1\%$ α -helix vs $3.4 \pm 0.1\%$ in $A\beta_{40}$, $56.2 \pm 0.1\%$ turn, which is 6.0% higher than in $A\beta_{40}$, and $33.2 \pm 0.1\%$ random coil.

Differences are also observed in the per residue secondary structure, as shown in Figure 3. For example, the β -strand propensity of residues Tyr10-His14 is higher in $A\beta_{40}(D23N)$ than in $A\beta_{40}$ causing a lower β -strand propensity at Gln15-Leu17. $A\beta_{40}(D23N)$ and $A\beta_{42}$ show similar β -strand propensities at positions Val18-Ala21 and Ala30-Ile32 that are absent in $A\beta_{40}$ (Figure 3). Overall, 5.0% of $A\beta_{40}(D23N)$ structures vs 1.4% of $A\beta_{40}$ structures display β -strands at the CHC and the C-terminal. From our simulations, we also see that $A\beta_{40}(D23N)$ has a negligible propensity for α -helical configurations, with a small signal at positions 22–27, a feature that may favor a faster appearance of fibril-compatible intermediate oligomers with β -strands at the CHC and the C-terminal in $A\beta_{40}(D23N)$ than in $A\beta_{40}$.

The turn propensity in the N-terminal, at His6-Gly9 and His13-Gln15, is very similar to $A\beta_{40}$ and $A\beta_{42}$. Differences are however observed for the CHC, the loop, region and the C-terminal. At the CHC, residues 17–20 have a propensity for turn of $\sim 50\%$, which is higher than in $A\beta_{40}$ and $A\beta_{42}$ by ~ 20 – 30% , resulting in less random coils. The C-terminal also

exhibits a higher propensity for turn in $A\beta_{40}$ and $A\beta_{40}(D23N)$ than in $A\beta_{42}$, particularly at residues 33–35, generating a longer turn extending from residues 33 to 39. Finally, the loop region has a higher propensity for turn in $A\beta_{40}(D23N)$ than in the wild-type alloforms, and the contacts that stabilize this region are very different from $A\beta_{40}$ as described below.

Tertiary Structure. While 16% of $A\beta_{42}$ conformations differ from any observed $A\beta_{40}$ conformations, cluster analysis reveals that 35% of $A\beta_{40}(D23N)$ conformations are distinct from those obtained for $A\beta_{40}$ (C_{α} -rmsd > 2 Å). Differences in the contact distribution between $A\beta_{40}$ and $A\beta_{40}(D23N)$ are shown in Figure 4. For instance, there is a shift of the contacts between the N-terminal and the residues 15–30 toward the C- or N-terminal. While this shift is qualitatively similar to what occurs in $A\beta_{42}$ as described previously, the overall reduction in contacts is more important in $A\beta_{40}(D23N)$. The N-terminal of $A\beta_{40}(D23N)$ loses many contacts with the region 15–30 without recovering them all with the C-terminal, as shown by Figure 4, allowing more conformational freedom. Another main difference is observed for the Asp(Asn)23-Lys28 contact propensity which is reduced by 54% in going from $A\beta_{40}(D23N)$ to $A\beta_{40}$ (Figure 4). This induces a rich polymorphism for the region between Ala21 and Ala30 as opposed to the other two alloforms studied. Noticeably, the side chains of Asn23 and Lys28 in $A\beta_{40}(D23N)$ are then unconstrained as they do not form other significant contacts (Figure 4).

The centers of the five dominant clusters of $A\beta_{40}(D23N)$ are displayed in the third column of Figure 5. These represent 27.3% of all generated conformations. As for $A\beta_{40}$, they are mostly disordered without any β -strands at the CHC or C-terminal as could be expected, since 65% of $A\beta_{40}(D23N)$ and $A\beta_{40}$

conformations are structurally similar (C_{α} -rmsd < 2 Å). For instance, the centers of the clusters 2 and 3 of $A\beta_{40}$ (D23N) are similar to the centers of the clusters 4 and 1 of $A\beta_{40}$, respectively. On the other hand, cluster 4 of $A\beta_{40}$ (D23N) is only similar to $A\beta_{40}$ marginal clusters (<1%), and its first cluster is not found in the structural ensemble of $A\beta_{40}$.

The $A\beta_{40}$ (D23N) unique clusters are shown in the fifth column of Figure 5. The first unique cluster, which has a population of 9%, is also the most dominant cluster for this sequence (see column 3). In term of secondary structure, this cluster is rather unstructured with β -strands only at residues 2–5 and 11–16. The fourth unique cluster displays a β -sheet between residues 18–20 and 30–32 and represents 1.6% of all visited structures. While the other $A\beta_{40}$ (D23N) unique clusters are as unstructured as $A\beta_{40}$, their contact distributions differ in two notable ways: (1) Their N-terminals can be isolated from the loop region and interact less with the CHC (unique clusters 1 and 3); their contact maps show very few contact between residues 1–15 and 22–29 and between residues 5–15 and 14–21 (data not shown), or (2) these unique clusters can have few contacts between residues 1–9 and 22–40 (unique clusters 2 and 5). For these two latter unique clusters, we note also that they exhibit electrostatic interactions between the positively charged Lys28 and the negatively charged Glu11, as Lys28 now interacts very weakly with Asn23 as opposed to wild-type. Overall, these unique morphologies do not have contacts between the N-terminal and the loop region allowing $A\beta_{40}$ (D23N) to be less compact than $A\beta_{40}$.

4. DISCUSSION

Knowledge of the $A\beta$ conformations at the monomer level in aqueous solution is of utmost importance since the monomer interacts with higher order oligomers⁶ and fibrils¹⁹ and is a building block of the cytotoxic dimer.²¹ Both $A\beta_{40}$ and $A\beta_{42}$ monomers have been studied by NMR, CD, and computer simulations. Results show that these peptides are described by a distinct ensemble of predominantly random coil structures. While many simulations were performed,²⁹ very few showed thorough sampling of the relevant morphological ensemble.^{34,35} The similarities and differences of our results with previous experiments and simulations on the $A\beta_{40}$ and $A\beta_{42}$ monomers can be summarized as follows.

Our contact distributions of $A\beta_{40}$ and $A\beta_{42}$ can be compared to the previous all-atom implicit solvent simulation results obtained by Yang and Teplow.²⁹ We note two differences. First, while we observe a fourth region between residues 30 and 40 in which intraregion contacts are important for both $A\beta_{40}$ and $A\beta_{42}$, the role of this region was only identified for $A\beta_{42}$ by Yang and Teplow. Second, there is a slight shift in the contacts between the N-terminal and the loop region. If these regions were seen to interact by contacts between residues 6–10 and 22–28 in the simulations of Yang and Teplow,²⁹ our results show interactions between residues 1–5 and 16–28 (Figure 4). On the other hand, both simulations agree on several points: the N-terminal interaction with the loop region, the increased number of contacts between the CHC and the C-terminal for $A\beta_{42}$, the presence of more localized contacts at the C-terminal for $A\beta_{42}$, and the fact that intraregion contacts are more dominant in the four regions corresponding to the turn distribution observed in Figure 3.

Inspection of the four sequence regions show interesting features. We find that the N-terminal (residues 1–16) of the

two wild-type alloforms is very similar with turns at His6-Gly9 and His13-Gln15, stabilizing extended morphologies having β -strands. A weak turn was also observed at Asp7-Glu11 by NMR,¹⁶ and two turns were predicted at His6-Gly9 and His14-Lys16 using all-atom REMD simulations.^{28,29} On the other hand, the presence of β -strands at the N-terminal of $A\beta_{40}$ is clearly a matter of debate from experimental and computational studies. While some experiments show that the N-terminal of $A\beta_{40}$ monomer may form β -strands,⁶³ others observe that it is extended and highly fluctuating^{14,17} without any β -strands.¹⁶ Three all-atom REMD simulations report either negligible²⁸ or low^{29,35} β -strand percentages. In contrast, coarse-grained DMD simulations report a propensity for β -strand of ~40% at Ala2-Phe4.⁶⁴ Taken together, all these studies suggest that the N-terminal may be in rapid exchange between an extended–turn–extended motif free of any H-bonds and β -stranded configurations, with a turn at His6-Gly9 being formed most of the time, as shown in our study by the propensities in Figure 3. We find that $A\beta_{42}$ has also a non-negligible probability to populate a β -hairpin at the N-terminal, a motif that has been predicted recently with two different all-atom potentials.^{34,35} This motif has however a lower probability in $A\beta_{42}$ than in $A\beta_{40}$. This motif, which was not observed experimentally yet due to its low population, might have consequences on the early formed $A\beta_{42}$ oligomers.

The formation of a loop region between residues 22–29 stabilized by a salt bridge between Asp23-Lys28 is thought to be one of the rate-limiting steps of $A\beta_{40}$ fibrillation.⁶⁵ Recently, the formation of this contact was shown to increase the population of competent fibril-like monomers for $A\beta_{10-35}$.²⁶ Here, we observe many morphologies with a turn between Ala21 and Lys28 stabilized by interactions between the side chains of Asp23 and Lys28 in the wild-type sequences (Figures 3 and 4), with a probability of ~65%. This turn was suggested from proteolysis experiment on the $A\beta_{40}$ monomer,¹⁸ and its morphology was resolved at the molecular level using solution NMR of the $A\beta_{21-30}$ peptide.¹⁸ The turn ensemble of residues 22–28 for $A\beta_{40}$, in our simulation, is marginally similar to the two NMR conformations of the $A\beta_{21-30}$ peptide with 20% of our conformations deviating by less than 1.5 Å. These differences can be explained by the presence of contacts between the N-terminal and the loop region with ~35% probabilities (Figure 4). Overall, our results on $A\beta_{40}$ are consistent with experiments on the $A\beta_{40}$ monomer showing turns at positions 23–29,¹⁴ 20–26,¹⁶ or 24–28.¹⁸ Our results on $A\beta_{40}$ also agree with previous REMD simulations on the $A\beta_{40}$ monomer²⁹ and MD simulations on the $A\beta_{10-35}$ monomer and dimer²⁶ that suggest that the loop region, residues 22–29, needs to undergo structural changes during fibrillation, leading to further stabilization of the Asp23-Lys28 contact.

The CHC (residues 17–21) and the C-terminal (residues 30–42) are recognized experimentally as the driving regions for aggregation through the formation of β -stranded and extended structures.^{66–68} Experiments revealed that the C-terminal of $A\beta_{42}$ monomer is more rigid than that of $A\beta_{40}$ ¹⁵ and displays β -strands at residues Val39-Ile41.^{14,16} Previous simulations also observed a more structured C-terminal in $A\beta_{42}$ than in $A\beta_{40}$.^{28,29,35,64} In agreement with these experiments and simulations, we observe in our study that the C-terminal residues Ala39-Ile41 modulate, by direct interactions, the appearance of distinct $A\beta_{42}$ morphologies with β -strands at the CHC and the C-terminal that are absent in the $A\beta_{40}$ ensemble. We isolated these morphologies which exhibit a β -sheet between residues 39–41

and 30–32 or between residues 39–41 and 18–20 (see fourth column in Figure 5). Our results emphasize the important role of Ile41-Ala42 in the monomer morphologies, and its role in the early oligomerization process as was observed experimentally.^{3,69}

Our findings on the aggregation prone mutant D23N can be summarized as follows. $A\beta_{40}$ (D23N) appears to form a more stable β -hairpin than the two wild-type peptides, as a significant portion of their N-terminal structures. As could be expected, $A\beta_{40}$ (D23N) monomer does not exhibit any significant interaction between Asn23 and Lys28. This region displays rather a high conformational flexibility giving the D23N alloform more freedom to adopt the loop conformations observed in its fibrillar states³⁹ and therefore reducing the free energy barriers. Interestingly, solid-state NMR indicates that $A\beta_{40}$ (D23N) forms fibrils with multiple morphologies, with a majority having anti-parallel β -sheets and a minority having parallel β -sheets.³⁹ From our simulations, we see that this wide polymorphism may originate in part from the unconstrained side chains of Asn23 and Lys28, the flexibility of the region 21–29, and their overall consequences on $A\beta_{40}$ (D23N) equilibrium ensembles.

Our results show that, while $A\beta_{40}$ (D23N) does not contain Ile41, the mutation D23N increases the β -strand content at the CHC and the C-terminal when compared to $A\beta_{40}$ (Figure 3). Specifically, one of our unique $A\beta_{40}$ (D23N) morphologies displays a β -sheet between residues 18–20 and 30–32 (fifth column in Figure 5). Even if its weight of 1.6% is small, this motif is interesting as it involves two regions known to be crucial during oligomerization.^{6,66,67}

The higher propensity of $A\beta_{40}$ (D23N) monomer to form β -strands at the CHC and the C-terminal and to prefer less collapsed topologies than $A\beta_{40}$ monomer suggests important consequences on its dock-and-lock mechanism when it binds to fibril edge. Previous simulations on the $A\beta_{10-40}$ monomer at the edge of a preformed fibril showed that the fibril edge induces a conversion of the CHC to β -stranded configurations in the monomer.^{31,70} In another computational study, interactions between the C-terminal peptide Met35-Val40 and a fibril were simulated showing a transition of the peptide from random coil to extended configuration upon binding to fibril edge.⁷¹ Taken together, our results suggest that $A\beta_{40}$ (D23N) may exhibit easier conversion to extended configuration upon binding to the fibril edge and therefore increases its fibril elongation rate.

Finally, it was observed using photoinduced cross-linking experiments that D23N promotes the formation of high-order oligomers when compared with wild-type $A\beta_{40}$.⁶ High-order oligomers are also formed by $A\beta_{42}$, which preferably forms pentamer and hexamer, in contrast to $A\beta_{40}$, which forms low-order oligomers, such as monomers to tetramers in rapid equilibrium.⁶ In our simulation, we observe that both $A\beta_{42}$ and $A\beta_{40}$ (D23N) monomers exhibit a noticeable reduction of contacts between the N-terminal and the residues 22–29 and between the N-terminal and the residues 18–21 of the CHC (Figure 4), a feature that could favor the appearance of high-order oligomers by freeing the segment 22–29 and the CHC.

5. CONCLUSION

In this study, we have revisited the structures of the $A\beta_{40}$ and $A\beta_{42}$ monomers using the OPEP force field and the HT-REMD technique and provided for the first time the structures of the peculiar $A\beta_{1-40}$ (D23N) monomer, which self-assembles

without any lag phase to predominant amyloid fibrils with antiparallel β -sheets.³⁹

We observe that even if these three alloforms are mostly disordered at the monomeric level, in agreement with experiments and previous simulations on $A\beta_{1-40}$ and $A\beta_{1-42}$, striking morphological differences exist. For instance, $A\beta_{1-42}$ has higher β -hairpin propensities at the C-terminal, residues 30–42, than that of $A\beta_{1-40}$. Moreover, 16% of all $A\beta_{1-42}$ conformations do not resemble to any $A\beta_{1-40}$ conformations.

The monomeric morphological ensemble of $A\beta_{40}$ (D23N) also presents significantly different from that of $A\beta_{40}$ with 35% of unique conformations. In particular, we observe a β -hairpin-like motif between residues 18–20 and 30–32 in $A\beta_{1-40}$ (D23N) that is not present in $A\beta_{1-40}$. Moreover, D23N enhances the conformational freedom of residues 22–30. It changes the network of contacts, the N-terminal becoming more independent from the rest of the protein, leading to less compact morphologies than the wild-type $A\beta_{1-40}$ peptide. These structural properties could explain why the kinetics and the final amyloid products vary so extensively between the $A\beta_{1-40}$ and $A\beta_{40}$ (D23N) peptides, by increasing the population of the amyloid-competent monomeric state.

AUTHOR INFORMATION

Corresponding Author

*E-mail: normand.mousseau@umontreal.ca.

ACKNOWLEDGMENT

This work was funded in part by the Canada Research Chairs program, the Fonds québécois de recherche sur la nature et les technologies (FQRNT), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Fonds de recherche en santé du Québec (FRSQ). Calculations were done on the computers of the Réseau québécois de calcul de haute performance. Financial support of CNRS and the Institut Universitaire de France is also acknowledged.

REFERENCES

- (1) Selkoe, J. D. *Nature* **2003**, *426*, 900–904.
- (2) Haass, C.; Selkoe, D. J. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 101–112.
- (3) Kirkitadze, M. D.; Condron, M. M.; Teplow, D. B. *J. Mol. Biol.* **2001**, *312*, 1103–1119.
- (4) Olofsson, A.; Lindhagen-Persson, M.; Sauer-Eriksson, A. E.; Ohman, A. *Biochem. J.* **2007**, *404*, 63–70.
- (5) Petkova, A. T.; Leapman, R. D.; Guo, Z.; Yau, W.-M.; Mattson, M. P.; Tycko, R. *Science* **2005**, *307*, 262–265.
- (6) Bitan, G.; Kirkitadze, M. D.; Lomakin, A.; Vollers, S. S.; Benedek, G. B.; Teplow, D. B. *PNAS* **2003**, *100*, 330–335.
- (7) Chen, Y.-R.; Glabe, C. G. *J. Biol. Chem.* **2006**, *281*, 24414–24422.
- (8) Lublin, A. L.; Gandy, S. *Mt. Sinai J. Med.* **2010**, *77*, 43–49.
- (9) Sakono, M.; Zako, T. *FEBS J.* **2010**, *277*, 1348–1358.
- (10) Faller, P. *ChemBioChem* **2009**, *10*, 2837–2845.
- (11) Bokvist, M.; Lindstrom, F.; Watts, A.; Grobner, G. *J. Mol. Biol.* **2004**, *335*, 1039–1049.
- (12) Selkoe, J. D. *Behav. Brain Res.* **2008**, *192*, 106–113.
- (13) Zhang, S.; Iwata, K.; Lachenmann, M. J.; Peng, J. W.; Li, S.; Stimson, E. R.; Lu, Y.-a.; Felix, A. M.; Maggio, J. E.; Lee, J. P. *J. Struct. Biol.* **2000**, *130*, 130–141.
- (14) Lim, K. H.; Collver, H. H.; Le, Y. T. H.; Nagchowdhuri, P.; Kenney, J. M. *Biochem. Biophys. Res. Commun.* **2007**, *353*, 443–449.
- (15) Yan, Y.; Wang, C. *J. Mol. Biol.* **2006**, *364*, 853–862.

- (16) Hou, L.; Shao, H.; Zhang, Y.; Li, H.; Menon, N. K.; Neuhaus, E. B.; Brewer, J. M.; Byeon, I.-J. L.; Ray, D.; Vitek, M. P.; Iwashita, T.; Makula, R. A.; Przybyla, A. B.; Zagorski, M. G. *J. Am. Chem. Soc.* **2004**, *126*, 1992–2005.
- (17) Zhang, A.; Qi, W.; Good, T. A.; Fernandez, E. J. *Biophys. J.* **2009**, *96*, 1091–1104.
- (18) Lazo, N. D.; Grant, M. A.; Condrón, M. C.; Rigby, A. C.; Teplow, D. B. *Protein Sci.* **2005**, *14*, 1581–1596.
- (19) Esler, W. P.; Stimson, E. R.; Jennings, J. M.; Vinters, H. V.; Ghilardi, J. R.; Lee, J. P.; Mantyh, P. W.; Maggio, J. E. *Biochemistry* **2000**, *39*, 6288–6295.
- (20) Giuffrida, M. L.; Caraci, F.; Pignataro, B.; Cataldo, S.; De Bona, P.; Bruno, V.; Molinaro, G.; Pappalardo, G.; Messina, A.; Palmigiano, A.; Garozzo, D.; Nicoletti, F.; Rizzarelli, E.; Copani, A. *J. Neurosci.* **2009**, *29*, 10582–10587.
- (21) Shankar, G. M.; Li, S.; Mehta, T. H.; Garcia-Munoz, A.; Shepardson, N. E.; Smith, I.; Brett, F. M.; Farrell, M. A.; Rowan, M. J.; Lemere, C. A.; Regan, C. M.; Walsh, D. M.; Sabatini, B. L.; Selkoe, D. J. *Nat. Med.* **2008**, *14*, 837–842.
- (22) Xu, Y.; Shen, J.; Luo, X.; Zhu, W.; Chen, K.; Ma, J.; Jiang, H. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 5403–5407.
- (23) Flock, D.; Colacino, S.; Colombo, G.; Di Nola, A. *Proteins* **2006**, *62*, 183–192.
- (24) Raffa, D. F.; Rauk, A. *J. Phys. Chem. B* **2007**, *111*, 3789–3799.
- (25) Triguero, L.; Singh, R.; Prabhakar, R. *J. Phys. Chem. B* **2008**, *112*, 7123–7131.
- (26) Reddy, G.; Straub, J. E.; Thirumalai, D. *J. Phys. Chem. B* **2009**, *113*, 1162–1172.
- (27) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (28) Sgourakis, N. G.; Yan, Y.; McCallum, S.; Wang, C.; Garcia, A. E. *J. Mol. Biol.* **2007**, *368*, 1448–1457.
- (29) Yang, M.; Teplow, D. B. *J. Mol. Biol.* **2008**, *384*, 450–464.
- (30) Anand, P.; Nandel, F. S.; Hansmann, U. H. E. *J. Chem. Phys.* **2008**, *129*, 195102.
- (31) Takeda, T.; Klimov, D. K. *Proteins* **2009**, *77*, 1–13.
- (32) Ono, K.; Condrón, M. M.; Teplow, D. B. *PNAS* **2009**, *106*, 14745–14750.
- (33) Ono, K.; Condrón, M. M.; Teplow, D. B. *J. Biol. Chem.* **2010**, *285*, 23186–23197.
- (34) Mitternacht, S.; Staneva, I.; Hard, T.; Irback, A. *Proteins* **2010**, *78*, 2600–2608.
- (35) Sgourakis, N. G.; Merced-Serrano, M.; Boutsidis, C.; Drineas, P.; Du, Z.; Wang, C.; Garcia, A. E. *J. Mol. Biol.* **2011**, *405*, 570–583.
- (36) Maupetit, J.; Tuffery, P.; Derreumaux, P. *Proteins* **2007**, *69*, 394–408.
- (37) Laghaei, R.; Mousseau, N.; Wei, G. *J. Phys. Chem. B* **2010**, *114*, 7071–7077.
- (38) Grabowski, T. J.; Cho, H. S.; Vonsattel, J. P. G.; Rebeck, G. W.; Greenberg, S. M. *Ann. Neurol.* **2001**, *49*, 697–705.
- (39) Tycko, R.; Sciarretta, K. L.; Orgel, J. P. R. O.; Meredith, S. C. *Biochemistry* **2009**, *48*, 6072–6084.
- (40) Petkova, A. T.; Ishii, Y.; Balbach, J. J.; Antzutkin, O. N.; Leapman, R. D.; Delaglio, F.; Tycko, R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 16742–16747.
- (41) Luhrs, T.; Ritter, C.; Adrian, M.; Riek-Loher, D.; Bohrmann, B.; Döbeli, H.; Schubert, D.; Riek, R. *PNAS* **2005**, *102*, 17342–17347.
- (42) Tycko, R. *Q. Rev. Biophys.* **2006**, *39*, 1–55.
- (43) Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24–34.
- (44) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (45) Chebaro, Y.; Dong, X.; Laghaei, R.; Derreumaux, P.; Mousseau, N. *J. Phys. Chem. B* **2009**, *113*, 267–274.
- (46) Lu, Y.; Derreumaux, P.; Guo, Z.; Mousseau, N.; Wei, G. *Proteins* **2009**, *75*, 954–963.
- (47) Laghaei, R.; Mousseau, N. *J. Chem. Phys.* **2010**, *132*, 165102.
- (48) Liang, C.; Derreumaux, P.; Mousseau, N.; Wei, G. *Biophys. J.* **2008**, *95*, 510–517.
- (49) Chen, W.; Mousseau, N.; Derreumaux, P. *J. Chem. Phys.* **2006**, *125*, 084911.
- (50) Dong, X.; Chen, W.; Mousseau, N.; Derreumaux, P. *J. Chem. Phys.* **2008**, *128*, 125108.
- (51) Chebaro, Y.; Mousseau, N.; Derreumaux, P. *J. Phys. Chem. B* **2009**, *113*, 7668–7675.
- (52) Maupetit, J.; Derreumaux, P.; Tuffery, P. *J. Comput. Chem.* **2010**, *31*, 726–738.
- (53) Maupetit, J.; Derreumaux, P.; Tuffery, P. *Nucleic Acids Res.* **2009**, *Jul 1*, 37 (Web Server issue), W498–503.
- (54) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116*, 9058–9067.
- (55) Frishman, D.; Argos, P. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 566–579.
- (56) Daura, X.; Suter, R.; van Gunsteren, W. F. *J. Chem. Phys.* **1999**, *110*, 3049–3055.
- (57) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 26–41.
- (58) Karplus, M.; Anderson, D. H. *J. Chem. Phys.* **1959**, *30*, 6–10.
- (59) Vuister, G. W.; Bax, A. *J. Am. Chem. Soc.* **1993**, *115*, 7772–7777.
- (60) Schmidt, J. M.; Blumel, M.; Lohr, F.; Ruterjans, H. *J. Biomol. NMR* **1999**, *14*, 1–12.
- (61) Newman, M. E. J.; Barkema, G. T. *Monte Carlo Methods in Statistical Physics*; Oxford University Press: Oxford, UK, 1999; 499 pages.
- (62) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.
- (63) Danielsson, J.; Jarvet, J.; Damberg, P.; GrÅd’slund, A. *FEBS J.* **2005**, *272*, 3938–3949.
- (64) Lam, A. R.; Teplow, D. B.; Stanley, H. E.; Urbanc, B. *J. Am. Chem. Soc.* **2008**, *130*, 17413–17422.
- (65) Sciarretta, K. L.; Gordon, D. J.; Petkova, A. T.; Tycko, R.; Meredith, S. C. *Biochemistry* **2005**, *44*, 6003–6014.
- (66) Liu, R.; McAllister, C.; Lyubchenko, Y.; Sierks, M. *J. Neurosci. Res.* **2004**, *75*, 162–171.
- (67) Bernstein, S. L.; Wyttenbach, T.; Baumketner, A.; Shea, J.-E.; Bitan, G.; Teplow, D. B.; Bowers, M. T. *J. Am. Chem. Soc.* **2005**, *127*, 2075–2084.
- (68) Melquiond, A.; Dong, X.; Mousseau, N.; Derreumaux, P. *Curr. Alzheimer Res.* **2008**, *5*, 244–250.
- (69) Bitan, G.; Vollers, S. S.; Teplow, D. B. *J. Biol. Chem.* **2003**, *278*, 34882–34889.
- (70) Takeda, T.; Klimov, D. K. *Biophys. J.* **2009**, *96*, 4428–4437.
- (71) O’Brien, E. P.; Okamoto, Y.; Straub, J. E.; Brooks, B. R.; Thirumalai, D. *J. Phys. Chem. B* **2009**, *113*, 14421–14430.

Structural Instability of the Active Site of T1 Lipase Induced by Replacement of Na⁺ with Water Complexed with the Phenylalanine Aromatic Ring

Yohsuke Hagiwara,[†] Jiyoung Kang,[†] and Masaru Tateno^{*,†}

[†]Graduate School of Pure and Applied Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba Science City, Ibaraki 305-8571, Japan

[‡]Graduate School of Life Science, University of Hyogo, 3-2-1 Kouto, Kamigohri, Akoh, Hyogo 678-1297, Japan

 Supporting Information

ABSTRACT: The crystallographic analysis of T1 lipase suggested an interaction between Na⁺ and the aromatic ring of Phe16 in the active site. However, experimental approaches could not dismiss the possible presence of water instead of Na⁺. Our previous molecular dynamics (MD) simulations suggested that the significantly large enthalpy gain of the Na⁺– π interaction was required to preserve the catalytic core structure of T1 lipase. In this study, to examine the effects of water, we performed further MD simulations of T1 lipase involving the water– π interaction, instead of the Na⁺– π interaction, exploiting various force fields, such as ff99, ff02, and an accurate potential field to describe the water– π interaction, which was generated using our recently developed scheme (referred to as the grid-based energy representation). The analyses revealed that the water– π complex was unstable in the catalytic core of T1 lipase even when the accurate potential of the water– π complex represented by the grid-based energy function was employed in the MD simulations and led to the disruption of the coordinated structure. In contrast, the catalytic core structure of T1 lipase involving the Na⁺– π complex was significantly stable in the 10 ns MD simulation using the grid-based energy representation of the Na⁺– π interaction. Thus, the possible presence of water may be excluded, and our previous proposal concerning the functional role of the structural element involving the Na⁺– π interaction in the catalytic site of T1 lipase has unambiguously been confirmed. Further, the strong coordination of Na⁺ and N _{ϵ} of His358 was also shown to be substantial to preserve the core structure of the catalytic site.

1. INTRODUCTION

The cation– π interaction is a noncovalent binding interaction, in which a cation is strongly attracted to the π electrons of an aromatic molecule. The interaction is commonly found in biological macromolecules and contributes to their structural stability and to drug–protein interactions.^{1–6} Even though only about 10 years have passed since such functional roles were proposed, cation– π interactions are now recognized as being as important as other conventional noncovalent interactions, i.e., hydrogen bonds, salt bridges, and hydrophobic interactions. The cations participating in the interactions include metal cations, such as Na⁺ and Mg²⁺, which are abundant in living organisms, as well as positively charged amino acid residues, including arginine (Arg) and lysine (Lys). Substantial stabilization energy has been detected in the gas phase^{7–12} and also in aqueous solutions^{2,13–15} upon the formation of metal– π interactions. However, only a few experimental structures containing complexes of metal cations and aromatic rings have been reported.^{16–20} Thus, the functional roles of the interactions remain to be clarified.

The X-ray crystallographic analysis of a thermoalkalophilic lipase from *Geobacillus zalihae* strain T1 (T1 lipase) revealed a characteristic feature; i.e., a spherical electron density was detected in the vicinity of the Phe16 side chain in the catalytic site (Figure 1).²¹ Since T1 lipase exhibits its catalytic activity under alkaline physiological conditions, which provide an abundance of Na⁺ ions, such as in palm oil mill effluent (POME),²² the electron density was considered to correspond to Na⁺.

However, the possibility that the electron density corresponded to a water molecule, rather than Na⁺, could not be ruled out.²¹ Actually, a survey of a set of 75 high resolution crystal structures of proteins (less than 1.1 Å) revealed the existence of such water– π interactions, and 18 water molecules were found facing the planes of the aromatic rings.²³

In our previous study, to investigate whether the electron density actually corresponds to Na⁺, we performed molecular dynamics (MD) simulations of T1 lipase, where Na⁺ was placed at the active site.²⁴ However, conventional force fields occasionally fail to provide an accurate description of the metal– π interactions. Accordingly, we developed a novel strategy, the “grid-based energy representation” scheme, which enables us to calculate their interaction energy with an accuracy corresponding to that of advanced *ab initio* post Hartree–Fock methods and a computational cost comparable to that of force field calculations.²⁴ A comparison of the MD simulation results obtained using the grid-based energy representation with those generated using the conventional force fields (polarizable and nonpolarizable ones) showed that an accurate estimation of the large binding energy (approximately 20 kcal/mol) was essential to correctly reproduce the experimental structure of the catalytic core of T1 lipase. In other words, a smaller interaction energy (\sim 10 kcal/mol, estimated by the nonpolarizable force field, AMBER ff99²⁵) caused serious structural disruptions.

Received: December 26, 2010

Published: June 14, 2011

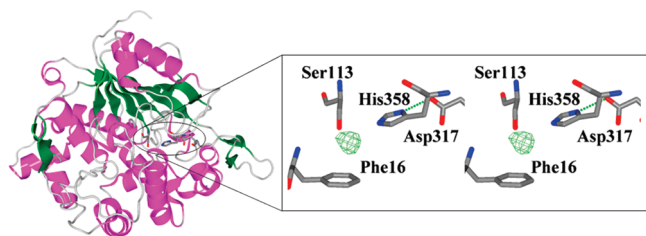


Figure 1. Crystal structure of the *Geobacillus zalihae* T1 lipase (PDB code: 2DSN). A close-up view of the configuration of the catalytic site is also shown in the right panel (stereo view). In the active site, a spherical electron density is observed, and the corresponding molecule is coordinated to the side chains of Phe16, Ser113, and His358.

Table 1. Summary of the MD Simulations Conducted in This Study

identify of the spherical electron density potential field	H ₂ O			Na ⁺
	ff99	ff02	GER ^a	GER ^a
single-protonated His358	5 ns	5 ns	5 ns	10 ns ^b
double-protonated His358	5 ns	5 ns	10 ns	– ^c

^aThe grid-based energy representation (GER) was exploited. ^bIn our previous study, a 5 ns MD simulation was performed,²⁴ in the present study, this MD simulation was extended up to 10 ns to confirm the structural convergence. ^cNot performed.

This excluded the possibility of the presence of water in the catalytic core, since the energy gain of the water– π interaction is ~ 4 kcal/mol.²⁶ Moreover, we have proposed the functional roles of the Na⁺–Phe16 interaction; i.e., it establishes a remarkably stable core structure by combining a hydrophobic aromatic ring (Phe16) and hydrophilic residues (Ser113 and His358), with the latter forming the catalytic triad. It was also implied that the rigid core structure formed by the Na⁺– π interaction may contribute to the large structural changes from the complex with ligands to the free form of the enzyme.

In the present study, to clarify the identity of the electron density, i.e., water or Na⁺, we performed additional MD simulations in which a water molecule was, in turn, assigned to the electron density (all of the MD simulations conducted in this study are summarized in Table 1). Since our previous study demonstrated the importance of an accurate description of the interaction energy,²⁴ in the present study, we also evaluated the feasibility of the conventional force fields, i.e., ff99 and ff02,²⁷ for the estimation of the water– π interaction energy and confirmed the poor accuracy of the force fields. Accordingly, we applied our grid-based energy representation scheme to conduct the MD simulation, which included the potential function for the accurate estimation of the interaction energy. The results of the MD simulation showed the serious instability of the water–Phe16 complex, leading to the structural disruption of the complex. These results provide further evidence that Na⁺, rather than water, is bound to Phe16 in the active site of T1 lipase.

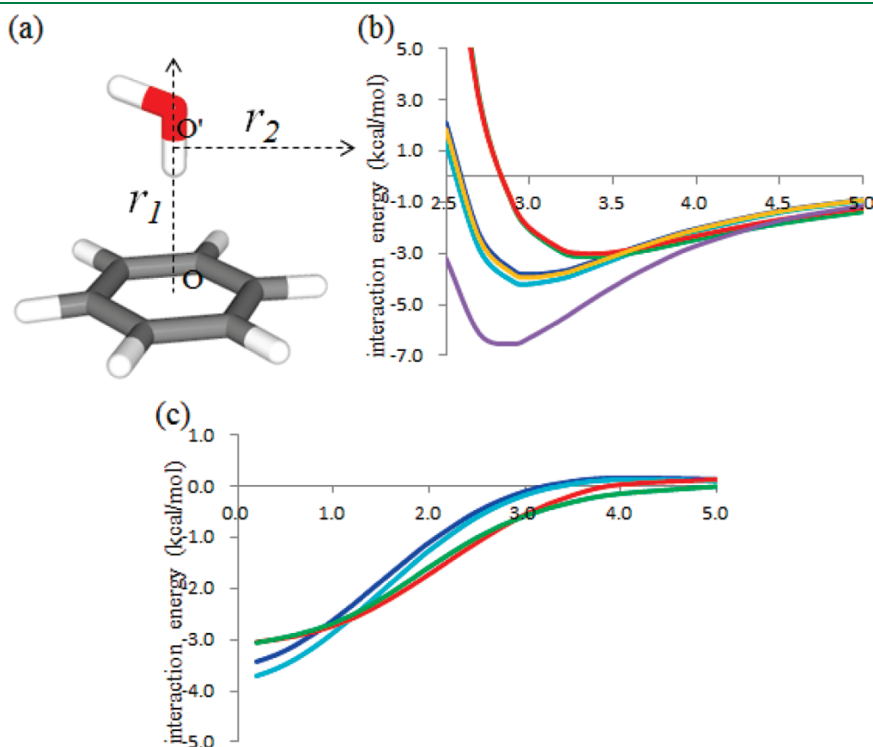


Figure 2. (a) Model structures for water– π interactions, where the oxygen atom of the water is placed on a line perpendicular to the benzene ring and passing through the center of mass of the benzene ring O. r_1 denotes the distance of the oxygen atom from O. O' is the point where r_1 is 3.3 Å (which is the optimal distance of r_1 in the potential curve), and r_2 denotes the distance of the oxygen atom from O' along a line parallel to the benzene ring and passing through O'. Energy profiles of the interaction energies of the water and the benzene ring with respect to (b) r_1 and (c) r_2 . The interaction energies of the water–benzene system obtained by MP2/aug-cc-pvtz and the molecular mechanics calculations using ff99 and ff02 are shown in red, blue, and light blue, respectively. The interaction energies obtained by ff02, in which the parameters responsible for the polarization are increased 8-fold (purple) and decreased 0.3-fold (orange) with respect to the original parameter values, to describe the polarizability. The energy potential curves obtained by the grid-energy representation are shown in green.

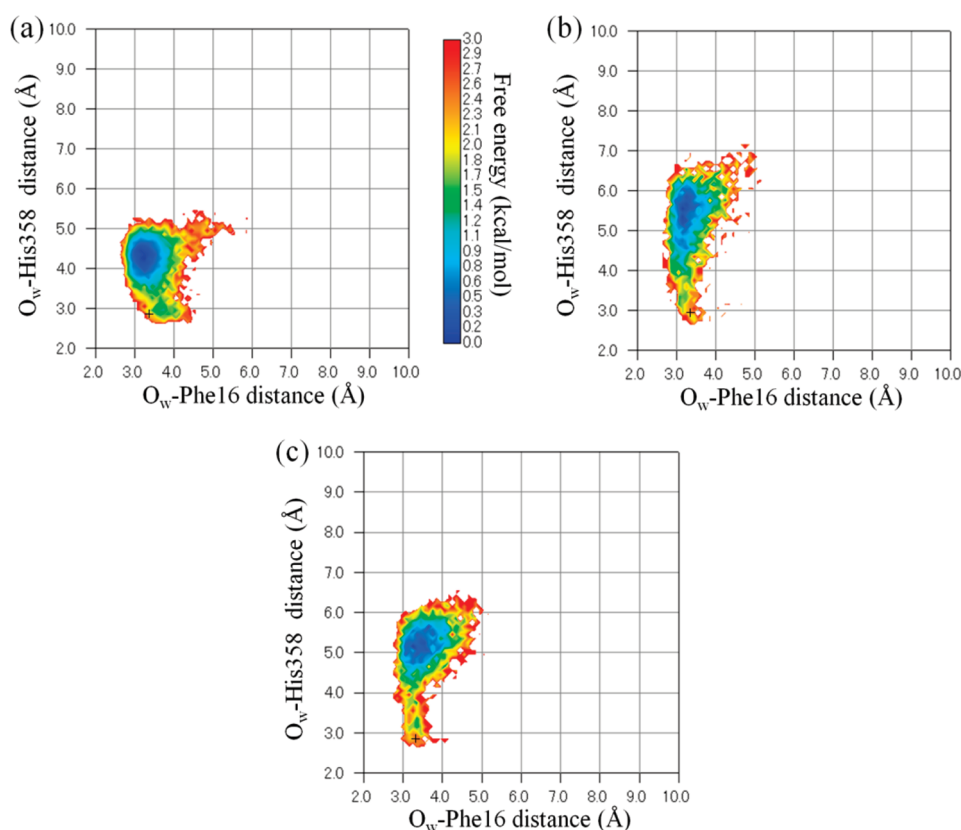


Figure 3. Free energy landscapes obtained by standard MD simulations (5 ns) of T1 lipase using ff99 (a), ff02 (b), and the grid-based energy representation (c). The vertical axis shows the distance between the oxygen atom of the water molecule (O_w) and N_ϵ of His358, and the horizontal axis is the distance between the oxygen atom and the center of mass of the Phe16 side chain. A cross (+) represents the crystal structure.

RESULTS AND DISCUSSION

Evaluation of Force Fields. First, we evaluated the feasibility of a nonpolarizable force field, ff99, and a polarizable force field, ff02, for the estimation of the interaction between a water molecule and an aromatic ring, benzene. For the calculations, we constructed model structures in which a hydrogen atom of the water was pointed toward the benzene (Figure 2a). The interaction that generates the binding force in such a conformation is referred to as an OH- π interaction, which is a hydrogen bond where the π electrons of the aromatic ring act as the hydrogen bond acceptor.²³

Then, we calculated the potential curves with respect to the distance between the oxygen atom of the water and the center of mass of the benzene, using three methodologies, ff99, ff02, and MP2/aug-cc-pvtz (Figure 2b). In this study, MP2 was employed to obtain the reference values; the correction of the MP2 results to estimate values calculated using the coupled-cluster with singles, doubles, and perturbative triples (CCSD(T)), as conducted in our previous study,²⁴ would not be required in this case for the following reason: The water- π interaction is dominated by a hydrogen bond property that is accurately described by MP2,^{26,28,29} whereas the correction is essential for the evaluation of the π - π interaction, which MP2 significantly overestimates. In fact, the MP2 results have been used as the reference in other recent studies.^{30–32}

According to the potential curve calculated using MP2, the energy minimum is located at 3.3 Å, where the stabilization energy is 3.11 kcal/mol. On the other hand, the energy minima in

the potential curves calculated using ff99 and ff02 are located at 3.0 Å in both cases, and the minimum energies were found to be overestimated by 0.71 and 1.08 kcal/mol, respectively. To consider other configurations, we constructed model structures in which the oxygen atom was placed on the axis that is parallel to the aromatic ring (Figure 2a). In the figure, O' denotes the cross point of the two axes (r_1 is 3.3 Å), and r_2 is defined as the distance between the oxygen atom and O' . With respect to the second configuration, the conventional force fields overestimated the minimum energy, as observed in Figure 2c, whereas the stabilization energies at larger distances ($r_2 > 1.0$ Å) were underestimated.

In this manner, the conventional force fields are not consistent with the reference potential curve calculated using MP2/aug-cc-pvtz. However, we performed the molecular dynamics (MD) simulations of T1 lipase using those potential fields to evaluate the effects of the differences in the energy potentials, since this strategy was informative in our previous study. The initial structure of the protein was obtained from its crystal structure (Figure 1), but the Na^+ in its active site was replaced with a water molecule (the oxygen atom of the water was placed at the coordinates of the Na^+).

MD Simulation Using ff99. First, we examined ff99 in a 5 ns MD simulation. Using the resultant MD trajectory, we calculated the two-dimensional (2-D) free energy profile with respect to the distance between the oxygen atom of the water molecule and the center of mass of the aromatic ring of Phe16, and that between the oxygen atom and N_ϵ of His358 (Figure 3a). It showed that the minimum corresponds to a structure that is not observed in

the crystal structure. In particular, the inconsistency of the water– N_e distance is remarkable, for the following reason: In the MD simulation, the N_e of His358 is stably hydrogen-bonded with the hydrogen atom of the hydroxyl group of the Ser113 side chain (Figure 4a). This hydrogen bond prevents the water from forming the hydrogen bond with N_e of His358, due to the steric clash between the hydrogen atoms of the water and the Ser113

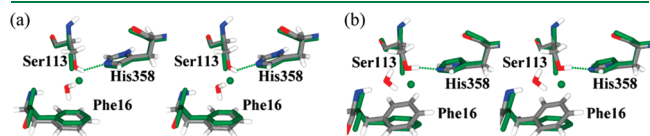


Figure 4. Conformations of the most stable state in Figure 3a (a) and Figure 3b (b). The crystal structure of the *Geobacillus zalihiae* lipase is colored green. The green sphere represents the position of Na^+ in the crystal structure. The snapshots of the MD simulations are shown in atom-specific colors.

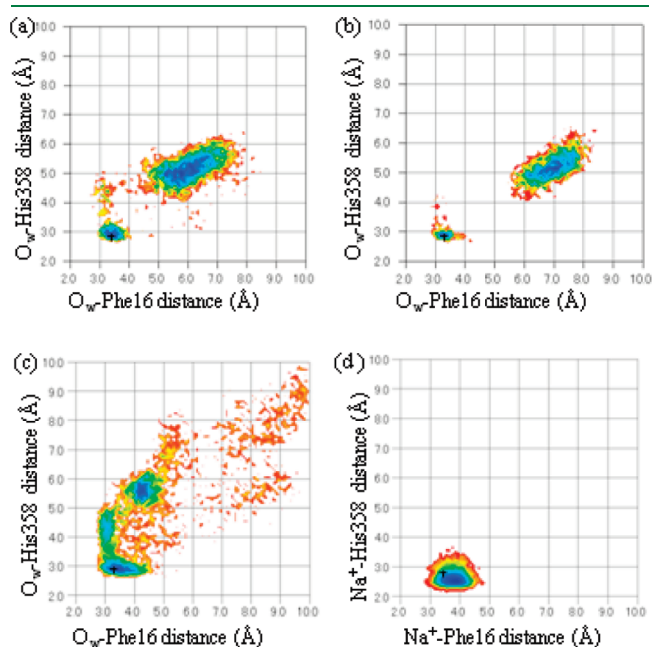


Figure 5. Free energy landscapes obtained by standard MD simulations of T1 lipase, in which the doubly protonated form is assigned to His358, using ff99 (5 ns) (a), ff02 (5 ns) (b), and the grid-based representation (10 ns) (c). The vertical axis shows the distance between the oxygen atom of the water molecule (O_w) and N_e of His358, and the horizontal axis is the distance between the oxygen atom and the center of mass of the Phe16 side chain. The cross (+) represents the crystal structure. Panel (d) shows the free energy profile obtained with the 10 ns MD simulation where Na^+ was assigned to the electron density instead of a water, using the grid-based representation. In this calculation, the state of His358 is the single protonated form.

side chain, which causes the increased water– N_e distance (the reasons for the structural changes are further discussed later).

It should be noted here that Asp317 is coordinated to His358 (Figure 1); its pK_a value was calculated as 3.85 using PROPKA 2.0.³³ Due to the presence of this aspartate residue, His358 is considered to be in the singly protonated form, where the N_δ is set to be protonated. This is also supported by the fact that N_e should be deprotonated for the catalysis, since the enzymatic reaction would be initiated by the abstraction of the hydrogen atom of the hydroxyl group from the catalytic Ser113 by the N_e of His358. However, we also examined the effects of the doubly protonated form of His358 as well as the singly protonated form. The resultant free energy profile and the representative structures corresponding to the metastable states of the profile revealed that the water– π coordination was not maintained in the MD simulation with ff99 (Figure 5a and 6a). This is consistent with the previous results that a water molecule cannot replace the Na^+ in the active site of T1 lipase.

MD Simulation Using ff02. We next examined the MD simulations using ff02, where His358 was set to the singly or doubly protonated form. These simulations revealed that the experimental configuration is also unstable in the presence of the water– π interaction, leading to a similar structural disruption to that of the MD simulations using ff99 (Figures 3b, 4b, 5b, and 6b). Note that this water molecule hydrogen-bonds with the Ser and His residues (Figure 1); therefore, it is possible that even in the MD simulations using ff99 and ff02, this water– π interaction is stabilized through those hydrogen bonds. Moreover, the interaction energies of ff99 and ff02 are larger than that of MP2. Nevertheless, this possibility was rejected in the present study, suggesting that the correct description of the interaction between the spherical electron density and the aromatic ring of Phe16 is essential to maintaining the structure of the catalytic site of T1 lipase.

Previous studies indicated that the contribution of the dispersion and electrostatic energies is most important for the description of the water– π interaction.^{32,34,35} In fact, when the parameters describing the polarization effect in ff02 are increased, the potential curve generated by ff02 did not match that of MP2 for $r \sim 0$, although it could be fitted for $r \gg 0$ (Figure 2b). This is consistent with the results of the previous studies. On the other hand, when the parameters are decreased, the potential curve obtained by ff02 converges to that of ff99, as shown in Figure 2b. Thus, we could not find the parameters that allowed the potential curve obtained by ff02 to completely match that of MP2; therefore, ff02 cannot reproduce the MP2 result. In this manner, the failure of ff02 is derived from an incorrect description of the dispersion energy. Various attempts to improve such problems found in the force fields have been reported.^{6,36–41}

Thus, the water– π coordination in T1 lipase is not possible in the MD simulations using the conventional force fields (the reasons for the structural changes are further discussed later).

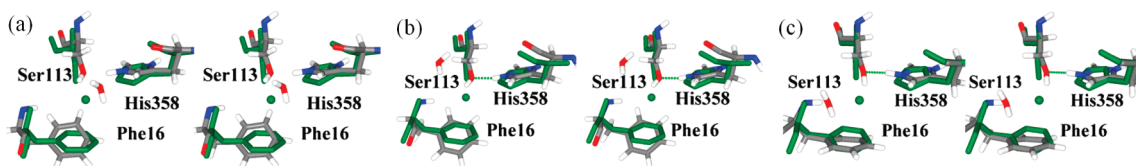


Figure 6. Conformations of the most stable state in Figure 5a (a), Figure 5b (b), and Figure 5c (c). The crystal structure of the *Geobacillus zalihiae* lipase is colored green, and the snapshots of the MD simulations are shown in atom-specific colors.

In the present study, we applied our general scheme, which may be applicable to correctly reproduce any potentials using effective functions, without detailed analyses of the defects of the potential terms in the total energy functions, as described below.

MD Simulation Involving the Grid-Based Energy Representation for the $\text{Na}^+-\pi$ Interaction. As in the case of our previous study, the structural disruptions could be caused by the differences in the potential curves provided by ff99 and ff02 from the correct potential function. Accordingly, we applied our grid-based energy representation scheme to obtain the effective potential that reproduces the potential energy curve at the level of MP2. In Figure 2b and c, the potentials generated by the grid-based energy representation show that the reference energy curves (obtained by the MP2) can be reproduced for both configurations of the water–benzene complex. Then, using the optimized parameter set, we performed MD simulations of T1 lipase, in which His358 was singly or doubly protonated. The resultant 2-D free energy profiles showed that the experimental configuration was not preserved in either case, as in the results of the MD simulations using the conventional force fields (Figure 3c, 5c, and 6c).

It should be noted here that, to evaluate the stability of the conformation involving the doubly protonated His358, we required 10 ns MD simulations for the structural convergence (Figure 5c; see Supporting Information in detail). Correspondingly, for the comparison, we also extended the MD simulation where Na^+ is assigned to the spherical electron density of the catalytic site, up to 10 ns in this study. The resultant free energy profile is well converged and thus unambiguously shows consistency with the experimental data (Figure 5d). Thus, the presence of Na^+ rather than water is confirmed again.

The structural transitions observed in the present MD simulations can be understood by comparison with our previous results: In our previous MD simulation involving the grid-based energy representation for the $\text{Na}^+-\text{Phe16}$ interaction, both distances concerning $\text{Na}^+-\text{N}_\epsilon$ and $\text{Na}^+-\text{Phe16}$ were consistent with the experimental values, thus preserving the experimental conformation of the catalytic core of T1 lipase.²¹ This would be accomplished by the strong interaction between Na^+ and the N_ϵ of His358. In contrast, in the present MD simulations, the interaction between the water molecule and the N_ϵ of His358 is not as strong as the $\text{Na}^+-\text{N}_\epsilon$ interaction, and so it competes for interactions with the Ser113 and His358 side chains. Namely, the weaker interaction between the water and His358 would cause the disruption, as observed in the present MD simulations using the ff99 and ff02 force fields. Conversely, this indicates that strong coordination with the surrounding amino acid residues is required to maintain the experimental structure of the catalytic core of T1 lipase, supporting the presence of Na^+ , rather than water, in the catalytic site.

Thus, we have concluded that the possible presence of a water molecule, instead of Na^+ , in the active site can be ruled out. Moreover, the results of our previous and present theoretical studies demonstrated that the core structure of the catalytic site of T1 lipase in the free state stably exists only in the MD simulation where Na^+ is assigned to the spherical electron density, using the correct potential fields. This further suggests that the Na^+-Phe interaction is essential for the formation of the stable core structure of the catalytic site through the formation of “ Na^+ bridges”, which establish the packing of the hydrophobic aromatic ring (Phe16) and the hydrophilic amino acid residues (Ser113 and His358).²⁴

In the crystallographic data of biological macromolecules, similar spherical electron densities to that of the $\text{Na}^+-\pi$ complex found in T1 lipase have been widely observed. Therefore, this may be a general issue in protein science. In the present study, we have indicated that MD simulation coupled with our scheme is useful in defining the spherical electron densities, i.e., cation, water, et cetera.

CONCLUSION

In the present study, using our grid-based energy representation, we have confirmed that the core structure of T1 lipase is established by a $\text{Na}^+-\pi$ interaction, rather than a water– π interaction. In addition to the large enthalpy gain of the $\text{Na}^+-\pi$ interaction, the strong coordination of Na^+ and N_ϵ of His358 was also shown to be substantial to preserve the core structure of the catalytic site. Since the cation– π interactions have been widely found in biological systems, the present study will be a solid platform to further investigate crucial roles of cation– π interactions involved also in other biological systems.

The grid-based energy representation scheme can be widely applied to perform long-time MD simulations in which accurate interaction energies must be calculated at advanced post-HF method levels, with reasonable computational costs, thereby enabling us to understand the dynamic properties and functional roles of those interactions in biological macromolecules.

MATERIALS AND METHODS

Grid-Based Energy Representation. The details of the grid-based energy representation are described in the Supporting Information.²⁴ The outline of the scheme is as follows. The first step is to calculate a density distribution function, $\rho(\mathbf{r})$, for the grid space defined by the coordinates of a π molecule (in this work, benzene was used), using a parameter set that regulates the shape of $\rho(\mathbf{r})$. The second step is to calculate the electrostatic energy, using an effective potential that includes $\rho(\mathbf{r})$ with respect to several configurations of the water– π complex. The third step is to calculate the interaction energy between the water and the π molecule for each configuration of the complex using the total energy function, in which the original electrostatic energy term is replaced with the ρ -containing energy term. The last step is to evaluate the obtained interaction energy values by calculating their deviations from the energy values obtained from higher-level *ab initio* calculations performed beforehand. For the MD simulations using this scheme, we employed the parameter set that provided the lowest value of the deviation. The formulations of the functions used in the first, second, and third steps correspond to eqs 1, 2, and 5 in the Supporting Information, respectively.

Quantum Mechanical Calculations. All *ab initio* molecular orbital calculations of the water–benzene system were performed using Gaussian 03.⁴² Geometry optimization for the isolated benzene was performed at the B3LYP^{43,44}/6-311+G(d,p) level, and the optimized structure was used to construct water–benzene complex structures. For single point calculations of those structures to obtain the reference potential energy curve, we employed the Møller–Plesset second order perturbation (MP2) method⁴⁵ with an aug-cc-pvtz basis set.⁴⁶ The basis set superposition error (BSSE) was corrected by the counterpoise method.

Molecular Dynamics Calculations. All calculations were performed using the AMBER 9 program package.⁴⁷ The MD simulation of the solvated system was performed under a constant pressure of 1.013×10^5 Pa, with a periodic boundary condition at 300 K. Temperature and pressure were controlled by the Berendsen algorithm.⁴⁸ The SHAKE algorithm was used to treat the bonds involving hydrogen,⁴⁹ and the time step for integration was set to 1 fs. Electrostatic interactions were calculated by the particle mesh Ewald (PME) method,⁵⁰ with a dielectric constant of 1.0. A cutoff of 12 Å was used to calculate the direct space sum for PME.

Initial coordinates of the protein were obtained from the crystal structure of T1 lipase (Protein Data Bank accession code 2DSN²¹), but the Na⁺ in the active site was replaced with a water molecule. First, hydrogen atoms were added to the crystal structure, using the LEAP module implemented in AMBER 9. The positions of the added hydrogen atoms were optimized by the steepest descent method, and then the optimization was performed for all protein atoms. The protein was subsequently immersed in a box of water molecules, consisting of 48 561 atoms modeled by TIP3P.⁵¹ Two Na⁺ ions were added to neutralize the system. Thus, the total atom number of the solvated protein system was 54 538. To relax the configuration of the solvent water molecules, the MD simulation was performed for 10 ps at 300 K, where a harmonic constraint was applied to all protein and Na⁺ atoms with a force constant of 500 kcal mol⁻¹ Å⁻². With respect to the density distribution function, the optimized values of a_x , a_y , and a_z for the carbon atoms were 0.198, 0.198, and 0.337 Å, respectively, and those for the hydrogen atoms were 0.225, 0.225, and 0.312 Å, respectively. The value of α in the cutoff function was set to 1.0×10^{-5} . A cube, with its center located on the center of mass of the aromatic ring, was generated; the volume of the cube was 10³ Å³, and the grid-point space inside the cube was set to 0.2 Å. Accordingly, the number of grid points involved in the cube was 125 000.

■ ASSOCIATED CONTENT

S Supporting Information. Further descriptions concerning the grid-based energy representation and the examination of the doubly protonated form of His358 are available. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: (+81)-791-58-0347. Fax: (+81)-791-58-0347. E-mail: tateno@sci.u-hyogo.ac.jp.

■ ACKNOWLEDGMENT

This work was partly supported by Grants-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), under contract Nos. 21340108 and 23654120. Computations were performed using the computer facilities under the “Interdisciplinary Computational Science Program” at the Center for Computational Sciences, University of Tsukuba, the Computer Center for Agriculture, Forestry, and Fisheries Research, MAFF, Japan, and the Supercomputer Center, Institute for Solid State Physics, University of Tokyo. J.K. was supported by a Research Fellowship of the Japan Society for the Promotion of Science.

■ REFERENCES

- (1) Chan, D. I.; Prenner, E. J.; Vogel, H. J. *Biochim. Biophys. Acta* **2006**, *1758*, 1184–1202.
- (2) Dougherty, D. A. *Science* **1996**, *271*, 163–168.
- (3) Dougherty, D. A. *J. Nutr.* **2007**, *137*, 1504S–1508S.
- (4) Gallivan, J. P.; Dougherty, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9459–9464.
- (5) Ma, J. C.; Dougherty, D. A. *Chem. Rev.* **1997**, *97*, 1303–1324.
- (6) Minoux, H.; Chipot, C. *J. Am. Chem. Soc.* **1999**, *121*, 10366–10372.
- (7) Dunbar, R. C. *J. Phys. Chem. A* **2000**, *104*, 8067–8074.
- (8) Gapeev, A.; Dunbar, R. C. *J. Am. Chem. Soc.* **2001**, *123*, 8360–8365.
- (9) Kish, M. M.; Ohanessian, G.; Wesdemiotis, C. *Int. J. Mass Spectrom.* **2003**, *227*, 509–524.
- (10) Ruan, C. H.; Rodgers, M. T. *J. Am. Chem. Soc.* **2004**, *126*, 14600–14610.
- (11) Ryzhov, V.; Dunbar, R. C.; Cerda, B.; Wesdemiotis, C. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 1037–1046.
- (12) Siu, F. M.; Ma, N. L.; Tsang, C. W. *J. Am. Chem. Soc.* **2001**, *123*, 3397–3398.
- (13) Costanzo, F.; Della Valle, R. G. *J. Phys. Chem. B* **2008**, *112*, 12783–12789.
- (14) Hu, J.; Barbour, L. J.; Gokel, G. W. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 5121–5126.
- (15) Meadows, E. S.; De Wall, S. L.; Barbour, L. J.; Gokel, G. W. *J. Am. Chem. Soc.* **2001**, *123*, 3092–3107.
- (16) Kooystra, P. J.; Kalk, K. H.; Hol, W. G. *Eur. J. Biochem.* **1988**, *177*, 345–349.
- (17) Labesse, G.; Ferrari, D.; Chen, Z. W.; Rossi, G. L.; Kuusk, V.; McIntire, W. S.; Mathews, F. S. *J. Biol. Chem.* **1998**, *273*, 25703–25712.
- (18) Liaw, S. H.; Kuo, I.; Eisenberg, D. *Protein Sci.* **1995**, *4*, 2358–2365.
- (19) Wouters, J. *Protein Sci.* **1998**, *7*, 2472–2475.
- (20) Wouters, J.; Maes, D. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2000**, *56*, 1201–1203.
- (21) Matsumura, H.; Yamamoto, T.; Leow, T. C.; Mori, T.; Salleh, A. B.; Basri, M.; Inoue, T.; Kai, Y.; Rahman, R. N. *Proteins* **2008**, *70*, 592–598.
- (22) Fischer, M.; Pleiss, J. *Nucleic Acids Res.* **2003**, *31*, 319–321.
- (23) Steiner, T. *Biophys. Chem.* **2002**, *95*, 195–201.
- (24) Hagiwara, Y.; Matsumura, H.; Tateno, M. *J. Am. Chem. Soc.* **2009**, *131*, 16697–16705.
- (25) Cheatham, T. E., III; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–862.
- (26) Ran, J.; Hobza, P. *J. Chem. Theory Comput* **2009**, *5*, 1180–1185.
- (27) Cieplak, P.; Caldwell, J.; Kollman, P. *J. Comput. Chem.* **2001**, *22*, 1048–1057.
- (28) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.
- (29) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104–112.
- (30) Jain, A.; Ramanathan, V.; Sankaramakrishnan, R. *Protein Sci.* **2009**, *18*, 595–605.
- (31) Li, S.; Cooper, V. R.; Thonhauser, T.; Puzder, A.; Langreth, D. C. *J. Phys. Chem. A* **2008**, *112*, 9031–9036.
- (32) Mishra, B. K.; Sathyamurthy, N. *J. Phys. Chem. A* **2007**, *111*, 2139–2147.
- (33) Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins* **2008**, *73*, 765–783.
- (34) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2000**, *122*, 11450–11458.
- (35) Riley, K. E.; Pitoňák, M.; Černý, J. i.; Hobza, P. *J. Chem. Theory Comput* **2010**, *6*, 66–80.
- (36) Soteras, I.; Orozco, M.; Luque, F. J. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2616–2624.
- (37) Chipot, C.; Maigret, B.; Pearlman, D. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1996**, *118*, 2998–3005.

- (38) Cubero, E.; Luque, F. J.; Orozco, M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5976–5980.
- (39) Wormer, P. E. S.; van der Avoird, A. *Chem. Rev.* **2000**, *100*, 4109–4144.
- (40) Mishra, B. K.; Sathyamurthy, N. *J. Phys. Chem. A* **2007**, *111*, 2139–2147.
- (41) Salonen, L. M.; Ellermann, M.; Diederich, F. *Angew. Chem., Int. Ed.* **2011**, *50*, 4808–4842.
- (42) Frisch, M. J. *Chem. Listy* **2006**, *100*, A9–A9.
- (43) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (44) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (45) Headgordon, M.; Pople, J. A.; Frisch, M. J. *Chem. Phys. Lett.* **1988**, *153*, 503–506.
- (46) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (47) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
- (48) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (49) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (50) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (51) Jorgensen, W. L.; Madura, J. D. *J. Am. Chem. Soc.* **1983**, *105*, 1407–1413.

The Ethidium—UA/AU Intercalation Site: Effect of Model Fragmentation and Backbone Charge State

Karol M. Langner,^{†,‡} Tomasz Janowski,[§] Robert W. Góra,^{||} Paweł Dziekoński,[⊥]
W. Andrzej Sokalski,^{*,†} and Peter Pulay[§]

[†]Institute of Physical and Theoretical Chemistry, Wrocław University of Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

[‡]Leiden Institute of Chemistry, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands

[§]Department of Chemistry and Biochemistry, Fulbright College of Arts and Sciences, University of Arkansas, Fayetteville, Arkansas 72701, United States

^{||}Theoretical Chemistry Group, Institute of Physical and Theoretical Chemistry, Wrocław University of Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

[⊥]Wrocław Center for Networking and Supercomputing, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

S Supporting Information

ABSTRACT: We report a systematic analysis of the intermolecular interactions of cationic ethidium intercalated into a UA/AU step of RNA for a single conformation based on crystallographic coordinates. Interaction energies at the MP2/6-31G** level were partitioned into electrostatic, exchange, delocalization, and correlation components. Various pairwise interaction models built from chemically intuitive fragments reproduce within a few percent values obtained when treating the intercalation site as a whole. Gas phase results are very sensitive to the charge state of the two phosphate groups, with the electrostatic term nearly tripling when the counterions are removed. But this is largely compensated by solvation, an effect represented here within the polarizable continuum model. In a few cases, more diffuse and larger basis sets as well as QCISD(T) corrections were applied in an effort to estimate plausible ethidium-nucleobase electron correlation effects.

1. INTRODUCTION

Since the earliest experimental reports by Lerman and Waring,^{1,2} intercalation has been defined by how it affects double helix structure: nucleic acid strands extend and unwind in order to accommodate aromatic ligands between base pairs. Alongside other modes by which small molecules reversibly attach to nucleic acids, such as groove binding and peripheral electrostatic interactions,³ intercalation is characterized by the parallel alignment of a small aromatic between nucleic acid base pairs. And while disrupting cellular metabolism is their most widely utilized property, the potency of intercalating ligands is now attributed to the inhibition of DNA-binding enzymes such as topoisomerases rather than intercalative binding itself.⁴ Nonetheless, the specific type of stacking interaction involved makes these ligands interesting in other ways, for example, as fluorescent probes⁵ and switchable ligands.⁶ The trademark binding mode of intercalators is even hypothesized to be rooted in an evolutionary role, such as promoting oligonucleotide polymerization during the prebiotic RNA world.⁷

It is clear that in all of these areas the interactions of intercalating molecules with the nearest fragments of their hosts are important. What is still not clear today, however, is how these interactions relate to the process of inserting a ligand into a nucleic acid strand, the molecular basis of which is just starting to be studied.⁸ Thermodynamic data concerning intercalation are quite abundant, and generally speaking the stabilizing effect of desolvation and intermolecular interactions balance unfavorable entropic contributions to intercalation arising from conformational changes in the

nucleic acid chain and from lost degrees of freedom.^{9–11} Such compensation can be used to differentiate between intercalation and groove binding, even if the binding energy is similar.¹² Unfortunately, this compensation also complicates the interpretation of experimental data, since the measured binding free energy is several times smaller than any of the single contributions it is typically divided into.

Here, we turn to the already mentioned local intermolecular interactions, denoted by ΔG_{mol} following Graves and Valea,¹⁰ interactions which in line with Lerman's original interpretation include π – π aromatic stacking between the intercalator's chromophore and nearest nucleic acid bases. Although various computational methods have been applied to study intercalation complexes already in the previous century,^{13–17} works that probe their quantum chemical nature have appeared only in the past decade. This is due to the sizes of the systems involved and also due to the recognized importance of dispersion forces (interactions between instantaneous electron density moments), prohibiting the use of standard density functional theory.

Probably the first to venture in the direction of *ab initio* chemical calculations were Bondarev et al.,¹⁸ who calculated second order Møller–Plesset (MP2) interaction energies separately between the intercalator amiloride and the four DNA bases using the 6-31++G(d,p) basis set. Soon afterward, Řeha et al.¹⁹ followed with a similar, more extensive investigation of four

Received: February 18, 2011

Published: June 18, 2011

intercalators, among them ethidium. The main conclusion of these first reports was that in electronic structure calculations of this kind it is indispensable to account for dispersion effects, as they constitute a substantial part of the interaction energy similarly to conventional stacks of nucleic acid bases.²⁰ More recently, Kubař et al.²¹ have combined molecular dynamics with quantum chemical calculations to estimate the various free energy terms involved in, among other things, a variant of the system studied here (their system comprised ethidium intercalated into a TA/AT step of DNA). They achieved a net binding free energy of -4.5 kcal/mol, close to the corresponding experimental value of -7 kcal/mol, with a stabilization energy component of -70 kcal/mol. This disparity of an order of magnitude again underscores the fact that various free energy contributions are compensated, and that intermolecular stabilization effectively plays a decisive role.

More MP2 calculations have followed, also by Langner et al.,²² for the same ethidium–UA/AU complex studied here, wherein they focused on the role of the first-order electrostatic component of the interaction energy, and on the horizontal alignment of the intercalator between base pairs. Dračinsky and Castaño²³ performed a similar study earlier, of the interaction energy between ellipticine and base pairs for various distances and twists. Single point energies have also been published by Xiao and Cushman²⁴ for camptothecin in an attempt to correlate them with experimental site selectivity in ternary cleavage complexes. Due to problems with the proper representation of dispersion interactions by the standard exchange–correlation functionals, density functional theory (DFT) has been adopted relatively late.²⁵ Time-dependent DFT and Car–Parrinello dynamics have also been employed, for example, by Fantacci et al.,²⁶ in order to characterize the influence of base pairs on the excited states of an intercalated ruthenium compound.

The MP2 method, despite being the method of choice for sizable systems that include π – π contacts, is also problematic as far as dispersion forces are concerned. It has been pointed out that it overestimates their magnitude significantly compared to CCSD(T) values,²⁷ and recently some effort has gone into correcting for this. In the context of nucleic acid intercalation, Hill and Platts²⁸ have applied local correlation methods and density fitting for several intercalators together with four surrounding nucleobases.

Another problem concerning π – π stacking interactions is that the dispersion energy saturates slowly with growing Gaussian basis sets. A practical measure applied in many previous studies has been to make polarization functions of small basis sets more diffuse by decreasing their exponents. For example, decreasing the exponent of d functions on heavy atoms from 0.8 to 0.25 in the 6-31G* basis set brings interaction energies for stacked nucleic acid bases closer to results obtained in a more saturated basis sets,^{29,30} and this protocol has already been used for studying intercalation complexes.^{19,21,23}

The focus in almost all of these studies has been on interactions between intercalators and nucleobases. While little has been written about the remaining parts of the nucleic acid strand and the physical nature of their interactions with intercalators, they must also play a role and lately have been garnering increased interest. For example, Horowitz et al.³¹ have been the first to compare the NMR structures of intercalated RNA linked at its 2',5' ends and previously known 3',5'-linked RNA intercalation geometries. They showed that while a modified backbone conformation does not change the characteristics of

intercalation binding (the unwinding angle and helical rise), it does influence the enthalpic part of the free binding energy. This result hints that intermolecular interactions within the intercalation site are affected by changes in nucleic acid backbone structure, and certainly also by the surroundings.

When they are included in *ab initio* calculations, the phosphate groups are normally capped with protons to simulate the proximity of counterions.²¹ Meanwhile, it is known that the behavior of counterions around nucleic acids is not trivial, with monovalent cations being highly diffusive³² and divalent ions binding more tightly.³³ Definitely in the native, dynamic environment of nucleic acids one would expect significant variations in the hydration shell and counterion positions and therefore also in the electron distribution around phosphate groups.

Geometry fluctuations are definitely also important for non-covalent interactions in intercalator binding sites, a point demonstrated by Svozil et al.³⁴ for dinucleotides in the gas phase. They find significant variability in stacking interaction energies along molecular dynamics trajectories, and the same may be true for intercalation complexes such as the one studied here. Molecular dynamics calculations suggest³⁵ that nucleic acid configurational entropy can change considerably upon intercalation, affecting the phosphate groups the most and depending strongly on the host sequence. It would be best to follow these examples and take multiple MD snapshots as input for subsequent interaction energy calculations, but such a systematic treatment is beyond the reach of current computational resources.

We take a first step by considering the various charge states of the phosphate groups. To this end, single point gas phase calculations were performed for ethidium intercalated into a AU/UA base pair step of RNA, in which the phosphate groups were protonated (capped with protons), quenched with counterions, hydrated with single water molecules, or simply left anionic. This selection of charge states gives an idea of the range of possible interaction energies (neglecting conformation effects). Further polarizable continuum calculations roughly correct these values for the more physiologically relevant setting of water solvent.

Another aim here is to probe the validity of dividing intercalation sites into nucleobases and other fragments within the minimal model proposed by Kubař et al.,²¹ which consists of the intercalator, four nearest nucleosides, and two phosphate groups between them. We also compare several theoretical approximations and interaction energy components, calculated between ethidium and the RNA fragment as a whole, with selected pairwise interaction schemes. Besides additivity and polarization effects, in this way, we gain information about the relative importance and independence of the molecular fragments within the intercalation site. Finally, a reasonable magnitude for the dispersion interaction is sought by evaluating QCISD(T) energies (quadratic configuration interaction including unlinked triples)³⁶ and using larger basis sets for the most sensitive part of the system, i.e., the π – π stacking interaction of ethidium with nucleobases.

2. COMPUTATIONAL METHODS

Despite recent advances and increasingly accurate quantum chemical descriptions of unprecedentedly large molecules,^{37,38} including aromatic stacking complexes,^{30,39–44} almost all systems relevant to biological processes are still simply too large to be treated *en masse* from first principles routinely. Studies of intercalated nucleic acids have been typically limited to two or four nearby nucleobases, and interactions are usually analyzed

pairwise between the intercalator and each base separately. This seemingly straightforward way to proceed, that is, to divide a system into smaller parts that are as chemically independent as possible, is also the conceptual starting point for many fragmentation strategies such as effective fragment potentials,⁴⁴ local correlation approaches,⁴⁵ the fragment molecular orbital (FMO) method,⁴⁶ and density fitting methods.^{28,47,48}

As already mentioned, here, we revisit the ethidium cation intercalated between AU/UA base pairs in a conformation obtained from the crystallography literature.⁴⁹ The system will be referred to shortly as Eth⁽⁺¹⁾-UA/AU throughout this report. In a previous study of the same complex,²² the ethidium cation was fragmented into its chromophore, ring, and side chain, and interactions with only the four nearest nucleobases were taken into account in a pairwise fashion—in the present work, such an approach defines one of the models used, namely, A4 (parts). In the previous report, interactions at the MP2 level between the ethidium chromophore and nearest four bases were shown to be sufficient in order to reproduce with crystallographic accuracy the alignment of the intercalator in the intercalation plane. This was repeated there, albeit with limited success, using electrostatic interactions based on atomic multipole moments.

In order to be consistent with previous results, the 6-31G** basis set was also used in this study, unless stated otherwise. For all of the models considered, the interaction energy between ethidium and the intercalation site was decomposed into several terms with generally understood physical meaning according to a hybrid variation–perturbation scheme,⁵⁰ and all calculations were done consistently in the dimer-centered basis set following the prescription of Boys and Bernardi.⁵¹

In a few cases, higher order correlation effects were evaluated by employing additional QCISD(T) calculations and larger basis sets. Only results for the aug-cc-pVDZ basis set (for all methods) are presented here, but several additional MP2 results can be found in the Supporting Information. It is appropriate to comment on our choice of the higher level method (QCI)—the major reason being that the program used⁵² is more efficient for QCISD than CCSD. On the other hand, it has been demonstrated that the quadratic corrections satisfactorily account for size-consistency and that there is little difference between the results of QCISD and CCSD, provided that the wave function is dominated by a single reference configuration.⁵³ Janowski and Pulay have recently confirmed this in the case of a qualitatively similar system (the benzene dimer), for which the intermolecular binding energies calculated at the QCISD(T) level are almost identical with results obtained by means of CCSD(T).³⁹

In the following two subsections, we provide a short introduction to the interaction energy decomposition approach used and describe the various interaction models to which it was applied.

2.1. Hybrid Variation–Perturbation Interaction Energy Analysis. Theoretical details for this type of analysis and its applications published to date can be found in previous articles and in the references cited therein.^{22,50,54} Essentially, a selection of interaction terms is obtained that is analogous to the ones found in state-of-the-art symmetry-adapted perturbation theory (SAPT).⁵⁵ Combined with an integral direct version of the SCF algorithm⁵⁶ and new implementation of parallelization within GAMESS (US),⁵⁷ this hybrid solution allows much larger problems to be tackled than was previously possible.

In its simplest form, a second-order Møller–Plesset (MP2) interaction energy calculation for an interacting dimer is broken

down in the following way:

$$\Delta E_{\text{MP2}} = \Delta E_{\text{el}}^{(1)} + \Delta E_{\text{ex}}^{(1)} + \Delta E_{\text{del}}^{(\text{R})} + \Delta E_{\text{corr}} \quad (1)$$

where these four components are usually interpreted as corresponding to certain quantum mechanical effects in the electronic wave function of the dimer, relative to the wave functions of isolated monomers in the dimer-centered basis set. These contributions are as follows:

$\Delta E_{\text{el}}^{(1)}$: first-order electrostatic interaction between monomer Hartree–Fock densities, unperturbed by their mutual influence

$\Delta E_{\text{ex}}^{(1)}$: the associated exchange repulsion attributed to the Pauli exclusion principle

$\Delta E_{\text{del}}^{(\text{R})}$: charge delocalization, which includes the induction and exchange-induction effects that complement the Hartree–Fock interaction energy

ΔE_{corr} : dynamic electron correlation at the MP2 level (in this case), which includes uncoupled dispersion effects and intra-molecular corrections

Summing the four terms in eq 1 successively from left to right, one obtains a series of interaction energies at different levels, rising in theoretical accuracy as well as computational cost:

$$\Delta E_{\text{el}}^{(1)} < E_{\text{HL}}^{(1)} < E_{\text{RHF}} < E_{\text{MP2}} \quad (2)$$

where $\Delta E_{\text{HL}}^{(1)} = \Delta E_{\text{el}}^{(1)} + \Delta E_{\text{ex}}^{(1)}$ is usually called the Heitler–London energy and ΔE_{RHF} is the interaction at the (restricted) Hartree–Fock level.

This analysis can be readily generalized to many-body systems, which leads to a partitioning of the various interaction energy terms into two-, three-, ..., *n*-body contributions, aside from the electrostatic term $\Delta E_{\text{el}}^{(1)}$ which is pairwise additive (see Góra et al.⁵⁴ for a detailed discussion and application of many-body calculations). Since many-body effects inherently include interactions between all subsets (or fragments) of a system, in the present case, they include interactions between nucleobases as well as dimers containing the intercalator. To avoid confusion, we will use ΔE when referring to the sum of all possible many-body interaction contributions and reserve ΔE for interactions involving only the intercalator.

Since several of the cases investigated in this study involve charged species, interaction energies estimated in the gas phase will be quite different from those in a polar solvent such as water. We attempt to estimate the magnitude of such a solvent effect using the polarizable continuum model (PCM),⁵⁸ which should adequately render the electrostatic shielding of the charged phosphate groups by a polarizable environment.

To this end, the hybrid variation–perturbation scheme outlined above can be generalized to complexes in an external potential and can readily be used to study intermolecular interactions in a PCM-type solvent. The particulars of the methodology can be found in a recent article by Góra et al.⁵⁹ and references cited therein. In short, the solute molecules are placed in a cavity inside an isotropic dielectric medium, represented by apparent charges on a solvent-accessible surface. These charges give rise to an external potential that modifies the one-electron Hamiltonians of the interacting species and their complex. Consequently, their wave functions and the resulting intermolecular interactions also change, and the final value is estimated in a self-consistent procedure. In passing to the solvent, however, the basic energetic quantity is changed⁵⁸ and an additional component $\Delta\Delta G(\text{AB})$ arises. This additional term represents indirect solvent-mediated interactions, or the

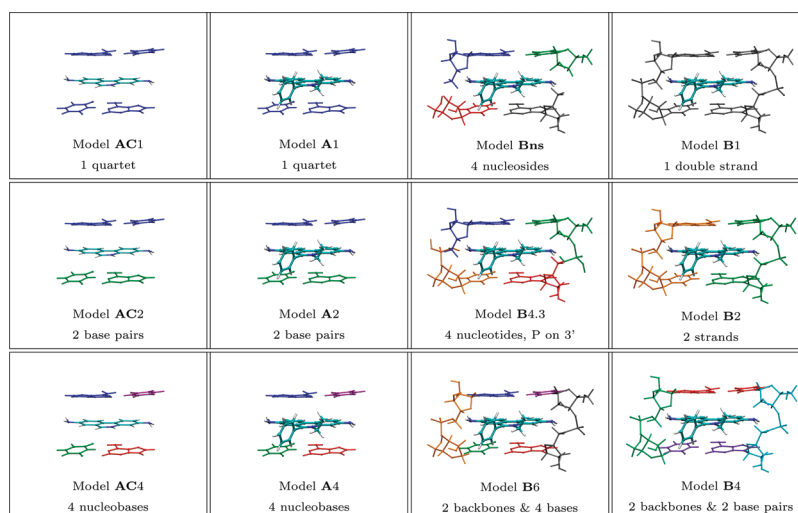


Figure 1. Schematic drawings of the models considered for the Eth⁽⁺¹⁾–UA/AU intercalation complex. Each color represents a piece of separately interacting RNA fragment around the binding site. The total interaction was constructed as the sum of all pairwise interactions among these pieces with the ethidium cation (see text for details).

differential free energy of solvation:

$$\Delta\Delta G(\text{AB}) = \Delta G_{\text{sol}}(\text{AB}) - \Delta G_{\text{sol}}(\text{A}) - \Delta G_{\text{sol}}(\text{B}) \quad (3)$$

where the terms on the right-hand side are the free energies of solvation of the complex and its constituents.

The last two quantities in eq 3 can be evaluated either in the cavity formed by a dimer or in their respective monomeric cavities. Using a dimeric cavity to evaluate $\Delta G_{\text{sol}}(\text{A})$ and $\Delta G_{\text{sol}}(\text{B})$ is in the spirit of the definition of the interaction energy,⁵⁹ and the corresponding differential free energy of solvation will therefore be referred to as $\Delta\Delta G^{\text{int}}$; the value obtained using monomeric cavities will be called $\Delta\Delta G^{\text{stab}}$ since conceptually it involves reorganization of the solvent due to dissociation. We also reserve the term *internal energy* for the expectation value of the gas-phase Hamiltonian using the solvent-modified wave function, following the usual PCM definition. All of the PCM results reported were estimated using the integral equation formalism version of the PCM approach (IEF-PCM), as implemented in GAMESS (US), including only electrostatic terms and assuming the standard parameters of a water solvent.

2.2. Interaction Models Used. All of the selected models are subsets of the intercalation site, which is based on an X-ray crystallographic geometry published by Jain and Sobell.⁴⁹ The crystallographic coordinates were treated here in the same way as in the previous study of this ethidium–RNA system by Langner et al.²² In short, hydrogen atoms were added with the Reduce code,⁶⁰ and their positions were optimized in GAMESS (US)⁵⁷ using a PM3 model Hamiltonian (non-hydrogen atoms were kept frozen). Presently, we reoptimized the positions of the hydrogen atoms using the UFF force field⁶¹ as well as DFT/B3LYP calculations. In both cases, one or several external hydrogen atoms on the RNA backbone changed their orientations, whereas no hydrogen atoms with close contacts moved more than 0.1 Å. This did not influence the interaction energies we are interested in by more than 0.2 kcal/mol (data not shown), and we thus retained the PM3 coordinates of hydrogen atoms for consistency.

The system was divided into fragments in several ways, most of which are illustrated by the schematic representations in Figure 1.

Names beginning with **A** were limited to the ethidium molecule and four neighboring base pairs. Model AC4 is the smaller, previously used version,²² which contains only the ethidium chromophore. The A4 variant labeled additionally *parts* considers the entire ethidium molecule but divides it into three parts—its chromophore, side chain, and ring. The versions with **B** in their names include the sugars and phosphate groups connecting the base pairs, whereas the intermediate model Bns disregards the phosphate groups and represents the RNA fragment only by nucleosides (adenosine or uridine).

There is an additional number (N_{int}) in each name (e.g., 1 or 4 in AC1 or AC4) that denotes the actual number of RNA–intercalator dimers evaluated when estimating the interaction energy ΔE in a pairwise fashion. In Figure 1, the subsystems in each case are indicated by different colors. For clarity, we also append this number to the name of the largest models, so that A1 stands for the case where the interaction energy of ethidium was evaluated with all four nucleobases in a single calculation (one dimer).

Whenever the number of interacting dimers N_{int} was more than one, the interactions of the intercalator with each RNA fragment were summed in a pairwise fashion. For instance, in the case of A2, the total pair interaction energy consists of two parts, corresponding to uncoupled contributions from the interaction of ethidium with two base pairs:

$$\Delta E_{\text{A2}} = \Delta E_{\text{Eth}\dots\text{A/U}} + \Delta E_{\text{Eth}\dots\text{U/A}} \quad (4)$$

where ΔE can be substituted by the interaction energy at any chosen level of theory, or by any term from eq 2.

The interaction energies on the right-hand side of eq 4 are estimated in different dimer-centered basis sets, and this rule was adopted for all of the models used. Therefore, the pairwise interaction energies of A2 and A4 differ from that of model A1 not only due to neglect of many-body interactions but also by virtue of basis set extension effects. The same is true when comparing B1 to B2 or B4 or to any of the other variants.

In contrast to the **A**-type models, differences can also arise due to capping effects when phosphate groups are included. For

example, in the B4- and B6-type models, four covalent bonds are severed, between each nucleobase and the sugar residue it is attached to, and the cut bonds are hydrogenated with protons. This means that in these two cases the intercalator effectively interacts with four extra nuclei and electrons compared to B1. In B4.3 and B4.5 on the other hand, two bonds were cut, between the phosphate groups and sugar residues at the 3' (closer to uracil) and 5' (closer to adenine) ends, respectively. Therefore, in B4.3, both phosphate groups were in the same monomer as adenosines, while in model B4.5, they were attached to uridines. In all of these cases where bonds were cut, the positions of the capping hydrogen atoms were optimized using the UFF force field.

Furthermore, the phosphate groups in the B class of interaction models were considered in several variants with respect to their charge state, specified by additional labels and illustrated in Figure 2. In models labeled as *neutral (H)*, the anionic phosphate groups were neutralized by attaching protons to the most anionic oxygen, as has been practiced in the literature.²¹ Variants quenched with counterions were also prepared, in which the positions of Na⁺ and K⁺ were optimized near each phosphate group—these are labeled *neutral (Na)* and *neutral (K)*. In the interaction models labeled *charged -2*, the charge on the RNA fragment was left entirely unbalanced, and *charged -2 (H₂O)* indicates that both anionic groups were hydrated with single water molecules (one per phosphate group). The position of the additional proton in the *neutral (H)* variant was optimized using the UFF force field, and counterions and water molecule positions were optimized at the DFT/B3LYP level in the vicinity of the relevant RNA backbone fragment, which was kept fixed.

3. RESULTS AND DISCUSSION

3.1. Pair-Wise Models. In all of the models tested (AC, A, Bns, B, and their variants), dividing the RNA fragment into nucleobases and backbone strands or into nucleotides is justified from the energetic point of view (Table 1). For the neutral models, partitioning of the system in this way changes the total interaction by no more than 2%. In particular, for the smallest AC-type models, the difference between the sum of pairwise interactions (the ethidium chromophore with each nucleobase separately, AC4) and the interaction calculated with all four nucleobases at once (AC1) was 0.4 kcal/mol. For the model including the ethidium side chain and ring (A), the analogous difference between A4 and A1 was only 0.2 kcal/mol. Such small differences indicate that, while the properties of adenine–uracil base pairs may be significantly changed from that of single nucleobases, this does not noticeably influence their interactions with ethidium.

The largest error associated with a pairwise approximation of the interaction energy was found in the case of the charged version (B·charged -2), around 8 kcal/mol or 6% for B4.5 compared to B1. This relatively large difference is understandable, since the delocalization of unbalanced charge is hindered more strongly when fragmentation occurs near the phosphate groups. Separating the two strands (as in B2) already increases the interaction by almost 5 kcal/mol.

All these conclusions hold for both the total interaction energy at the MP2 level as well as the particular components into which it is divided, although the electrostatic part $\Delta E_{el}^{(1)}$ seems to be the most affected by fragmentation. The good performance of such a pairwise approximation provoked us to calculate the nonadditive

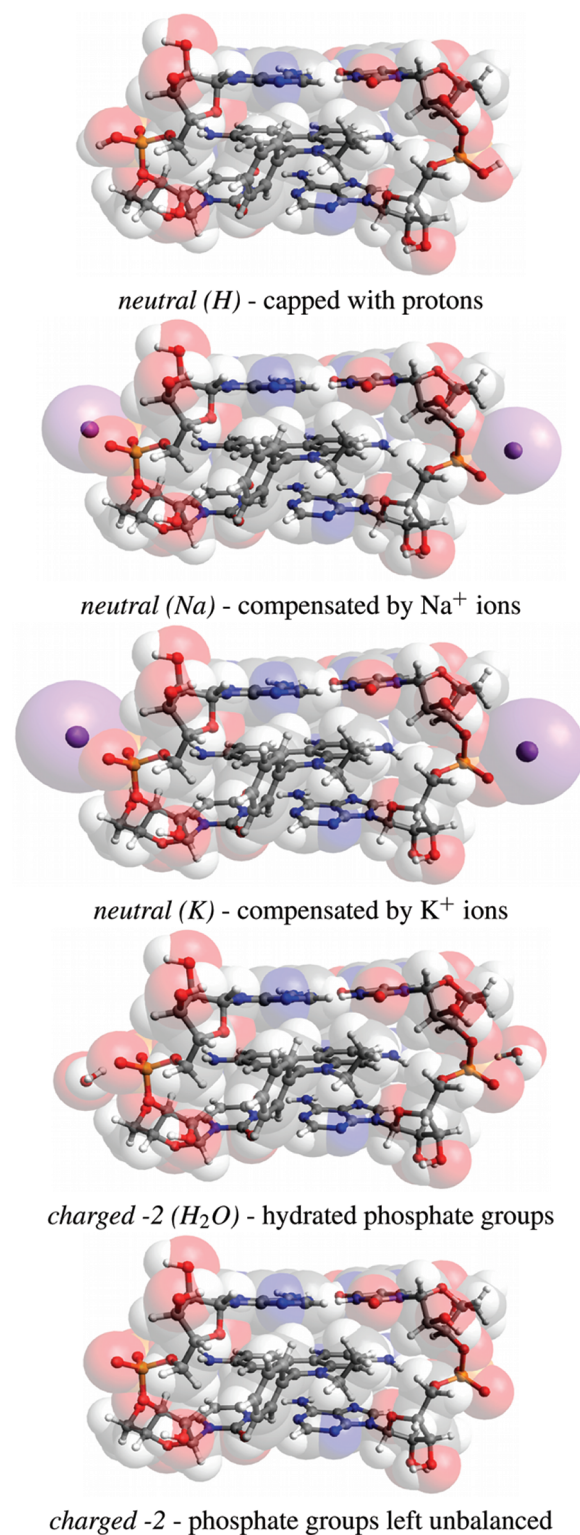


Figure 2. The five phosphate group variants considered in this work, in the order of increasing interaction energy (from top to bottom) for the intercalated ethidium cation in the gas phase (according to Table 1).

contributions to the interaction energy at the Hartree–Fock level and its components for the models AC and A (containing ethidium or its chromophore and four nucleobases) as well as for selected B variants (ethidium with two dinucleotides). Since in

Table 1. Components of the Interaction Energy for the Eth⁽⁺¹⁾–UA/AU Intercalation Complex in the Gas Phase Following eq 1 in the Text^a

	N_{int}	N_{AO}	$\Delta E_{\text{el}}^{(1)}$	$\Delta E_{\text{ex}}^{(1)}$	$\Delta E_{\text{del}}^{(R)}$	ΔE_{corr}	ΔE_{MP2}
model AC1	1	878	−24.8	31.8	−4.8	−33.4	−31.3
model AC2	2	615	−25.3	32.0	−4.7	−33.5	−31.5
model AC4	4	475	−26.0	32.3	−4.7	−33.3	−31.7
model A1	1	1030	−28.6	40.2	−6.1	−39.7	−34.3
model A2	2	733	−28.9	40.2	−6.0	−39.4	−34.2
model A4	4	601	−29.8	40.7	−6.2	−39.2	−34.5
model A4 (parts)	12	475	−30.1	41.6	−6.4	−39.6	−34.5
model Bns	4	753	−35.7	51.7	−10.6	−51.9	−46.6
model B1·neutral (H)	1	1776	−41.9	54.9	−10.9	−54.7	−52.6
model B2·neutral (H)	2	1106	−43.0	55.4	−11.1	−54.4	−53.1
model B4·neutral (H)	4	829	−42.6	55.7	−11.1	−54.7	−52.7
model B6·neutral (H)	6	829	−43.5	56.2	−11.3	−54.4	−52.9
model B4.3·neutral (H)	4	799	−43.5	56.7	−12.1	−54.7	−53.6
model B4.5·neutral (H)	4	832	−40.3	46.2	−10.5	−49.4	−54.0
model B1·neutral (Na)	1	1802	−50.7	55.1	−10.9	−55.0	−61.5
model B1·neutral (K)	1	1830	−52.7	55.1	−10.9	−55.1	−63.6
model B1·charged −2 (H ₂ O)	1	1814	−117.1	55.2	−11.0	−56.4	−129.3
model B1·charged −2	1	1766	−121.1	55.2	−11.1	−56.5	−133.5
model B2·charged −2	2	1101	−123.5	56.0	−13.9	−56.6	−138.0
model B4·charged −2	4	824	−122.5	56.5	−15.8	−57.1	−138.9
model B6·charged −2	6	824	−123.3	57.0	−16.0	−56.8	−139.2
model B4.3·charged −2	4	794	−126.0	56.7	−15.2	−57.0	−141.6
model B4.5·charged −2	4	827	−120.5	46.5	−15.5	−51.8	−141.2

^a The symbols in the left column correspond to various molecular interaction models as illustrated in Figure 1 and described in the text. The column N_{int} contains the number of pair-wise calculations comprising the interaction, and N_{AO} denotes the maximum number of atomic orbitals used in any pair-wise calculation within a particular model. All energies were obtained using the 6-31G** basis set, and are given in kcal/mol.

these cases there is no capping or overlap between fragments, the interaction energy can be expressed as a sum of pairwise and nonadditive components evaluated consistently in the basis set of the entire system. The results of these calculations (Table 2) show that all of the nonadditive terms are virtually negligible and that the observed consistency of the full and fragmented models in Table 1 is in fact due to the small magnitude of these effects and not to a fortuitous cancellation of errors. The nonadditivities are also small in the case of B-type models, in which the charged phosphate groups are either compensated by counterions or hydrated. Only in the latter case, the nonadditive polarization effects are larger, since hydration does not quench the excess charge as effectively as compensation by protons or counterions. And in this case, the pairwise approximation overestimates the extent of delocalization effects by approximately 12%, while for the other studied systems, this error is between 2% and 5%.

It needs to be stressed that the nonadditivities discussed above include interactions between all nucleobases (or strands) alongside ethidium–nucleobase (or strand) interactions and thus are not directly comparable to the primary results of this work, which include only the latter. To illustrate this point explicitly, consider ΔE_{RHf} in the AC-type models, that is, the energy obtained by bringing the intercalator from infinity into its position between the base pairs, assuming there are no relaxation or deformation effects. On the other hand, $\Delta \epsilon_{\text{RHf}}$ is the energy needed to simultaneously dissociate all five molecules of the intercalation site. Nonetheless, the nonadditivities that exist in $\Delta \epsilon_{\text{RHf}}$ ($\Delta^3 \epsilon_{\text{ex}}^{(1)}$,

Table 2. Many-Body Partitioning of the Total Interaction Energy for the Eth⁽⁺¹⁾–UA/AU Intercalation Complex^a

	$\Delta \epsilon_{\text{el}}^{(1)}$	$\Delta^2 \epsilon_{\text{ex}}^{(1)}$	$\Delta^3 \epsilon_{\text{ex}}^{(1)}$	$\Delta^2 \epsilon_{\text{del}}^{(R)}$	$\Delta^3 \epsilon_{\text{del}}^{(R)}$	$\Delta \epsilon_{\text{RHf}}$
model AC	−90.6	108.0	0.0	−32.4	0.8	−14.2
model AC/aug-cc-pVDZ	−87.5	109.1	0.0	−35.2	1.3	−12.5
model A	−94.5	116.6	−0.2	−33.9	1.1	−10.9
model B·neutral (Na)	−117.1	132.1	−0.2	−38.9	1.7	−22.4
model B·neutral (K)	−119.0	132.2	−0.2	−39.0	1.8	−24.2
model B·charged −2 (H ₂ O)	−161.8	134.1	−0.2	−42.0	4.8	−65.1

^a Δ^n , where $n = 2, 3, 4+$, denotes two-, three- and higher-order many-body effects, respectively. In contrast to ΔE_{RHf} and other values given in Table 1, these interaction energies include not only ethidium–nucleobase (or ethidium–strand) interactions but also nucleobase–nucleobase or strand–strand interactions; hence, the two are not directly comparable. In the B-type models, counterions (or water molecules) together with the corresponding strands were treated as single monomers. In all cases, energies are in kcal/mol, and four-body and higher contributions, namely, $\Delta^{(4+)} \epsilon_{\text{ex}}^{(1)}$ and $\Delta^{(4+)} \epsilon_{\text{del}}^{(R)}$, were insignificantly small and are therefore omitted here. Unless otherwise noted, all results were obtained using the 6-31G** basis set.

$\Delta^3 \epsilon_{\text{del}}^{(R)}$, and higher order terms) are inherent also to ΔE_{RHf} , so it is worthwhile to show that they are small. Our goal here, however, is not to provide precise values but rather to confirm that these effects are in fact small, since this in turn confirms that the effects of nucleobase polarization are similarly small.

It is known that in order to accurately evaluate many-body effects beyond the Hartree–Fock level correlated methods higher than MP2 need to be employed. This point has been recently demonstrated for a uracil tetramer,⁶² although the magnitude of these effects remains a fraction of the total interaction energy. This may be attributed to an intrinsically local nature of electron correlation effects, and therefore we expect similar magnitudes for the systems studied here. Furthermore, Ghosh et al.⁴⁴ have demonstrated that for oligomers of nucleic acid bases pairwise interactions between only nearest neighbors are a sufficient approximation. Combined with our results, this opens a potential efficient route for calculations of intercalators bound to entire nucleic acid strands.

Recently, Hill and Platts reported similar considerations for three intercalators, ethidium being among them.²⁸ They compared interaction energies calculated for the intercalator between two base pairs separately and for the entire base pair step, which respectively correspond to A2 and A1 in our case. Interestingly, they also found only small differences between the two for ethidium, although in their case, the molecule was intercalated in the AT/AT step and slightly different methods and basis sets were used. They reported larger deviations in the case of daunomycin situated between GC/GC base pairs, namely, around +5 kcal/mol or 15% of the interaction energy with both base pairs. Therefore, the question remains whether these deviations originate from different properties of GC base pairs compared to AT or AU, or rather from differences between the chromophores of these intercalators. In any case, the accuracy of this type of fragmentation will depend heavily on the molecular details and cannot be recommended for all scenarios.

3.2. Electron Correlation Corrections. It is well established that the MP2 method overestimates the dispersion part of stacking interaction energies, due to missing response effects,²⁷ and it is now quite common to correct for this. [In general, perturbative corrections are doomed to fail for conductors where the HOMO–LUMO gap goes to zero and energy denominators accordingly diverge. This effect is partially responsible for the deteriorating quality of MP2 results as the sizes of aromatic systems increase and orbitals become more delocalized.] Here, we perform additional QCISD(T) calculations for the smallest, nonfragmented (AC1) and fragmented (AC4) models in order to assess the extent of these effects (Table 3). A correction of around 9 kcal/mol is obtained (about 30% of both ΔE_{corr} and ΔE_{MP2} for models AC and A). Surprisingly, the contribution of triple excitations is relatively minor compared to polycyclic aromatic hydrocarbons such as the recently studied coronene dimer,⁴³ although this difference might be caused by the limited basis set used here. Another reason might be the fact that the system under study consists of separate molecules, smaller than coronene. In any case, the description provided by MP2 is quite accurate here, closer in quality to the benzene dimer and other small aromatic systems. Although the largest model studied here (AC1) is not small, the π – π stacking interaction of the ethidium cation decomposes easily into smaller, pairwise components between the intercalator and separate nucleobases.

On the other hand, dispersion interactions are also quite difficult to saturate with respect to the basis set,^{30,38} and in our case, an additional –17 kcal/mol are added to the MP2 interaction energy for model AC1 when 6-31G** is replaced with aug-cc-pVDZ (Table 3). Combined with the 9 kcal/mol correction discussed above (connected to the overestimation of dispersion by MP2 that is relatively unaffected by basis set extension), this

Table 3. Interaction Energy at Various Levels of Theory for Models AC1 and AC4 of the Eth⁽⁺¹⁾–UA/AU Intercalation Complex in the Gas Phase (All Values in kcal/mol)^a

model/basis set	$\Delta E_{\text{el}}^{(1)}$	$\Delta E_{\text{HL}}^{(1)}$	ΔE_{RHF}	ΔE_{MP2}	ΔE_{QCISD}	$\Delta E_{\text{QCISD(T)}}$
AC1/6-31G**	–24.8	7.0	2.1	–31.3	–18.1	–22.0
AC4/6-31G**	–26.0	6.3	1.7	–31.3	–18.4	–22.4
AC1/6-31G** ($d = 0.25$)	–24.5	6.4	0.8	–44.3	–28.3	–33.9
AC4/6-31G** ($d = 0.25$)	–25.5	5.7	0.0	–44.7	–29.1	–34.7
AC4/aug-cc-pVDZ	–25.1	7.6	1.0	–48.8	–32.2	–39.2

^a The triples contribution, and consequently QCISD(T), for AC1/aug-cc-pVDZ is expected to be within 1 kcal/mol of AC4/aug-cc-pVDZ and was judged too expensive to calculate (see the Supporting Information for details).

gives an overall correction of about –8 kcal/mol. Since these effects are essentially local in nature, we shall assume that this correction is also applicable to B-type models. This gives a corrected gas phase interaction energy of roughly –61 kcal/mol for model B·neutral (H) and –142 kcal/mol for B·charged –2. For comparison, Kubař et al.²¹ also reported a correlated (RI-MP2) interaction energy of –70 kcal/mol for ethidium between an AT/TA step in a system corresponding to the present B·neutral (H). We also tested the modified basis set 6-31G** ($d = 0.25$), where d functions were made more diffuse by lowering their exponents from 0.8 to 0.25, as well as the smaller basis set 6-31G* ($d = 0.25$) that is more widespread (see the Supporting Information), and both of these produced correlation energies close to the aug-cc-pVDZ basis set. Although electron correlation effects are dominant and in fact are the origin of stabilization in the case of the smaller neutral models, where ΔE_{RHF} is repulsive, it is the electrostatic component that differentiates between different charge states of the phosphate groups in the gas phase. An analogous observation has been made for 16 stacked DNA bases,⁶³ where relative stabilities correlate more strongly with the electrostatic term than with the dominant dispersion component.

3.3. Phosphate Group Charge and Solvation Effects. As already mentioned, several variants of the extended model B were considered. In the first case, *neutral* (H), the anionic phosphate groups were neutralized by protonation. Quenching by adding one of two counterions (Na^+ or K^+) to each phosphate group was designated as *neutral* (Na) and *neutral* (K). The *charged* –2 model on the other hand leaves the phosphate groups charged or hydrated by a single water molecule. In order to consider the relevance of these different situations, one should keep in mind the significant fluctuations an intercalation site and its surroundings may undergo in solution, and that it has been shown that the movement of counterions is diffusive around DNA.^{32,33} Therefore, the quenching in models B (Na^+) and B (K^+) is at best a temporary configuration. At times when no counterion is present in the immediate vicinity of a phosphate group, it will simply be hydrated.

Our results show that monohydration damps the interaction compared to bare charged phosphate groups only slightly (about +5 kcal/mol). Hydration with further water molecules should increase this effect, and counterions that are farther away may still quench them indirectly by polarizing adjacent water molecules. In an attempt to estimate the effect of solvation, we performed PCM calculations at the Hartree–Fock level in selected cases

Table 4. The Effect of Polarizable Continuum Model (IEF-PCM) Solvation on the Interaction Energy at Various Levels of Theory for Neutral and Charged Variants of Model B1^a

		$\Delta E_{\text{el}}^{(1)}$	$\Delta E_{\text{ex}}^{(1)}$	$\Delta E_{\text{del}}^{(R)}$	$\Delta \Delta G_{\text{RHF}}^{\text{int}}$	ΔE_{corr}	ΔE_{MP2}
model B1 · neutral (Na)	in vacuum	−50.7	55.1	−10.9		−55.0	−61.5
	in solvent	−58.2	56.6	−6.7	29.0 (21.2)		−34.3(−42.1)
model B1 · neutral (K)	in vacuum	−52.7	55.1	−10.9		−55.1	−63.6
	in solvent	−60.5	56.6	−6.7	31.2 (23.5)		−34.5(−42.2)
model B1 · charged −2 (H ₂ O)	in vacuum	−117.1	55.2	−11.0		−56.4	−129.3
	in solvent	−122.3	56.6	−7.2	92.7 (84.8)		−36.6(−44.5)

^aThe $\Delta \Delta G_{\text{RHF}}^{\text{int}}$ values in parentheses are the corresponding $\Delta \Delta G_{\text{RHF}}^{\text{stab}}$ energies; ΔG_{MP2} values in italics were estimated assuming the gas phase ΔE_{corr} energies. All energies were obtained using the 6-31G** basis set, and are given in kcal/mol. See the text for further details.

(Table 4). The individual components of the internal interaction energy are not substantially modified when passing to the solvent. In particular, one notes a slight increase in $\Delta E_{\text{el}}^{(1)}$, which is balanced by an increase in exchange-repulsion and a decrease in $\Delta E_{\text{del}}^{(R)}$. The resulting internal interaction energy has almost the same magnitude as its gas phase counterpart.

The polarizable environment, however, shields the electrostatic interactions to a large extent, and in effect the Hartree–Fock interaction free energies ΔG_{RHF} are repulsive, close to 20 kcal/mol for all three chosen models. It should be noted that more than 90% of the $\Delta \Delta G_{\text{RHF}}^{\text{int}}$ coupling term comes from the corresponding Heitler–London component which includes electrostatic interactions with the polarized solvent. Therefore, $\Delta \Delta G_{\text{RHF}}^{\text{int}}$ can be interpreted as an effective correction to the $\Delta E_{\text{el}}^{(1)}$ term. Although we did not compute the corresponding solvent-adjusted ΔE_{corr} terms, these should not differ much from their gas phase counterparts. We therefore roughly estimate the interaction free energy ΔG_{MP2} for all of the cases presented in Table 4 to be −35 kcal/mol and the alternate $\Delta \Delta G^{\text{stab}}$ calculated using monomeric cavities in eq 3 to be about −43 kcal/mol. It should be stressed that the approximations made here are quite crude, and we do not aim to establish quantitative results. This is impossible, since the PCM method was not parametrized for such systems and the result will depend heavily on the choice of the solute cavities. We also neglect the effects of geometry relaxation here and account only for the electrostatic contribution to the free energy required to bring the interacting molecules from an infinite separation to the complex in a dilute solvent.

Nonetheless, a particularly intriguing observation is that the polarizable environment seems to equate all of the considered charge states of the phosphate groups. What relevance then does the large magnitude of the gas phase interaction energy in the case of the charged models have? This is an especially pertinent question since the overall binding free energies for various intercalators are typically estimated around −10 kcal/mol by both experimental⁹ and theoretical¹⁶ methods. Moreover, the final phase of intercalation—insertion of the planar chromophore between two base pairs—facilitates significant changes in the hydration shell and entails a significant free energy barrier.⁸ If the charge density around the phosphate groups fluctuates as a result, the instantaneous interactions may intensify in a way reminiscent of the charged models studied here. Therefore, our results can be viewed as identifying the broadest range of intermolecular interaction strengths possible for this conformation of the intercalation site with **B · neutral** (*H*) and **B · charged** −2 giving the approximate lower and upper limits. It would be interesting to confront the present static scenario with a dynamic range that

takes conformational changes of the binding site and surrounding solvent molecules into account on equal footing.

4. CONCLUSIONS

In this contribution we have taken a detailed look at the intermolecular interactions in the intercalation site Eth⁽⁺¹⁾–UA/AU, which consists of the ethidium cation and the RNA AU/UA base pair step, built on the basis of crystallographic data. We consider the effect of dividing the system into fragments and calculating *ab initio* interaction energies in a pairwise fashion, as well as many-body effects and the influence of phosphate group quenching by various surroundings. We also explore the effects of solvation on these interactions by assuming a polarizable continuum model of water around the intercalation site.

Our results can be summarized in three basic conclusions, the first pertaining to system fragmentation into parts interacting pairwise with the intercalator. Namely, the various ways chosen of partitioning the system reproduce the supermolecular interaction energy within a few percent, even when the anionic phosphate groups are left unbalanced. But not all of the interaction models are equal in this regard. For example, breaking the RNA strand into backbone and/or nucleobase fragments (as in **B2** and **B6**) results in interaction energies closer to model **B1** than when the strands are split near the phosphate groups (as for example in **B4.3** and **B4.5**). It is important to remember that the success of such fragmentation schemes varies on a case-by-case basis, so we cannot recommend it as a general approach. Specifically, the present exploration of interaction models allows us to conclude that the previously adopted partitioning scheme²² was satisfactory, as far as the energetics of the intercalator itself are concerned.

A second outcome, concerning the electron correlation part of the π – π stacking interaction with the nucleobases, is the addition of almost −10 kcal/mol to the correlated interaction energy in model **AC**. This result was obtained by additional calculations at the QCISD(T) level of theory and employing more saturated basis sets.

Finally, for differently quenched phosphate groups, we estimate the possible range of interactions experienced by ethidium in the gas phase for this particular intercalation site conformation, roughly 60–140 kcal/mol. On the one hand, capping the phosphate group with a hydrogen atom mimics an overly tightly bound counterion and represents the weakest possible interaction of the cationic intercalator with the RNA backbone. Conversely, leaving the charge on the phosphate groups unbalanced represents the strongest possible interaction. Quenching with counterions and hydrating with water molecules leads to

intermediate interaction energies, and in fact we find that all of these cases produce interaction energies between the two extremes. Including solvation via a PCM method has the intuitive effect of shielding electrostatic interactions and bringing the interaction energies in these different models to comparable values, namely, around -35 kcal/mol. The disparity between this and the wide range found in the gas phase suggests that the forces acting on the intercalator during insertion may depend heavily on conformational changes and fluctuations in the hydration shell.

It is important to keep in mind that these results were obtained for a single crystallographic binding site conformation and could likely be changed by dynamic effects. In this study, we focused on the intermolecular interaction energy ΔG_{mol} , calculated for structures that are fully prepared for the insertion of the intercalator. Intercalation inevitably involves the untwisting of the helix and other deformations that present a significant endothermic contribution to the total free energy, as well as hydration and potentially other effects. Therefore, the impact of our results, and ΔG_{mol} in general, on intercalation energetics should be considered in the context of these other contributions that make up the total binding free energy.

■ ASSOCIATED CONTENT

S Supporting Information. More detailed interaction energy decomposition results (all components of pairwise interaction models), as well as the Cartesian coordinates of the intercalation site used in our calculations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Andrzej.Sokalski@pwr.wroc.pl.

■ ACKNOWLEDGMENT


K.M.L. and R.W.G. are indebted to the staff at Wrocław Center for Networking and Supercomputing for a generous allotment of computing resources and technical assistance and Wrocław University of Technology for support. This work was also supported by the National Science Foundation under award number CHE-0911541 and by the Mildred B. Cooper Chair at the University of Arkansas. Acquisition of the Star of Arkansas supercomputer was supported in part by the National Science Foundation under award number MRI-0722625.

■ REFERENCES

- (1) Lerman, L. S. *J. Mol. Biol.* **1961**, *3*, 18–30.
- (2) Waring, M. J. *J. Mol. Biol.* **1965**, *13*, 269–282.
- (3) Strekowski, L.; Wilson, B. *Mutat. Res.* **2007**, *623*, 3–13.
- (4) (a) Brano, M. F.; Cacho, M.; Gradillas, A.; de Pascual-Teresa, B.; Ramos, A. *Curr. Pharm. Des.* **2001**, *7*, 1745–1780. (b) Ferguson, L. R.; Denny, W. A. *Mutat. Res.* **2007**, *623*, 14–23. (c) Neto, B. A. D.; Lapis, A. A. M. *Molecules* **2009**, *14*, 1725–1746.
- (5) (a) Guttman, A.; Cooke, N. *Anal. Chem.* **1991**, *63*, 2038–2042. (b) Chen, W.; Turro, N. J.; Tomalia, D. A. *Langmuir* **2000**, *16*, 15–19.
- (6) (a) Starčević, K.; Karminski-Zamola, G.; Piantanida, I.; Žinić, M.; Šuman, L.; Kralji, M. *J. Am. Chem. Soc.* **2005**, *127*, 1074–1075. (b) Stojković, M. R.; Marcz, S.; Glavaš-Obrovac, L.; Piantanida, I. *Eur. J. Med. Chem.* **2010**, *45*, 3281–3292.
- (7) Horowitz, E. D.; Engelhart, A. E.; Chen, M. C.; Quarles, K. A.; Smith, M. W.; Lynn, D. G.; Hud, N. V. *Proc. Natl. Acad. Sci.* **2010**, *107*, 5288–5293.
- (8) Mukherjee, A.; Lavery, R.; Bagchi, B.; Hynes, J. T. *J. Am. Chem. Soc.* **2008**, *130*, 9747–9755.
- (9) Breslauer, K. J.; Remeta, D. P.; Chou, W.-Y.; Ferrante, R.; Curry, J.; Zaunckowski, D.; Snyder, J. G.; Marky, L. A. *Proc. Natl. Acad. Sci.* **1987**, *84*, 8922–8926.
- (10) Graves, D. E.; Velea, L. M. *Curr. Org. Chem.* **2000**, *4*, 915–929.
- (11) Kostjukov, V. V.; Khomytova, N. M.; Evstigneev, M. P. *Biopolymers* **2009**, *91*, 773–790.
- (12) Chaires, J. B. *Arch. Biochem. Biophys.* **2006**, *453*, 26–31.
- (13) Misra, V. K.; Honig, B. *Proc. Natl. Acad. Sci.* **1995**, *92*, 4691–4695.
- (14) Elcock, A. H.; Rodger, A.; Richards, W. G. *Biopolymers* **1996**, *39*, 309–326.
- (15) Resat, H.; Mezei, M. *Biophys. J.* **1996**, *71*, 1179–1190.
- (16) Bagiński, M.; Fogolari, F.; Briggs, J. M. *J. Mol. Biol.* **1997**, *274*, 253–267.
- (17) Medhi, C.; Mitchell, J. B. O.; Price, S. L.; Tabor, A. B. *Biopolymers* **1999**, *52*, 84–93.
- (18) Bondarev, D. A.; Skawinski, W. J.; Venanzi, C. A. *J. Phys. Chem. B* **2000**, *104*, 815–822.
- (19) Řeha, D.; Kabeláč, M.; Ryjáček, F.; Šponer, J.; Šponer, J. E.; Elstner, M.; Suhai, S.; Hobza, P. *J. Am. Chem. Soc.* **2002**, *124*, 3366–3376.
- (20) Hobza, P.; Šponer, J. E. *Chem. Rev.* **1999**, *99*, 3247–3276.
- (21) Kubař, T.; Hanus, M.; Ryjáček, F.; Hobza, P. *Chem.—Eur. J.* **2006**, *12*, 280–290.
- (22) Langner, K. M.; Kędzierski, P.; Sokalski, W. A.; Leszczyński, J. *J. Phys. Chem. B* **2006**, *110*, 9720–9727.
- (23) Dračinský, M.; Castaño, O. *Phys. Chem. Chem. Phys.* **2004**, *6*, 1799–1805.
- (24) Xiao, X.; Cushman, M. *J. Am. Chem. Soc.* **2005**, *127*, 9960–9961.
- (25) (a) Tuttle, T.; Kraka, E.; Cremer, D. *J. Am. Chem. Soc.* **2005**, *127*, 9469–9484. (b) Barone, G.; Guerra, C. F.; Gambino, N.; Silvestri, A.; Lauria, A.; Almerico, A. M.; Bickelhaupt, F. M. *J. Biomol. Struct. Dyn.* **2008**, *26*, 115–129. (c) Li, S.; Cooper, V. R.; Thonhauser, T.; Lundqvist, B. I.; Langreth, D. C. *J. Phys. Chem. B* **2009**, *113*, 11166–11172.
- (26) Fantacci, S.; Angelis, F. D.; Sgamellotti, A.; Marrone, A.; Re, N. *J. Am. Chem. Soc.* **2005**, *127*, 14144–14145.
- (27) Cybulski, S. M.; Lytle, M. L. *J. Chem. Phys.* **2007**, *127*, 141102.
- (28) Hill, J. G.; Platts, J. A. *Chem. Phys. Lett.* **2009**, *479*, 279–283.
- (29) Šponer, J.; Jurečka, P.; Marchan, I.; Luque, F. J.; Orozco, M.; Hobza, P. *Chem.—Eur. J.* **2006**, *12*, 2854–2865.
- (30) Pitoňák, M.; Janowski, T.; Neogrady, P.; Pulay, P.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1761–1766.
- (31) Horowitz, E. D.; Lilavivat, S.; Holladay, B. W.; Germann, M. W.; Hud, N. V. *J. Am. Chem. Soc.* **2009**, *131*, 5831–5838.
- (32) Varnai, P.; Zakrzewska, K. *Nucleic Acids Res.* **2004**, *32*, 4269–4280.
- (33) Kirmizialtin, S.; Elber, R. *J. Phys. Chem. B* **2010**, *114*, 8207–8220.
- (34) Svozil, D.; Hobza, P.; Šponer, J. *J. Phys. Chem. B* **2010**, *114*, 1191–1203.
- (35) Kolar, M.; Kubař, T.; Hobza, P. *J. Phys. Chem. B* **2010**, *114*, 13446–13454.
- (36) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968–5975.
- (37) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
- (38) Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P. *Chem. Rev.* **2010**.
- (39) Janowski, T.; Pulay, P. *Chem. Phys. Lett.* **2007**, *447*, 27–32.
- (40) Fiethen, A.; Jansen, G.; Hesselmann, A.; Schütz, M. *J. Am. Chem. Soc.* **2008**, *130*, 1802–1803.
- (41) ElSohly, A. M.; Hopkins, B. W.; Copeland, K. L.; Tschumper, G. S. *Mol. Phys.* **2009**, *108*, 923–928.
- (42) Singh, N. J.; Min, S. K.; Kim, D. Y.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 515–529.
- (43) Janowski, T.; Ford, A. R.; Pulay, P. *Mol. Phys.* **2010**, *108*, 249–257.

- (44) Ghosh, D.; Kosenkov, G.; Vanovschi, V.; Williams, C. F.; Herbert, J. M.; Gordon, M. S.; Schmidt, M. W.; Slipchenko, L. V.; Krylov, A. I. *J. Phys. Chem. A* **2010**, *114*, 12739–12754.
- (45) Pulay, P. *Chem. Phys. Lett.* **1983**, *100*, 151.
- (46) Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904–6914.
- (47) Hohenstein, E. G.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 184111.
- (48) Platts, J. A.; Hill, J. G. *Mol. Phys.* **2010**, *108*, 1497–1504.
- (49) Jain, S. C.; Sobell, H. M. *J. Biomol. Struct. Dyn.* **1984**, *1*, 1161–1177.
- (50) Sokalski, W. A.; Roszak, S.; Pecul, K. *Chem. Phys. Lett.* **1988**, *153*, 153–159.
- (51) Boys, S. F.; Bernardi, F. *Mol. Phys.* **2002**, *100*, 65–73.
- (52) (a) Janowski, T.; Ford, A. R.; Pulay, P. *J. Chem. Theory Comput.* **2007**, *3*, 1368–1377. (b) Janowski, T.; Pulay, P. *J. Chem. Theory Comput.* **2008**, *4*, 1585–1592.
- (53) Lee, T. J.; Rendell, A. P.; Taylor, P. R. *J. Phys. Chem.* **1990**, *94*, 5463–5468.
- (54) Góra, R. W.; Sokalski, W. A.; Leszczyński, J.; Pett, V. B. *J. Phys. Chem. B* **2005**, *109*, 2027–2033.
- (55) Jeziorski, B.; Moszyński, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.
- (56) (a) Almlöf, J.; Faegri, K.; Korsell, K. *J. Comput. Chem.* **1982**, *3*, 385–599. (b) Haser, M.; Ahlrichs, R. *J. Comput. Chem.* **1989**, *10*, 104–111.
- (57) 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*, 1347–1363.
- (58) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3093.
- (59) Góra, R. W.; Bartkowiak, W.; Roszak, S.; Leszczyński, J. *J. Chem. Phys.* **2004**, *120*, 2802–2813.
- (60) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. *J. Mol. Biol.* **1999**, *285*, 1735–1747.
- (61) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
- (62) Pitoňák, M.; Neogrády, P.; Hobza, P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 1369–1378.
- (63) Langner, K. M.; Sokalski, W. A.; Leszczyński, J. *J. Chem. Phys.* **2007**, *127*, 111102.

Hybrid QM/MM Calculations on the First Redox Step of the Catalytic Cycle of Bovine Glutathione Peroxidase GPX1

Juraj Kóna^{*†} and Walter M. F. Fabian[‡][†]Institute of Chemistry, Center for Glycomics, Slovak Academy of Sciences, Dúbravská cesta 9, 845 38 Bratislava, Slovak Republic[‡]Institute of Chemistry, Karl Franzens University, Graz, Heinrichstrasse 28, A-8010 Graz, Austria Supporting Information

ABSTRACT: The first reaction step of the redox cycle of bovine erythrocyte glutathione peroxidase from class 1 (GPX1) was investigated using hybrid quantum mechanics/molecular mechanics (QM/MM) calculations using the ONIOM methodology. The reduction of hydrogen peroxide by the active-site selenocysteine in selenolate form assisted by the Arg177 residue was modeled based on a proposal from previous molecular dynamics simulations and pK_a calculations (*J. Chem. Theory Comput.* **2010**, *6*, 1670–1681). The redox reaction is predicted as a concerted S_N2 nucleophilic substitution with a concomitant proton transfer from Arg177 onto leaving hydroxide ion upon reduction of hydrogen peroxide. The height of the reaction barrier was predicted in range of 6–11 kcal mol⁻¹, consistent with an experimental rate constant of ca. 10^7 M⁻¹ s⁻¹. The proposed GPX1-Se⁻-Arg177H⁺ mechanism for GPX1 is compared with the GPX3-SeH-Gln83 one proposed for human glutathione peroxidase from class 3 (GPX3) and with the solvent-assisted proton exchange mechanism proposed for GPX-like organic selenols. The structural and energetic parameters predicted by various density functional theory methods (B3LYP, MPW1PW91, MPW1K, BB1K, M05-2X, M06-2X, and M06) are also discussed.

1. INTRODUCTION

Glutathione peroxidase from class 1 (GPX1, EC 1.11.1.9) is a selenoprotein¹ which protects cells from oxidative damage by catalyzing the reduction of H₂O₂, lipidhydroperoxides, and other organic hydroperoxides using a selenocysteine residue (Sec) and glutathione (GSH, γ -glutamylcysteinylglycine) as a reducing substrate. A general scheme of the overall catalytic cycle is shown in Figure 1a. It has been experimentally suggested that the catalytically active form of the enzyme is the selenolate anion (E-Se⁻).^{2,3} In the first redox step, E-Se⁻ is oxidized to the selenenic acid (E-SeOH) with the accompanying reduction of a hydroperoxide substrate to a corresponding alcohol (or water in the case of hydrogen peroxide) with an experimentally measured⁴ rate constant of 4.1×10^7 M⁻¹ s⁻¹. In the second step, the E-SeOH reacts with GSH to produce a selenyl sulfide adduct (E-SeSG). In the third step, a second molecule of GSH attacks E-SeSG to regenerate the active reduced form of the enzyme, and the oxidized form of GSH (GSSG) is formed as a byproduct. This step is the rate-determining step of the overall mechanism with a rate constant of $2.3\text{--}5.1 \times 10^5$ M⁻¹ s⁻¹.^{4,5} As it can be seen in the reaction scheme (Figure 1a), a proton must be supplied for the first redox step and be abstracted in the third reaction step to maintain overall stoichiometry of the catalytic process. The mode of action of these proton transfers is not clear because no ionizable amino acid residues were found in the proximity of the Sec reaction center in available X-ray structures of GPX1^{2,6} and other glutathione peroxidases.^{7–10} The active site is characterized by the highly conserved Sec-Trp-Gln triad, however, five ionizing residues (Arg50, 96, 177, and 178 and His79) do place in the active site with a radius of 6–12 Å from Sec45 in GPX1 (PDB ID: 1GP1).² A similar situation occurs in other glutathione peroxidases, e.g., for human plasma glutathione peroxidase 3 (GPX3), six

ionizing residues (Arg123, 168, 180, and 201, Lys106, and His200) can be found within a radius of 12 Å around Sec73 (PDB ID: 2R37).^{7,10} It is supposed that these residues influence the catalytic process and are involved in specific binding of GSH.^{2,7,11} Molecular dynamics (MD) simulations on GPX1 indicated that side chains of Arg177 or His79 can approach to proximity of Sec45, thus, possibly acting as the acid/basic catalyst in reduction of peroxides by GPX1 (the GPX1-Se⁻-Arg177H⁺ mechanism in Figure 1c).¹¹ A chemical role of arginine as an acid catalyst in the redox reactions of GPX1 (or GPX3) has not been verified yet. Recently, a mechanism based on the solvent-assisted proton exchange (SAPE) was proposed for redox reactions of GPX-like organoselenium compounds^{12–15} and for human plasma GPX3^{16–18} using quantum mechanics (QM) calculations. In the proposed mechanism for GPX3¹⁶ (Figure 1b) water molecules participate in proton exchange between the selenol (and thiol groups) of selenocysteine [and thioredoxin (TRX)] and the amide groups of protein amino acids. It is based on the assumption that nonionizing Gln83 (side chain amide nitrogen) and Leu51 (backbone amide nitrogen) play roles of catalytic bases, and the active reduced form of Sec is selenol rather than selenolate. The calculated reaction barriers of the GPX3 mechanism (GPX3-Se⁻-H₂O-Gln83H⁺, SeH-Gln83, and GPX3-Se⁻-H₂O-Leu51H⁺) are in the range of 17–22 kcal mol⁻¹ and correspond to chemical reactions of ca. 10^7 times slower, as it was found in the experimental measurements.^{4,5} The calculated barriers in that work^{16,17} are in good agreement with benchmark calculations¹⁴ for GPX-like organoselenium compounds, indicating that a more favorable scenario of the redox mechanism of GPX3 and other glutathione

Received: February 22, 2011

Published: June 14, 2011

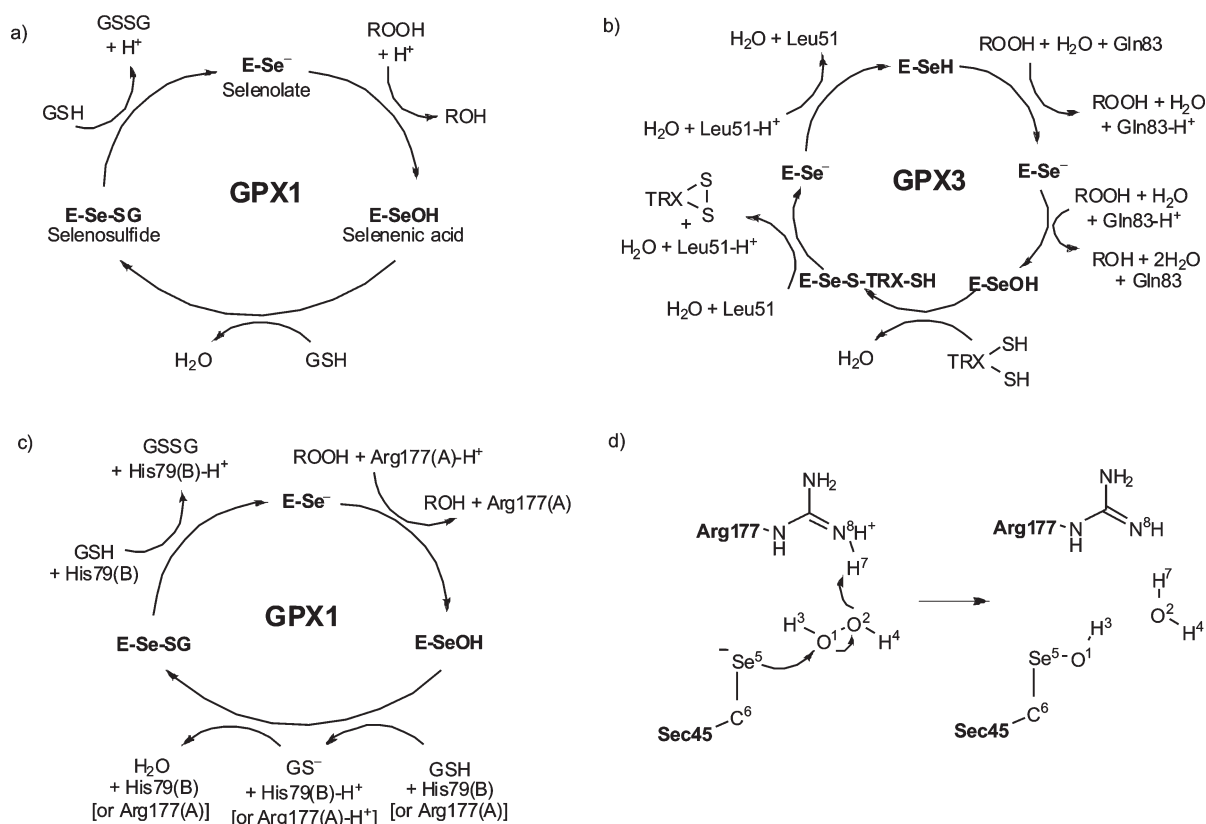


Figure 1. (a) A general scheme for the catalytic cycle of GPX1 after Epp et al.² The resting state of Sec is selenolate. A proton donor and acceptor in the first and third redox steps are not known. (b) A catalytic mechanism proposed for GPX3 by Prabhakar et al.¹⁶ based on DFT calculations. The resting state of Sec is selenol. The concomitant proton transfers occur between Sec and amide nitrogen of Gln83 and between TRX and amide nitrogen of Leu51 facilitated by solvent water molecules. (c) A proposed catalytic mechanism of GPX1 by Ali et al.¹¹ of which the first redox step is calculated at the hybrid QM/MM level in this work. (d) The calculated reaction step with the atom numbering.

peroxidases, consistent with the experimental rate^{4,5} of $2.3\text{--}5.1 \times 10^5 \text{ M}^{-1} \text{ s}^{-1}$, could be operative. The proposed GPX3 mechanism^{16,17} has the reaction barriers comparable to those calculated for the redox reactions in the gas phase and in aqueous solution rather than in the enzyme.^{12–15,19–23} Indeed, the QM calculations on the redox reactions of selenols and thiols without assistance of the proton exchange catalyst predicted unreasonably high reaction barriers.^{24–26}

Here we modeled the first reaction step of the redox cycle of bovine GPX1 based on the proposed GPX1-Se⁻-Arg177H⁺ mechanism (Figure 1c) from the previous MD simulations and pK_a calculations¹¹ using density functional theory (DFT) methods. We will show that the mechanism with the selenolate form of Sec45 and a direct participation of a weak acid or base, in our case Arg177, in the reduction of hydrogen peroxide has a reaction barrier consistent with an experimental rate constant measured for this reaction step in human GPX1 and is energetically favorable over the GPX3-Se⁻-H₂O-Gln83H⁺,²⁷ GPX3-SeH-Gln83,²⁸ and the SAPE mechanisms.^{12–15} We will also briefly discuss inaccuracies in the prediction of the reaction barrier calculated with three hybrid (B3LYP, MPW1PW91, MPW1K) and four meta-hybrid functionals (BB1K, M05-2X, M06-2X, and M06).

2. COMPUTATIONAL DETAILS

The reaction mechanism of the first redox step was calculated using an enzyme model, which consisted of the structure of

GPX1, enzyme substrate (H₂O₂), and explicit solvent molecules (water cap around the active-site Sec45 consisted of 246 water molecules). The structure of GPX1 and water was built from a snapshot selected from a trajectory calculated in the previous MD simulations¹¹ based on the crystal structure of bovine GPX1 (PDB ID: 1GP1).² The GPX1 embedded in a box with water molecules was equilibrated 4 ns in the MD simulation with selenolate redox state (E-Se⁻) without a substrate and a glutathione cofactor. Then, in a selected snapshot all water molecules except those located around Sec45 (up to ca. 15 Å) were deleted, and the structure was subsequently minimized without any constraints at the molecular mechanics (MM) level prior to quantum mechanics/molecular mechanics (QM/MM) calculations. The ionization states of the ionizing amino acid residues of the enzyme were predicted by the PropKa program²⁷ considering an *in vivo* pH of 7.

Transition states and corresponding intermediates of the reaction were optimized at the hybrid QM/MM level with the ONIOM methodology²⁸ with the mechanic embedding scheme (ONIOM-ME-DFT/6-31G(d,p):Amber)^{17,18} using the Gaussian 09 package²⁹ (Gaussian inputs of the MPW1PW91 optimized geometries are available in Supporting Information). The electrostatic effects of the enzyme were evaluated with the electronic embedding (ONIOM-EE) scheme by single point energy calculations on optimized structures in the ONIOM-ME scheme (ONIOM-EE-DFT/6-311+G(3df,2p):Amber//ONIOM-ME-DFT/6-31G(d,p):Amber). We used for the optimization the

ONIOM-ME scheme since optimization of GPX3 with the ONIOM-EE scheme did not improve the results and gave a slightly larger rms deviation when the ONIOM-EE geometries were compared with the X-ray structure of the enzyme.¹⁸ For the QM part of the ONIOM model (Sec45, Gly46, side chain of Thr47 without a methyl group, side chains of Gln80, Trp158, and Arg177, 3 molecules of H₂O located around the Sec45 residue, and the substrate H₂O₂), various DFT functionals were used: The hybrid generalized gradient approximation (GGA) Becke exchange functional with the Lee–Yang–Parr correlation functional (B3LYP),^{30,31} the modified Perdew–Wang exchange functional combined with the Perdew–Wang–1991 correlation functional (MPW1PW91),^{32,33} MPW1PW91 modified for kinetics (MPW1K),³⁴ the meta-hybrid GGA Becke88–Becke95 1-parameter for kinetics (BB1K),³⁵ and Minnesota functionals with double the amount of nonlocal exchange (M05-2X³⁶ and M06-2X)³⁷ as well as meta-hybrid GGA M06 functional.³⁷ In addition, the second-order Møller–Plesset theory (MP2)^{38,39} method was used for single point energy calculations on the optimized B3LYP, MPW1PW91, and MPW1K structures using the ONIOM-EE and ONIOM-ME schemes and the Pople’s split valence triple- ζ basis set augmented by diffuse and polarization functions [6-311+G(3df,2p)]. The MM part of the system was treated with the standard AMBER⁴⁰ (enzyme) and TIP3P⁴¹ (water) force fields. For H₂O₂, selenolate (Sec45–Se[–]) and selenenic acid (Sec45–SeOH) redox forms of selenocysteine parameters were derived either from the AMBER and GAFF force fields or built from data obtained by ab initio Hartree–Fock calculations⁴² [HF/6-31G(d)] using the Gaussian 09 package.²⁹ Atom types, charges, and added force field parameters are provided in Supporting Information. The electrostatic potential fitting algorithm of the Merz–Singh–Kollman scheme⁴³ was used to estimate atomic charges from the HF/6-31G(d) calculations.

All stationary points were characterized as minima or as transition states by vibrational frequency calculations. Transition states were verified as having one and only one imaginary frequency. In addition, for each transition state, intrinsic reaction coordinate (IRC) calculations were performed to confirm that it connects the correct reactant and product minima. Thermodynamic quantities were calculated at 298 K and 101 325 kPa using standard rigid rotor and harmonic oscillator partition function expressions. Zero-point and thermal corrections to enthalpy and Gibbs free energies were calculated from unscaled frequencies obtained at the same level as the geometry optimizations.

The stability of wave function of the transition state in calculations with the restricted HF formalism was tested with the unrestricted MP2 and B3LYP calculations. We found for both methods no differences in energies; therefore, all calculations in this work were performed with the restricted formalism.

3. RESULTS AND DISCUSSION

In the first part we describe and discuss results obtained by the B3LYP and MPW1PW91 functionals in order to compare structural and energetic parameters with those calculated by the same functionals for GPX3^{16,17} and organic selenols.^{12–14} Then, we will discuss structural and energetic differences found among seven functionals.

3.1. Redox Mechanism GPX1–Se[–]–Arg177H⁺. Based on scan calculations along the Se⁵–O¹ reaction coordinate (for atom numbering see Figure 1d) at the B3LYP/6-31G(d,p) level, a preliminary structure of a transition state was obtained. By

subsequent transition-state optimization, the structure of **TS12** was localized at a distance $d(\text{Se}^5\text{–O}^1) = 2.52 \text{ \AA}$ (B3LYP) and 2.43 \AA (MPW1PW91) (Table 1 and Figure 2). The **TS12** was verified by vibrational frequency calculations with one large imaginary frequency [$\nu_i = 286i$ (B3LYP) and 355i cm^{-1} (MPW1PW91)] belonging to the breaking–forming bonds (O¹–O², Se⁵–O¹, and O²–H⁷). The IRC calculations started from **TS12** resulted in a Michaelis complex **1** and a selenenic acid intermediate **2**. The reaction path **1** \rightarrow **TS12** \rightarrow **2** is predicted by the hybrid QM/MM methodology as a general acid-catalyzed nucleophilic substitution proceeding by the S_N2 mechanism, in which addition of the nucleophile (selenolate ion of Sec45–Se[–]), elimination of the leaving group (hydroxide ion from the hydrogen peroxide substrate), and proton transfer from the acid catalyst (Arg177–H⁺) onto the leaving group take place in a concerted manner [$d(\text{Se}^5\text{–O}^1) = 2.52 \text{ \AA}$, $d(\text{O}^1\text{–O}^2) = 1.71 \text{ \AA}$, $d(\text{O}^2\text{–H}^7) = 1.87 \text{ \AA}$, and $d(\text{H}^7\text{–N}^8) = 1.04 \text{ \AA}$ (B3LYP)]. The proton transfer significantly lags behind the breaking of the O¹–O² bond of the substrate, indicating mainly its role in stabilization of the formed byproduct (H₂O) rather than in facilitating the departure of the leaving group in the transition state. When we compared pK_a values of arginine (12.5), the forming selenenic acid intermediate (12.0),²² and the byproduct H₂O (15.7), the proton transfer Arg177–H⁺ + OH[–] \rightarrow Arg177 + H₂O would be only slightly preferred compared to Sec45–SeOH + OH[–] \rightarrow Sec45–O[–] + H₂O. However, the geometry of **TS12** disfavors the direct proton transfer from Sec45–SeOH onto OH[–]. In **TS12**, the value of the angle between the nucleophile selenium atom, the reaction oxygen center, and the leaving group approaches 180° [$\phi(\text{Se}^5\text{–O}^1\text{–O}^2) = 172.59^\circ$ (B3LYP) and 173.1° (MPW1PW91)]. Thus, the position of the atoms is more optimal for transfer of H⁷ from Arg177 than for H³ from H₂O₂ to OH[–] [also $d(\text{O}^2\text{–H}^7) = 1.87 \text{ \AA}$ is shorter than $d(\text{O}^2\text{–H}^3) = 2.06 \text{ \AA}$ (B3LYP)]. It is known from the other QM calculations^{19,21,24–26} performed for the reaction of selenols with hydroperoxides without a proton exchange catalyst, that the leaving hydroxide ion has to abstract a proton directly from the oxygen of the forming selenenic acid. Thus, in the transition state the Se–O–O linear structure is deformed to a more bent geometry [for example, $\phi(\text{Se–O–O}) = 73\text{–}112^\circ$,²⁴ $82\text{–}138^\circ$,²⁶ or 164°].¹⁹

The longer distance $d(\text{Se}^5\text{–O}^1) = 2.52 \text{ \AA}$ (B3LYP) and 2.43 \AA (MPW1PW91) compared with TS geometries of selenols with the SAPE mechanism [$d(\text{Se–O}) = 2.14 \text{ \AA}$ (MPW1PW91)]¹² indicates an earlier transition state when Sec45 is in the selenolate form, and the acid catalyst directly participates in the redox process.

Our best prediction of the thermochemical kinetics parameters, including larger basis set [6-311+G(3df,2p)] and electrostatic effects of the enzyme (ONIOM-EE), is 11.0 kcal mol^{–1} (MPW1PW91), 11.2 kcal mol^{–1} (M06), and 6.0 kcal mol^{–1} (B3LYP) for the activation free energy (ΔG^\ddagger) and –24.2, –31.5, and –23.8 kcal mol^{–1} for the reaction energy (ΔG_r , Table 3). These values were further validated using the MP2 method as an ab initio reference method. Under assumption that DFT methods used in this work have an average error of 2–5 kcal mol^{–1} for the reaction barriers,^{34–37,44–47} the results correlate quite well with the MP2 ones, where values of $\Delta G^\ddagger = 8.4 \text{ kcal mol}^{-1}$ (MP2//MPW1PW91) and 8.3 kcal mol^{–1} (MP2//B3LYP) were predicted (Figure 3). They also reasonably converge to an experimentally measured rate constant for the first redox step of human GPX1 ($\Delta G_{\text{exp}}^\ddagger$ of ca. 7 kcal mol^{–1} can be estimated from $k_{\text{exp}} = 4.1 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$).⁴

Table 1. Selected Geometry Parameters, Interatomic Distances (d) (in Å), Valence (φ), and Torsion angles (Φ) ($^\circ$) of the Optimized Transition States and Intermediates at the Various DFT Levels Using the ONIOM-ME-DFT/6-31G(d,p):Amber Methodology.^a

	$d(\text{Se}^5-\text{O}^1)$	$d(\text{O}^1-\text{O}^2)$	$d(\text{O}^2-\text{H}^7)$	$d(\text{H}^7-\text{N}^8)$	$d(\text{O}^2-\text{H}^3)$	$\varphi(\text{C}^6-\text{Se}^5-\text{O}^1)$	$\varphi(\text{Se}^5-\text{O}^1-\text{O}^2)$	$\Phi(\text{C}^6-\text{Se}^5-\text{O}^1-\text{O}^2)$
B3LYP								
1	3.085	1.449	2.133	1.021	1.940	117.21	121.14	-125.71
TS12	2.523	1.713	1.868	1.035	2.060	107.14	172.59	162.69
2	1.833	2.631	1.028	1.655	1.846	98.02	140.26	-127.81
MPW1PW91								
1	3.057	1.427	2.107	1.018	1.923	117.40	120.81	-126.96
TS12	2.426	1.729	1.812	1.036	2.043	107.29	173.09	166.43
2	1.814	2.609	1.031	1.613	1.813	98.15	139.46	-128.75
MPW1K								
1	3.063	1.406	2.064	1.013	1.898	117.12	120.37	-129.27
TS12	2.339	1.752	1.776	1.032	2.019	107.18	173.17	166.09
2	1.818	2.631	1.000	1.698	1.848	94.51	139.79	-131.26
BB1K								
1	3.117	1.411	2.069	1.013	1.894	114.74	119.11	-135.31
TS12	2.350	1.756	1.802	1.029	2.030	107.23	173.83	163.74
2	1.821	2.664	0.995	1.736	1.871	94.67	137.68	-128.32
M05-2X								
1	3.132	1.422	2.041	1.016	1.906	113.02	119.78	-137.91
TS12	2.310	1.781	1.785	1.036	2.034	107.35	172.62	151.39
2	1.827	2.636	1.037	1.600	1.776	94.798	131.71	-130.60
M06-2X								
1	3.142	1.422	1.998	1.022	1.908	111.03	119.47	-141.71
TS12	2.317	1.778	1.760	1.041	2.043	107.08	172.11	142.55
2	1.825	2.636	1.028	1.626	1.796	94.791	132.95	-129.57
M06								
1	3.127	1.423	2.035	1.024	1.908	115.13	119.77	-132.41
TS12	2.461	1.724	1.837	1.036	2.054	109.73	169.28	147.61
2	1.827	2.676	1.010	1.727	2.811	95.44	139.81	-125.58

^a For atom numbering see Figure 1d.

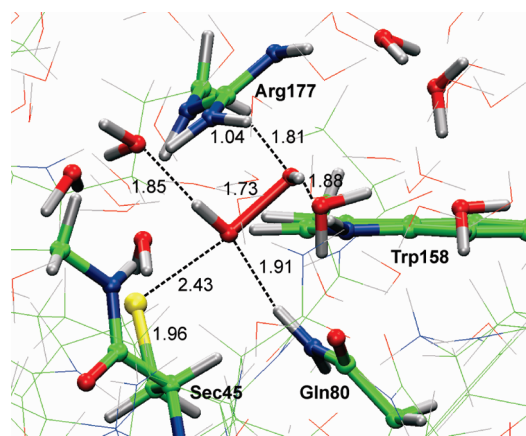


Figure 2. The MPW1PW91 optimized transition state TS12 with selected distances (in Å) (selenium is in yellow, carbons in green, nitrogens in blue, oxygens in red, and hydrogens in gray).

The calculated mechanism with the GPX1-Se⁻-Arg177H⁺ dyad has the reaction barrier ca. 7–8 kcal mol⁻¹ lower compared

with those calculated for the first redox step of GPX3 with the GPX3-SeH-Gln83 dyad [18.0 kcal mol⁻¹, ONIOM-ME-B3LYP/6-311+(d,p):Amber//ONIOM-ME-B3LYP/6-31G(d):Amber]¹⁷ or the SAPE mechanism E-Se⁻-H₂O-Gln83-H⁺ [17.1 kcal mol⁻¹, IEF-PCM-B3LYP/6-311+G(d,p)//B3LYP/6-31G(d)].¹⁶ For the sake of clarity, we compare the ONIOM-ME results (Table 2) since the mechanism with the GPX3-SeH-Gln83 dyad¹⁷ was not performed with the ONIOM-EE methodology. Based on the previous experimental study³ and pK_a calculations for selenocysteine in GPX1¹¹ and in water,⁴⁸ Sec45 can be in either selenol or selenolate form, with the preferable selenolate form in physiological pH. Our calculations on the first redox step of GPX1 further support selenolate as the resting reduced state of GPX1. We show that if a weak acid or base (roughly said a catalyst with pK_a > 7) can directly participate in proton exchange with a hydroperoxide substrate, the redox step can proceed very fast with a rate of ca. 10⁷ s⁻¹.

The SAPE mechanism proposed by Bayse and Antony^{12–15} has not been tested yet for any enzyme GPX system at the QM/MM level, but it is plausible mainly for the redox reaction of organic selenol converted to selenenic acid in absence of an acidic

or basic catalyst in aprotic solvents, e.g., the reaction of selenol in tetrahydrofuran⁴⁹ or dichloromethane⁵⁰ where a catalytic amount of water needed for the SAPE process is produced as a byproduct in the reduction of hydroperoxide substrate. Indeed, the reaction barriers for the SAPE mechanism^{12–15} in the range of 19–22 kcal mol⁻¹ are comparable to the proposed GPX3-SeH-Gln83 mechanism but 2–3 times higher compared with the GPX1-Se⁻-Arg177H⁺ mechanism investigated in this work. Clearly, the GPX1 could catalyze the reduction of hydroperoxides in the resting selenol form, as an alternative when the selenolate form is not available, but this reaction would proceed only slowly with a rate of ca. 10⁻² s⁻¹.

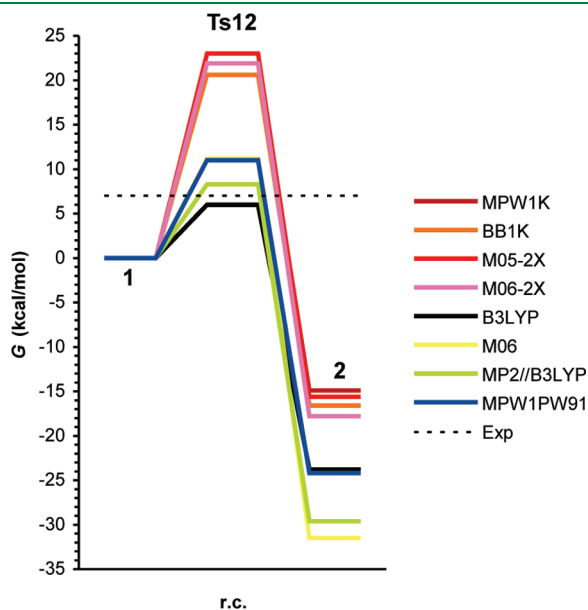


Figure 3. A reaction profile (ΔG) of the first redox step of GPX1 calculated with the MP2 and various DFT methods at the ONIOM-EE-QM/6-311+G(3df,2p):Amber//ONIOM-ME-DFT/6-31G(d,p):Amber level. $\Delta G_{\text{exp}}^{\ddagger}$ (dashed line) of 7 kcal mol⁻¹ is estimated from $k_{\text{exp}} = 4.1 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$.

3.2. DFT Benchmark on Selenium Reactivity. We tested seven functionals for the structural prediction of transition states and reaction barriers. Two functionals fitted for thermochemical kinetics (MPW1K and BB1K) and new hybrid meta-GGA functionals (M05-2X,³⁶ M06-2X,³⁷ and M06³⁷), which all have an average error on the reaction barriers in kinetics benchmarks^{34–37,44–47} less than 2 kcal mol⁻¹, together with the MPW1PW91^{32,33} functional used for modeling redox reactivity of selenium compounds with the SAPE mechanism^{12–15} and with the B3LYP^{30,31} functional used for the modeling of the redox cycle of GPX3^{16,17} were applied. The MPW1PW91 and B3LYP functionals were less accurate in benchmarks on thermochemical kinetics parameters, and they underestimated the reaction barriers ca. 3–5 kcal mol⁻¹, from which MPW1PW91 performs slightly better.^{34–37,44–47}

As can be seen in Table 1 [ONIOM-ME-DFT/6-31G(d,p):Amber] the most important structural parameters, the distances $d(\text{Se}^5-\text{O}^1)$ and $d(\text{O}^1-\text{O}^2)$, differ in the transition state more significantly for the $d(\text{Se}^5-\text{O}^1)$ (0.007–0.213 Å) and less apparently for the $d(\text{O}^1-\text{O}^2)$ (0.016–0.068 Å). The functionals with a smaller portion of the Hartree–Fock (HF) exchange (25%, MPW1PW91; 27%, M06 and 20%, B3LYP) predicted an earlier transition state, i.e., with the longer $d(\text{Se}^5-\text{O}^1)$ and shorter $d(\text{O}^1-\text{O}^2)$ distances compared with others (MPW1K, BB1K, M05-2X, and M06-2X; 42–56% of the HF exchange). For such TS geometries the reaction barriers in a large range of 13–28 kcal mol⁻¹ [ΔG^{\ddagger} , ONIOM-ME-DFT/6-31G(d,p):Amber] were found (Table 2). The results calculated with MPW1K, BB1K, M05-2X, and M06-2X converged to each other with $\Delta G^{\ddagger} = 27\text{--}28$ kcal mol⁻¹ and $\Delta G_{\text{r}} = -50$ to -55 kcal mol⁻¹. The MPW1PW91, M06, and B3LYP predicted reaction barriers with significantly lower values of 17.0, 18.6, and 13.0 kcal mol⁻¹ and comparable reaction energies of -50.6 , -60.4 , and -53.3 kcal mol⁻¹ with the other DFT methods. The relatively high values of ΔG^{\ddagger} for the enzymatic chemical step predicted by the former functionals were further validated using more flexible basis set. When the 6-311+G(3df,2p) basis set was applied on the 6-31G(d,p) optimized structures, a dramatic decrease of the reaction barrier by about

Table 2. Relative Gibbs Free Energies (ΔG) Calculated for the Optimized Transition States and Intermediates at the Various DFT Levels^a

ONIOM-ME ^b	B3LYP	MPW1PW91	MPW1K	BB1K	M05-2X	M06-2X	M06
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TS12	13.0 (286i)	17.03 (355i)	27.1 (481i)	26.8 (461i)	28.2 (499i)	28.3 (490i)	18.6 (405i)
2	-53.3	-50.6	-51.9	-54.1	-50.4	-54.9	-60.4
ONIOM-ME ^c	B3LYP	MPW1PW91	MPW1K	BB1K	M05-2X	M06-2X	M06
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TS12	10.3	14.4	24.1	22.5	23.7	22.8	13.5
2	-57.0	-56.6	-56.9	-59.5	-56.6	-60.4	-69.8
ONIOM-ME ^c	M06 ^d	M06 ^e	M06 ^f	MP2 ^g	MP2 ^h	MP2 ⁱ	
1	0.0	0.0	0.0	0.0	0.0	0.0	
TS12	13.8	13.7	9.3	13.8	12.0	10.9	
2	-66.3	-67.6	-69.2	-62.2	-64.9	-67.8	

^a Using the ONIOM-ME-DFT/6-31G(d,p):Amber methodology (in kcal mol⁻¹), single point calculations energies with the 6-311+G(3df,2p) basis set applied with the MP2 and DFT methods, and imaginary vibrational frequencies (values in round brackets) for the transition states (ν^{\ddagger} , in cm⁻¹).

^b ONIOM-ME-DFT/6-31G(d,p):Amber. ^c ONIOM-ME-QM/6-311+G(3df,2p):Amber//ONIOM-ME-DFT/6-31G(d,p):Amber. ^d M06//B3LYP.

^e M06//MPW1PW91. ^f M06//M06-2X. ^g MP2//B3LYP. ^h MP2//MPW1PW91. ⁱ MP2//MPW1K.

Table 3. Single Point Energy Calculations of Relative Gibbs Free Energies (ΔG) Calculated at the MP2 and Various DFT Levels^a

ONIOM-EE	B3LYP	MPW1PW91	MPW1K	BB1K	M05-2X	M06-2X	M06
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TS12	6.0	11.0	21.9	20.6	23.0	21.9	11.2
2	-23.8	-24.2	-14.9	-16.6	-15.6	-17.8	-31.5
ONIOM-EE	M06 ^b	M06 ^c	M06 ^d	MP2 ^e	MP2 ^f	MP2 ^g	
1	0.0	0.0	0.0	0.0	0.0	0.0	
TS12	9.6	10.8	9.6	8.3	8.4	9.6	
2	-34.8	-34.8	-26.7	-29.6	-32.7	-25.8	

^a Using the ONIOM-EE-QM/6-311+G(3df,2p):Amber scheme (in kcal mol⁻¹). ^b M06//B3LYP. ^c M06//MPW1PW91. ^d M06//M06-2X. ^e MP2//B3LYP. ^f MP2//MPW1PW91. ^g MP2//MPW1K.

3–5 kcal mol⁻¹ to values of 10–23 kcal mol⁻¹ was found (Table 2). This indicates that at least medium-size basis set augmented by diffuse functions should be used for more accurate prediction for the kinetics parameters of the GPX-like systems. The differences in the prediction of the reaction barrier by the B3LYP, MPW1PW91, and M06 functionals are still extremely high (more than 8 kcal mol⁻¹) compared to the MPW1K, BB1K, M05-2X, and M06-2X functionals. The MPW1PW91 functional predicted the barrier (14.4 kcal mol⁻¹) in the best agreement with the MP2 method (12.0 kcal mol⁻¹, MP2//MPW1PW91), which was used as a reference QM method. The MPW1PW91 functional was also used by Bayse and Antony^{12–15} for modeling redox reactivity of selenium compounds and showed the best performance in a DFT benchmark¹⁴ on the reaction barriers of the redox reaction of methaneselenol with hydroperoximethane via the SAPE mechanism. The reaction barrier for reduction of hydrogen peroxide by selenocysteine and methaneselenol in their selenolate forms in gas phase, and an aqueous solution with a height of 15–21 kcal mol⁻¹ was also predicted by MP2 and QCISD(T) methods.^{19,21}

To evaluate electrostatic effects of enzyme surrounding and water environment on the reactivity of Sec45, the ONIOM methodology with electronic embedding was used [ONIOM-EE-DFT/6-311+G(3df,2p):Amber//ONIOM-ME-DFT/6-31G(d,p):Amber]. The ONIOM-EE methodology together with the MPW1PW91, M06, and B3LYP functionals further decrease the reaction barrier ca. 2–4 kcal mol⁻¹ to values of 11.0, 11.2, and 6.0 kcal mol⁻¹, indicating the importance of electrostatic effects of the enzyme (Table 3). However, for other functionals the ONIOM-EE barrier decreases only ca. 1–2 kcal mol⁻¹, predicting still high activation energies (22–23 kcal mol⁻¹) compared with the experimentally measured rate constant⁴ of 4.1×10^7 M⁻¹ s⁻¹ (Figure 3). We conclude that the MPW1PW91, M06, and B3LYP functionals predict the activation energies in the best agreement with the experimental results (estimated $\Delta G_{\text{exp}}^\ddagger$ is ca. 7 kcal mol⁻¹) and with our reference MP2 calculations (8.3–9.6 kcal mol⁻¹, Table 3). The other DFT methods, which have the higher portion of the HF exchange, seem to overestimate the reaction barrier for the first redox step of the GPX1 cycle by at least 10 kcal mol⁻¹. When the new hybrid meta-GGA M06 functional, with a similar portion of the HF exchange as MPW1PW91 and B3LYP have, was applied changing M06-2X (54% HF) with M06 (27% HF),³⁷ the reaction barrier dramatically decreases from 21.9 to 9.6–11.2 kcal mol⁻¹ (Table 3, M06, M06//M06-2X, M06//B3LYP, and M06//MPW1PW91). The M06 values are now very close to the MPW1PW91 and MP2 results.

4. CONCLUSION

Using the QM/MM methodology in this work as well as the MD simulations and pK_a calculations in the previous work,¹¹ we demonstrated that the ionized selenolate state of Sec45 in GPX1 is the favorable form for the reduction of hydroperoxide substrates. The efficient reduction of hydroperoxides by GPX1 can only proceed in the enzyme environment, with a general acid/base catalyst which directly participates in proton exchange with the substrate. Such a role could play a weak acid or base, i.e., the catalyst with pK_a > 7. In GPX1 it could be Arg177. The GPX1-Se⁻-Arg177H⁺ mechanism has the activation barrier of 6–11 kcal mol⁻¹ consistent with the experimentally measured rate constant for the first redox step of GPX1⁴ and is 2–3 times lower compared with those calculated for the GPX3-Se⁻-H₂O-Gln83H⁺,¹⁶ the GPX3-SeH-Gln83¹⁷ of glutathione peroxidase from class 3, and the SAPE mechanism^{12–15} of GPX-like organoselenium compounds. The acidic strength and the proper position of the catalytic acid/base in the active-site for direct proton exchange with the hydroperoxide substrate seem to be crucial factors for a rate of ca. 10^7 M⁻¹ s⁻¹ of the redox enzymatic reaction. This can explain why the QM calculations¹⁹ on the redox reaction of selenocysteine without the assistance of the proton exchange catalyst predicted the higher reaction barrier corresponding to a slow chemical process of a rate of ca. 0.25 M⁻¹ s⁻¹.

Our DFT benchmark calculations indicate that for the more accurate prediction of the selenium reactivity, at least a medium-size basis set augmented by diffuse function and proper description of surrounding electrostatic effects are required. The DFT methods, which were designed for kinetics (MPW1K and BB1K) as well as the new Minnesota functionals with double the amount of nonlocal exchange (M05-2X and M06-2X), all with a higher portion of the HF exchange, seem to significantly overestimate the reaction barrier for GPX1 by more than 10.0 kcal mol⁻¹ compared with the experimental results,⁴ the MP2 method, or with the MPW1PW91, M06, and B3LYP functionals. We recommend for modeling the selenium reactivity the functionals with a lower portion of the HF exchange.

■ ASSOCIATED CONTENT

S Supporting Information. Total and relative electronic energies, thermal corrections to Gibbs free energies, force field parameters used for the Sec45 residue and H₂O₂, and Gaussian input files of the MPW1PW91 optimized structures. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: chemkona@savba.sk. Telephone: +421-2-59410322.

ACKNOWLEDGMENT

Financial support for this research was granted by the Scientific Grant Agency of the Ministry of Education of Slovak Republic and Slovak Academy of Sciences (project VEGA-02/0176/09), the Slovak Research and Development Agency (project APVV-0117-06), the Austrian Federal Ministry of Science and Research (Action Austria-Slovakia, ICM-2010-02933), and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 212043.

REFERENCES

- (1) Flohé, L.; Gunzler, W. A.; Schock, H. H. *FEBS Lett.* **1973**, *32*, 132.
- (2) Epp, O.; Ladenstein, R.; Wendel, A. *Eur. J. Biochem.* **1983**, *133*, 51.
- (3) Gettins, P.; Crews, B. C. *J. Biol. Chem.* **1991**, *266*, 4804.
- (4) Takebe, G.; Yarimizu, J.; Saito, Y.; Hayashi, T.; Nakamura, H.; Yodoi, J.; Nagasawa, S.; Takahashi, K. *J. Biol. Chem.* **2002**, *277*, 41254.
- (5) Roy, G.; Nethaji, M.; Mugesh, G. *J. Am. Chem. Soc.* **2004**, *126*, 2712.
- (6) Kavanagh, K. L.; Johansson, C.; Smees, C.; Gileadi, O.; Von Delft, F.; Weigelt, C. J.; Sundstrom, M.; Edwards, A.; Oppermann, U. *Crystal Structure of the Selenocysteine to Glycine Mutant of Human Glutathione Peroxidase 1*; The Research Collaboratory for Structural Bioinformatics (RCSB) : RCSB-Rutgers, RCSB-San Diego Supercomputer Center, and RCSB-University of Wisconsin-Madison; <http://www.rcsb.org/>. Accessed December 09, 2010; PDB ID: 2F8A.
- (7) Ren, B.; Huang, W. H.; Akesson, B.; Ladenstein, R. *J. Mol. Biol.* **1997**, *268*, 869.
- (8) Dimastrogiovanni, D.; Anselmi, M.; Miele, A. E.; Boumis, G.; Petersson, L.; Angelucci, F.; Di Nola, A.; Brunori, M.; Bellelli, A. *Proteins* **2010**, *78*, 259.
- (9) Alphey, M. S.; Konig, J.; Fairlamb, A. H. *Biochem. J.* **2008**, *414*, 375.
- (10) The Research Collaboratory for Structural Bioinformatics : RCSB-Rutgers, RCSB-San Diego Supercomputer Center, and RCSB-University of Wisconsin-Madison; <http://www.rcsb.org/>. Accessed December 09, 2010; PDB ID: 2R37, 2P31, 2I3Y, 2HE3, 2G3S, 3CYN.
- (11) Ali, S. T.; Jahangir, S.; Karamat, S.; Fabian, W. M. F.; Nawara, K.; Kóna, J. *J. Chem. Theory Comput.* **2010**, *6*, 1670.
- (12) Bayse, C. A. *J. Phys. Chem. A* **2007**, *111*, 9070.
- (13) Bayse, C. A.; Antony, S. *J. Phys. Chem. A* **2009**, *113*, 5780.
- (14) Bayse, C. A.; Antony, S. *Main Group Chem.* **2007**, *6*, 185.
- (15) Bayse, C. A. *J. Inorg. Biochem.* **2010**, *104*, 1.
- (16) Prabhakar, R.; Vreven, T.; Morokuma, K.; Musaev, D. G. *Biochemistry* **2005**, *44*, 11864.
- (17) Prabhakar, R.; Vreven, T.; Frisch, M. J.; Morokuma, K.; Musaev, D. G. *J. Phys. Chem. B* **2006**, *110*, 13608.
- (18) Prabhakar, R.; Musaev, D. G.; Khavrutskii, I. V.; Morokuma, K. *J. Phys. Chem. B* **2004**, *108*, 12643.
- (19) Cardey, B.; Enescu, M. *J. Phys. Chem. A* **2007**, *111*, 673.
- (20) Kóna, J.; Brinck, T. *Org. Biomol. Chem.* **2006**, *4*, 3468.
- (21) Cardey, B.; Enescu, M. *ChemPhysChem* **2005**, *6*, 1175.
- (22) Bachrach, S. M.; Walker, C. J.; Lee, F.; Royce, S. *J. Org. Chem.* **2007**, *72*, 5174.
- (23) Bachrach, S. M.; Demoin, D. W.; Luk, M.; Miller, J. V. *J. Phys. Chem. A* **2004**, *108*, 4040.
- (24) Benková, Z.; Kóna, J.; Gann, G.; Fabian, W. M. F. *Int. J. Quantum Chem.* **2002**, *90*, 555.
- (25) Pearson, J. K.; Boyd, R. J. *J. Phys. Chem. A* **2006**, *110*, 8979.
- (26) Pearson, J. K.; Boyd, R. J. *J. Phys. Chem. A* **2007**, *111*, 3152.
- (27) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704.
- (28) Vreven, T.; Byun, K. S.; Komáromi, I.; Dapprich, S.; Montgomery, J. A., Jr.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 815.
- (29) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Scalmani, G.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.
- (30) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (31) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (32) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.
- (33) Perdew, J. P.; Burke, K.; Wang, Y. *Phys. Rev. B* **1996**, *54*, 16533.
- (34) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811.
- (35) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715.
- (36) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.
- (37) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (38) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. *Chem. Phys. Lett.* **1990**, *166*, 275.
- (39) Möller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
- (40) Cheatham, T. E.; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845.
- (41) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (42) Roothaan, C. C. *J. Rev. Mod. Phys.* **1951**, *23*, 69.
- (43) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.
- (44) Zhao, Y.; Pu, J. Z.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2004**, *6*, 673.
- (45) Zhao, Y.; Gonzalez-Garcia, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012.
- (46) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A* **2007**, *111*, 10439.
- (47) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, S656.
- (48) Byun, B. J.; Kang, Y. K. *Biopolymers* **2011**, *95*, 345.
- (49) Goto, K.; Nagahama, M.; Mizushima, T.; Shimada, K.; Kawashima, T.; Okazaki, R. *Org. Lett.* **2001**, *3*, 3569.
- (50) Back, T. G.; Dyck, B. P. *J. Am. Chem. Soc.* **1997**, *119*, 2079.

Continuous Constant pH Molecular Dynamics in Explicit Solvent with pH-Based Replica Exchange

Jason A. Wallace and Jana K. Shen*

Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma 73019, United States

ABSTRACT: A computational tool that offers accurate pK_a values and atomically detailed knowledge of protonation-coupled conformational dynamics is valuable for elucidating mechanisms of energy transduction processes in biology, such as enzyme catalysis and electron transfer as well as proton and drug transport. Toward this goal we present a new technique of embedding continuous constant pH molecular dynamics within an explicit-solvent representation. In this technique we make use of the efficiency of the generalized-Born (GB) implicit-solvent model for estimating the free energy of protein solvation while propagating conformational dynamics using the more accurate explicit-solvent model. Also, we employ a pH-based replica exchange scheme to significantly enhance both protonation and conformational state sampling. Benchmark data of five proteins including HP36, NTL9, BBL, HEWL, and SNase yield an average absolute deviation of 0.53 and a root mean squared deviation of 0.74 from experimental data. This level of accuracy is obtained with 1 ns simulations per replica. Detailed analysis reveals that explicit-solvent sampling provides increased accuracy relative to the previous GB-based method by preserving the native structure, providing a more realistic description of conformational flexibility of the hydrophobic cluster, and correctly modeling solvent mediated ion-pair interactions. Thus, we anticipate that the new technique will emerge as a practical tool to capture ionization equilibria while enabling an intimate view of ionization coupled conformational dynamics that is difficult to delineate with experimental techniques alone.

INTRODUCTION

Solution pH has a profound effect on the stability and the function of proteins by changing the protonation states of titratable groups. Proteins can become denatured under extreme pH conditions. Enzymes are often catalytically active in a narrow pH range.¹ Protein–protein interactions² and protein–ligand binding³ are also modulated by the protonation states of titratable groups. Accurate determination of active-site pK_a values informs about the catalytic mechanism of proteins.⁴ Knowledge of the native- and denatured-state pK_a values can be used to quantify electrostatic effects on protein stability.⁵

Although the importance of solution pH has long been recognized, molecular simulation techniques have traditionally neglected it. In a standard molecular dynamics (MD) simulation the protonation states of ionizable side chains are set at the beginning of the simulation based on the comparison of the desired pH condition and the solution or model compound pK_a values. This fixed protonation scheme can be a source of error in several instances. For example, if the pK_a values are near the pH of interest, then the protonated and deprotonated states should coexist, which obviously is not reflected in simulation with fixed protonation states. Additionally, even when reasonable protonation states may be set for the initial conformation, conformational rearrangement may favor an entirely new set of protonation states.

In recent years, considerable effort has been made to develop methodologies that explicitly include pH as an external parameter in MD simulations, similar to temperature, allowing protonation states of ionizable groups to respond to changes in the chemical environment and the external pH.^{6–11} These constant pH techniques differ in the way protonation states are updated. In the discrete methods, protonation states are periodically updated using Monte Carlo sampling, while in the continuous approach

titration coordinates are introduced and propagated simultaneously with the spatial coordinates (see a most recent review).¹² One of the most promising constant pH techniques, termed continuous constant pH molecular dynamics (CPHMD)^{9,13} is based on the λ dynamics approach to free energy calculations,¹⁴ allowing ionizable groups to switch continuously between protonated and unprotonated forms. Protonation and deprotonation are accomplished in a manner similar to many free energy simulation techniques, where an alchemical coordinate, λ , is introduced. The novelty of the λ dynamics approach lies in the fact that the alchemical coordinate is assigned to a fictitious λ particle, and the force on the particles is derived analytically. CPHMD has been shown to give accurate and robust predictions for protein pK_a values¹² and has opened a door to theoretical studies of pH-dependent protein dynamics and folding.^{15–17}

In the aforementioned CPHMD method, the generalized Born (GB) implicit-solvent model is used to calculate forces on both spatial and titration coordinates. The major advantage of using GB models in constant pH methodologies is that convergence of pK_a 's can be achieved with a reasonable amount of sampling time, which has not been demonstrated feasible with explicit-solvent models (see more discussions later). Another benefit of using GB models within the CPHMD framework is that forces on the titration coordinates can be computed analytically. However, as CPHMD and other GB-based constant pH techniques are maturing into practical tools, problems inherited from the underlying GB models are becoming the limiting factor for further improvement of accuracy. Recent GB simulation studies have revealed several problems that seem

Received: February 28, 2011

Published: June 22, 2011

difficult to overcome. Specifically, attractive electrostatic interactions are overestimated,^{18,19} and improvement through adjustment of GB input radii that define dielectric boundary^{18,19} is limited.²⁰ Also, due to the lack of solvent granularity, GB simulations cannot reproduce the solvation peaks seen in the interaction free energy profiles from explicit-solvent simulations.¹⁸ Furthermore, there have been noted problems with the stability of hydrophobic interactions^{21,22} and the overly compact and rigid unfolded states,^{23,24} which are likely due to the approximate nature of the nonpolar solvation term based on solvent-accessible surface area (SA model). Finally, the inaccuracies of the GB/SA model in the representation of electrostatic and nonpolar energetics result in a more favorable sampling of helical relative to extended states.^{25,24}

The limitations of GB models affect the accuracy and the applicability of the CPHMD method in several ways. First, a small error in the electrostatic solvation energy calculated by the GB model alters the relative deprotonation free energy in reference to solution and therefore the pK_a shift. This type of “electrostatic” error is typically small for solvent-exposed residues, because the GB model, in particular GBSW used in this work, has been tuned to reproduce the explicit-solvent data of solvent-exposed polar or charged interactions.¹⁸ However, the “electrostatic” error becomes significant for deeply buried residues^{26,12} because the inaccuracy in the desolvation energies of deeply buried atoms remains an unsolved problem in GB models. Nonetheless, the electrostatic error is systematic (Wallace and Shen, unpublished data), and a post correction may be introduced if necessary. The second type of GB-related error which affects the accuracy of $\Delta\Delta G^{\text{deprot}}$ arises from the small distortion in the conformation or distribution of conformations. The impact of this “conformational” error on the protonation-state sampling is typically not systematic, and the extent of the error is unpredictable. Finally, the dependence of conformational sampling on the GB model also hinders the application of the CPHMD method to polyionic systems, such as DNA and RNA, for which GB models are not well suited.

In light of the above considerations, we introduce here a method to extend the CPHMD framework to explicit-solvent simulations. In principle, forces on both spatial and titration coordinates can be derived from explicit-solvent sampling. However, the latter is not practical because a lengthy simulation time is required to accurately compute solvation-related forces based on explicit-solvent sampling. Consequently, we devise a method which takes advantage of the efficiency of the GB model to compute solvation forces on titration coordinates while propagating conformational dynamics via all-atom interactions in explicit solvent. Additionally, we implement a replica-exchange protocol based on the pH biasing energy to significantly accelerate the convergence of the simultaneous sampling of protonation and conformational space. Thus, by making use of the more accurate explicit-solvent sampling, the new method aims to improve the accuracy of CPHMD by reducing the aforementioned “conformational” error and to allow applications to many problems where implicit-solvent models are not feasible.

The rest of the paper is organized as follows. First, we describe in detail the CPHMD method in explicit solvent and the pH-based replica-exchange protocol. We then examine potential artifacts due to the use of both explicit- and implicit-solvent schemes and the response of solvent molecules to titration. Next, we present and discuss results of model compound titrations and analyze the convergence behavior with the new sampling protocol. Finally, we benchmark the accuracy of the new method by

calculating pK_a values of five proteins including HP36, NTL9, BBL, HEWL, and SNase. We compare the results with the GB-based CPHMD simulations and experiment. We find that the explicit-solvent CPHMD offers slightly more accurate pK_a predictions but significantly deeper physical insights. Surprisingly, convergence of the explicit-solvent CPHMD titrations is achieved for all proteins with a simulation length of 1 ns per replica, suggesting that the new method will emerge as a powerful and practical tool for theoretical studies of electrostatic phenomena.

METHODS

Continuous Constant pH Molecular Dynamics in Implicit Solvent. To prepare for the discussion of the explicit-solvent based CPHMD method, we first briefly review the CPHMD method employing the generalized Born (GB) implicit-solvent model; more details can be found elsewhere.^{9,13,26} Based on the λ dynamics technique,¹⁴ CPHMD utilizes an extended Hamiltonian to simultaneously propagate spatial (real) and titration (virtual) coordinates. Thus, the total Hamiltonian of the system can be written as

$$\mathcal{H}(\{\mathbf{r}_a\}, \{\boldsymbol{\theta}_i\}) = \sum_a \frac{m_a}{2} \dot{\mathbf{r}}_a^2 + U^{\text{int}}(\{\mathbf{r}_a\}) + U^{\text{hybr}}(\{\mathbf{r}_a\}, \{\boldsymbol{\theta}_i\}) + \sum_i \frac{m_i}{2} \dot{\boldsymbol{\theta}}_i^2 + U^*(\{\boldsymbol{\theta}_i\}) \quad (1)$$

where $a = 1, N_{\text{atom}}$ is the index for atomic coordinates, and $i = 1, N_{\text{titr}}$ is the index for the continuous variables θ_i which is related to the titration coordinate λ_i by

$$\lambda_i = \sin^2(\theta_i) \quad (2)$$

Boundaries are naturally imposed on the titration coordinate through the sine function, where $\lambda_i = 0$ corresponds to the protonated state and $\lambda_i = 1$ corresponds to the deprotonated state. For residues with two competing titration sites, a second continuous variable can be included to allow interconversion between proton tautomeric states. This is indicated in eq 1 by bold $\boldsymbol{\theta}$.

In eq 1 the first term is the kinetic energy of the real system (atoms), U^{int} is the internal potential energy which is independent of titration, and U^{hybr} is a hybrid effective energy term which enables the coupling between conformational and protonation degrees of freedom. In the GB-based CPHMD method, it is written as a sum of the van der Waals, Coulombic, and GB electrostatic solvation free energies:

$$U^{\text{hybr}}(\{\mathbf{r}_a\}, \{\boldsymbol{\theta}_i\}) = U^{\text{vdW}}(\{\mathbf{r}_a\}, \{\boldsymbol{\theta}_i\}) + U^{\text{Coul}}(\{\mathbf{r}_a\}, \{\boldsymbol{\theta}_i\}) + U^{\text{GB}}(\{\mathbf{r}_a\}, \{\boldsymbol{\theta}_i\}) \quad (3)$$

where the latter is given as^{27,28}

$$U^{\text{GB}} = -\frac{1}{2} \sum_{a,b} \left(1 - \frac{e^{-\kappa r_{ab}}}{\epsilon_w} \right) \frac{q_a q_b}{\sqrt{r_{ab}^2 + \alpha_a \alpha_b \exp(-r_{ab}^2 / F \alpha_a \alpha_b)}} \quad (4)$$

Here r_{ab} is the distance between two atoms, q_a and q_b are the respective partial charges, ϵ_w is the dielectric constant for water, and α_a and α_b are the effective Born radii, which can be interpreted as the spherically averaged distance from the atom to the dielectric boundary. To approximately account for the effect of salt screening, a Debye–Hückel term^{29,28} is included in the above equation, where $\kappa^2 = 8\pi q^2 I / \epsilon_k T$, and I is the ionic

strength. The dependence of U^{Coul} and U^{GB} on λ is realized through linear scaling of the partial charges on titratable residues between the protonated and the unprotonated forms. In an analogous fashion, the van der Waals interactions involving titratable hydrogens are also attenuated by the titration coordinates. The last term in eq 1 represents the biasing energy acting on the titration coordinates:

$$U^*(\{\theta_i\}) = \sum_i -U^{\text{mod}}(\theta_i) + U^{\text{pH}}(\theta_i) + U^{\text{barr}}(\theta_i) \quad (5)$$

where U^{mod} is a potential of mean force function for deprotonating a model compound in solution along the titration coordinate, U^{barr} is a harmonic potential which suppresses the residence time of unphysical intermediate values of λ , and U^{pH} provides the additional free energy for the protonation equilibrium due to solution pH:

$$U^{\text{pH}}(\lambda_i) = \log(10)k_b T(\text{p}K_a^{\text{mod}} - \text{pH})\lambda_i \quad (6)$$

where $\text{p}K_a^{\text{mod}}$ is the experimentally determined $\text{p}K_a$ of a model compound in solution.

Continuous Constant pH Molecular Dynamics in Explicit Solvent. The key to CPHMD and other continuous titration methods is to simultaneously derive forces on the spatial and titration coordinates. While it is straightforward to compute forces on spatial coordinates in explicit-solvent simulations, there is inherent difficulty in the latter due to the need for very accurate estimate of the electrostatic desolvation free energy (see U^{GB} in eq 3). In fact, attempts to directly calculate the free energy of charging titratable residues repeatedly during molecular dynamics by considering explicit interactions between solvent molecules and solute have encountered severe convergence problems in the context of both discrete⁷ and continuous constant pH MD methods.^{6,30} Our own tests revealed that the variance in the instantaneous forces on the titration coordinates is up to an order of 100 kcal/mol per λ unit, whereas the forces exerted from the pH biasing energy 1 pH unit away from the model compound $\text{p}K_a$ is only 1.3 kcal/mol per λ unit. Therefore, we decided to use a “mixed-solvent” scheme, where the GB model is used to derive forces on the titration coordinates, while the explicit-solvent model is used to propagate the spatial coordinates. To enable a direct coupling between solvent dynamics and proton titration of solute, we retain the λ -dependent scaling of van der Waals interactions involving titrating hydrogens and solvent molecules. An analogous “mixed solvent” scheme has been developed by Baptista and co-workers and applied in the context of the discrete constant pH MD for protein titration studies.⁸ One important difference is that their scheme does not include a direct (van der Waals) coupling between solvent dynamics and solute titration.

The caveat of the “mixed solvent” scheme is that no formal Hamiltonian exists and that potential artifacts may occur. Since the solvation-related force on titration coordinates is treated in a mean-field manner without explicitly accounting for the electrostatic interactions with nearby water molecules, inadequate or lagged response of solvent to the change in the charge state of the titrating site may occur. We expect this undesirable side effect to be minimal because of the aforementioned van der Waals coupling between solute protonation and solvent dynamics, and because in continuous evolution of titration coordinates, the energy change is small at each time step. Nevertheless, a preventive fix is to increase the time step for λ coordinates

(currently the same as spatial coordinates), thereby allowing relaxation of surrounding solvent molecules. Such a strategy has been demonstrated to be very effective in the discrete constant pH molecular dynamics simulations using the “mixed-solvent” scheme.⁸ Another source for potential artifacts in this and other “mixed-solvent” simulations is related to the fact that the total energy is no longer strictly conserved, which may result in a drift or pronounced fluctuation in temperature and energy of the system. We will examine these potential artifacts later in detail.

pH-Based Replica-Exchange Sampling Protocol. It has been noted previously^{9,13,10} that in constant pH molecular dynamics, the convergence of protonation-state sampling and resulting $\text{p}K_a$ values is slow due to the tight coupling of conformational dynamics and protonation equilibria. To address this issue, the temperature-based replica-exchange (T-REX) protocol³¹ was applied to enhance conformational sampling in the GB-based continuous²⁶ and discrete³² constant pH methods, which has led to significant improvement in the convergence of calculated $\text{p}K_a$ values. A straightforward implementation of the T-REX protocol in explicit-solvent simulations is however not effective because of the large number of replicas needed to account for the solvent degrees of freedom.³³ Recently, Simmerling and co-workers have proposed a mixed-solvent scheme to reduce the number of replicas,³⁴ which may be incorporated into the explicit-solvent CPHMD presented in this work. One issue that was noted,³⁴ and is currently being addressed,¹⁹ is the distorted conformational distribution due to inaccuracy of the underlying implicit-solvent model. To avoid this problem, we decided to enhance the sampling of protonation space directly by making use of a REX protocol based on the pH-biasing energy (eq 6). This protocol is a specific application of the reaction-coordinate replica-exchange method.³⁵

In the pH-based REX protocol, simulations of independent replicas are run at the same temperature but different pH conditions. The exchange of pH conditions between a pair of replicas adjacent in pH is periodically attempted according to the Metropolis criterion, which gives the exchange probability as

$$P = \begin{cases} 1 & \text{if } \Delta \leq 0 \\ \exp(-\Delta) & \text{otherwise,} \end{cases} \quad (7)$$

where Δ represents the change in the total pH-biasing energy defined as

$$\Delta = \beta(U^{\text{pH}}(\{\theta_i\}; \text{pH}') + U^{\text{pH}}(\{\theta'_i\}; \text{pH}) - U^{\text{pH}}(\{\theta_i\}; \text{pH}) - U^{\text{pH}}(\{\theta'_i\}; \text{pH}')) \quad (8)$$

Here β is the inverse temperature, the first two terms are the pH-biasing potential energies (eq 6) for the first and second replicas after the exchange, and the last two terms are the corresponding energies before the exchange.

SIMULATION DETAILS

Model Compounds. As in the previous work,^{13,26} model compounds for Asp, Glu, His, and Lys side chains are single amino acids acetylated at N-terminus (ACE), and N-methylamidated at C-terminus (CT3). The model $\text{p}K_a$ values (used in eq 6) were 4.0, 4.4, and 10.4 for Asp, Glu, and Lys, respectively.³⁶ The model $\text{p}K_a$ of His was taken as 6.6 and 7.0 for the N δ and N ϵ sites, respectively.³⁷ The model compound for the C-terminus attached to phenylalanine (CT-Phe) in HP36 was the acetylated

C-terminal hexapeptide (KEKGLF) from HP36 with a measured pK_a of 3.2.³⁸ The parameters in the potential of mean force function U^{mod} were determined using thermodynamic integration (TI) in explicit solvent.¹³ Parameterization simulations at each combination of λ and α for double-site titratable residues were run for 1 ns. In the TI procedure, the protonation states of other titratable residues in the model peptide for CT-Phe were fixed because their pK_a 's are at least 1 pH unit higher than the C-terminus. Except for CT-Phe, the ionic strength in the GB calculation was set to zero during the TI simulations following the previous protocol.¹³ For CT-Phe the ionic strength was 150 mM in accord with experiment.³⁸

Proteins. Five proteins were studied in this work: the 45-residue binding domain of 2-oxoglutarate dehydrogenase multi-enzyme complex, BBL (PDB: 1W4H), the 36-residue subdomain of villin headpiece, HP36 (PDB: 1VII), the 56-residue N-terminal domain of ribosomal L9 protein, NTL9 (PDB: 1CQU), the 149-residue, of which 129 residues were resolved in the crystal structure, hyperstable variant of staphylococcal nuclease Δ +PHS, SNase (PDB: 3BDC), and the 129-residue hen egg white lysozyme, HEWL (PDB: 2LZT). For all structures, the HBUILD facility of CHARMM³⁹ was used to add hydrogens. Unless otherwise specified, no explicit ions were added in the pH-REX simulation because of the small simulation box and the low ionic strengths used in experiment. See later discussions. The ionic strengths in the GB calculations were set to 200, 150, 100, 100, and 50 mM for BBL, HP36, NTL9, SNASE, and HEWL, respectively, consistent with the experimental conditions.^{38,40–43} Unless otherwise noted, both N- and C-termini of proteins were left in the free, charged form. For SNase, the published crystal structure was missing residues 1–6 and 142–149. To avoid potential errors, the structure was acetylated at N-terminus and amidated at C-terminus. For NTL9, the C-terminus was amidated in accord with experiment.⁴¹

Simulation Protocol. We have implemented the explicit-solvent CPHMD method in a developmental version of CHARMM (c35b3)³⁹ and the pH-REX sampling scheme in the MMTSB tool set.⁴⁴ All of the simulations described in this work were performed with the all-atom CHARMM22/CMAP force field for proteins⁴⁵ and the CHARMM modified version of the TIP3P water model.⁴⁶ The solvation forces on the titration coordinates were calculated using the GBSW implicit-solvent model²⁸ with the refined¹⁸ atomic input radii of Nina et al.⁴⁷ The SHAKE algorithm was applied to all hydrogen bonds and angles to allow a 2 fs time step. Nonbonded electrostatic interactions were calculated using the particle-mesh Ewald summation with a charge correction to reduce pressure and energy artifacts for systems with a net charge.⁴⁸ In the GB calculation, all input parameters were identical to the previous work.²⁶

All simulations were performed under ambient pressure and temperature conditions using the Hoover thermostat⁴⁹ with Langevin piston pressure coupling algorithm.⁵⁰ Proteins and model compounds were built and then placed in a truncated octahedron water box of a specified size such that the distance between the solute and the edges of the box was at least 14 Å. Water molecules within 2.6 Å of any heavy atom of the solute were deleted. Energy minimization was carried out in three stages. First, a harmonic restraint with a force constant of 50 kcal/mol·Å was applied to solute heavy atoms, and the structure was energy minimized with 50 steps of the steepest descent (SD) and 200 steps of the adoptive basis Newton–Raphson (ABNR) methods. Then the force constant was reduced to 25 kcal/mol·Å, and the

same minimization protocol was applied. Finally, the force constant was reduced to 10 kcal/mol·Å, and the structure was energy minimized with 5 SD and 20 ABNR steps.

In the pH-REX simulation of a model compound, three pH replicas, one at the reference pK_a and two at 1 pH unit above and below the reference value, were used. Three independent pH-REX simulations were conducted, where each REX simulation lasted 1.2 ns per replica and the first 200 ps was discarded in the pK_a calculation. For proteins, one pH-REX simulation was performed. In the pH-REX protocol, the pH spacing was 1 pH unit, and the pH range extended at least 1 unit above and below the highest and lowest experimentally determined pK_a value for the protein. Specifically, for BBL the pH range is 2–9, for HP36, NTL9, and SNase it is 0–7, and for HEWL it is 0–9. Each pH replica was subjected to 4 ps of restrained equilibration without pH exchange, where a harmonic potential with the force constant of 10 kcal/mol·Å was applied to all solute heavy atoms. Following equilibration, unrestrained simulation with the pH-REX protocol was performed. The exchange in pH was attempted every 100 dynamic steps or 0.2 ps for model compound and 500 steps or 1 ps for protein simulations. The success rate for exchanges was at least 40%. Protein simulations lasted 2 ns, and the first 0.25 ns was discarded in the analysis and the pK_a calculation. Simulation of HP36 was run for 4 ns in order to observe pK_a behavior at longer simulation times.

Calculation of pK_a Values. To calculate the pK_a of a titratable site, we first recorded the population of protonated ($\lambda < 0.1$, N^{prot}) and unprotonated ($\lambda > 0.9$, N^{unprot}) states from simulations of different pH replicas. The resulting unprotonated fractions S at multiple pH values were then fitted to the following modified Hill equation, in accord with the commonly used model for fitting pH-dependent NMR chemical shifts:⁴²

$$S(\text{pH}) = \frac{s_{A^-} + s_{HA} 10^{n(pK_a - \text{pH})}}{1 + 10^{n(pK_a - \text{pH})}} \quad (9)$$

where n is the Hill coefficient, which represents the slope of the transition region of the titration curve,⁴² s_{A^-} and s_{HA} are fitting parameters, which represent the extrapolated S values at extreme acidic and basic pH conditions for the observed titration event. Equation 9 becomes the Hill equation when protonation or unprotonation is complete in the simulated pH range, e.g., $s_{A^-} = 1$ and $s_{HA} = 0$, which was the case for nearly all residues. Occasionally, for acidic residues with significant negative pK_a shifts, s_{HA} deviated significantly from 0 as a result of incomplete protonation at the lowest pH condition. Finally, to account for the small systematic deviations of calculated pK_a 's of model compounds relative to the reference values, we made the following postcorrections: Asp (+0.2), Glu (+0.3), and His (−0.3) to the pK_a values of proteins.

RESULTS AND DISCUSSION

Stability of Trajectory. Before applying the explicit-solvent CPHMD to titration simulations, it is important to examine potential artifacts due to caveats in the mixed scheme and the change in total net charge. As mentioned earlier, the proposed method does not conserve energy because the protonation states of titratable groups are changed using an implicit description of the electrostatic interactions with solvent, which may lead to drift or increased fluctuation in temperature and energy of the simulated system. Another source for potential artifacts is related

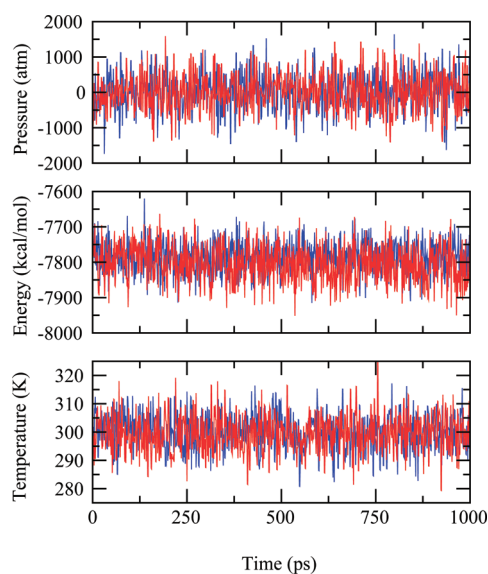


Figure 1. Instantaneous pressure, potential energy, and temperature in the explicit-solvent CPHMD simulation of lysine at pH of 10.4 (blue) and in the simulation of the neutral lysine with fixed protonation state (red).

to the fluctuating net charge on the system during proton titration. In the default implementation of Ewald summation a neutralizing plasma, which is a uniform distribution of a charge equal and opposite to the net charge, is added to the summation to avoid divergence in Coulomb energy for periodic systems.⁵¹ This background plasma has been noted to introduce pressure artifacts for small net-charged systems, which could dramatically affect the dynamics of simulations at constant pressure.⁴⁸ Brooks and co-workers showed that the artifacts are drastically reduced by invoking a charge correction term.⁴⁸ We applied this correction term in all of our simulations.

To assess the extent of the spurious effects, we examined the temperature, pressure, and total potential energy of the system along the trajectory using two protocols. In the first protocol, a blocked lysine was subjected to CPHMD titration at pH of 10.4. In the second protocol, a fixed-charge simulation was conducted using a neutral lysine with an otherwise identical simulation setup. As shown in Figure 1, the time series for temperature, pressure, and potential energy in the CPHMD titration of lysine (with 1:1 protonated and deprotonated states) is virtually indistinguishable from the conventional simulation of neutral lysine with fixed protonation state. The pressure fluctuations are quite large for both systems, but this is expected because of the small size of the simulation box. Also, any energy leaking into or out of the system due to the nonconservative change in protonation state is not readily apparent, as there is no visible drift in the total potential energy for this system. To further verify the stability of pressure, temperature, and potential energy, we performed CPHMD titrations for other model compounds and proteins. No systematic drift or increased fluctuation was observed in any of the three quantities at the simulation time scales (several nanoseconds) for either model compounds or proteins. Thus, we conclude that, with the net charge correction and the Hoover thermostat, potential artifacts in pressure, temperature, and potential energy are negligible.

Response of Explicit Solvent to Titration. Although the van der Waals interactions between titratable hydrogen atoms

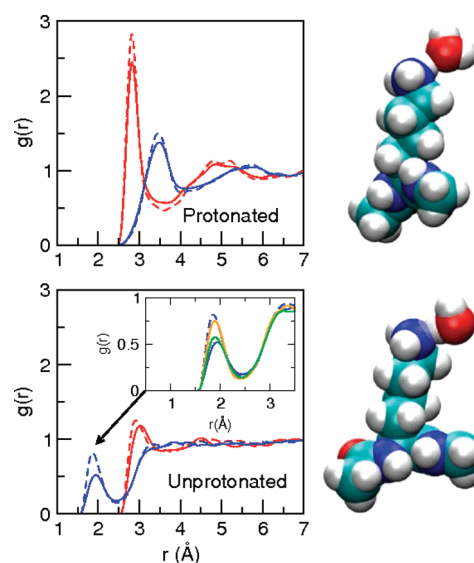


Figure 2. Response of explicit solvent to titration. Radial distribution function for the titratable nitrogen atom of lysine to the hydrogen (blue) or oxygen (red) atom of water. Dashed lines are from the simulation with the fixed protonation state; solid lines are from the CPHMD titration with protonated (charged) and deprotonated (neutral) states coexisting. Snapshots of the interacting water and lysine are shown. The charged lysine donates a hydrogen bond to water (upper), while the neutral lysine accepts a hydrogen bond from water (lower). Simulations with fixed protonation states were run for 1 ns. The CPHMD titration time was 2 ns, and the deprotonated fraction was about 0.5. The inset gives radial distribution functions when a very stringent cutoff ($\lambda > 0.99$) is used to define the deprotonated state (green) and when λ values are updated every 10 MD steps in addition to the stringent cutoff (orange).

and solvent molecules are explicitly described, the lack of explicit treatment of electrostatic interactions may have an undesirable effect such that water molecules cannot adjust quickly to a low-energy position following a change in the titration coordinate, resulting in an unrealistic arrangement of solvent around the titrating site. To examine the response of explicit water molecules to solute titration, we calculated the radial distribution functions (RDF) for the charged (protonated) and neutral (unprotonated) lysine from the (conventional) simulations (one for charged and one for neutral) and compared them with the RDF's from one CPHMD titration simulation. The latter simulation was conducted at a pH condition such that the charged and neutral populations are almost equal. As seen in Figure 2, the positions of maxima and minima in the RDF's of the charged and neutral forms of Lys are identical in the conventional simulations and the CPHMD titration, which demonstrates that the water structure is qualitatively indistinguishable. To further investigate the reorientation of water molecules in response to titration, we took a closer look at the solute–solvent interactions that give rise to the peaks of the RDF's. Interestingly and reassuringly, the relative orientation of lysine and the nearby water is identical in the conventional simulations and the CPHMD titration. Figure 2 also shows the representative snapshots of the charged and neutral lysines interacting with an adjacent water molecule. When lysine is charged, it acts as a hydrogen-bond donor, interacting with the oxygen atom of water. When lysine is neutral, it acts as a hydrogen-bond acceptor, interacting with the hydrogen atom of water.

Despite the remarkable agreement in the positions of maxima and minima of the RDF's, the amplitude of the peaks from the

Table 1. Calculated and Experimental pK_a Values of Model Compounds

residue	calcd ^a	calcd ^b	ref ^c	pace lab ^d
Asp	3.79 ± 0.09	3.77 ± 0.02	4.0	3.67 ± 0.04
Glu	4.09 ± 0.11	4.05 ± 0.01	4.4	4.25 ± 0.05
His	6.89 ± 0.08	6.89 ± 0.01	6.6/7.0	6.54 ± 0.04
Lys	10.21 ± 0.02	10.41 ± 0.02	10.4	10.40 ± 0.08
CT-Phe	3.38 ± 0.06		3.2 ^e	—

^a Results using the standard simulation protocol where the λ value was updated every MD step and the simulation length was 1.2 ns per pH replica. The average pK_a 's obtained by fitting S data from three independent pH-REX titrations are listed along with one-half of the difference between the highest and lowest calculated values. ^b Results from test simulations where the λ value was updated every 10 MD steps and the simulation length was 10 ns per pH replica. ^c Measured pK_a 's based on the blocked single amino acids from Nozaki and Tanford.³⁶ These model pK_a 's were used in the pH biasing energy (eq 6). For His, the listed pK_a 's are the microscopic values for δ and ϵ sites. The resulting macroscopic pK_a is 6.45.¹³ Errors in the measurements are typically ±0.1–0.2.⁵⁷ ^d The most recent data from Pace lab based on potentiometric titrations of alanine pentapeptide Ac-AA-X-AA-NH₂, where X denotes the titrating residue.⁵⁷ ^e Measured pK_a of the C-terminal carboxylic acid in the C-terminal peptide of HP36 (sequence KEKGLF) based on the NMR titration data from Raleigh lab.³⁸

CPHMD titration is reduced as compared to those from conventional simulations. This reduction in the amplitude of RDF can be mainly attributed to the slight lagging in water equilibration following a switch in protonation state and to a lesser extent the cutoff chosen in our definition of protonated and deprotonated states. The inset in Figure 2 shows that with a very stringent cutoff ($\lambda > 0.99$) there is small improvement in the amplitude of the RDF. If we use the stringent cutoff combined with a λ update of every 10 MD steps, the amplitude of the RDF is dramatically increased to nearly superimpose on the result from the simulation with fixed protonation state. If the frequency of switching protonation state is much slower, then the RDF's would exactly match those calculated from the simulations at fixed charge. Baptista and co-workers showed that in the MD simulation, the reorganization time of water following the most dramatic protonation event from the fully neutral to doubly charged state of succinic acid is 1–3 ps.⁸ Considering the average residence time at either protonation state in our simulation was on average about 1 ps and the transition between protonation states is continuous, water molecules have sufficient time to rotate to a favorable position following titration. Nevertheless, the data of lysine titration shows that the update frequency or time step for propagation of titration coordinates (currently set to be the same as the propagation of conformational dynamics) can be increased to ensure the full extent of water relaxation. A potential drawback is the slow down of protonation-state sampling.

Convergence and Accuracy of Model Compound Titrations. Before attempting to perform titration simulations of proteins, it is important to assess the required simulation time to reach converged values for the unprotonated fraction (S) of model compounds as well as the accuracy and the precision of the calculated pK_a 's. We first examine titration simulations conducted at a single pH value. Explicit-solvent CPHMD titration of a blocked lysine was performed at the pH equal to the reference pK_a of 10.4. The S values stabilized at about 5 ns, and there was little change over the remainder of the 10 ns

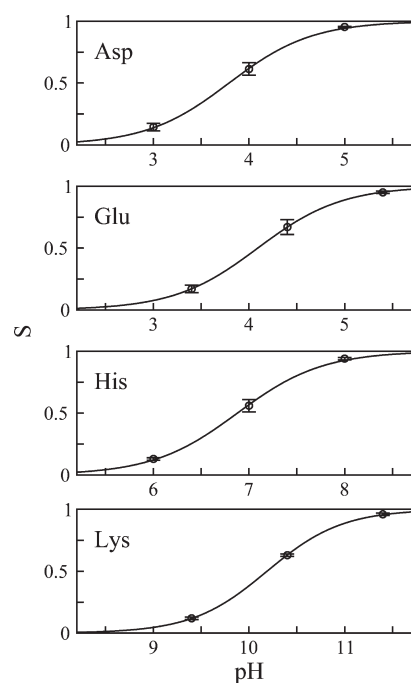


Figure 3. Titration curves for the blocked model compounds from the explicit-solvent CPHMD simulations. Three independent pH-REX simulations were performed. Each REX simulation utilized three pH replicas with each replica running for 1 ns. The average unprotonated fractions S (calculated from the three runs and shown as circles) at three pH values were fit to the Henderson–Hasselbach equation and shown as lines. At each pH, an error bar indicates the range of the calculated S values, which is the largest at the pH closest to the pK_a value. These ranges are 0.10, 0.12, 0.10, and 0.02 for Asp, Glu, His, and Lys, respectively.

simulation. We repeated the simulation twice with different randomly assigned velocities and observed a similar convergence time. Similar results were also found for the blocked Asp, Glu, and His which have two titration sites. The lengthy simulation time (5 ns) required for the convergence of pK_a values for single amino acids indicates the need for accelerated sampling. To directly enhance the protonation-state sampling, we applied the pH-based replica-exchange protocol with three replicas placing at pH values of 9.4, 10.4, and 11.4 in the lysine titration. The S values were converged within 1 ns for all model compounds, demonstrating significant acceleration over the single pH simulation. We summarize these results in Table 1. The uncertainty or random error in the calculated model compound pK_a 's ranges from 0.02 to 0.11, which is similar to the range found in potentiometric and NMR titration experiments (see Table 1). To further assess convergence, we examine the reproducibility of S values and the quality of fitting to the Henderson–Hasselbach equation. In Figure 3 results of three independent pH-REX simulations (1 ns/replica) for Asp, Glu, His, and Lys are shown. The error in the S value ranges from 0.02 to 0.12, and the χ -square value of the fitting is virtually 0. Thus, the above data demonstrate that 1 ns pH-REX titrations offer converged sampling for protonation equilibria.

Next we examine the accuracy of the calculated pK_a 's of model compounds. As compared to the target reference values, the pK_a 's of Asp, Glu and Lys are underestimated by 0.2–0.3 pH units, while that of His is overestimated by 0.3 pH units

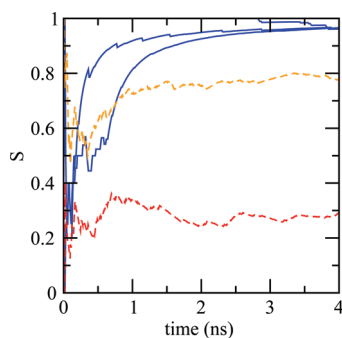


Figure 4. Enhancement of protonation-state and conformational sampling for protein titrations. Cumulative unprotonated fraction of Asp44 of HP36. Data from the pH-REX simulations are shown in red for replica at pH of 2 and orange for replica at pH of 3. Data from three independent single pH simulations at pH 2.3 are shown in blue.

(Table 1). There are two possible sources for the systematic deviations. The first possibility has to do with artifacts in simulations of net charged periodic systems using Ewald potential. Even with the net charge correction, Brooks and co-workers noted that the charged form may be slightly favored in the free energy simulation of a single ion, and this deviation depends on the size of the simulation box.⁴⁸ Our tests however showed that increasing the box size did not affect the pK_a results for model compounds. We further ruled out the net charge related artifact because the same systematic errors, e.g., underestimation of the pK_a 's for Asp and Glu and overestimation of the pK_a for His, were also observed in the GB-based CPHMD simulations.²⁶

The systematic errors in pK_a 's indicate that the deprotonation free energy based on the potential of mean force function, which is determined by the thermodynamic integration (TI) procedure, does not exactly match that in the titration simulation. One possible reason for the discrepancy is the difference in water relaxation because in the TI simulation water has more time to relax at a specific λ value than in the titration simulation. To investigate this issue, we repeated the titrations with slower λ dynamics, updating λ every 10 MD steps. Interestingly, the deviation for the pK_a of Lys is abolished, but the deviation for Asp, Glu and His remained. Examination of the λ and x trajectories revealed that the two degenerate protonation states (doubly deprotonated in the case of Asp or Glu and doubly protonated in the case of His) occasionally experience prolonged residence time. In the absence of extensive analysis and consideration, we suggest that one route for correcting this bias is to make the barrier in the x (tautomeric) dimension a function of λ such that when λ approaches the degenerate protonation state, interconversion becomes increasingly difficult. This is clearly a limitation that needs to be addressed in our future work. Nevertheless, since this bias is present in both model compound and protein titrations, the effect on the calculated pK_a shifts is negligible. To correct for the systematic deviations, we added post corrections for all the calculated pK_a values of proteins (see Simulation Details Section).

Enhanced Sampling of Protonation and Conformational States of Proteins. We have demonstrated that the pH-REX protocol significantly accelerates the pK_a convergence for model compounds. Now we show that the pH-REX protocol significantly enhances sampling in both protonation and conformational space for proteins. Take the titration of HP36 as an example. Figure 4 displays the time series of the unprotonated

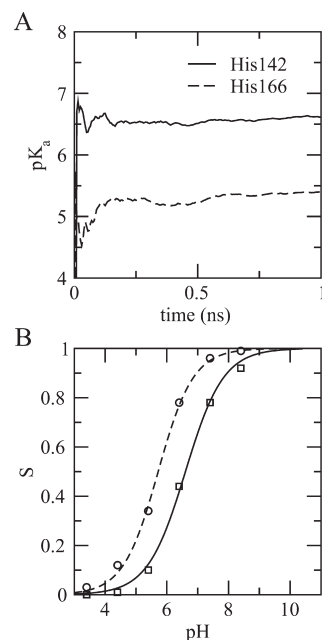


Figure 5. Convergence of protein titrations. (A) Time series of the calculated pK_a 's for BBL from the explicit-solvent CPHMD simulations with pH-REX protocol. The S values at pH of 7 and 6 are used for His142 and His166, respectively. (B) Titration data based on the 1 ns simulation and best fits to the modified HH equation (eq 9).

fraction for Asp44 from one pH-REX simulation and three single pH simulations. In the single pH simulations, Asp44 was trapped in the unprotonated form at pH 2.3, as a result of the persistent salt-bridge interaction with Arg15. In the pH-REX simulation, however, both protonated and unprotonated forms of Asp44 were sampled at pH 2 and 3, because the simulation was able to capture both formation and disruption of the salt bridge. Thus, by making use of the direct coupling between protonation events and conformational dynamics, the pH-REX protocol allows the protein to overcome local energy barriers, while retaining the correct thermodynamic distribution. In this regard, pH-REX has a similar effect as the temperature-based REX protocol, which significantly accelerates the sampling convergence of both protonation and conformational states in the GB-based CPHMD simulations.²⁶

Convergence and Overall Accuracy of Protein Titrations.

In order for titration simulations to be practical, protonation-state sampling needs to converge within a reasonable amount of time. While we have shown that 1 ns of pH-REX titration is sufficient for obtaining converged pK_a 's for model compounds, we also observed that 1 ns titration also yields converged pK_a 's for proteins, despite the fact that the degrees of freedom in a protein system may be orders of magnitude greater as compared to a model compound. This seemingly surprising observation is consistent with data from the GB-based CPHMD simulations^{26,12} and can be attributed to the fact that pK_a 's are mainly determined by local environment. To illustrate the rapid convergence in protein titrations, we monitor the times series of the S value and pK_a as well as the quality of fitting. In Figure 4 we can see that the S values for HP36 stabilize at 1 ns. The small fluctuation after 1 ns does not cause noticeable change in the pK_a value because of the logarithmic relationship between S and pK_a . Figure 5A shows that, after only a few hundred ps, the calculated pK_a 's of the two histidines in BBL become stable and do not change in the remaining simulation

Table 2. Calculated and Experimental pK_a Values in HP36, BBL, and NTL9

residue	explicit solvent ^b	GB	expt ^a
BBL			
His142	6.94 ± 0.06 (6.83)	6.47 ± 0.03	6.47 ± 0.04
His166	5.78 ± 0.04 (5.90)	4.84 ± 0.19	5.39 ± 0.02
HP36			
Asp44	2.66 ± 0.09 (2.77)	3.17 ± 0.11	3.10 ± 0.01
Glu45	3.36 ± 0.31 (3.28)	3.49 ± 0.09	3.95 ± 0.01
Asp46	3.03 ± 0.09 (3.12)	3.51 ± 0.03	3.45 ± 0.12
Glu72	3.50 ± 0.21 (3.45)	3.53 ± 0.10	4.37 ± 0.03
CT-Phe	3.31 ± 0.20 (3.16)	3.16 ± 0.14	3.09 ± 0.01
			3.24 ± 0.12
NTL9			
Asp8	2.83 ± 0.07 (2.80)	3.19 ± 0.20	2.99 ± 0.05
Glu17	3.57 ± 0.14 (3.50)	3.67 ± 0.13	3.57 ± 0.05
Asp23	2.75 ± 0.16 (2.82)	2.11 ± 0.11	3.05 ± 0.04
Glu38	3.38 ± 0.30 (3.40)	3.70 ± 0.19	4.04 ± 0.05
Glu48	3.47 ± 0.17 (3.42)	3.74 ± 0.20	4.21 ± 0.08
Glu54	3.65 ± 0.22 (3.49)	3.64 ± 0.08	4.21 ± 0.08
avg abs dev	0.44 (0.45)	0.36	
rmsd	0.50 (0.52)	0.47	
max abs dev	0.87 (0.92)	0.99	

^a pK_a 's determined by NMR titration for BBL,⁴⁰ HP36,³⁸ and NTL9.⁴¹

^b Values in parentheses were obtained from the 2 ns simulation.

time. This is encouraging given the fact that one of the histidines is buried and as such may require more sampling. Another indication of convergence is the quality of fitting to the HH equation. Figure 5B shows nearly perfect fits ($R^2 > 0.95$) for both residues based on the 1 ns titration data.

To assess the overall accuracy of the explicit-solvent CPHMD method, we performed titration on five test proteins, HP36, BBL, NTL9, SNase, and HEWL and compared the calculated pK_a 's with experiment as well as the GB-based simulations, where the latter used the same pH-REX protocol and salt as well as temperature conditions. The results are presented in Tables 2, 4, and 5 along with the estimates of statistical uncertainty, which were calculated as half of the difference between the pK_a 's calculated from the first and last half of the 750 ps simulation. The total simulation length was 1 ns, and the data from the first 250 ps were discarded. As a validation of convergence, the pK_a 's calculated using 2 ns simulations are also listed. In reference to experimental data, the overall root-mean-squared deviation (rmsd) from the explicit-solvent titrations is 0.74, which is slightly lower than the rmsd from the GB-based titrations (0.82). As a more informative measure of calculation accuracy, linear regressions of the calculated vs measured pK_a shifts are shown in Figure 6 for the explicit-solvent and GB simulations. While the R^2 value and the slope are 0.48 and 0.61, respectively, from the explicit-solvent titrations, they are 0.23 and 0.36 from the GB titrations. Since the correlations are relatively low, we repeated the regression analysis by removing the data points with the four largest absolute pK_a shifts. The R^2 value from the explicit-solvent simulations dropped from 0.48 to 0.25, while R^2 from the GB simulations also dropped dramatically, from 0.23

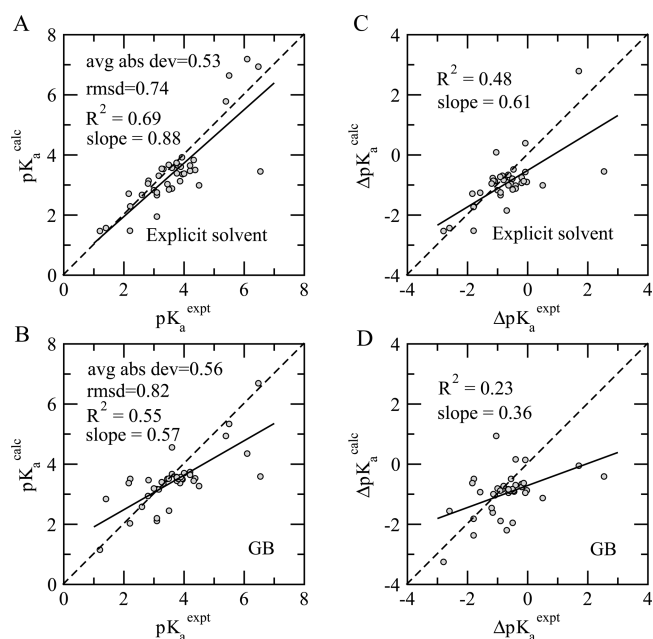


Figure 6. Comparison between calculated and experimental pK_a values and pK_a shifts relative to model values. Calculated pK_a values from the explicit-solvent and GB-based titrations are shown in A and B, respectively. Calculated pK_a shift from explicit-solvent and GB-based titrations is shown in C and D, respectively. Regression line (solid), slope, and R^2 value are shown on each plot as well as $y = x$ line (dashed) to facilitate visual comparison.

to 0.06. Thus, the results show that the improvement due to explicit solvent is robust. Since the data set comprised of mainly acidic residues, the slopes being below 1 suggest that both simulations overestimate the negative pK_a shifts or underestimate the pK_a 's. A close examination of the correlations reveals that the significantly improved agreement with experiment in the explicit-solvent titrations is due to reduction of relatively large errors for several groups. Thus, overall the explicit-solvent simulations offer an increased accuracy for predicting protein pK_a 's. The reasons in specific cases will be delineated next.

Small Proteins BBL, HP36, and NTL9. We first examine the performance of the explicit-solvent CPHMD titrations for three small proteins with 36–56 residues and all α as well as mixed α – β topologies. The results are listed in Table 2 along with the GB titration data. The convergence of both explicit- and implicit-solvent titrations is excellent. The largest difference between the pK_a 's calculated from the first and the last half of the simulation is 0.3 units. Extending the explicit-solvent simulations to 2 ns leads to a maximum pK_a change of 0.15 units and does not improve the agreement with experiment. Overall, the explicit-solvent data are similar to the GB data. The rms as well average absolute and maximum deviations from experiment in the explicit-solvent titration are 0.50, 0.44 and 0.87, respectively, similar to the GB titration. The deviations from experiment arise from the overestimation of the negative pK_a shifts of acidic residues in both explicit- and implicit-solvent titrations.

We examine two cases where the pK_a 's from the explicit-solvent titration are at least 0.6 pH units different from the GB titration. In both cases, the explicit-solvent titration improves agreement with experiment. Asp23 is a residue where the explicit-solvent titration reduces the overestimation of the pK_a downshift of Asp23 from 0.9 to 0.3 units. This is because the salt-bridge

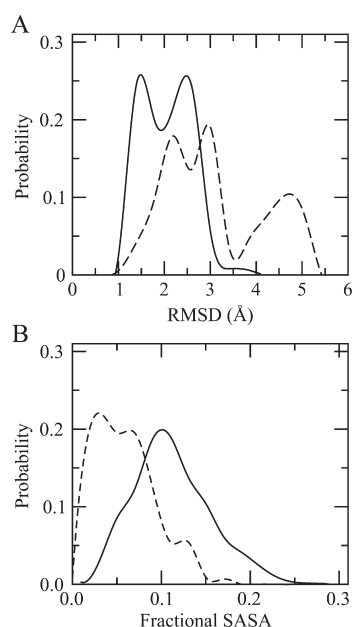


Figure 7. Structural comparison of BBL from explicit-solvent (solid) and GB (dashed) simulations at pH 5. (A) Probability distributions of backbone rmsd. (B) Ratio of the solvent accessible surface area (SASA) of His166 in BBL relative to the solvent-exposed value.

interaction with the nearby amino terminus was overstabilized in the GB simulation, a known problem in GB models.¹³ His166 is the only buried residue in this data set. While being excluded from solvent, it also interacts with three nearby lysines. Thus, both desolvation and electrostatic repulsion destabilize the protonated or charged form of His166, leading to a downward pK_a shift relative to the model value. This is reflected in the experimental pK_a of 5.39, about 1.1 pH units lower than the model value. In the explicit-solvent titration the pK_a shift is underestimated by 0.41 pH units, while it is overestimated in the GB titration by 0.55 pH units. Detailed analysis of the trajectories reveals the major cause of the difference to be structural. Figure 7A shows that in the explicit-solvent simulation, the conformations stayed close to the starting structure with the backbone rmsd centered at 2 Å. In the GB simulation, however, a conformational cluster developed that significantly deviates from the initial structure with the backbone rmsd centered at 4.9 Å. Figure 7B shows that while His166 is slightly exposed to solvent in the explicit-solvent simulation, it is fully enclosed in the GB simulation. Examination of the average distances to the nearby lysines reveals that the Coulomb interactions in both explicit-solvent and GB simulations are similar. Therefore, we suggest that the overestimation of the pK_a shift for His166 in the GB simulation is mainly due to the overestimation of desolvation penalty as a result of exaggerated cloistering of His166. Reduced mobility, especially of buried sites, has been also observed in other GB simulations.²⁴

Although for these small proteins the explicit-solvent pK_a calculations are quite accurate, it is important to further discuss another issue concerning the explicit-solvent CPHMD method. Since the net charge is changing and may become large depending on the protonation state of the protein, we examined the effect of adding an approximate number of counterions to minimize the net charge of the system in all pH conditions. Because of the large number of basic residues of NTL9 and the

Table 3. Effects of Adding Explicit Ions on Calculated pK_a Values in NTL9

residue	calcd ^b	ions ^c	expt ^a
Asp8	2.83 ± 0.07	2.91 ± 0.31	2.99 ± 0.05
Glu17	3.57 ± 0.14	3.38 ± 0.19	3.57 ± 0.05
Asp23	2.75 ± 0.16	2.98 ± 0.16	3.05 ± 0.04
Glu38	3.38 ± 0.30	3.48 ± 0.04	4.04 ± 0.05
Glu48	3.47 ± 0.17	3.42 ± 0.34	4.21 ± 0.08
Glu54	3.65 ± 0.22	3.52 ± 0.25	4.21 ± 0.08
avg abs dev	0.41	0.40	
rmsd	0.48	0.49	
max abs dev	0.73	0.78	

^a pK_a 's determined by NMR titration.⁴¹ ^b Calculated pK_a 's from explicit-solvent titrations without counterions (as listed in Table 2). ^c Calculated pK_a 's from simulations with an identical set up except for the addition of Cl^- ions such that the net charge of the protein at all pH conditions was minimized.

Table 4. Calculated and Experimental pK_a Values in SNase

residue	explicit solvent ^c	GB	expt ^a
Glu10	3.14 ± 0.09 (3.33)	3.47 ± 0.01	2.82 ± 0.07
Asp19	2.29 ± 0.15 (2.49)	3.51 ± 0.02	2.21 ± 0.07 ^b 6.54 ± 0.06
Asp21	3.45 ± 0.28 (3.55)	3.59 ± 0.00	3.01 ± 0.01 6.54 ± 0.02 ^b
Asp40	3.13 ± 0.23 (3.35)	3.37 ± 0.09	3.87 ± 0.09
Glu43	3.83 ± 0.08 (3.76)	3.45 ± 0.00	4.32 ± 0.04
Glu52	3.92 ± 0.01 (3.88)	3.52 ± 0.02	3.93 ± 0.08
Glu57	3.67 ± 0.16 (3.64)	3.52 ± 0.01	3.49 ± 0.09
Glu67	3.66 ± 0.06 (3.67)	3.45 ± 0.06	3.76 ± 0.07
Glu73	3.53 ± 0.11 (3.54)	3.36 ± 0.13	3.31 ± 0.01
Glu75	3.54 ± 0.27 (3.58)	3.40 ± 0.06	3.26 ± 0.05
Asp77	<0.0 (<0.0)	3.14 ± 0.03	<2.2
Asp83	2.54 ± 0.12 (2.84)	3.50 ± 0.04	<2.2
Asp95	2.71 ± 0.57 (2.97)	3.37 ± 0.06	2.16 ± 0.07
Glu101	3.64 ± 0.11 (3.67)	3.51 ± 0.01	3.81 ± 0.10
Glu122	3.61 ± 0.03 (3.75)	3.57 ± 0.01	3.89 ± 0.09
Glu129	3.74 ± 0.11 (3.71)	3.57 ± 0.12	3.75 ± 0.09
Glu135	3.39 ± 0.20 (3.44)	3.56 ± 0.03	3.76 ± 0.08
avg abs dev	0.46 (0.48)	0.63	
rmsd	0.86 (0.85)	0.96	
max abs dev	3.09 (3.00)	2.95	

^a pK_a determined by NMR titration.⁴² ^b The major transition when the experimental data was fit to a two pK_a model. ^c Values in parentheses were obtained from the 2 ns simulation.

resulting net positive charge, NTL9 is an ideal test case to quantify the magnitude of the effect. As shown in Table 3 the calculated pK_a values in the simulations with neutralizing counterions are virtually identical to those where no net charge neutralizing ions were added. Thus, at least for the short simulation time required to obtain converged pK_a values, the data indicate that it is not necessary to include neutralizing ions.

SNase. The calculated pK_a 's for a larger protein, a hyperstable variant of the 149 residue SNase, are summarized in Table 4.

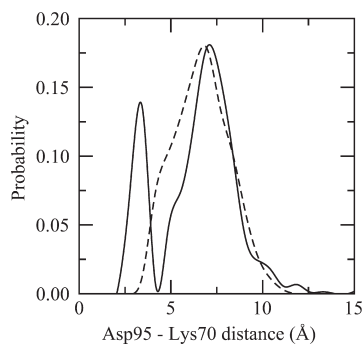


Figure 8. Probability distribution of the minimum distance between the carboxylate oxygens of Asp95 and amino nitrogen of Lys70 of SNase from the explicit-solvent (solid) and GB (dashed) titrations at pH of 3.

SNase is a good test system because the structure-based continuum calculations gave very poor agreement with experiment presumably due to the lack of explicit treatment of protein flexibility.⁴² Overall, the explicit-solvent titration offers a better agreement with experiment. The rms as well as the average absolute and maximum deviations in the explicit-solvent titration are 0.86, 0.46, and 3.09, respectively, while they are 0.96, 0.63, and 2.95 in the GB titration. Extending the explicit-solvent simulations to 2 ns gives results that are very similar.

We first examine Asp95, for which the explicit-solvent titration was able to reduce the overestimation of pK_a from the GB-based titration from 1.21 to 0.55 units. The major reason for the improvement is related to the strength of the interaction with Lys70. In the crystal structure obtained at pH of 8, the minimum distance between the charge centers on Asp95 and Lys70 is 4.7 Å, which suggests a salt-bridge interaction. Figure 8 gives the probability distribution of the minimum distance between the charge centers from the explicit-solvent and the GB simulations. Although the average distance is identical at 6.1 Å, the difference lies in the distribution. The GB simulation sampled a unimodal distribution centered around 7 Å. By contrast, the explicit-solvent simulation sampled two distinct populations, one centered at 2.8 Å, representing the conformations where Asp95 and Lys70 are closely associated, and another one centered at 7.1 Å, representing the conformations where the two side chains are rotated away from each other. The bimodal distribution seen in the explicit-solvent simulation is a direct result of including discrete solvent molecules and reflects a more realistic description of the ion pair interaction. However, the GB simulation neglects solvent granularity and models the ion pair interaction in a mean-field manner, which results in a less tight salt-bridge pairing and an underestimation of the pK_a shift for Asp95.

Another case where the inclusion of explicit solvent resulted in the more accurate pK_a calculation is for Asp77. The experimental measurement provides an upper bound of 2.2 for the pK_a . In the explicit-solvent simulation, the pK_a was calculated to be in the correct range, but in the GB simulation, the pK_a shift was underestimated by at least 1 pH unit. Asp77 is within a hydrogen-bond distance of two backbone amide hydrogens of Asn119 and Thr120, which are located in a loop connecting a β -sheet motif to an α -helix (Figure 9, upper left snapshot). In Figure 9 we monitor the minimum distance between the carboxylate oxygens of Asp77 and the backbone amide hydrogen of Asn119 or Thr120. In the explicit-solvent simulation the distance was stable,

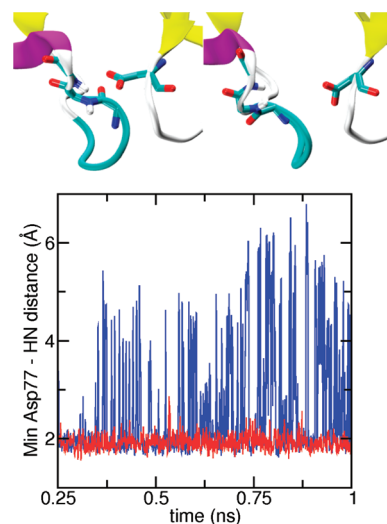


Figure 9. Comparison of the local environment of Asp77 of SNase from the explicit- and implicit-solvent titrations at pH 3. Upper panel: In the initial structure Asp77 forms backbone hydrogen bonds with Asn119 and Thr120 (left snapshot). These interactions were broken in the GB simulation (right snapshot). Lower panel: Time series of the minimum distance between the carboxylate oxygens of Asp77 and the backbone amide hydrogen of Thr119 or Asn120 from the explicit-solvent (red) and GB (blue) simulations at pH of 3.

fluctuating around 2 Å during the entire trajectory, revealing that the backbone hydrogen bonding between Asp77 and Asn119/Thr120 was intact. However, in the GB simulation, this interaction was disrupted as a result of the high mobility of the aforementioned loop (see Figure 9, upper right snapshot). This analysis suggests that the underestimation of the pK_a shift for Asp77 in the GB simulation is due to the distortion of local structure.

The largest pK_a error from the explicit- and implicit-solvent titrations is for Asp21, which interacts with Asp19 on the other end of the β -hairpin. NMR titration data showed two distinct transitions for the two residues.⁴² The major transitions have the pK_a of 2.21, assigned to Asp19, and 6.54, assigned to Asp21.⁴² The latter is the only upward shifted pK_a relative to the model value for SNase. Both the explicit- and implicit-solvent titrations were not able to reproduce the direction of the pK_a shift for Asp21 and underestimated the pK_a by about 3 pH units, although the explicit-solvent simulation was able to differentiate between the two pK_a 's. During the explicit-solvent simulation at pH of 3, the average distance between the carboxylate oxygens of both residues was 3.7 Å. This close proximity was stabilized by a persistent hydrogen bond between the carboxylate oxygen of Asp19 and the backbone amide nitrogen of Asp21. However, the coupled titration behavior with two transitions was not observed when fitting the data for either Asp19 or Asp21. The only indication of coupling was a low Hill coefficient (0.56) for Asp19, which indicates anticooperativity, consistent with experiment.⁴² We also examined the GB titration data. The interaction between Asp19 and Asp21 was very strong but both residues titrated with the same pK_a , and the Hill coefficients were about 1. Thus, compared to the GB titration, the explicit-solvent simulation was able to provide, to some extent, the description of the coupled proton binding events for Asp19 and Asp21. However, the explicit-

Table 5. Calculated and Experimental pK_a Values in HEWL

residue	explicit solvent ^b	GB	expt ^d
Glu7	2.67 ± 0.01 (2.69)	2.58 ± 0.06	2.6 ± 0.2
His15	6.64 ± 0.10 (6.60)	5.34 ± 0.47	5.5 ± 0.2
Asp18	3.05 ± 0.13 (3.15)	2.94 ± 0.01	2.8 ± 0.3
Glu35	7.19 ± 0.15 (6.83)	4.35 ± 0.18	6.1 ± 0.4
Asp48	1.57 ± 0.48 (1.77)	2.84 ± 0.15	1.4 ± 0.2
Asp52	2.88 ± 0.08 (3.21)	4.56 ± 0.02	3.6 ± 0.3
Asp66	1.47 ± 0.60 (0.46)	1.15 ± 0.43	1.2 ± 0.2
Asp87	1.48 ± 0.41 (1.46)	2.03 ± 0.07	2.2 ± 0.1
Asp101	2.99 ± 0.09 (3.06)	3.27 ± 0.32	4.5 ± 0.1
Asp119	2.85 ± 0.05 (2.98)	2.45 ± 0.13	3.5 ± 0.3
CT-Leu	1.95 ± 0.37 (1.89)	2.20 ± 0.14	2.7 ± 0.2
avg abs dev	0.70 (0.70)	0.72	
rmsd	0.84 (0.80)	0.93	
max abs dev	1.50 (1.44)	1.75	

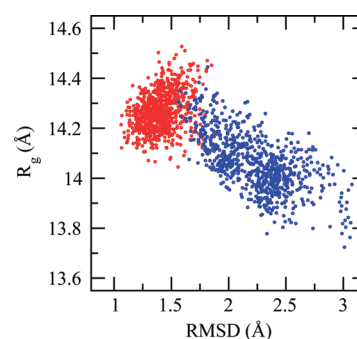
^a Consensus pK_a 's based on NMR titration using multiple nuclei.⁴³

^b Values in parentheses were obtained from the 2 ns simulation.

solvent simulation was not able to fully capture the negative cooperativity, which may be due to insufficient sampling.

HEWL. The last protein we consider is HEWL, which has been used as a standard test system for many pK_a prediction methods.^{52,53} Also, the most recent study of Nielsen and co-workers, where a consensus set of pK_a 's was derived from pH-dependent chemical shifts of different nuclei, makes HEWL the most vetted protein pK_a benchmark system available.⁴³ Table 5 lists the calculated pK_a 's from the explicit- and implicit-solvent titrations. Overall, the calculated pK_a 's from the explicit-solvent titration are closer to experiment than the GB titration. The rms as well as the average absolute and maximum deviations in the explicit-solvent titration are 0.84, 0.70, and 1.50, respectively, while they are 0.93, 0.72, and 1.75, respectively, in the GB titration. Below we examine the cause for the significant differences between the explicit- and implicit-solvent titration data for residues Glu35 and Asp52.

The catalytic residues of HEWL are Glu35 and Asp52 reside at the interface between two domains and have the consensus pK_a 's of 6.1 and 3.6, respectively. The experimental range of pK_a 's calculated from chemical shifts of different nuclei was 6.0–6.8 for Glu35 and 3.4–4.0 for Asp52.⁴³ The pK_a 's from the explicit-solvent simulation are 7.19 and 2.88, while those from the GB simulation are 4.35 and 4.56, respectively. Thus, considering the model values of 4.4 and 4.0 for Glu and Asp, the calculated pK_a shifts are in the correct direction in the explicit-solvent simulation but wrong in the GB simulation. Since the optimum pH for the activity of HEWL is around 5,⁵⁴ the pK_a calculation using the explicit-solvent CPHMD method is able to offer the correct protonation or charge states for the catalytic residues, which is not the case with the GB-based method. We note that the previous GB-based CPHMD simulations with the temperature-based replica-exchange protocol gave a correct direction of the pK_a shift for Glu35.²⁶ We examined the trajectory to delineate the cause for the significantly different pK_a 's. In the GB simulation, there is a significant rearrangement of the native structure. We plot the radius of gyration vs the heavy atom rmsd using the explicit- and implicit-solvent simulation data (Figure 10). The conformations in explicit solvent have rmsd

**Figure 10.** Comparison of the conformational states of HEWL sampled in the explicit-solvent (red) and GB (blue) simulations at pH 6.

values, with respect to the crystal structure, ranging from 1.1 and 1.6 Å, and R_g values ranging from 14.1 to 14.4 Å. However, the conformations in the GB simulation have much larger rmsd (1.6–2.8 Å) and much smaller R_g (13.8–14.2 Å), which suggests a significant compaction and global deviation from the crystal structure. This global rearrangement of structure is propagated to the local conformational environment around the active-site residues, which can be seen from the differences in the solvent exposure of side chains. At pH 6, Glu35 has a SASA of 18.9 Å² in the explicit-solvent simulation, which is similar to the value of 10 Å² based on the crystal structure but much smaller than the value of 38.9 Å² from the GB simulation. The significant increase in solvent exposure for Glu35 in the GB simulation leads to an overestimation of the self-solvation energy of Glu35 and thus an underestimation of the upward pK_a shift. For Asp52 the story is exactly reversed. The solvent exposure of Asp52 is underestimated in the GB as compared to the explicit-solvent simulation. At pH 4, the average SASA of Asp52 is 2.4 Å² in the GB simulation, whereas it is 25.4 Å² in the explicit-solvent simulation, which is much closer to the initial value of 26.6 Å². Therefore, the self-solvation energy of Asp52 is underestimated in the GB simulation leading to a calculated pK_a value that is too high. Thus, HEWL is a clear case where elimination of the “conformational” error introduced by GB can dramatically improve the accuracy of pK_a calculations for residues of biological significance.

DISCUSSION

Like temperature and pressure, solution pH is another important experimental condition that needs to be taken into account in molecular simulations in order to accurately capture physical reality. Motivated by the recent success of the GB implicit-solvent based CPHMD method in the accurate pK_a predictions and mechanistic studies of pH-dependent conformational dynamics of proteins, we have developed a robust approach to extend the CPHMD framework to explicit-solvent molecular dynamics simulations. In this approach, the explicit-solvent force field is used to drive conformational dynamics, while the GB model is used to efficiently estimate the role of solvent in modulating the cost of electrostatic free energy for protonation/deprotonation. The resulting explicit-solvent CPHMD method offers an increased accuracy and a wider applicability as compared to the GB-based CPHMD method while retaining the efficiency and the robustness of the capability for proton titration. To overcome a critical hurdle related to the slow convergence of pK_a calculations, which has plagued

CPHMD and other constant pH methodologies, we have implemented a replica-exchange protocol based on the pH-biasing energy to directly accelerate protonation-state sampling. Remarkably, due to the tight coupling between titration and conformational degrees of freedom, this protocol also led to significant enhancement in conformational sampling, allowing pK_a 's to converge within 1 ns for small model compounds and large proteins. The random errors in the calculated pK_a 's for model compounds were about or below 0.1 pH units.

To benchmark the accuracy of the explicit-solvent based CPHMD method, we have calculated pK_a 's for five proteins and compared with results from the GB-based method and experiment. We found that the explicit-solvent titrations resulted in an average absolute error of 0.53 and rmsd of 0.74, on par with those from the GB-based titrations. However, by bringing the outliers closer to experimental values, the explicit-solvent method offers significantly improved correlation with experiment as compared to the GB-based method. Detailed analysis revealed that this improvement is due to more accurate conformational sampling in explicit solvent. For example, the explicit-solvent simulation preserved the structural integrity of the loop region, bringing the calculated pK_a of Asp77 from SNase closer to experiment. Compaction of HEWL in the implicit-solvent simulation caused distortion of the active site and large deviations in the calculated pK_a values for Glu35 and Asp52, while explicit-solvent simulation preserved the native conformation leading to a correct prediction of the protonation states at the optimum pH value for catalytic activity. Including solvent granularity enabled a more realistic description of ion-pair interactions, as was the case for Asp95 of SNase, where the explicit-solvent simulation gave a bimodal distribution representing both the close-range and solvent-separated interactions with Lys70, which resulted in a more accurate estimate of pK_a . Finally, in the explicit-solvent simulation the hydrophobic cluster in BBL showed an increased mobility relative to the GB simulation, allowing His166 to be partially exposed to solvent, which resulted in a reduction in the pK_a shift due to desolvation penalty. The latter aspect is somewhat surprising, but is compatible with previous GB simulation studies revealing overly rigid hydrophobic assemblies.^{23,24} It is also consistent with the experimental evidence⁵⁵ and previous simulation study⁵⁶ suggesting water penetration into the hydrophobic core of SNase. Although in the presented cases, the differences between the explicit-solvent and GB-based pK_a results are small (all within 0.5 pH units), our unpublished data shows that the explicit-solvent method offers improvement as high as 4 pH units for the worst prediction cases in the engineered mutants of SNase (Wallace and Shen, unpublished data).

While the results demonstrated in this work are encouraging, we note that several potential issues merit attention. First, a potential delay in the response of solvent reorganization to protonation/deprotonation may lead to unfavorable interactions or inaccuracy in the solvation energetics of the titration site. This problem can be effectively avoided by allowing a few additional dynamics steps between titration updates to allow relaxation of solvent around the titrating site, as has been demonstrated in the discrete constant pH techniques.¹¹ Also, we identified a small bias toward the charged form in the titration of Asp, Glu and His residues due to the occasionally prolonged residence time of the two degenerate protonation states (doubly deprotonated in the case of Asp or Glu and doubly protonated in the case of His). Although the effect of this systematic error on the calculated pK_a

shifts is likely minimal, it is clearly a limitation that needs to be addressed in the future. Finally, the accuracy of pK_a calculations is still limited by the accuracy of the GB model to determine the deprotonation free energy. The largest deviation and the single outlier found in this work is Asp21 in SNase, where both explicit- and implicit-solvent simulations were not able to reproduce the direction of the positive pK_a shift, and underestimated the pK_a by 3 pH units. NMR data showed that the titration of Asp21 is coupled to that of Asp19, which has a negative pK_a shift. Although the explicit-solvent simulation was able to differentiate between the two pK_a 's, it could not quantitatively reproduce the extent of the negative cooperativity in proton binding. One possible cause is that more exhaustive sampling may be required to fully capture coupled titration events. This issue deserves further investigation in our future studies. Another aspect that deserves further investigation is related to the effect due to ions. In the current work and previous GB-based CPHMD studies, an approximated Debye–Hückel model is applied in the GB electrostatic calculation to account for the bulk effect of salt screening, which may not be accurate for highly charged systems such as nucleic acids where local charge density can be very high. Finally, in order to apply the explicit-solvent CPHMD to studies of large-scale conformational changes, it may become necessary to combine with a method for global enhancement of conformational sampling such as the temperature-based replica-exchange scheme. Despite these remaining limitations, the current accuracy and precision of the explicit-solvent based CPHMD technique are encouraging, considering the fact that experimentally determined pK_a 's can deviate by 0.5–1 pH units depending on the nuclei monitored.⁴³ Thus, we anticipate that explicit-solvent CPHMD simulations will emerge as a practical tool for gaining novel insights into protonation-related phenomena that are ubiquitous in biology and chemistry. Examples include the mechanism of proton channels, drug-efflux pumps, pH-dependent catalytic reactions of ribozymes, as well as titration behavior of mixed micelle systems.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jana.k.shen@ou.edu. Telephone: (405) 325-0458.

ACKNOWLEDGMENT

Financial support provided by University of Oklahoma and the American Chemical Society Petroleum Research Fund.

REFERENCES

- (1) Warshel, A. *Biochemistry* **1981**, *20*, 3167–3177.
- (2) Sheinerman, F. B.; Norel, R.; Honig, B. *Curr. Opin. Struct. Biol.* **2000**, *10*, 153–159.
- (3) Warshel, A. *Acc. Chem. Res.* **1981**, *14*, 284–290.
- (4) Nielsen, J. E.; Mccammon, J. A. *Protein Sci.* **2003**, *12*, 1894–1901.
- (5) Shen, J. K. *J. Am. Chem. Soc.* **2010**, *132*, 7258–7259.
- (6) Börjesson, U.; Hünenberger, P. H. *J. Chem. Phys.* **2001**, *114*, 9706–9719.
- (7) Bürgi, R.; Kollman, P. A.; van Gunsteren, W. F. *Proteins* **2002**, *47*, 469–480.
- (8) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J. Chem. Phys.* **2002**, *117*, 4184–4200.
- (9) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III. *Proteins* **2004**, *56*, 738–752.

- (10) Mongan, J.; Case, D. A.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 2038–2048.
- (11) Machuqueiro, M.; Baptista, A. M. *Proteins* **2008**, *72*, 289–298.
- (12) Wallace, J. A.; Shen, J. K. *Methods Enzymol.* **2009**, *466*, 455–475.
- (13) Khandogin, J.; Brooks, C. L., III. *Biophys. J.* **2005**, *89*, 141–157.
- (14) Kong, X.; Brooks, C. L., III. *J. Chem. Phys.* **1996**, *105*, 2414–2423.
- (15) Khandogin, J.; Chen, J.; Brooks, C. L., III. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18546–18550.
- (16) Khandogin, J.; Brooks, C. L., III. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 16880–16885.
- (17) Khandogin, J.; Raleigh, D. P.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2007**, *129*, 3056–3057.
- (18) Chen, J.; Im, W.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2006**, *128*, 3728–3736.
- (19) Okur, A.; Wickstrom, L.; Simmerling, C. *J. Chem. Theory Comput.* **2008**, *4*, 488–498.
- (20) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 115–127.
- (21) Chen, J.; Brooks, C. L., III. *Phys. Chem. Chem. Phys.* **2008**, *10*, 471–481.
- (22) Wallace, J. A.; Shen, J. K. *Biochemistry* **2010**, *49*, 5290–5298.
- (23) Voelz, V. A.; Singh, V. R.; Wedemeyer, W. J.; Lapidus, L. J.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 4702–4709.
- (24) Shen, J. K. *Biophys. J.* **2010**, *99*, 924–932.
- (25) Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.
- (26) Khandogin, J.; Brooks, C. L., III. *Biochemistry* **2006**, *45*, 9363–9373.
- (27) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1348–1356.
- (28) Im, W.; Lee, M. S.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1691–1702.
- (29) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. *Theor. Chem. Acc.* **1999**, *101*, 426–434.
- (30) Börjesson, U.; Hünenberger, P. H. *J. Phys. Chem. B* **2004**, *108*, 13551–13559.
- (31) Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E* **2007**, *75*, 026109.
- (32) Meng, Y.; Roitberg, A. E. *J. Chem. Theory Comput.* **2010**, *6*, 1401–1412.
- (33) Nadler, W.; Hansmann, U. H. E. *J. Phys. Chem. B* **2008**, *112*, 10386–10387.
- (34) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 420–433.
- (35) Okamoto, Y. *J. Mol. Graphics Modell.* **2004**, *22*, 425–439.
- (36) Nozaki, Y.; Tanford, C. *Methods Enzymol.* **1967**, *11*, 715–734.
- (37) Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E. *Biochemistry* **1993**, *32*, 8045–8056.
- (38) Bi, Y. *Studies of the folding and stability of the villin headpiece subdomain*, Ph.D. Thesis, Stony Brook University: Stony Brook, NY, 2008.
- (39) Brooks, B. R.; et al. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
- (40) Arbely, E.; Rutherford, T. J.; Sharpe, T. D.; Ferguson, N.; Fersht, A. R. *J. Mol. Biol.* **2009**, *387*, 986–992.
- (41) Kuhlman, B.; Luisi, D. L.; Young, P.; Raleigh, D. P. *Biochemistry* **1999**, *38*, 4896–4903.
- (42) Castañeda, C. A.; Fitch, C. A.; Majumdar, A.; Khangulov, V.; Schlessman, J. L.; García-Moreno, E., B. *Proteins* **2009**, *77*, 570–588.
- (43) Webb, H.; Tynan-Connolly, B. M.; Lee, G. M.; Farrell, D.; O'Meara, F.; Søndergaard, C. R.; Teilum, K.; Hewage, C.; McIntosh, L. P.; Nielsen, J. E. *Proteins* **2011**, *79*, 685–702.
- (44) Feig, M.; Karanicolas, J.; Brooks, C. L., III. *J. Mol. Graphics Modell.* **2004**, *22*, 377–395.
- (45) Mackerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1400–1415.
- (46) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (47) Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239–5248.
- (48) Bogusz, S.; Cheatham, T. E., III; Brooks, B. R. *J. Chem. Phys.* **1998**, *108*, 7070–7084.
- (49) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.
- (50) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613–4621.
- (51) Hummer, G.; Pratt, L. R.; García, A. E. *J. Phys. Chem.* **1996**, *100*, 1206–1215.
- (52) Mongan, J.; Case, D. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 157–163.
- (53) Khandogin, J. Modeling protonation equilibria in biological macromolecules. In *Multi-scale quantum models for biocatalysis* York, D. M., Lee, T.-S., Eds.; Springer: New York, 2009; Chapter 10, pages 261–284.
- (54) Bartik, K.; Redfield, C.; Dobson, C. M. *Biophys. J.* **1994**, *66*, 1180–1184.
- (55) Denisov, V. P.; Schlessman, J. L.; García-Moreno, E., B.; Halle, B. *Biophys. J.* **2004**, *87*, 3982–3994.
- (56) Damjanović, A.; García-Moreno, B.; Lattman, E. E.; García, A. E. *Proteins* **2005**, *60*, 433–449.
- (57) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. *Protein Sci.* **2006**, *15*, 1214–1218.

Microsolvation Effect on the Twist of β -sheets

Joel Ireta*

Departamento de Química, División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana-Iztapalapa, A.P. 55-534, México D. F. 09340

ABSTRACT: Under certain circumstances β -sheets prefer to be twisted instead of flat. To get insight into the reasons of such preference, bare and microsolvated parallel and antiparallel two-strand polyalanine β -sheets are investigated using density functional theory. Full geometry optimizations show that microsolvation increases interstrand twisting and promotes a flat to twist transition. It is found that the latter behavior is connected to compressive strain resulting from microsolvation. Residues in flat β -sheets adjust the sense of its local intrastrand twist, which leads to the appearance of interstrand twist, to release strain and to favor water–water hydrogen bonding. The predicted microsolvation effect is corroborated analyzing the geometry of residues forming β -sheet motifs in protein crystals.

1. INTRODUCTION

β -sheets are characteristic structural motifs of proteins and polypeptides. Such motifs consist of two or more parallel or antiparallel hydrogen-bonded peptide strands.¹ There is a renewed interest in studying β -sheets since it was discovered that its aggregation plays a central role in the mechanism underlying several diseases, including Alzheimer's, Huntington's, Parkinson's, and the prion encephalopathies.^{2–6} A large number of theoretical studies have been devoted to investigate formation, stability, and geometric peculiarities of β -sheets.^{7–17} Nevertheless basic questions regarding their structure are still open. For example, polypeptides, like polyalanine, segments from proteins that form amyloid-like fibrils, fibrous proteins, like silk and β -keratin, crystallize as flat β -sheets, i.e., the chain axes of the strands forming the sheet lay on the same plane. β -sheet motifs in globular protein crystals, however, show large and systematic structural deviations with respect to the classical (flat) geometry.^{18–20} They show an interstrand right-handed twist coupled to an intrastrand left-handed twist,¹⁸ i.e., chain axes are not coplanar, and strand peptide units are converted into the next along the backbone by a rotation that systematically differs from 180°. There is no consensus on the causes that originate the twist of β -sheets. It has been proposed that the twist is either intrinsic to isolated strands or induced by interchain interactions. The emergence of the twist as an intrinsic feature of isolate β -strands has been suggested to be connected to entropic factors that lower the free energy of left-hand twisted chains with respect to straight or right-handed ones,¹⁸ to the tendency of the backbone C–C single bond to eclipse the lone pair of the backbone N atom,¹² to the electrostatic attraction between the carbonyl carbon of one peptide unit and the carbonyl oxygen of an adjacent peptide unit,⁷ and to intrastrand steric hindrances between the oxygen of the carbonyl group and the side chain.^{20,8} The emergence of the twist as a consequence of interstrand interactions has been attributed to constraints imposed by hydrogen-bond (hb) formation,⁹ to intra- and interstrand interactions involving side groups¹⁰ and to interstrand electrostatic interactions.¹¹ Recently it was found that segments from proteins forming amyloid-like fibrils crystallize as flat sheets.^{2,3} Studies using molecular dynamics simulations and

empirical force fields, however, reported that these systems tend to twist in solution,^{21,22} suggesting that β -sheets twist as a response to external conditions and that β -sheets adopt both flat and twisted conformations. Theoretical studies considering the full role of the self-consistent electronic structure and a full optimization of the geometry have found that both twisted and flat conformations are stable structures and that the energy related to twisting is negligible.¹³ Other study based on electronic structure calculations found that breaking side chain–backbone hbs promotes the formation of a twisted β -sheet.²³ In this work the origin of twisting is investigated by constructing atomistic models for flat and twisted two-strand parallel and antiparallel polyalanine β -sheets. Their geometries were fully optimized using density functional theory (DFT). The influence of hydration on twisting is investigated microsolvating the models with explicit water molecules and fully optimizing their geometries. Moreover, hydration effect on twisting is also investigated analyzing preferred conformations of residues in contact with water molecules in β -sheet motifs in protein crystal structures.

As detailed below, it was found that both twisted and flat structures are intrinsic conformations of two-strand antiparallel and parallel polyalanine β -sheets in vacuum. It is shown that strands in all the structures are under compressive strain, owing primarily to interstrand hydrogen bonding. The backbone response to strain provokes local intrastrand twisting along the strands. Nevertheless flat β -sheets in vacuum did not develop noticeable interstrand twisting because residues locally twist in both senses. It was also found that flat β -sheets transform to twisted ones upon hydration. The latter behavior is connected to further compressive strain resulting from microsolvation. Residues in flat β -sheets adjust the sense of its local intrastrand twist, which leads to the appearance of interstrand twist, to release strain and to favor water–water hydrogen bonding. Furthermore comparing average geometries of residues (except glycine and proline) forming β -sheet motifs in protein crystals, it is corroborated the predicted effect of hydration on the twist of β -sheets.

Received: March 29, 2011

Published: June 30, 2011

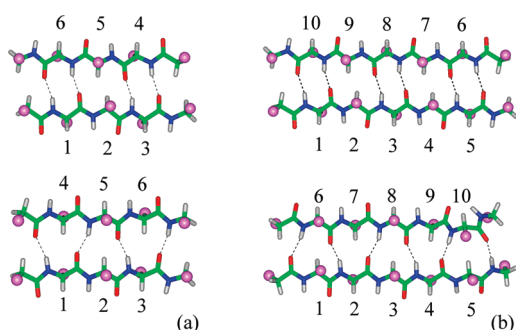


Figure 1. (a) Short polyaniline β -sheet models constructed with three residues per strand: upper structure, antiparallel β -sheet model and lower structure, parallel β -sheet model. Residue labels go from 1 to 6. (b) Large polyaniline β -sheet models constructed with five residues per strand: upper structure, antiparallel β -sheet model and lower structure, parallel β -sheet model. Residue labels go from 1 to 10. Nitrogen atoms are in blue, carbon atoms in green, oxygen atoms in red, and hydrogen atoms in light gray. For clarity methyl groups are represented as pale magenta balls. Dotted lines indicate hbs.

The latter result also suggests that the peptide aminoacid composition (excluding glycine and proline) does not determine if a β -sheet adopts a flat or a twisted conformation.

2. METHOD

Full geometry optimizations were carried out using DFT together with the generalized gradient approximation of Perdew, Burke, and Ernzerhof (PBE) for the exchange–correlation functional.²⁴ Conformations of the structures here investigated are primarily dictated by the formation of hbs, and such hbs are described by DFT using PBE (thereafter called DFT-PBE) within an error of 1 kcal/mol with respect to Møller–Plesset second-order perturbation theory (MP2) results.²⁵ It has been shown that geometric parameters obtained with DFT-PBE for an extensive set of hydrogen-bonded dimers, including dimers formed with peptide-like molecules, are in good agreement with those obtained with MP2, except for cases in which hbs in the dimer are highly bent.²⁵ In the full optimized β -sheet models intrastrand hbs are close to linearity, the largest deviation found for the N–H \cdots O angle is of 25° with respect to linearity (i.e., respect to 180°). This gives us confidence on the DFT-PBE geometries obtained here. DFT-PBE does not describe van der Waals interactions properly. However it is expected minimal alteration of the energy trends obtained here as the spatial distribution of atoms around a given one in flat and twisted conformations is very similar. Troullier–Martins pseudopotentials,^{26,27} plane-wave basis sets, and periodic boundary conditions are used for solving the DFT Kohn–Sham equations as it is incorporated in the abinit code.^{28,29} Here it is worth noting that the use of plane waves does not introduce a basis set superposition error.²⁵ Eight structural models consisting of two strands were constructed for investigating favorable β -sheet conformations. Four models were constructed in a parallel β -sheet conformation and four in an antiparallel β -sheet conformation. These models differ in their number of residues per strand. The thereafter called short models (Figure 1a) were built with three residues per strand, and those built with five residues per strand are called large models (Figure 1b). Structures called flat were built as flat β -sheets, and structures called twisted were built as twisted β -sheets. Antiparallel short β -sheets

are labeled as structures I (flat) and II (twisted). Parallel short β -sheets are labeled as structures III (flat) and IV (twisted). Antiparallel large β -sheets are labeled as structures V (flat) and VI (twisted). Finally parallel large β -sheets are labeled as structures VII (flat) and VIII (twisted). Strands were capped with the CH₃CH₂CO– group in the N terminus and with the CH₃CH₂NH– group in the C terminus. As periodic boundary conditions were used for solving the DFT Kohn–Sham equations, the calculations were carried out using a supercell sufficiently large enough to ensure that interactions with the periodic images were negligible. Here an orthorhombic supercell was used to fully optimize the geometry. For all the calculations the energy cutoff of the plane wave basis set was 70 Ry, and the sampling of the Brillouin zone was replaced by the Γ -point.

Conformations of residues are given here in terms of local cylindrical coordinates instead of the standard ϕ and ψ coordinates. The latter is found convenient because cylindrical coordinates give a direct measurement of intrastrand twisting, while using dihedral angles only an approximate estimation can be done for twisting. However, the procedure to determine cylindrical coordinates is rather involved, and it is explained next. The relative position of the i -th residue respect to its neighbors along the polypeptide chain is given in terms of a certain rotation θ_i (the twist) and a certain translation L_i (the rise) along a local chain axis z_i . These parameters define a screw symmetry operation that superimposes the coordinates of the peptide bonds flanking the α -carbon of the i -th residue. For left-handed rotations θ is defined to be $\theta > 180^\circ$. Here is more convenient to report the degree of local intrastrand twist as $\tau_i = \theta_i - 180^\circ$, thus $\tau > 0$ for left-handed intrastrand twist. The local interstrand twist, T_{ij} , is calculated as the smallest angle formed between the local chain axis vectors z_i and z_j . These vectors are connected to the i - and j -th interstrand hydrogen-bonded residues in the β -sheet. For right-handed interstrand twist T_{ij} is defined to be $T_{ij} > 0$. The quaternion-based superposition fit method^{30–33} is used to determine the cylindrical coordinates θ_i , L_i , and z_i for each residue. In this method θ and L are such that minimize the root-mean-square deviation of distances between the set of superimposed coordinates (for further details see, e.g., ref 33). It has been shown that the use of cylindrical coordinates is advantageous for mapping the conformational space of residues in finite and infinite polypeptides as well as in protein structures.^{33–36}

For building the initial structure of the strands forming the β -sheet models, the equilibrium geometry of a residue in an infinitely long polyaniline chain is used. For flat sheets, the corresponding geometry of a residue in a fully extended structure (FES, with $\phi = -159.7^\circ$ and $\psi = 164.4^\circ$), in which $L = 3.57 \text{ \AA}$ and $\theta = 180^\circ$ ($\tau = 0^\circ$) is used.³⁵ As compressed polyaniline FES develop a left-handed twist (see Figure 3b in ref 33), for twisted sheets the geometry used for each residue corresponds to that in a compressed FES (with $\phi = -120.4^\circ$ and $\psi = 150.4^\circ$), in which $L = 3.39 \text{ \AA}$ and $\theta = 196.4^\circ$ ($\tau = 16.4^\circ$). This local left-handed intrastrand twist forces to build the β -sheet with an interstrand right-handed twist to keep reasonable interstrand hb distances. To estimate the influence of hydration on the relative stability and the structure of twisted and flat conformations, the short-antiparallel β -sheet models are microsolvated with explicit water molecules. Hydrated β -sheet models were built using 8, 16, and 20 water molecules, half of them for microsolvating one strand and the rest for microsolvating the other one. Certainly there are many ways water molecules self-arrange around peptide strands, for example, forming water clusters at certain positions along the

Table 1. Local Geometry Parameters (L_i , τ_i) for Residues in Short β -Sheet Structures I–IV

residue label	I		II		III		IV	
	L	τ	L	τ	L	τ	L	τ
1	3.53	8.9	3.47	24.3	3.47	2.4	3.47	1.1
2	3.52	2.2	3.49	8.3	3.45	0.2	3.42	19.1
3	3.52	-7.9	3.43	-7.2	3.41	4.9	3.42	9.6
4	3.51	11.2	3.46	23.9	3.46	9.5	3.37	36.4
5	3.51	3.6	3.50	8.7	3.44	1.6	3.43	-0.4
6	3.50	-8.1	3.45	-5.6	3.43	-3.6	3.35	5.0

strands, and some of them may be lower in energy than the arrangements investigated here. A full exploration of the conformational space of such microsolvated systems is out of the scope of this work. Here, water molecules were adsorbed such that they formed hbs with the $-N-H$ amide groups and the $-C=O$ carbonyl groups that were not interstrand hydrogen bonded. The rest of water molecules, if any, were absorbed such that an uninterrupted network of hbs was formed along the water molecules. The purpose of the latter is to investigate the effect of water–water hb cooperativity on the twist of the β -sheets. It is expected that hydrated models constructed in such way resemble better a fully solvated peptide. Full geometry optimizations of each of the microsolvated models were carried out using DFT-PBE.

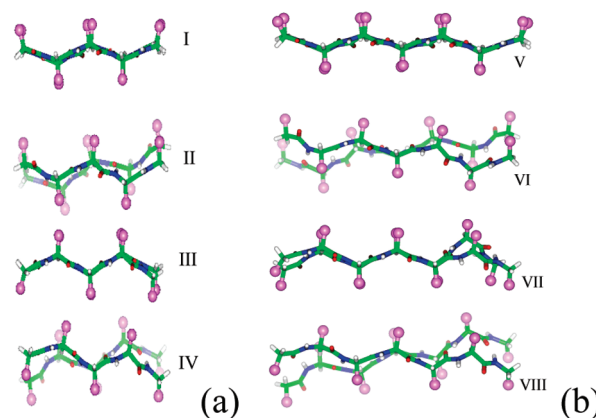
Preferred conformations that residues adopt in β -sheet motifs in proteins are investigated by analyzing an extended set of protein structures resolved by X-ray crystallography and obtained from the protein data bank (PDB).³⁷ The set of 1591 structures was formed with protein structures determined at 1.5 Å resolution or better and a sequence identity of less than 50%. It was corroborated that the result of the analysis is independent from the X-ray resolution and the percentage of sequence identity. The set of analyzed residues does not contain glycines and prolines, as the conformational space of these two residues is known to be markedly different from others, owing to the lack of a side group in glycine and the constraints imposed by the ring formation in the proline backbone. Residues flanking glycines and prolines were also excluded from the analyzed set. A residue was considered to be part of a β -sheet motif if it was in an extended conformation, i.e., if $L > 2.8$ Å³³ and if peptide bonds flanking its α -carbon were hydrogen bonded to peptide bonds flanking the α -carbon of a residue also in an extended conformation and separated at least three positions along the chain. Two peptide bonds were considered to be hydrogen bonded if the distance between the O atom (hb acceptor atom) and the N atom (hb donor atom) was smaller than 3.3 Å and the angle $C=O \cdots N$ bigger than 120°. It was corroborated that the main characteristics of the distributions for L , τ , and T thus obtained were not dependent on the latter two parameters. A residue was taken into account for the analysis only if it belonged to a part of the structure where at least three consecutive residues were found in a β -sheet conformation. In this way ending effects and β -bridges were avoided.

3. RESULTS

Let me first discuss the fully optimized geometries for the bare flat and twisted short β -sheet models. The values for the local

Table 2. T_{ij} Values for Residues in Short β -Sheet Structures I–IV

residue labels	I	II	residue labels	III	IV
1, 6	-6.5	13.8	1, 4	-5.7	21.3
2, 5	0.9	11.1	2, 5	-0.2	15.9
3, 4	6.0	14.8	3, 6	4.9	13.9

**Figure 2.** Side view of the fully relaxed geometries for the (a) short and (b) large polyaniline β -sheet models. For the color code see caption of Figure 1. For clarity methyl groups are represented as pale magenta balls.

geometric parameters (L_i , τ_i) are given in Table 1. First, it is noticeable that all L values are smaller than $L_{FES} = 3.57$ Å (the rise of a residue in FES), indicating that all residues in structures I–IV are under compressive strain. Recalling that the geometry of an optimized isolated strand is the FES, a residue under compressive strain is then characterized by $L < L_{FES}$, i.e., its length is shorter than the corresponding one in the FES. Likely, such compressive strain is originated by interstrand hydrogen bonding, i.e., longer hbs than the observed ones are expected if strand compression is not allowed. All residues present some degree of intrastrand twist, i.e., τ significantly deviates from zero. Residues at the end of each strand (those labeled with numbers 1, 3, 4, and 6 in Figure 1a) tend to be more twisted and more compressed than the central residue. To avoid ending effects on the structural trends, only central residues are considered for analysis. The average values of τ for the central residue in structures I and III, the flat antiparallel and parallel short β -sheet models, are 2.9° and 0.9°, respectively. However the corresponding average values of τ for the central residue in structures II and IV, the twisted antiparallel and parallel short β -sheet models, are 8.5° and 9.3° respectively, i.e., intrastrand left-handed twist is at least three times larger in the fully relaxed twisted models than in the fully relaxed flat models. Local interstrand twist values, T_{ij} , are listed in Table 2. According to these values all residues present some degree of interstrand twist, however, focusing only on central residues T values indicates that structures I and III are flat, i.e., $T < 1^\circ$, but structures II and IV are significantly twisted, $T > 11^\circ$. Side views of the fully relaxed structures I to IV are shown in Figure 2a. There is evidence that fully relaxed structures II and IV present interstrand twist, however, fully relaxed structures I and III are flat, as indicated by the T values corresponding to their central residues.

Table 3. Local Geometry Parameters (L_i , τ_i) for Residues in Large β -Sheet Structures V–VIII

residue label	V		VI		VII		VIII	
	L	τ	L	τ	L	τ	L	τ
1	3.51	11.4	3.4	30.1	3.48	3.35	3.48	3.0
2	3.53	1.1	3.5	9.6	3.42	0.1	3.46	15.6
3	3.55	-1.0	3.54	7.6	3.45	8.1	3.45	7.9
4	3.52	0.6	3.46	12.4	3.5	12.3	3.44	18.2
5	3.49	-8.4	3.36	-6.6	3.49	7.68	3.45	10.5
6	3.51	12.7	3.43	31.3	3.39	20.61	3.35	39.5
7	3.53	2.8	3.5	10.2	3.47	3.3	3.48	3.0
8	3.54	-1.4	3.55	6.3	3.48	-5.0	3.5	9.8
9	3.52	2.1	3.49	9.3	3.29	-9.7	3.43	4.2
10	3.48	-10.8	3.39	-4.3	2.88	1.75	3.27	7.7

Table 4. T_{ij} Values for Residues in Large β -Sheet Structures V–VIII

residue labels	V	VI	residue labels	VII	VIII
1, 10	-8.6	18.4	1, 6	10.1	21.3
2, 9	-1.0	14.4	2, 7	3.5	14.3
3, 8	-0.4	12.7	3, 8	3.6	16.0
4, 7	1.6	14.7	4, 9	17.2	20.0
5, 6	7.6	19.0	5, 10	6.1	14.1

To verify independence of intra- and interstrand twist with respect to strand length, four polyaniline β -sheet models built using five residues per strand (the large models) were also fully optimized using DFT-PBE. The corresponding values for the local geometric parameters (L_i , τ_i) are listed in Table 3. As for the short models, L values indicate that all residues are under compressive strain. Also all residues present some degree of intrastrand twist, and those at strand endings (residues labeled with the numbers 1, 5, 6, and 10) are more compressed and more twisted than those in the middle of the strands. Focusing on central residues, labeled with the numbers 2–4 and 7–9 in Figure 1b, it is found that intrastrand twist is significantly smaller for structures V and VII than for structures VI and VIII. Average values for τ in structures V and VII, the flat large models, are 0.7° and 1.5° , respectively. However, average values for τ in structures VI and VIII, the twisted large models, are 9.2° and 9.8° , respectively, i.e., the fully relaxed twisted large models present an intrastrand twist at least six times larger than the fully relaxed flat large models. Local interstrand twist values, T_{ij} , for the large models are listed in Table 4. Average values (considering only central residues) for T are 0.1° and 8.1° for structures V and VII, respectively. The average T value for structure VII is significantly larger than the average T value for structure V, owing to a large local interstrand twist in residues close to the right end of the sheet. Such distortion is due to the formation of an intrastrand hb between residues 9 and 10 in structure VII. Still interstrand twisting is not noticeable in such structures (Figure 2b). In the short flat parallel β -sheet, model III, an intrastrand hb similar to that in model VII is not formed. Likely the potential energy surface associate to strand endings is quite flat, small fluctuations in the forces may thus drive strand endings to conformations noticeably different among the models. This may be also the reason for the

Table 5. Energetic Stability (ΔE in kcal/mol) of Twisted β -Sheets with Respect to the Corresponding Flat Ones

ΔE	antiparallel		parallel	
	small	large	small	large
ΔE	0.3	0.7	0.1	2.1

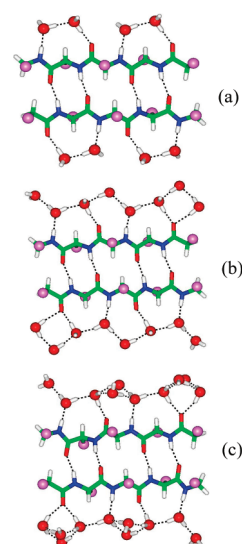


Figure 3. Small antiparallel flat β -sheet model microsolvated with (a) 8, (b) 16, and (c) 20 water molecules. In the small antiparallel twisted β -sheet model the relative orientation of water molecules respect to the backbone is the same as depicted here. For the color code see caption of Figure 1. For clarity methyl groups are represented as pale magenta balls. Dotted lines indicate hbbs.

discrepancy between residue conformations at the endings and at the middle of the β -sheet models. Average values for T in structures VI and VIII, the twisted large models, are 13.9° and 16.8° , respectively. These values are still two times larger, at least, than interstrand twist in the corresponding flat models. Interstrand twisting in the fully relaxed twisted large models is evident in Figure 2b. According to these results, strand lengthening does not change our observations regarding the development of intra- and interstrand twisting in the flat and twisted models.

The relative stability of twisted β -sheets with respect to the flat ones, calculated as $\Delta E = E_{\text{twist}} - E_{\text{flat}}$ where E_{twist} is the total energy of twisted sheets and E_{flat} the total energy of the corresponding flat ones, is listed in Table 5. These ΔE values are small in all the cases, indicating that both flat and twisted conformations can be adopted by parallel and antiparallel polyaniline β -sheets, though flat conformations may be preferred over twisted ones in vacuum. The stability trend may change if van der Waals interactions are taken into account. However it is expected a minimal alteration of such a trend as the spatial distribution of atoms in the flat and in the corresponding twisted conformation is very similar, hence similar van der Waals contributions to the stability of both flat and twisted structures.

Next the effect of hydration on the structure and the relative stability of β -sheets is investigated fully optimizing the geometry of flat and twisted β -sheets in which 8, 16, and 20 water molecules were adsorbed along the backbone. Only the small antiparallel β -sheet models, structures I and II, were microsolvated. Water

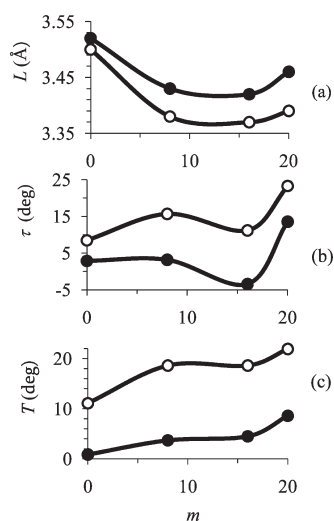


Figure 4. Change in the geometric parameters for the central residue with respect to the number of water molecules, m , in microsolvated antiparallel β -sheet models: (a) change in the rise, (b) change in the intrastrand twist, and (c) change in the interstrand twist. Closed circles stand for flat β -sheets. Open circles stand for twisted β -sheets. Solid lines are added for guiding the eye.

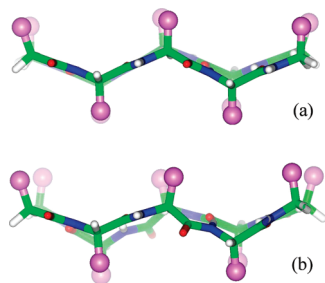


Figure 5. Flat to twist transition upon microsolvation with 20 water molecules: (a) flat conformation hydrated with 8/16 water molecules and (b) flat conformation hydrated with 20 water molecules. For the color code see caption of Figure 1. For clarity methyl groups are represented as pale magenta balls, and water molecules are not depicted.

molecules were absorbed such that an uninterrupted network of hbs, if possible, was formed along water molecules. The purpose of the latter is to investigate the effect of water–water hb cooperativity on the twist of β -sheets. It is expected that hydrated models constructed in such way resemble better a fully solvated peptide. A full scan of all possible water molecule orientations is beyond scope of the present work.

Among the several water molecule orientations examined, those shown in Figure 3 are the most stabilizing ones, and their orientations are equivalent in both flat and twisted β -sheets. Focusing on the geometric parameters for the central residues to avoid ending effects (see discussion above), it is found that L is systematically smaller than the corresponding values for vacuum structures (Figure 4a). The rise is reduced, on average, 2% and 3% due to hydration, i.e., hydration applies further compressive strain along the strands. Intrastrand twist increases as a response to hydration (Figure 4b). In particular average values for τ in the microsolvated structures with 20 water molecules are significantly larger (around three to four times larger) than the corresponding

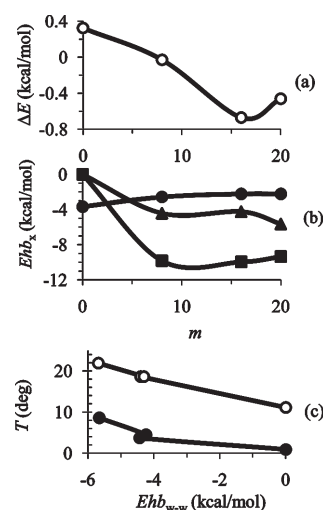


Figure 6. (a) Relative stability of the twisted antiparallel β -sheet respect to the flat one as the number of water molecules microsolvating the system, m , is increased. (b) Change in the hb strength in the microsolvated antiparallel β -sheets with respect to m . Circles stand for strand–strand hbs, $E_{hb_{s-s}}$. Triangles stand for water–water hbs, $E_{hb_{w-w}}$. Squares stand for strand–water hb, $E_{hb_{w-s}}$. (c) Change in the interstrand twist, T , of microsolvated antiparallel β -sheets with respect to the strength of $E_{hb_{w-w}}$, solid circles stand for the flat β -sheet and open circles for the twisted one. Solid lines connecting the points are added solely to guide the eye.

τ values for the structures in vacuum. The average τ value for structure I (the flat model) hydrated with 20 water molecules is even larger than the average τ value for the nonhydrated structure II (the twisted model), suggesting that the former structure is not longer flat. Interstrand twist also increases in hydrated structures (Figure 4c). The T value in the flat model (structure I) microsolvated with 20 water molecules is significantly larger, around 9 times larger, than the corresponding T value for the flat model (structure I) in vacuum, thus illustrating that the flat β -sheet transformed into a twisted one upon hydration. The latter is corroborated by the side view of the fully relaxed hydrated structures corresponding to the flat model (Figure 5). It is evident that the structure microsolvated with 20 water molecules is not longer flat.

Regarding the relative stability of twisted structures with respect to the flat ones upon hydration, it is shown in Figure 6a that ΔE becomes negative for hydrated structures. This trend indicates that twisted conformations are slightly favored over flat ones upon hydration. For the β -sheets hydrated with 20 water molecules, the stability of twisted respect to flat is reduced owing to the fact that the latter is not longer flat but twisted, i.e., the structure of the reference state changes and gets similar to that of the twisted state. Obviously the discussion presented above with respect to the influence of van der Waals interactions on the energy trend also applies here.

To get insight into why the β -sheet models microsolvated with 20 water molecules largely change their intra- and interstrand twist, hydrogen-bonding contributions to the total association energy have been analyzed. These contributions are the strand–strand hb contribution calculated as $E_{hb_{s-s}} = [E_{s-s}(\text{bare}) - 2E_{FES}]/hb_{s-s}$, where $E_{s-s}(\text{bare})$ is the total energy of the bare β -sheet in the geometry corresponding to the hydrated one, E_{FES} is the energy of the fully relaxed strand forming the β -sheet, and

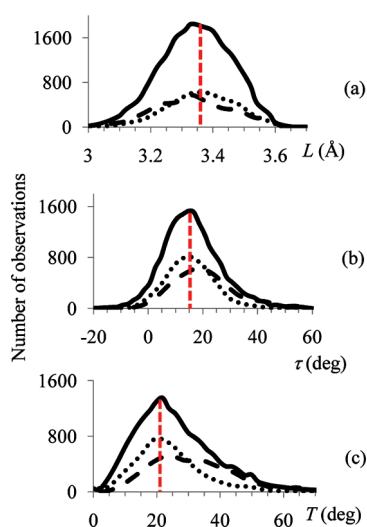


Figure 7. Distribution of values for the local geometric parameters of residues in antiparallel β -sheet motifs in protein crystals: (a) rise, L , (b) intrastrand twist, τ , and (c) interstrand twist, T . Solid line stands for all residues. Dotted lines stand for residues with no contact with water. Dashed lines stand for residues in contact with water. Dashed vertical lines mark the position of the peak for the distribution for all residues. For comparison purposes the distribution values for residues in contact with water are scaled five times.

hb_{s-s} is the number of $-N-H\cdots O$ hbs between strands. The water–water hb contribution calculated as $Ehb_{w-w} = [E_{hy} - mE_w]/hb_{w-w}$, where E_{hy} is the total energy of the hydrating water molecules in the geometry corresponding to the hydrated system but in the absence of the β -sheet, E_w is the total energy of a fully relaxed water molecule, m is the number of water molecules hydrating the β -sheet, and hb_{w-w} the number of $-OH\cdots OH-$ hbs. And the strand–water hb contribution calculated as $Ehb_{w-s} = [E_{sheet} - E_{s-s}(\text{bare}) - E_{hy}]/hb_{w-s}$, where E_{sheet} is the total energy of the hydrated β -sheet and hb_{w-s} is the number of $-NH\cdots OH-$ and $-CO\cdots OH-$ hbs. The strength of these hbs with respect the number of hydrated water molecules is depicted in Figure 6b. Only hbs of the bare/hydrated structure I are considered in plot of Figure 6b because the difference in the strength with respect to the corresponding hbs in the bare/hydrated structure II is small, around 0.2 kcal/mol or less. According to Figure 6b, Ehb_{w-s} is significantly stronger than Ehb_{w-w} and Ehb_{s-s} , being the latter the weakest one. In the bare structure the strength of these strand–strand hbs is -3.7 kcal/mol, which is in agreement with values reported in the literature.^{13,38} This strength is reduced up to -2.2 kcal/mol upon hydration. For β -sheets hydrated with 20 water molecules Ehb_{w-w} becomes ~ 1.4 kcal/mol stronger than that of Ehb_{w-w} in structures hydrated with 8 and 16 water molecules. On the contrary Ehb_{w-s} becomes weak by ~ 0.6 kcal/mol. Thus the large twist for the β -sheets hydrated with 20 water molecules seems to be related to noticeable changes in Ehb_{w-w} and Ehb_{w-s} . Correlating the strength of the different hbs with the interstrand twist, it is found that T increases as Ehb_{w-w} become stronger (Figure 6c). Such trend thus indicates that a flat to a twist transition is primarily driven by the strengthening of Ehb_{w-w} . Strands adjust its twist to favor stronger water–water interactions even that it does weaken Ehb_{w-s} and Ehb_{s-s} .

Real β -sheets are usually twisted in globular proteins. Then it is desirable to know how the structural parameters predicted by

DFT-PBE compare with that of real β -sheets in protein crystals. This comparison gives information on the influence of hydration and strand amino acid composition on the tendency to twist. An extensive set of protein structures derived from X-ray crystallography is analyzed to obtain the corresponding distribution of values for the local geometric parameters (L_i , τ_i and T_i). The criteria to determine if a residue belongs to a β -sheet are described in the Method Section. Only results for residues in antiparallel β -sheet motifs are presented, however, similar results were found for residues in parallel β -sheets. It is found that 14% of the residues in the investigated set are in β -sheet like motifs, 10% in antiparallel, and 4% in parallel conformation. The corresponding (L_i , τ_i , T_i) distributions of values for residues in antiparallel β -sheets are shown in Figure 7. The values for L , τ , and T corresponding to the peak of the distributions are $L \approx 3.35$ Å, $\tau \approx 15^\circ$, and $T \approx 22^\circ$. These values are in good agreement with those predicted by DFT-PBE for antiparallel β -sheets, particularly for microsolvated ones. Considering that the set of analyzed residues in β -sheet conformation contains only 8% of alanine, the good agreement between predicted and observed values for τ and T indicates that strand amino acid composition has little influence on inter- and intrastrand twist in real β -sheets. To investigate the effect of hydration on the twist of β -sheet motifs in protein crystals, the (L_i , τ_i , T_i) distribution of values has been obtained considering: (i) only residues in contact with water molecules (hydrated) in the set of studied proteins and (ii) only residues with no contact at all with water (nonhydrated). A residue was considered to be in contact with water if the distance between the N or O atoms of the peptide unit and the O atoms of water was less than 3 Å. The number of hydrated residues forming β -sheets was found to be significantly smaller than nonhydrated ones forming β -sheets. For comparison purposes the distribution of values for residues in contact with water is scaled five times. The (L_i , τ_i , T_i) distribution of values (Figure 7) shows a tendency that agrees with DFT results. The peak of the L distribution for hydrated residues is shifted toward smaller values of L with respect to the corresponding peak for nonhydrated ones (Figure 7a). Thus hydration applies compression along the backbone chain, as predicted by DFT-PBE. The peak of the τ distribution for the hydrated residues is shifted to larger values with respect to the peak for nonhydrated ones (Figure 7b). The latter indicates that hydration induces larger intrastrand twist as predicted by DFT-PBE. Also the trend observed for the distribution of T values is in agreement with the DFT-PBE results. The position of the peaks of the distribution indicates that hydration favors a larger interstrand twist (Figure 7c).

4. DISCUSSION

To discuss the sufficiency of the models here studied, average standard geometrical parameters (ϕ , ψ dihedral angles and interstrand hydrogen-bonding distances) for the optimized models are compared to experimental values from the literature. The reported (ϕ , ψ) mean values for alanine in antiparallel β -sheet motifs are $\phi_{exp} = -130.2 \pm 21.4^\circ$ and $\psi_{exp} = 143.8 \pm 14.6^\circ$.³⁹ Average (ϕ , ψ) values for our bare antiparallel β -sheet models are $(-144.7^\circ, 151.6^\circ)$, $(-136.9^\circ, 150.4^\circ)$, $(-149.3^\circ, 153.8^\circ)$, and $(-139.2^\circ, 153.3^\circ)$ for models I, II, V, and VI, respectively. The corresponding (ϕ , ψ) average values for the flat models hydrated with 8, 16, and 20 water molecules are $(-132.2^\circ, 137.0^\circ)$, $(-136.0^\circ, 131.2^\circ)$, and $(-128.2^\circ, 149.0^\circ)$, respectively. And for the corresponding twisted hydrated models are $(-118.2^\circ,$

136.8°), (−120.4°, 132.9°), and (−114.1°, 147.6°). All these theoretical values fall inside the range of the standard deviation for (ϕ_{exp} , ψ_{exp}). The same is found for the parallel β -sheet models. The reported (ϕ , ψ) mean values for alanine in parallel β -sheet motifs are $\phi_{\text{exp}} = -122.0 \pm 22.0^\circ$ and $\psi_{\text{exp}} = 136.6 \pm 16.8^\circ$.³⁹ The average (ϕ , ψ) values for the parallel β -sheet models investigated here are (−135.5°, 137.7°), (−127.4°, 140.7°), (−135.2°, 137.6°), and (−131.4°, 144.3°) for the models III, IV, VII, and VIII, respectively. Furthermore the average N–H···O hydrogen-bonded distances for the bare (2.0 Å) and the hydrated (1.9 Å) β -sheet models are in agreement with theoretical values reported in the literature.^{15,17,38,40} Moreover these values also fall in the range of the standard deviation for the corresponding values in crystal structures in proteins (1.96 ± 0.16 Å for antiparallel and 1.97 ± 0.15 Å in parallel β -sheets).⁴¹ These results give us confidence in the models used for describing β -sheet conformations.

Despite the compelling evidence indicating that β -sheets adopt both flat and twisted conformations, there is no consensus on the reasons underlying such behavior. Several studies have pointed out that isolated strands have an intrinsic tendency to twist.^{7,8,12,18,20} Indeed the structure connected to the minimum in the DFT-PBE potential energy surface for isolated polyalanine in extended conformation is left-hand twisted by 3° (see, e.g., Figure 2 in ref 33).^{33,34} However the latter value is around three times smaller than a typical intrastrand twist in a real β -sheet (see, e.g., Figure 7b). Moreover if the observed twist in real β -sheets is due to the intrinsic intrastrand twist of isolated strands, then one should explain why some β -sheets adopt flat conformations. Thus effects not connected to the intrinsic asymmetry of the minimum of the potential energy surface must be considered to explain the large intra- and interstrand twist in real β -sheets. In this work it is shown that strands in β -sheets are under compressive strain. According to ref 33, polyalanine backbone tends toward left-hand twist under compressive strain. Here it is found that β -sheets in which all residues locally twist to the left as response to compressive strain develop a right-handed interstrand twist. In flat β -sheets, however, residues are not as compressed as in twisted β -sheets and develop both left- and right-handed local intrastrand twist. The latter leads to an average intrastrand twist that is not large enough to force a noticeable interstrand twist in vacuum. Likely, local interstrand repulsive contacts prevent that all residues in flat sheets twist to the left upon interstrand hydrogen bonding, i.e., the response of the backbone to the compressive strain applied by hydrogen bonding is frustrated by local interstrand contacts. Microsolvation applies further compressive strain along strands favoring twisting. Residues in hydrated flat sheets adjust the sense of its intrastrand twist as a response to the extra compressive strain. Therefore flat sheets transform to twisted ones upon hydration. Compressive strain arising from microsolvation is likely exerted by water–water hydrogen bonding. In a network of aligned hbs the interaction is strengthened, and the hydrogen-bonding distance gets shorter due to a cooperative effect. The water–water hbs formed upon hydration with eight water molecules promote a reduction between 2% and 3% in the rise of the residues (see Figure 4a) and increase their interstrand twist by more than 60% with respect to residues in the bare β -sheets. Increasing the number of hydrating water molecules up to 16 does not change the water–water hydrogen-bonding strength (Figure 6b), therefore the rise and the interstrand twist do not change significantly (Figure 4). A further increment in the number of hydrating water molecules up to 20 strengthens the

water–water hydrogen bonding around 30%. Such an increment in the water–water strength, however, does not longer compress the residues; probably intra- and/or interstrand contacts prevent it. Therefore residues release the extra strain with a strong change in their intrastrand twist followed by a noticeable change in the interstrand twist. Thus in microsolvated β -sheets, strands tend to further reduce its rise and/or to change its twist to favor the development of an optimal network of hbs, i.e., to maximize $E_{\text{hb}_{\text{w-w}}}$, along the water molecules.

The analysis of the geometry of residues in real β -sheet motifs from an extended set of protein crystals supports the tendencies predicted by our calculations. Moreover, predicted conformations correspond to alanine, and the observed ones to all residues except glycine and proline, hence twisting is independent of the amino acid composition of the strands. Furthermore, the agreement between the predicted and observed twist in real β -sheet motifs gives confidence to the results obtained with DFT-PBE.

5. CONCLUSIONS

In conclusion the results presented in this work indicate that twisted and flat structures are intrinsic conformations of both parallel and antiparallel β -sheets. It is shown that twisting primarily results as a response of the backbone to compressive strain applied by interstrand interactions and hydration. Moreover it is shown that the strengthening of water–water hydrogen bonding along the solvation shell of β -sheets promotes a flat to twist transition. The good agreement between predicted and observed conformations of residues in β -sheet motifs in protein crystals corroborates the effect of hydration on the twist of β -sheets here described. Therefore the models here investigated are realistic, hence useful for elucidating more accurate atomistic models, for e.g., amyloid-like fibrils.

AUTHOR INFORMATION

Corresponding Author

*E-mail: irect@xanum.uam.mx.

ACKNOWLEDGMENT

The author thanks the Laboratorio de Visualización y Cómputo en Paralelo at UAM-Iztapalapa.

REFERENCES

- (1) Lehninger, A. L.; Nelson, D. L.; Cox, M. M. *Principles of Biochemistry*, 2nd ed.; Worth: New York, 1993; Chapter 4, pp 123.
- (2) Nelson, R.; Sawaya, M. R.; Balbirnie, M.; Madsen, A. Ø.; Riek, C.; Grothe, R.; Eisenberg, D. *Nature* **2005**, *435*, 773–778.
- (3) Sawaya, M. R.; Sambashivan, S.; Nelson, R.; Ivanova, M. I.; Sievers, S. A.; Apostol, M. I.; Thompson, M. J.; Balbirnie, M.; Wiltzius, J. J. W.; McFarlane, H. T.; Madsen, A. Ø.; Riek, C.; Eisenberg, D. *Nature* **2007**, *447*, 453–457.
- (4) Sunde, M.; Serpell, L. C.; Bartlam, M.; Fraser, P. E.; Pepys, M. B.; Blake, C. C. F. *J. Mol. Biol.* **1997**, *273*, 729–739.
- (5) Nelson, R.; Eisenberg, D. *Adv. Protein Chem.* **2006**, *73*, 235–282.
- (6) Dobson, C. M. *Nature* **2005**, *435*, 747–749.
- (7) Maccallum, P. H.; Poet, R.; Milner-White, E. J. *J. Mol. Biol.* **1995**, *248*, 374–384.
- (8) Yang, A. S.; Honig, B. *J. Mol. Biol.* **1995**, *252*, 366–376.
- (9) Weatherford, D. W.; Salemme, F. R. *Proc. Natl. Acad. Sci. U.S.A.* **1979**, *76*, 19–23.
- (10) Chou, K. C.; Némethy, G.; Scheraga, H. A. *J. Mol. Biol.* **1983**, *168*, 389–407.

- (11) Wang, L.; O'Connell, T.; Tropsha, A.; Hermans, J. *J. Mol. Biol.* **1996**, *262*, 283–293.
- (12) Shamovsky, I. L.; Ross, G. M.; Riopelle, R. J. *J. Phys. Chem. B* **2000**, *104*, 11296–11307.
- (13) Rossmeisl, J.; Nørskov, J. K.; Jacobsen, K. W. *J. Am. Chem. Soc.* **2004**, *126*, 13140–13143.
- (14) Zhao, Y.-L.; Wu, Y.-D. *J. Am. Chem. Soc.* **2002**, *124*, 1570–1571.
- (15) Scheiner, S. *J. Phys. Chem. B* **2006**, *110*, 18670–18679.
- (16) Perczel, A.; Gáspári, Z.; Csizmadia, I. G. *J. Comput. Chem.* **2005**, *26*, 1155–1168.
- (17) Viswanathan, R.; Asensio, A.; Dannenberg, J. J. *J. Phys. Chem. A* **2004**, *108*, 9205–9212.
- (18) Chothia, C. *J. Mol. Biol.* **1973**, *75*, 295–302.
- (19) Salemme, F. R. *Prog. Biophys. Mol. Biol.* **1983**, *42*, 95–133.
- (20) Ho, B. K.; Curmi, P. M. G. *J. Mol. Biol.* **2002**, *317*, 291–308.
- (21) Periole, X.; Rampioni, A.; Vendruscolo, M.; Mark, A. E. *J. Phys. Chem. B* **2009**, *113*, 1728–1737.
- (22) Esposito, L.; Pedone, C.; Vitagliano, L. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11533–11538.
- (23) Plumley, J. A.; Dannenberg, J. J. *J. Am. Chem. Soc.* **2010**, *132*, 1758–1759.
- (24) Perdew, J.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (25) Ireta, J.; Neugebauer, J.; Scheffler, M. *J. Phys. Chem. A* **2004**, *108*, 5692–5698.
- (26) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
- (27) Fuchs, M.; Scheffler, M. *Comput. Phys. Commun.* **1999**, *119*, 67–98.
- (28) ((a)) Gonze, X.; Beuken, J. M.; Caracas, R.; Detraux, F.; Fuchs, M.; Rignanese, G. M.; Sindic, L.; Verstraete, M.; Zera, G.; Jollet, F.; Torrent, M.; Roy, A.; Ghosez, P.; Raty, J. Y.; Allan, D. C. *Comput. Mater. Sci.* **2002**, *25*, 478–492. ((b)) Gonze, X.; Rignanese, G. M.; Verstraete, M.; Beuken, J. M.; Pouillon, Y.; Caracas, R.; Jollet, F.; Torrent, M.; Zerah, G.; Mikami, M.; Ghosez, P.; Veithen, M.; Raty, J. Y.; Olevano, V.; Bruneval, F.; Reining, L.; Godby, R.; Onida, G.; Hamann, D. R.; Allan, D. C. *Z. Kristallogr.* **2005**, *220*, 558–562.
- (29) Torrent, M.; Jollet, F.; Bottin, F.; Zerah, G.; Gonze, X. *Comput. Mater. Sci.* **2008**, *42*, 337–351.
- (30) Kearsley, S. K. *Acta Cryst. A* **1989**, *45*, 208–210.
- (31) Quine, J. R. *J. Mol. Struct. (THEOCHEM)* **1999**, *460*, 53–66.
- (32) Kneller, G. R.; Calligari, P. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62*, 302–311.
- (33) Ireta, J.; Scheffler, M. *J. Chem. Phys.* **2009**, *131*, art-085104.
- (34) Penev, E.; Ireta, J.; Shea, J.-E. *J. Phys. Chem. B* **2008**, *112*, 6872–6877.
- (35) Ireta, J.; Neugebauer, J.; Scheffler, M.; Rojo, A.; Galván, M. *J. Am. Chem. Soc.* **2005**, *127*, 17241–17244.
- (36) Calligari, P. A.; Kneller, G. R.; Giansanti, A.; Ascenzi, P.; Porrello, A.; Bocedi, A. *Biophys. Chem.* **2009**, *141*, 117–123.
- (37) 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.
- (38) Plumley, J. A.; Tsai, M. I.-H.; Dannenberg, J. J. *J. Phys. Chem. B* **2011**, *115*, 1562–1570.
- (39) Hovmöller, S.; Zhou, T.; Ohlson, T. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 768–776.
- (40) Rossmeisl, J.; Hinnemann, B.; Jacobsen, K. W.; Nørskov, J. K. *J. Chem. Phys.* **2003**, *118*, 9783–9794.
- (41) Baker, E. N.; Hubbard, R. E. *Prog. Biophys. Mol. Biol.* **1984**, *44*, 97–179.

Entropy Calculations of Single Molecules by Combining the Rigid–Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations

Ernesto Suárez, Natalia Díaz, and Dimas Suárez*

Julián Clavería 8, Departamento de Química Física y Analítica, Universidad de Oviedo, Oviedo, 33006 Spain

S Supporting Information

ABSTRACT: As shown by previous theoretical and computational work, absolute entropies of small molecules that populate different conformers can be predicted accurately on the basis of the partitioning of the intramolecular entropy into vibrational and conformational contributions. Herein, we further elaborate on this idea and propose a protocol for entropy calculations of single molecules that combines the rigid rotor harmonic oscillator (RRHO) entropies with the direct sampling of the molecular conformational space by means of classical molecular dynamics simulations. In this approach, the conformational states are characterized by discretizing the time evolution of internal rotations about single bonds, and subsequently, the mutual information expansion (MIE) is used to approach the full conformational entropy from the converged probability density functions of the individual torsion angles, pairs of torsions, triads, and so on. This RRHO&MIE protocol could have broad applicability, as suggested by our test calculations on systems ranging from hydrocarbon molecules in the gas phase to a polypeptide molecule in aqueous solution. For the hydrocarbon molecules, the ability of the RRHO&MIE protocol to predict absolute entropies is assessed by carefully comparing theoretical and experimental values in the gas phase. For the rest of the test systems, we analyze the advantages and limitations of the RRHO&MIE approach in order to capture high order correlation effects and yield converged conformational entropies within a reasonable simulation time. Altogether, our results suggest that the RRHO&MIE strategy could be useful for estimating absolute and/or relative entropies of single molecules either in the gas phase or in solution.

INTRODUCTION

A common assumption that lies at the heart of many entropy calculations is that the absolute entropy of a single molecule can be separated into several meaningful contributions.^{1,2} Perhaps the most straightforward and useful division of entropy is into the whole-body translational and rotational ($S_{\text{trans}} + S_{\text{rot}}$) and the intramolecular configurational (S_{config}) contributions, the latter one accounting for the entropy of internal degrees of freedom. Interestingly, the development of reliable and cost-effective strategies for computing the configurational entropy of complex molecular systems is a topic that has received much attention during recent years given that, for instance, the ability to compute accurate ΔS_{config} values could be very useful in understanding both the experimental and theoretical data on folding^{3–8} and/or association^{9–14} of biomolecular systems. Unfortunately, there are still many open questions about S_{config} , concerning its relationship to molecular structure and the importance of correlation among internal motions given that, for relatively large molecular systems, such correlations have been studied only through linear approximations or low-order truncated mutual information expansions.^{15–23}

According to Grubmüller and co-workers,^{24,25} methods that can compute entropy values can be classified into three broad categories: (a) methods based on the computation of free energy differences using thermodynamic integration,²⁶ (b) the hypothetical scanning approach developed by Meirovitch and co-workers,^{27,28}

and (c) an array of direct methods that extract the entropy of a single molecule from configurations generated by carrying out a conventional molecular dynamics (MD) or Monte Carlo simulation. In this work, we are basically interested in the latter category since our approach aims to estimate biomolecular entropies directly from MD simulations.

In what follows, we will briefly review a family of direct methods for entropy calculations that are more relevant to our approach taking into account that these methods can be divided into *parametric* (quasi-harmonic) and *nonparametric* methods depending whether or not they assume a functional form for the probability density function of the internal degrees of freedom. In addition, we also distinguish a third category of direct methods, the hybrid methods that combine elements of the two kinds (i.e., parametric and nonparametric).

Quasi-Harmonic Methods. The original quasiharmonic analysis was the first example of a direct method applied to biomolecular systems. It was first introduced by Karplus and Kushick, showing that the difference in configurational (i.e., non kinetic) entropy between two molecular conformations can be estimated from their respective covariance matrices.¹⁵ The basic idea is to consider the underlying configurational density function $P(\mathbf{q})$ in the classical configurational entropy, $S_{\text{config}} = -k_B \int_C P(\mathbf{q}) \ln$

Received: March 30, 2011

Published: June 30, 2011

$P(\mathbf{q}) d\mathbf{q}$, as a multivariate normal distribution, leading thus to the following entropy expression:

$$S_{\text{config}} = \frac{1}{2} k_{\text{B}} [n + \ln((2\pi)^n \det(\boldsymbol{\sigma}))] \quad (1)$$

where $\det(\boldsymbol{\sigma})$ is the determinant of the covariance matrix of the n internal coordinates. This equation can be applied to estimate only relative entropy values. However, the implementation of the quasi harmonic (QH) method suffers from some practical drawbacks due to the required transformation to internal coordinates and the consequent approximations made in the Jacobian. The need for this transformation comes from the fact that removal of the center of mass translation (unavoidable for convergence reasons) makes $\boldsymbol{\sigma}$ singular if Cartesian coordinates are used.¹⁷

To overcome the limitations of the original QH method, Schlitter has proposed an *ad hoc* approximation to the entropy in which the Cartesian covariance matrix is modified by adding a diagonal matrix so that the resulting $\boldsymbol{\sigma}$ matrix is nonsingular. Starting with the formula for the entropy of a one-dimensional quantum-mechanical harmonic oscillator (HO), Schlitter proposed the following heuristic entropy expression for a system of multiple particles:

$$\begin{aligned} S'_{\text{HO}} &= \frac{1}{2} k_{\text{B}} \sum_i \ln \left(1 + \frac{k_{\text{B}} T e^2}{\hbar^2} \langle q_i^2 \rangle_c \right) \\ &= \frac{1}{2} k_{\text{B}} \ln \left[\prod_i \left(1 + \frac{k_{\text{B}} T e^2}{\hbar^2} \langle q_i^2 \rangle_c \right) \right] \end{aligned} \quad (2)$$

where $\langle q_i^2 \rangle_c$ is the classical variance of the eigenvectors of the mass-weighted covariance matrix $\boldsymbol{\sigma}' = \mathbf{M}\boldsymbol{\sigma}$. Of particular importance is the fact that by adding both translational and rotational entropy contributions to the Schlitter's expression for the configurational entropy, it is possible to estimate absolute rather than relative entropies that are directly comparable to the rigid-rotor harmonic-oscillator (RRHO) entropies obtained from normal-mode analysis and standard statistical thermodynamic formulas.

Following the introduction of Schlitter's method, the QH analysis has been upgraded in order to compute absolute entropies.¹⁸ To this end, the reformulated quasiharmonic approximation (QHA) constructs a pseudoHessian matrix (\mathbf{H}) of the molecular system directly from the Cartesian covariance matrix ($\mathbf{H})_{ij} = k_{\text{B}} T (\boldsymbol{\sigma}^{-1})_{ij}$. The corresponding eigenvectors of \mathbf{H} can be seen to represent motional modes around the average system configuration. By associating each of the quasi-harmonic modes with a one-dimensional harmonic oscillator, the total configurational entropy can be approximated as a sum of harmonic contributions.

Despite the improvements introduced since the original formulation of the QH method, either Schlitter's approach or the renovated QHA method suffer from three potential flaws: (a) only linear correlations are taken into account, and therefore, supralinear correlations among the system variables are ignored; (b) formally, the multim minima potential energy surface is exceedingly smoothed by defining only one minimum and ignoring any anharmonicity (including multimodality) of the essentially multimodal probability density function; (c) there is no clear and unambiguous way to separate the overall rotation from the internal motions (e.g., there are uncertainties up to 80 J/mol K due to the arbitrariness in the choice of reference atoms for the preliminary structure superposition).²⁹ Accordingly, these methods provide an upper limit to the true absolute entropy S_{tot} .³⁰

Furthermore, given a covariance matrix, the function that maximizes entropy is precisely a Gaussian distribution function.³¹

To mitigate their well-known limitations, other authors have proposed several refinements of the QH methods that estimate the importance of the anharmonicity and/or supralinear correlation effects.^{16,19,32–34} For example, three years after Karplus introduced the QH analysis, Berendsen et al. proposed a simple strategy to account for the anharmonicity in the configurational probability density function.^{16,32} In this approach, the configurational part of the classical entropy $S_{\text{config}}(\mathbf{q})$, where $\mathbf{q} = (q_1, q_2, \dots, q_n)$ is an array of internal coordinates, is computed by combining the sum of marginal configurational entropies of the individual q_i variables with the correlation contributions captured by the Karplus model:

$$S_{\text{config}}(q) = -k_{\text{B}} \sum_i \int P(q_i) \ln P(q_i) dq_i + [S_{\text{config}}^{\text{QH}} - S_{\text{config,diag}}^{\text{QH}}] \quad (3)$$

where $S_{\text{config}}^{\text{QH}}$ is the configurational entropy computed by eq 1 and $S_{\text{config,diag}}^{\text{QH}}$ is the uncorrelated or diagonal Gaussian contribution, calculated by zeroing the nondiagonal elements of the covariance matrix.

Very recently, Baron et al. have proposed a more refined method that estimates both the effects of anharmonicity and supralinear correlations on the classical entropy (i.e., not only configurational entropy).^{19,33} The starting point is provided by the renovated QHA approach that uses only Cartesian coordinates.¹⁸ Subsequently, the entropy corrections are estimated from the quasiharmonic coordinates or modes and their corresponding probability densities obtained from the simulations. The classical entropy correction for the nonharmonic behavior, $\Delta S_{\text{cl}}^{\text{ah}} = S_{\text{cl}}^{\text{ah}} - S_{\text{cl}}^{\text{ho}}$, is estimated as the difference between the sum of the marginal entropies obtained directly for the individual modes, $S_{\text{cl}}^{\text{ah}}$, and the entropic term obtained considering the quasiharmonic coordinates as harmonic oscillators, $S_{\text{cl}}^{\text{ho}}$. On the other hand, the pairwise supralinear correction proposed by Baron et al., $\Delta S_{\text{cl}}^{\text{pc}} = \sum_{n>m} (S_{\text{cl,mn}}^{\text{ah}} - S_{\text{cl,m}}^{\text{ah}} - S_{\text{cl,n}}^{\text{ah}})$, is obtained from the classical entropies $S_{\text{cl,mn}}^{\text{ah}}$ computed using the joint probability distribution of the modes m and n , and the corresponding marginal entropies $S_{\text{cl,m}}^{\text{ah}}$ and $S_{\text{cl,n}}^{\text{ah}}$.¹⁹ The estimated entropy finally reads

$$S \approx S_{\text{QHA}} + \Delta S_{\text{cl}}^{\text{ah}} + \Delta S_{\text{cl}}^{\text{pc}} \quad (4)$$

In principle, as remarked upon by the authors, this method can be generalized for the inclusion of higher order correlations.³³ Computational results following this approach have been reported for a microsecond MD trajectory of a peptide model,³³ showing the importance of sufficient phase-space sampling to estimate entropic contributions. It has also been shown that, in accordance with previous studies,^{13,19} the pairwise supralinear correlation is normally large while the effect of the anharmonicity on the entropy calculations is relatively small.

Nonparametric Methods. On the basis of the mathematical tools of information theory and advanced statistics, it is possible to estimate the full-dimensional configurational probability density function $P(\mathbf{q})$ without resorting to any analytical approximation, unlike the QHA and Schlitter methods. This is the case of the method of Hnizdo et al.²⁰ that is based on the use of a series of k th nearest-neighbor (NN) entropy estimators,³⁵ \hat{S}_k , with k being large enough to make a smooth estimation but small enough to make the estimation as local as possible (e.g., $k \in \{1, 2, \dots, S\}$). In

this approach, each molecular configuration is represented as a vector with d components (e.g., the number of internal degrees of freedom). The NN estimation rests on the simple assumption that the configurational probability density function can be approximated *locally* in a nonparametric manner around each sample point q_i using the volume of a d -dimensional sphere centered at q_i and with a radius chosen such that it contains k neighbor data points. Given that the NN \hat{S}_k estimators yield asymptotically unbiased and consistent entropies as the number of data points N increases, they can provide accurate results for any probability distribution provided that sufficiently large samples of molecular simulation data are available. In practice, however, this method has been applied to small molecules due to the large computational cost of the NN searching algorithms when the dimensionality of the problem, d , is larger than 10–15.

Another nonparametric approach has been proposed by Gilson and co-workers based on the mutual information expansion (MIE),^{14,21} which is a systematic expansion of the entropy of a multidimensional system in mutual-information terms of increasing order n that capture the n -body correlations among the molecular internal coordinates. The size of the problem can be reduced up to manageable limits by neglecting all fourth- and higher-order MIE terms, thus allowing the calculation of S_{config} values for several small molecules as well as the change in S_{config} upon binding for protein–ligand systems,^{14,21} but at the cost of sampling millions of molecular configurations for reaching converged results. A combination of the NN and MIE techniques³⁶ has also been proposed.

Hybrid Approaches. Very recently, Hensen et al. have developed a new direct method that combines and improves different techniques with the specific aim of making entropy calculations for relatively large biomolecules feasible.^{24,25} These authors distinguish three parts or blocks in their method. First, they replace the k th NN entropy estimators by adaptive anisotropic ellipsoidal kernels that capture the configurational density in sufficient detail for up to 45-dimensional spaces. Second, they generate minimally coupled subspaces of internal degrees of freedom by applying a linear orthogonal transformation to Cartesian coordinates in such a way that the mutual information among the resulting coordinates is minimized. The new coordinates are subsequently clustered according to their degree of correlation (correlation among different clusters is neglected). Each oversized cluster ($d > 15$) is subdivided into smaller groups with maximum dimensionality $d = 15$, and its configurational entropy is computed as a sum of the estimated entropy of its components (subclusters) and then corrected by means of mutual information functions. For the stiffest degrees of freedom resulting from the orthogonal transformation, Hensen et al. also propose to employ a generalized quasiharmonic Schlitter formula that accounts for their quantum mechanical nature. Thus, the basic idea behind the method of Hensen et al. is that the combination of parametric and sophisticated nonparametric approaches could help overcome many limitations of the QH methods, but without seriously compromising its applicability to relatively large systems.

Clearly, a key element in the adaptive strategy pursued by Hensen et al. is their statistically based clustering of internal degrees of freedom into subsets that are weakly correlated and the separation into *softer* and *stiffer* degrees of freedom. In this respect, other authors have also designed direct methods that assume (*a priori*) a separation between internal degrees of freedom. For example, the method of Thorpe and Ohkubo³⁷ takes advantage of the fact that typical molecular mechanics (MM) Hamiltonians in

implicit solvent are easily separable into *vibrational* (identified as stretching, bending, and improper torsional motions) and *non-vibrational* (identified as torsions) degrees of freedom. These authors show that entropic contributions with respect to an arbitrary temperature for medium-sized systems can be obtained from standard thermodynamical statistical formulas and molecular partition functions which, in turn, can be derived from the density of state functions computed with the weighted histogram analysis method and replica-exchange MD simulations. In this way, it turns out that entropy differences for small- and medium-sized systems can be obtained from conformational and vibrational energy terms. More recently, Brüschweiler and Li²³ have claimed that the configurational entropy is separable into contributions from *hard* and *soft* degrees of freedom (the latter ones identified again with torsion angles) and that correlation effects among the *soft* variables cancel in good approximation on the basis of the results of test calculations on some dipeptide systems. These authors propose then to assess entropy changes between two states of a system at the same temperature, assuming that the entropic contributions of the hard variables do not change. As a first order approximation for conformational entropies is used, this approach is computationally very efficient and can be applicable to protein systems.

Combining the Rigid–Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations.

In previous works, we have investigated the role of conformational entropy in the absolute and relative stability of collagen model peptides²² as well as in the binding of small peptides to the active site of matrix metalloproteases.³⁸ However, the methodological scope of these previous articles was very narrow, as they are focused on the properties of the biomolecular systems being considered. Hence, in this work, we deal with the methodological details of our protocol for entropy calculations of single molecules whose molecular conformational space is sampled by means of classical MD simulations. Following the original proposal by Karplus and co-workers,^{12,39,40} we assume that the total entropy (S_{tot}) of a single molecule (excluding translation and rotation) can be partitioned into a vibrational (\bar{S}_{vib}) and a pure conformational contribution (S_{conform}):

$$S_{\text{tot}} = \bar{S}_{\text{vib}} + S_{\text{conform}} \quad (5)$$

This simple partitioning scheme can be shown to be formally exact as long as one neglects the entropic contributions from the high energy regions between any pair of wells on the potential energy surface of the molecular system.¹² In addition to the entropy partitioning, our approach is further characterized by the two following features:

- First, we employ the harmonic approximation to compute the mean value of S_{vib} over a time series of representative MD snapshots. The required energy minimization and normal-mode calculations can be done even for relatively large molecular systems provided that molecular mechanics or low level quantum mechanical methods are used. Of course, these normal mode and entropy calculations can be carried out within the framework defined by the conventional RRHO approximations and the standard statistical thermodynamic formulas,⁴¹ thus allowing one to estimate absolute entropies.
- Second, the conformational states along the MD trajectory are determined by means of the discretization of the time evolution of internal rotations about single bonds. This transformation, which does not require any *a priori* separation between softer or stiffer degrees of freedom, implies

that the conformational entropy of the whole MD trajectory (S_{conform}) should be naturally computed using the Shannon informational entropy:^{36,42,43}

$$S_{\text{conform}} = k_{\text{B}} \sum_j^{N_{\text{conf}}} (-p_j \ln p_j) \quad (6)$$

where p_j is interpreted as the statistical weight of the j th conformer. However, at this point, we resort to the above-mentioned MIE method in order to approach the full conformational entropy from the converged probability density functions of the individual torsion angles, pairs of torsions, triads, and so on.

The assumption of the entropy partitioning expressed in eq 5 discriminating between vibrational and conformational entropies; the computation of the mean values of the RRHO entropy contributions accounting for the translational, rotational, and vibrational contributions to the absolute entropy; the discretization of the torsional angles; and the concomitant use of MIE for estimating the conformational entropy from data provided by classical MD simulations constitute, altogether, the basic features of the RRHO&MIE entropy method examined in this work. Nevertheless, it must be noticed that some of these methodological ingredients have been employed in previous calculations of the gas-phase entropy of flexible molecules.^{44–49} Thus, it has been shown that the absolute entropy of a *mixture of conformers* can be computed with reasonable accuracy by averaging the RRHO entropy of all of the conformers present at a given temperature and then adding an entropy of *mixing* (ΔS_{mix}) that accounts for the entropic gain in the *mixture* of conformers:

$$S = \bar{S} + \Delta S_{\text{mix}} = \sum_{\alpha} p_{\alpha} S_{\alpha} + R \sum_{\alpha} (-p_{\alpha} \ln p_{\alpha}) \quad (7)$$

where p_{α} , the molar fraction of the α conformer, is typically estimated with the Maxwell–Boltzmann distribution formula in terms of quantum chemical enthalpies or approximate free energies. All of the distinctive conformers including enantiomeric conformers of the same energy are identified using either a direct counting method or automatic conformational search algorithms. Obviously, p_{α} and ΔS_{mix} in the “mixture of conformers” method are equivalent, respectively, to the statistical weight of the α conformer and the conformational entropy S_{conform} in the framework defined by the RRHO&MIE protocol, and consequently, the two approaches would be essentially identical. However, we will see that conceptual and practical differences subsist in the way that the S_{conform} and ΔS_{mix} terms are handled and that the RRHO&MIE approach is more suitable for dealing with relatively large systems.

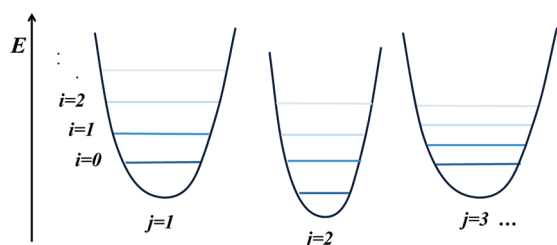
To introduce further details of the RRHO&MIE method and illustrate its potential benefits and limitations, the rest of this paper is organized as follows. First, we will comment on the partitioning of the total entropy (excluding translation and rotation) into the vibrational and conformational components that forms the basis of the RRHO&MIE approach. The separation between vibrational and conformational effects relies heavily on the discretization of the torsional degrees of freedom, and therefore, we will describe this transformation in a detailed manner. Then, after having restated a few definitions about mutual information functions, we will point out that the original MIE expression can be reformulated in such a way that all redundancy in the calculation of the n -order terms is removed. The discretization process combined with the reformulated MIE equation allow us to compute

S_{conform} values including higher order terms beyond the second- or third-order terms that have been considered in most of the previous works. The performance of the RRHO&MIE protocol will be critically discussed on the basis of a series of test calculations on different systems ranging from hydrocarbon molecules in the gas phase to a polypeptide molecule in aqueous solution. For the hydrocarbon molecules (three C_6H_{14} and five C_7H_{16} isomers), their conformational entropies derived from classical MD simulations are combined with their RRHO entropies obtained by carrying out quantum chemical frequency calculations, and the resulting absolute entropies are then compared with experimental data. In this way, we will examine to what extent the assumptions made in the formulation of the RRHO&MIE protocol and the use of classical MD simulations affect the quality of the computed absolute entropies. Second, we will focus on more complex test systems: a series of dipeptide molecules in the gas phase. Although experimental absolute entropies for these test systems are not available, they constitute important cases of study in our validation calculations because they present larger correlation effects among internal degrees of freedom due to the presence of intramolecular H-bond and polar interactions. On the basis of the results obtained for the dipeptide molecules in the gas phase, we will see that the combined discretization process and the MIE approximations are able to capture high order correlation effects and simultaneously yield converged conformational entropies within a reasonable simulation time. Finally, we will analyze the results obtained for a polypeptide molecule in aqueous solution, which can be a representative of the kind of molecular systems for which the RRHO&MIE entropy calculations could be particularly interesting. Besides analyzing the source of conformational correlations, we will also compare the convergence properties and absolute values of S_{conform} as provided by the RRHO&MIE (using an implicit solvent model and MM normal mode calculations) and the QHA methods. Overall, we hope that the methodological proposals and the results of our test calculations could be useful to extend the range of applicability of approximate entropy methods for studying more challenging systems involving polypeptide-folding or molecular association processes.

THEORY AND COMPUTATIONAL METHODS

a. Decomposition of the Intramolecular Entropy. As shown in previous works by Gilson and Zhou,⁴⁰ the entropy of a system with multiple potential energy wells can be formally decomposed into two parts: the entropy that arises from vibrational motions within a single well and the entropy due to conformational transitions between different energy wells. In the case of a single molecule either in the gas phase or in the condensed phase, this entropy decomposition is clearly an approximation whose goodness would depend on the actual molecular structure being studied, temperature, environmental effects, etc. It is true that other entropy methods have been formulated that can derive the total configurational entropy directly from Cartesian coordinates, thus avoiding any assumption about the additivity or lack thereof of the S_{conform} and S_{vib} contributions¹⁹ (although there may remain the problem of separating the overall rotation from the internal motions²⁹). However, although we recognize that the total entropy is strictly a global property, its formal decomposition into the conformational and vibrational terms constitutes the basis for the present work. As a matter of fact, many experimental and theoretical methods have provided meaningful results by accepting that free

Scheme 1. Schematic Example of a Multiminima Potential Energy Surface



energies or entropy contributions can be attributed to particular degrees of freedom and/or physical interactions.^{1,2}

In the Supporting Information, we provide an alternative derivation of the formal decomposition of the single molecule entropy assuming that the translational and rotational degrees of freedom have been removed and that the resulting potential energy surface in terms of the remaining internal degrees of freedom can be approximated by a set of *distinguishable* energy basins, as shown in Scheme 1. The final entropy decomposition formula is

$$S_{\text{tot}} = \sum_j p_j S_{\text{vib}}^j + S_{\text{conform}} \quad (8)$$

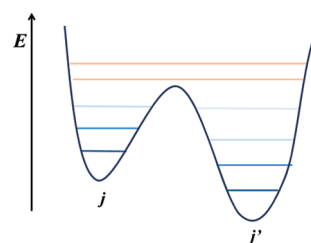
where p_j and S_{vib}^j stand for the probability and vibrational entropy associated with the j th energy basin, respectively, while S_{conform} is the conformational entropy that arises from the populations of the different energy basins.

The practical implementation of the entropy decomposition (eq 8) implies that each molecular configuration of a single molecule (e.g., one MD snapshot) employed in the entropy calculations should be associated with one molecular j conformer (or j energy basin) representing its conformational state. In our protocol, this assignment is only relevant for the evaluation of the conformational entropy S_{conform} and can be achieved by discretizing the time evolution of the torsion angles (see below). The average value of S_{vib} is obtained separately by means of energy minimization calculations followed by normal-mode analyses within the context of the harmonic oscillator model. In terms of accuracy, the formal entropy decomposition should provide nearly exact results for small molecules in the gas phase (ideal conditions) at room temperature given that, in this case, only the lowest vibrational levels would be populated and they could be tagged univocally to a single conformer.

Could the entropy decomposition provide meaningful results in other more complicated situations such as that depicted in Scheme 2? As temperature increases, high energy vibrational levels that can be simultaneously assigned to different conformers would become populated, partially blurring the distinction between vibrational and conformational motions. Although the practical implementation of eq 8 would not be impeded in this case, the entropy given by eq 8 would be overestimated due to the double-counting of the contributions of those high energy vibrational levels accessible from the two energy basins. Therefore, we propose that, in general, the decomposition of the single molecule entropy could be useful, as it would provide an upper bound to the actual value. Nonetheless, its performance needs to be assessed by carrying out test calculations.

b. On the Use of the Harmonic Oscillator Model for Vibrational Entropies. We think that one remarkable advantage of discriminating between vibrational and conformational entropies is the utilization of the HO model for computing the average

Scheme 2. High Energy Vibrational Levels Can Be Assigned to Different Energy Basins



vibrational contribution over a series of representative structures. Clearly, a straightforward computational protocol can be applied for obtaining the mean values of S_{vib} . Starting with a set of representative MD snapshots, optimization calculations relax the internal geometry of each molecular structure to that corresponding to a particular j energy basin, whose vibrational entropy, S_{vib}^j , is subsequently estimated by means of conventional normal mode calculations. In fact, the combination of the HO approximation, the rigid rotor model, and the standard formulas of statistical thermodynamics based upon canonical partition functions can be used to estimate absolute entropies that in some cases admit a direct comparison with experimental data (e.g., in the gas phase).⁵⁰ We can also benefit from many efficient implementations for performing either geometry optimizations and second derivative calculations using molecular mechanics⁵¹ and/or quantum mechanical methods depending on the size of the molecular system. Part of the limitations of the HO model (e.g., the lack of anharmonic effects, errors arising from the level of theory, etc.) could be mitigated by using empirical corrections in the form of scaling factors.^{52,53} We also note that the RRHO entropies (complemented with the conformational contributions) could be useful within the context of approximate free energy methods like the so-called molecular mechanics Poisson–Boltzmann method.⁵⁴ Moreover, previous computational experience with these approaches has shown that the average normal mode entropy converges quite well in terms of the length of simulations and the number of molecular configurations required.^{54,55}

c. Discretization of the Torsional Degrees of Freedom. To take into account the S_{conform} contribution, which arises from the population distribution of the molecular conformers, we propose to discretize the probability density functions of those torsion angles that are commonly used to define the molecular conformational state. To this end, we apply the following protocol.

First, we collect the time series with the values of the torsion angles along the MD simulation. For each torsion angle θ , we have a sample of size N , $\{\theta_1, \dots, \theta_N \mid \theta_i \in [0, 2\pi)\}$, where N is the number of MD snapshots ($\sim 10^5$ – 10^6). To obtain an analytic representation of the underlying probability density function for the torsion θ , we employ the von Mises kernel estimator (the von Mises density is the circular analogue of the Gaussian distribution). Specifically, the probability density function of a given torsion, $\rho(\theta)$, is then approximated by the arithmetic mean of N von Mises distributions centered on the θ_i values:

$$\hat{\rho}(\theta; v) = \frac{1}{2\pi N I_0(v)} \sum_{i=1}^N \exp\{v \cos(\theta - \theta_i)\}$$

where $I_r(v)$ is the modified Bessel function of order r , and v , which is the so-called concentration parameter, is the inverse of the smoothing parameter of the kernel estimator. The value of v

is obtained by applying the recently derived “von Mises-scale plug-in rule”³⁶ for the smoothing parameter, which results in the following expression that depends on the number N of data points:

$$v = [3N\hat{\kappa}^2 I_2(2\hat{\kappa}) \{4\pi^{1/2} I_0(\hat{\kappa})^2\}^{-1}]^{2/5}$$

with $\hat{\kappa}$ being an estimate of the concentration parameter of the global data, for which we take $\hat{\kappa} = 1$, as this value leads to a slightly oversmooth distribution that is more convenient for our purposes. By setting an empirical value for $\hat{\kappa}$ rather than for v , we keep the dependence on N .

The advantage of using the von Mises kernel estimator instead of a normalized histogram method is that we can characterize the analytical properties of $\rho(\theta) \approx \hat{\rho}(\theta;v)$ in order to automatically optimize the location of the maximum and minimum values of $\rho(\theta)$, which is a prerequisite for discretizing the time evolution of the torsion angle θ . This task is performed by the analytical evaluation of the first and second derivatives of $\hat{\rho}(\theta;v)$ over a grid of $\phi_k = k(\pi/(180))$ points with $k = 0, 1, \dots, 359$. On one hand, the approximate positions of the $\hat{\rho}(\theta;v)$ critical points are first determined by averaging two consecutive grid points k and $k + 1$ for which $\hat{\rho}'(\phi_k;v) \hat{\rho}'(\phi_{k+1};v) \leq 0$. For the torsional distributions, the maximum critical points are easily identified thanks to their largely negative second derivative, and therefore, we first locate the maxima of $\hat{\rho}(\theta;v)$ and use the intermediate position between two consecutive maxima as the initial guess for searching the minima. The minima of $\hat{\rho}(\theta;v)$ are found by means of a steepest descent search that adopts a convergence threshold for the residual gradient of 10^{-4} and employs a linear interpolation approximation for evaluating the gradient on the basis of the $\hat{\rho}'(\phi_k;v)$ values.

Once the $\theta_{\min,i}$ values corresponding to the $\hat{\rho}(\theta;v)$ minima are found (let us suppose that there are m minima and $\theta_{\min,i} < \dots < \theta_{\min,m}$), the configurational space of θ defined by the $[0, 2\pi)$ interval is divided into m nonoverlapping intervals ($[\theta_{\min,1}, \theta_{\min,2})$, \dots , $[\theta_{\min,m-1}, \theta_{\min,m})$, $[\theta_{\min,m}, 2\pi) \cup [0, \theta_{\min,1})$) that, in turn, define the different conformational states accessible to θ . In this way, the initial time series containing N data points, $\{\theta_1, \dots, \theta_N\}$, is easily transformed into a set of N integer numbers $\{a_1, \dots, a_N\}$ labeling the conformational states populated by the torsion angle. For example, if θ corresponds to an internal rotation about a $C(\text{sp}^3) - C(\text{sp}^3)$ bond, its associated a_i variable could have values of 1, 2, and 3 representing the $g+$, $g-$, and $anti$ conformations, respectively. Therefore, the continuous variable θ characteristic of the torsion angle becomes a discrete random variable A , whose probability mass function, $P(A)$, can be estimated by the maximum likelihood method fed with its corresponding outcomes $\{a_1, \dots, a_n\}$. Finally, we note that the loss of entropy during the $\theta \rightarrow A$ transformation can be expected to be *vibrational*, and that we assume that such a contribution can be reasonably accounted for by normal mode calculations. In other words, the entropy due to the fluctuations of the torsion angles around a local minimum should be recovered by the S_{vib} calculations. This means that, in our approach, it is not necessary to distinguish between *soft* or *hard* degrees of freedom (or between *vibrational* or *nonvibrational* ones) because the conformational entropy turns out to be purely *informational* as a consequence of the discretization process. The probability density function for the torsion angle θ about the C2–C3 bond of 2-methyl-hexane is shown in Figure 1.

d. Mutual Information Expansion: Application to Conformational Entropy Calculations and Reformulation into a Computationally Efficient Scheme More Suitable for Very Large Systems. As shown in the previous section, the conforma-

tional state of a torsion angle can be associated with a one-dimensional random variable A . Analogously, the conformational state of a set of M torsion angles can be described by an M -dimensional random vector (A_1, \dots, A_M) , or alternatively, the conformational state can also be associated with an ordered set $\{A_1, \dots, A_M\}$, where A_i specifies the conformational state of the i th torsion and M is the size in terms of the number of torsion angles of our system $\mathcal{A} = \{A_1, \dots, A_M\}$. Of course, for medium-sized and large molecules, the number of potentially accessible conformers is huge ($\sim 3^M$), and in this case, obtaining the underlying probability mass function $P(A)$ is practically impossible due to sampling limitations. As other authors have done previously,^{21,36} we will use the mutual information expansion as a workaround to this problem. The basic idea here is that if we are able to obtain converged values of the probability mass functions of the individual torsion angles $p(A_i)$, pairs of torsions $p(A_i, A_j)$, triads $p(A_i, A_j, A_k)$, and so on, then we can approach the full-dimensional informational entropy of the (A_1, A_2, \dots, A_M) variables (i.e., the conformational entropy) by including systematically n -order correlations among the A_i variables as measured by mutual information functions.⁴² More specifically, the mutual information expansion leading to the total entropy can be written as

$$\begin{aligned} S(A_1, \dots, A_M) &= \sum_{i=1}^M S(A_i) - \sum_{i < j} I_2(A_i, A_j) \\ &+ \sum_{i < j < k} I_3(A_i, A_j, A_k) \\ &- \sum_{i < j < k < l} I_4(A_i, A_j, A_k, A_l) + \dots \end{aligned}$$

where $S(A_i)$ is the informational entropy of the i th torsion angle along the MD simulation while $I_2(A_i, A_j)$, $I_3(A_i, A_j, A_k)$, and so forth are the corresponding mutual information functions that capture the general dependence among the (A_1, A_2, \dots, A_M) variables (unlike the covariance function used in the QH methods that only measures linear correlations). More specifically, the mutual information shared by two variables, A_i and A_j , is computed by combining the informational entropies of the single- and two-variable probability mass functions of the torsion angles:

$$I_2(A_i, A_j) = S(A_i) + S(A_j) - S(A_i, A_j)$$

where $S(A_i, A_j)$ is the joint entropy of A_i and A_j . Both $S(A_i)$ and $S(A_i, A_j)$ can be computed using a Shannon-type expression, $-k_B \sum p_\alpha \ln p_\alpha$ where p_α is the corresponding probability mass function and the sum runs over the possible states accessible to the individual torsion angles A_i or to the pair of torsions (A_i, A_j) . Similarly, the third-order function $I_3(A_1, A_2, A_3)$ includes correlation effects among three torsion angles:

$$\begin{aligned} I_3(A_i, A_j, A_k) &= S(A_i) + S(A_j) \\ &+ S(A_k) - S(A_i, A_j) - S(A_i, A_k) - S(A_j, A_k) \\ &+ S(A_i, A_j, A_k) \end{aligned}$$

In general, the mutual information shared among k variables $\mathcal{F} = \{A_1, A_2, \dots, A_k\}$ can be generalized by

$$I_k(\mathcal{F}) = \sum_{l=1}^k (-1)^{l+1} \sum_{\substack{\mathcal{F} \subset \mathcal{J} \\ |\mathcal{J}|=l}} S(\mathcal{J})$$

where for every value of l , the inner sum runs over all possible $\binom{k}{l}$ subsets of \mathcal{F} with l elements, $|\mathcal{J}|$ being the cardinality of \mathcal{J}

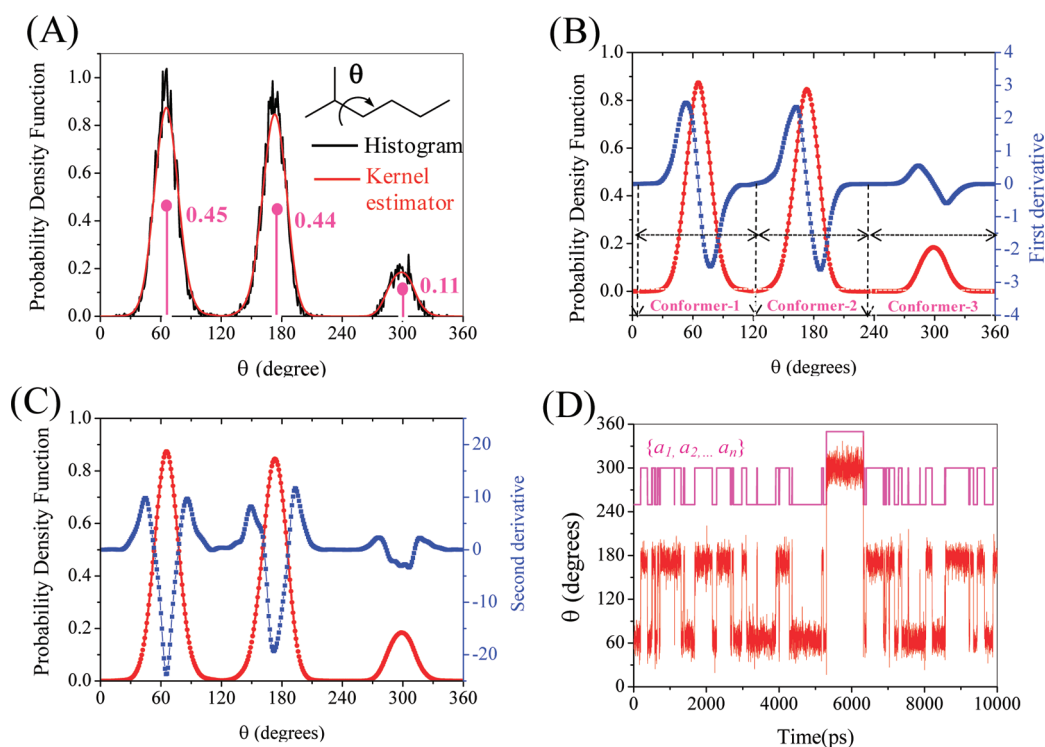


Figure 1. (A) Probability density function for the torsion angle θ about the C2–C3 bond of 2-methyl-hexane as obtained from a histogram representation and a Von Mises kernel estimator. The probability mass function of the three conformational states is also indicated by the vertical bars. (B and C) Superposition of the probability density function and its first and second derivatives as estimated by the Von Mises kernel. (D) Time evolution of the torsion angle θ and its associated discrete variable A (see text for details).

(i.e., the number of elements). A special case arises when $k = 1$, that is, $I_1(J_1) = S(J_1)$. The total entropy of the M variables $\mathcal{A} = \{A_1, \dots, A_M\}$ is usually approximated by a truncated mutual information expansion up to order n ,^{21,42} which becomes an exact expression when $n = M$:

$$S^{(n)}(\mathcal{A}) = \sum_{k=1}^n (-1)^{k-1} \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=k}} I_k(\mathcal{J}) \quad (9)$$

However, we note again that in previous calculations of configurational entropies based on the MIE approach, only low order (i.e., $n \leq 3$) approximations have been used.

A computational shortcoming of the usual form of the MIE is that, for any subset \mathcal{J} of \mathcal{A} with cardinality lower than n , its entropy must be evaluated more than once. If the order of the expansion is not too high and a reasonable amount of rapid access memory is available, all of the required $S(\mathcal{J})$ terms can be stored and used repeatedly in the calculation of the high order terms. For large and strongly correlated systems (e.g., $M > 100$, $n = 3-5$), however, this simple approach could become unfeasible due to memory depletion, while a direct implementation in which the $S(\mathcal{J})$ terms would be recomputed on-the-fly as needed would be prohibitively expensive. In order to remove all redundancy in the calculation of the MIE terms, the original expression can be reformulated in the following manner:

$$S^{(n)}(\mathcal{A}) = \sum_{k=1}^n \left[\sum_{i=0}^{n-k} (-1)^i \binom{M-k}{i} \right] \sum_{\substack{\mathcal{J} \subset \{A_1, \dots, A_M\} \\ |\mathcal{J}|=k}} S(\mathcal{J}) \quad (10)$$

where the entropy of each subset \mathcal{J} is computed and used only once. The formal proof of the equivalence between eqs 10 and 9 is given in the Supporting Information. To the best of our knowledge, the above expression has not been reported so far, although its implementation could greatly simplify the MIE computational problem for very large systems with hundreds or even more torsion angles. Finally, we note that all of the MIE calculations reported in this work were carried out using a FORTRAN90 code that has been developed in our laboratory and that will be reported elsewhere.⁵⁷

e. Molecular Dynamics Simulation Settings. *Small Systems: Gas-Phase MD Simulations.* MD calculations for each individual molecule were carried out by means of the Amber10 package.⁵⁸ The generalized AMBER force field⁵⁹ was used for the alkane molecules, while the dipeptides (Ace–X–X–Nme, with X = Ala, Ser, Asn, Leu, and Lys) were represented by the AMBER03 force field.⁶⁰ To derive the atomic charges of the alkane molecules, we performed HF/6-31G(d) geometry optimizations of the fully extended conformers followed by single-point B3LYP/cc-pVTZ calculations using the Gaussian 03 program.⁶¹ Other parameters were generated automatically using the antechamber module included in the Amber10 package. Subsequently, 2.0 μ s MD trajectories for all of the alkane and dipeptide compounds were run in the gas phase at 298 K and 1.0 atm using a 1.0 fs time step. Coordinates were saved every picosecond of simulation time (2×10^6 structures).

Polypeptide System: MD Simulation in Solution. We simulated the following hexapeptide sequence, Ace–Pro–Phe–Glu–Leu–Arg–Ala–NH₂ (termed the PFG peptide), which corresponds to one of the peptide sequences that has been selected from a peptide library mixture for probing the cleavage

site motifs of matrix metalloproteinases (MMPs).⁶² Starting coordinates were obtained from conformational search calculations using the LMOD program⁶³ linked to the Amber package. In the LMOD calculations, we employed the AMBER03 force field coupled with the Hawkins–Cramer–Truhlar pairwise generalized-Born (HCT-GB) model.⁶⁴ The lowest energy LMOD structure of PFG was then surrounded by a periodic truncated octahedral box of TIP3P water molecules that extended ~ 12 Å from the protein atoms (~ 1400 water molecules). The solvent molecules were initially relaxed by means of energy minimizations and 50 ps of MD. Subsequently, the full system was minimized and heated gradually to 300 K during 50 ps of MD. During the MD simulation, the system remains coupled to a thermal and a hydrostatic bath at $T = 300$ K and $P = 1.0$ atm, the time step of integration was 2.0, the SHAKE procedure on the X–H bonds was applied, and the particle-mesh-Ewald approach was used for nonbonded interactions. A 2.0 μ s trajectory was computed, and coordinates were saved for analysis every picosecond.

f. Normal Mode Calculations. *Alkane Molecules: Quantum Mechanical Calculations.* For the alkane molecules, we evaluated the average value of their rotational and vibrational entropies by carrying out quantum mechanical (QM) calculations on selected MD snapshots. To this end, 2000 equally spaced snapshots were extracted from the MD trajectories of each compound. All of the MD snapshots were minimized and scored in terms of their relative MM energies. The relative energies of the relaxed structures were compared in order to filter out all of the energetically equivalent structures, obtaining thus a relatively small set of energetically distinguishable molecular conformers (i.e., this means that only one structure for each pair of enantiomeric conformers is retained). The statistical weight of each energetically unique conformer was estimated by its relative abundance in the initial data set containing the 2000 structures. These conformers were subsequently minimized at the B3LYP/cc-pVTZ level of theory^{65,66} and further characterized by analytical frequency calculations. Then, thermal contributions to the gas-phase entropy of the translational, rotational, and vibrational degrees of freedom were obtained within the context of the RRHO approximation and using the B3LYP/cc-pVTZ moments of inertia and vibrational frequencies. Entropy contributions of overall rigid-body rotation take into account the corresponding external symmetry number for each conformer.

Polypeptide Systems: Normal Mode MM Calculations. Taking into account the relatively large size of the polypeptide system, we decided to use MM methodologies for carrying out the required normal mode calculations. Moreover, since the solute and solvent fluctuations are coupled to each other during the MD simulation of the TIP3P water box, we also used the HCT-GB implicit solvent model for removing the explicit consideration of solvent degrees of freedom. Thus, we extracted 10 000 equally spaced snapshots from the 2.0 μ s trajectories of PFG. These structures were postprocessed through the removal of all solvent and counterion molecules. Then, solute entropic contributions were estimated for each structure using the NAB package.⁶⁷ Prior to the normal mode calculations, the geometries of the system described by the AMBER03 force field were minimized until the root-mean-squared deviation of the elements in the gradient vector was less than 10^{-5} kcal/(mol Å). It may be interesting to note that the large majority of the minimized structures (>98%) corresponded to different minima on the potential energy surface, as expected from the relatively large size of this system. These minimizations and the subsequent normal mode calculations⁵¹ were carried out

using the HCT-GB solvent model. Finally, the RRHO entropic contributions were averaged over the 10 000 snapshots.

RESULTS AND DISCUSSION


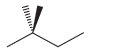
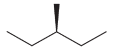
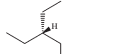
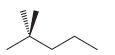
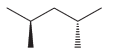
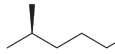

Absolute Entropies of Alkanes. Table 1 and Figures 2 and 3 collect the results of our entropy calculations for the C_6H_{14} (1–3) and C_7H_{16} (4–8) isomers. Of particular interest in these test calculations is the fact that the experimentally available gas-phase entropies of the alkane molecules can be directly compared with our theoretical estimates, since we employ both the RRHO approximation and standard statistical formulas for obtaining the translational–rotational–vibrational entropies of polyatomic molecules,⁴¹ which are subsequently complemented by adding the conformational entropy contributions arising from the discretization of the time evolution of the torsion angles during the classical MD simulations. Furthermore, consideration of these molecules, which are still small enough for exhaustively exploring their conformational space, allows us to show more clearly the relationship of our method with the closely related protocol of the “mixture of conformers” model, which has provided accurate entropy values of small molecules.^{44,45}

Automatic checks of the relaxed snapshots show that the number of energetically distinguishable conformers (N_E in Table 1) that are populated at 298 K in the gas phase vary between 4 and 15 for the various alkane molecules, except for 2,2-dimethyl-butane (2), which shows a single energy level because torsional motions of the methyl groups and about the C2–C3 bond interconnect structures that are degenerate in terms of their potential energy. Precisely, the RRHO entropy value of 2,2-dimethyl-butane at the B3LYP/cc-pVTZ level, 357.77 J/(K mol), is only less than 1 J/(K mol) below the experimental value (378.65), thus showing that the B3LYP/cc-pVTZ level of theory accounts well for the majority of the gas-phase entropy of alkane molecules having a single conformer. This also suggests that frequency scaling, which has been commonly used to correct for systematic errors in the ab initio computation of vibrational frequencies⁵² is probably not required at the B3LYP/cc-pVTZ level. For the rest of the test compounds, however, several energy-distinguishable conformers are significantly populated at 298 K. Their individual RRHO entropy values are similar, but they exhibit non-negligible differences as large as 5 J/(K mol).

The mean values of the RRHO absolute entropies (\bar{S} in Table 1) underestimate the experimental data up to tenths of a J/(K mol) for the flexible hydrocarbon molecules. The correlation plot shown in Figure 2A shows more clearly the differences between the experimental and the average theoretical entropies: the resulting squared correlation coefficient is quite low, $R^2 = 0.70$ (a unit slope is imposed), and a large offset at the intercept arises (-14 J/(K mol)). Moreover, the relative entropy values among the C_6H_{14} or C_7H_{16} isomers are poorly described by the average RRHO values: for example, the ΔS_{exp} value between 1 and 2 is -30.17 J/(K mol), whereas the computed $\Delta \bar{S}$ value is only -7.89 J/(K mol). Therefore, it is clear that neglecting conformational entropy contributions significantly affects the quality of the entropy calculations even for simple hydrocarbon molecules.

Concerning the conformational entropies of the alkane molecules estimated from the classical MD simulations, we first assess their statistical convergence by plotting the S_{conform} values vs simulation time at various expansion orders for 2-methyl-hexane (7) and 3,3-dimethyl-pentane (8; see Figure 3). Thus, all of the S_{conform} profiles in Figure 3 converge to nearly zero-slope curves

Table 1. Average Value of the Translation-Rotational-Vibrational Entropy (\bar{S}) and Converged Values of the Conformational Entropy at 298 K (values in J/mol K) for the Alkane Compounds Studied in This Work^a

Alkane molecules	N_E	S_{exp}^b	\bar{S}^c	$S_{conform}^d$	Ω_{term}	$S_{conform}^{term}$	$\bar{S} + \Delta S_{conform}^e$
1 hexane 	9	388.82 ±0.84	365.94	32.44	3 ²	18.27	380.11
2 2,2-dimethyl-butane 	1	358.65 ±0.84	357.77	45.67	3 ⁵	45.67	357.77
3 3-methyl-pentane 	5	382.88 ±0.67	365.27	43.11	3 ³	27.40	380.98
4 3-ethyl-pentane 	8	411.50	393.78	48.7	3 ³	27.40	415.08
5 2,2-dimethyl-pentane 	4	392.88	388.56	46.13	3 ⁵	45.67	389.02
6 2,4-dimethyl-pentane 	4	396.73	386.38	43.56	3 ⁴	36.54	393.40
7 2-methyl-hexane 	15	419.99	397.25	45.12	3 ³	27.40	414.96
8 3,3-dimethyl-pentane 	4	398.02	381.90	52.35	3 ⁴	36.54	397.71

^a The number of energetically indistinguishable conformers at $T = 0$ (Ω_{term}) and the number of energetically distinguishable conformers (N_E) at 298 K are also indicated. ^b From references: hexane,⁶⁸ 2,2-dimethyl-butane,⁶⁹ 3-methyl-pentane,⁷⁰ 3-ethyl-pentane,⁷¹ 2,2-dimethyl-pentane,⁷¹ 2,4-dimethyl-pentane,⁷¹ 2-methyl-hexane,⁷¹ and 3,3-dimethyl-pentane.⁷⁰ ^c $\bar{S} = S_{trans} + S_{rot}^{RR} + S_{vib}^{HO}$ and using the B3LYP/cc-pVTZ geometries and frequencies. The S_{rot} and S_{vib} contributions are averaged according to the relative frequency of the energetically distinguishable conformers during the classical AMBER MD simulations. ^d Including the fifth-order MIE conformational entropy derived from the classical MD trajectories. ^e $\Delta S_{conform} = S_{conform} - S_{conform}^{term}$ (see text for details).

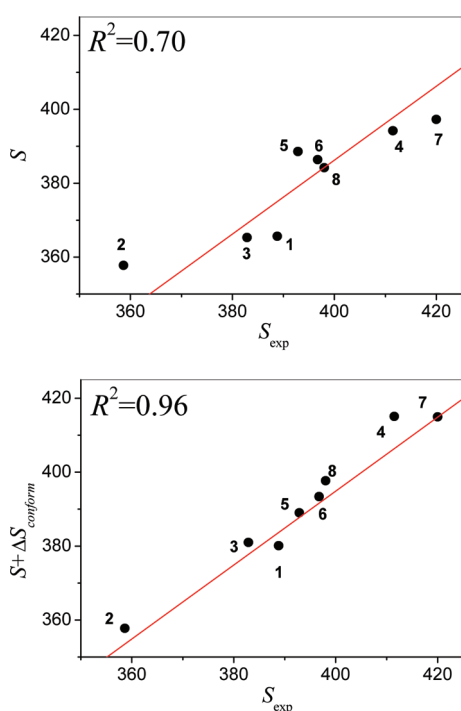


Figure 2. Correlation plots between theoretical absolute entropies (in J/(K mol)) for the eight alkane molecules considered in this work before (A) and after (B) adding the conformational entropy contribution to the mean RRHO entropies.

with respect to simulation time after ~ 1.5 and ~ 1.0 μs for 7 and 8, respectively, regardless of the expansion order. An even

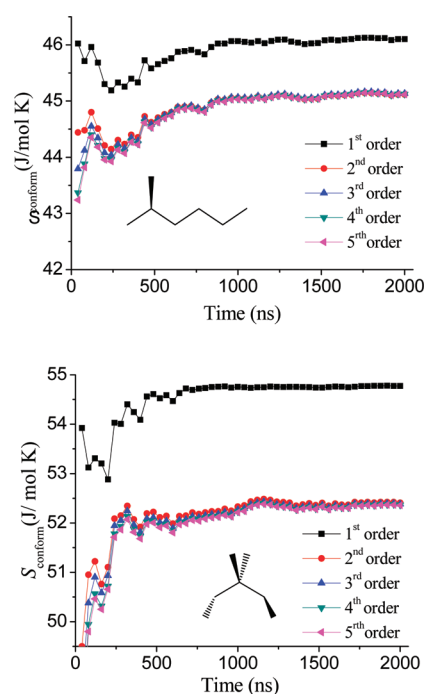


Figure 3. Convergence plots of the gas-phase conformational entropy (J/mol K) for 2-methyl-hexane and 3,3-dimethyl-pentane.

faster convergence was observed for the rest of the alkane molecules (see Figure S1 in the Supporting Information), and therefore, we conclude that the reported $S_{conform}$ values for these small alkane molecules essentially lack any statistical uncertainty. Although the

RRHO entropy is clearly the largest part, the S_{conform} values, which range between ~ 30 and ~ 50 J/(K mol), represent a significant contribution. Analyzing now the relative importance of correlation effects in S_{conform} , we find that the first-order MIE, which assumes that all torsion angles are independent variables, overestimates the conformational entropy. This behavior is not entirely unexpected because, for these small systems, the convergence plots in Figure 3 and Figure S1 (Supporting Information) strongly suggest that all correlation effects are effectively taken into account by our calculations, and as a consequence, the total entropy diminishes with respect to the first order value. However, we see in Figure 3 and Figure S1 that the entropy curves at first order recover most of the total conformational entropy; that is, entropy reduction due to correlation effects is rather small, only $1-5$ J/(K mol). Moreover, it turns out that this reduction is basically due to pair correlation, as the converged second order values are practically indistinguishable from the rest of the higher order S_{conform} entropies in all cases. Therefore, from these test calculations on the selected alkane molecules, we conclude that (a) S_{conform} is a significant entropic contribution and (b) the degree of correlation among the torsional degrees of freedom is rather low so that the second order MIE approximation gives sufficiently accurate values for the examined alkane molecules.

Before comparing our theoretical absolute entropies for the various C_6H_{14} and C_7H_{16} isomers with experimental data, it is necessary to realize that our S_{conform} calculations, which are based on classical MD simulations, imply that individual atoms in a covalently bound molecule are distinguishable particles. Therefore, in order to compare our data with experimental third-law entropies, we have to remove from the S_{conform} values the conformational entropy that arises from the number of possible rearrangements (Ω_{term}) that a single molecule can formally undergo through internal rotations about bonds to terminal symmetrical groups (e.g., $-\text{CH}_3$) without altering any molecular property. For instance, in the case of 2,2-dimethyl-butane, internal rotation of each of the four terminal methyl groups as well as around the C2–C3 bond generates three conformers per rotatable bond of identical energy and molecular properties, so that 2,2-dimethyl-butane has a total of $\Omega_{\text{term}} = 3^5$ possible intramolecular arrangements that would result in an entropy contribution $S_{\text{conform}}^{\text{term}} = R \ln \Omega_{\text{term}} = 45.67$ J/(K mol) (the $S_{\text{conform}}^{\text{term}}$ values for the rest of the alkane molecules can be likewise computed). The addition of the entropy differences, $\Delta S_{\text{conform}} = S_{\text{conform}} - S_{\text{conform}}^{\text{term}}$, to the RRHO mean values \bar{S} allow us to properly compare between theoretical and experimental data. Note also that entropic effects due to the presence of enantiomeric conformers are taken into account automatically by the S_{conform} calculations. We see in Figure 2 that the linearity of the correlation plot between the experimental and RRHO-based entropies is quite improved after having added the $\Delta S_{\text{conform}}$ term to \bar{S} . Thus, the squared correlation coefficient is now 0.96, and the intercept is around -2.6 J/(K mol).

The acceptable goodness of the linear fit between experimental and theoretical data in Figure 2 confirms the importance of S_{conform} and supports the usefulness of decomposing the entropy into its vibrational and conformational parts. In the Supporting Information, we present more calculations of the absolute entropies of the eight alkane molecules by using the above-mentioned “mixture of conformers” model.^{44,45} These calculations, which are based entirely on QM data, suggest that a significant fraction of the observed error in the RRHO&MIE conformational entropy calculations can emerge from small unbalances in the probability density functions of torsion angles. On the other hand, further

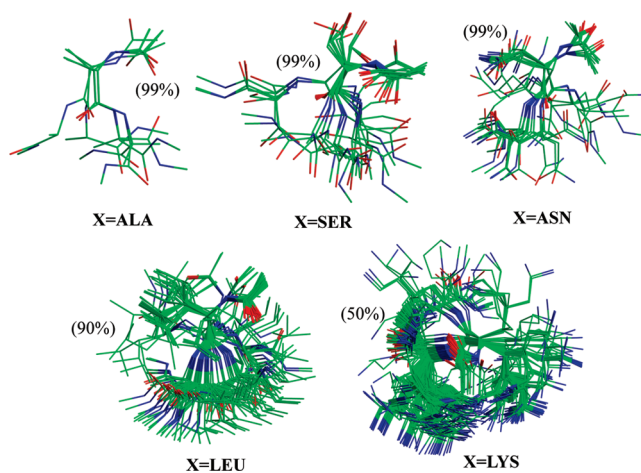


Figure 4. Superposition of the energetically distinguishable conformers of the Ace–X–X–Nme systems. The percentage of the MD snapshots represented by the conformers is given in parentheses.

support for applying the entropy decomposition can be found in previous works,^{44–49} following the “mixture of conformers” strategy. Therefore, it may also be relevant to comment here about the quality of the results of the “mixture of conformers” model. For example, the mean unsigned difference (MUD) between experimental and theoretical gas-phase entropies reported by Guthrie⁴⁷ for 128 organic compounds with up to 10 carbon atoms is 3.7 J/(K mol) (data derived from unscaled B3LYP/6-31G** frequencies and a semiempirical estimation of the ΔS_{mix} term). A similar MUD (4.4 J/(K mol)) was observed in our calculations on the eight alkane compounds.

Conformational Entropy of Dipeptides. The purpose of introducing the MIE approach within the context of the conformational entropy calculations is to capture correlation among torsional motions. For the alkane molecules studied in this work, it turns out that such correlation effects are almost negligible, and in any case, they are entirely accounted for by the pairwise approximation (i.e., second order correlation). Therefore, it is necessary to analyze more complex molecules in order to assess the ability of our approach for estimating higher order entropic contributions and its dependence on the dimensionality of the problem as defined by the number M of rotatable bonds. To this end, we examined five dipeptide molecules (Ace–X–X–Nme) of increasing size with X = Ala, Ser, Asn, Leu, and Lys (i.e., the peptides are capped by acetyl (Ace) and N-methyl (Nme) groups). The potential ability of these molecules to dynamically form and break intramolecular interactions through direct H-bond contacts or through-space electrostatic forces can introduce a significant correlation in their torsional motions and simultaneously maintain an important flexibility.

Figure 4 shows the superposition of the most populated conformers of the Ace–X–X–Nme systems that were obtained from minimizing 2000 equally spaced MD snapshots in each system and selecting the energetically distinguishable structures. As expected, all of the dipeptides are flexible molecules in the gas phase that experience frequent conformational transitions along the MD simulations. For the X = Ala system with $M = 8$ rotatable bonds, 99% of conformational variability is represented by only four structures, but the number of populated conformers grows up rapidly with M : the X = Lys system with $M = 16$ rotatable bonds populates more than 1200 different structures. Of course,

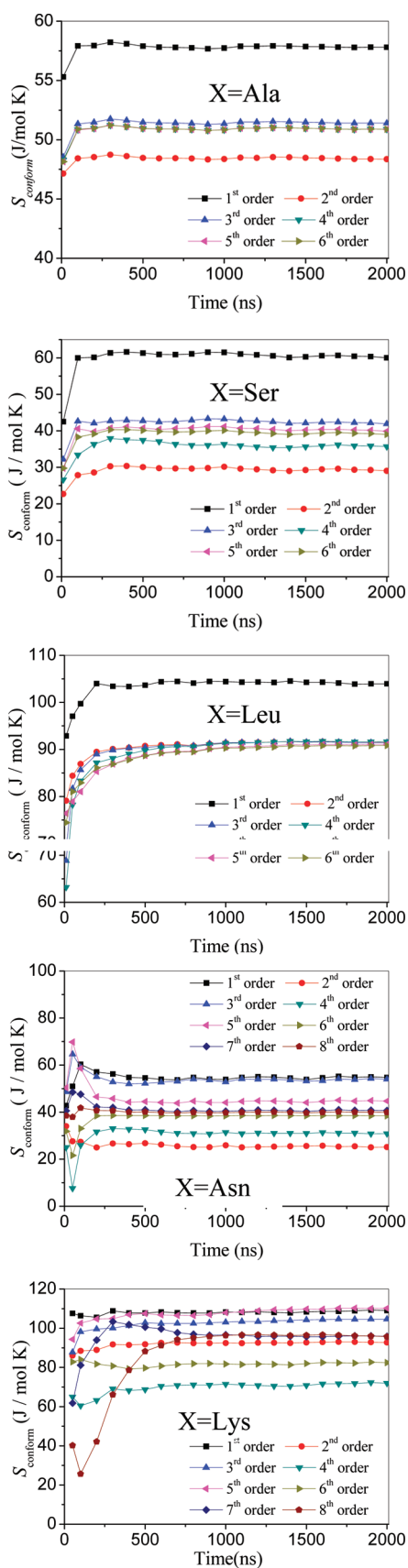


Figure 5. Convergence plots of the gas-phase conformational entropy ($J/(K \text{ mol})$) for the Ace–X–X–Nme molecules with X = Ala, Ser, Asn, Leu, and Lys.

the structure and relative abundance of the conformers are influenced by H-bond interactions that interconnect main chain groups (e.g., between the Ace–C=O carbonyl and the H–N–X amide at the second residue position) and/or side chain groups. Thus, we found that, in general, at least two or three H-bond contacts have abundances of about 55–65% so that they are quite stable interactions.

The conformational entropies at various orders of the Ace–X–X–Nme molecules are plotted along the simulation time in Figure 5. For the majority of the systems, it turns out that their S_{conform} profiles converge to nearly zero-slope curves after $\sim 0.5 \mu\text{s}$, thus suggesting that the MD simulations have exhaustively explored their phase space. However, the limiting S_{conform} values show a clear dependence on the expansion order n of the MIE method (see also Table 2). In consonance with expectations, the magnitude and relative weight of the n -order corrections increase with M on going from X = Ala ($M = 8$) to X = Lys ($M = 16$). Thus, with respect to the unimportance of high order correlation effects in Figure 3, we see now that both the intramolecular interactions and the partial rigidity introduced by the amidic bonds in the dipeptide molecules can have an impact on the conformational entropy thanks to the appearance of correlation among torsional motions. For the Ace–Ala–Ala–Nme system, which has only one more rotatable bond than the C_7H_{16} isomers and populates just a few conformers (see Figure 4), the second and third order corrections to S_{conform} are noticeable: -9.4 and $3.0 J/(K \text{ mol})$. These correlation effects are more clearly seen in the rest of the Ace–X–X–Nme systems as their S_{conform} curves span a range of tenths of a $J/(K \text{ mol})$. In order to achieve convergence in the S_{conform} values with respect to the expansion order n , MIE calculations up to sixth to eighth order were required for several systems (see Table 2). Moreover, we see in Figure 5 that n -order corrections to the conformational entropy fluctuate in both sign and magnitude depending on the molecular system, revealing thus the true complexity of the mutual information expansion. Nevertheless, the extended sampling (2×10^6 configurations) and the ability of the MIE calculations to estimate high order correlation effects allow us to obtain reasonably well-converged S_{conform} values for all of the Ace–X–X–Nme systems. In the case of the X = Asn and X = Lys systems, the difference between the S_{conform} estimations up to seventh and eighth order amounts to only $\sim 0.2 \text{ kJ/mol}$ in terms of free energies at 300 K. Similarly, the corresponding free energy differences for the S_{conform} values of the rest of the Ace–X–X–Nme dipeptides at the fifth and sixth orders are also negligible.

An interesting comparison can be made between the X = Leu ($M = 14$) and Asn ($M = 13$) systems. On one hand, the hydrophobic system has a considerable conformational entropy ($\sim 90 J/(K \text{ mol})$), which is in consonance with its relatively large number of accessible conformers. Curiously, high order ($n > 3$) corrections to S_{conform} are very small, which is very probably related to the fact that the Ace–Leu–Leu–Nme molecule forms only two intramolecular H-bond interactions with moderate abundances ($< 35\%$). On the other hand, much fewer conformers are populated during the MD trajectory of the X = Asn system, but correlation effects are now rather important, most likely because three to four H-bond interactions with abundances between 30% and 65% involving either side chain or backbone groups are constantly being formed and broken during the simulation, thus coupling the conformational changes of torsions separated by several covalent bonds. For the X = Lys system with $M = 16$ rotatable bonds, in addition to the relevance of correlation effects,

Table 2. Limiting Values of the Conformational Entropy at Various Orders (in J/(mol K)) for Dipeptides in the Gas Phase after 2.0 μ s of Simulation Time^a

	N_E	M	S_{conform}							
			1	2	3	4	5	6	7	8
Ace-(Ala) ₂ -Nme	14	8	57.80	48.36	51.40	50.88	50.88	50.93		
Ace-(Ser) ₂ -Nme	25	10	60.03	29.02	41.92	35.66	39.90	39.06		
Ace-(Asn) ₂ -Nme	69	13	54.76	25.14	54.00	30.86	44.70	38.41	40.66	39.90
Ace-(Leu) ₂ -Nme	232	14	103.96	91.49	91.44	91.64	91.10	90.84		
Ace-(Lys) ₂ -Nme	1357	16	109.08	92.70	104.64	71.97	110.18	82.41	95.91	95.74

^a The number of rotatable bonds considered in the conformational entropy calculations (M) and the number of energetically distinguishable conformers (N_E) observed in a sample of 2000 MD snapshots are also indicated. A total of 2×10^6 MD snapshots were used in all of the calculations.

its S_{conform} curves at high orders exhibit a slower convergence with respect to the simulation time (see Figure 5) as a consequence of the larger dimensionality of the problem, which demands a greater sampling effort to obtain reliable probability mass functions for all of the torsion angles and their combinations. Nevertheless, as mentioned above, the S_{conform} values for $X = \text{Lys}$ were converged to $\sim 96 \text{ J}/(\text{K mol})$ at orders $n = 7$ and 8, a value which happens accidentally to be quite close to the second order one (~ 93 ; see Table 2).

Absolute Entropy of the PFG Peptide. On the basis of the test calculations on the alkane molecules, it seems that there is no clear preference between the “mixture of conformers” and the conformational entropy frameworks for estimating the total entropy of small molecules. However, a sharp difference between the two formally equivalent approaches appears in larger biomolecules for which it is almost impossible to explore all their conformational space as required by the “mixture of conformers” approach. But as mentioned in the Introduction, the partitioning of the single molecule entropy, the discretization of the torsional degrees of freedom, and the adoption of the MIE technique can provide altogether a workaround to the sampling limitation for estimating the total entropy of complex molecular systems from the output provided by extensive MD simulations. To better illustrate the potential applicability of such an approach, we present here the results obtained for the PFG hexapeptide molecule, which is known to be a ligand of the MMP enzymes.

The 2.0 μ s MD trajectory of the PFG peptide in explicit solvent, which was started from the initial structure favored by the LMOD algorithm, populates two different conformational regions of the solute molecule characterized by radii of gyration of ~ 5.0 and $\sim 6.0 \text{ \AA}$, respectively (Figure S3 in the Supporting Information). Secondary structure analyses assign a helical conformation to the central (i.e., 2–6) residues in $\sim 55\%$ of the analyzed snapshots (we note in passing that the fact that PFG tends to adopt a helical structure seems in consonance with the ability of the MMPs to bind and hydrolyze collagen peptide chains that have also a helical structure). However, although PFG possesses some secondary structure, it still exhibits a relatively large dynamical flexibility through either its backbone or side chain motions, and therefore, it is conceivable that conformational entropy could be large enough to play a significant role in the free energy change upon binding of PFG to MMPs.³⁸

Most likely, the full theoretical understanding of the role played by entropy in the activity of PFG would require the computation of its absolute entropy in aqueous solution, as suggested by our previous calculations on the complexes formed between the MMP-2 enzyme and small peptide substrates.³⁸ In our approach,

such calculations could be achieved by combining the RRHO entropy of PFG using an implicit solvent model with the solute conformational entropy derived from the MD simulation so that the resulting entropy would be combined with other free energy terms as defined by approximate methodologies like the MM-PBSA protocol.⁵⁴ As the PFG hexapeptide contains 111 atoms and virtually all of the MD snapshots correspond to energetically distinguishable conformers, we decided to perform energy minimization and normal mode calculations on a subset of 10 000 MD snapshots using the NAB package and the AMBER03 force field (note that DFT calculations would be computationally too expensive). As shown in Figure 6A, although the RRHO entropies significantly fluctuate, the resulting time series over 2.0 μ s is rather stable and the corresponding mean value of S (1415 J/(K mol)) was estimated to within a standard error of only 0.3 J/(K mol). This small uncertainty suggests that the average RRHO entropy of PFG can be considered sufficiently converged for most purposes.

Turning our attention to the convergence plots of conformational entropy in Figure 6B derived from 10^6 MD snapshots, we note first that the S_{conform} values show a large dependence on the MIE order. In fact, the sum of marginal entropies leads to a limiting value of 137 J/(K mol) that is $\sim 60\%$ above the entropy estimations made from second to fifth order approximations. Given that PFG is a flexible molecule that contains a significant number of rotatable bonds ($M = 25$) and has several polar groups capable of forming H-bond interactions, the entropy reduction caused by correlation effects is well understood. However, for the same reasons, the S_{conform} calculations at the fourth and fifth orders now have a poor convergence with respect to simulation time: there remains an uncertainty of a few J/(K mol) in their limiting values after 2.0 μ s. Besides the sampling limitations at high order, it is also clear that the S_{conform} values of PFG have a non-negligible uncertainty with regard to the n order employed in the calculations. For example, the limiting value of S_{conform} at second order is $\sim 3 \text{ kJ/mol}$ in terms of free energies below that at third order.

Segregation of the conformational entropy into backbone and side chain contributions allows us to further analyze the origin of the large correlation effects in the dynamics of PFG. We see in Figures 6C,D that the backbone ($M = 11$) and side chain ($M = 14$) S_{conform} curves show an acceptable convergence at the various orders. Curiously, the entropy curves reflecting the conformational changes of the ψ and ϕ torsion angles present marked oscillations and converge more slowly than those of the amino acid side chains, indicating thus that the solvent-exposed side chains move faster than the backbone chain, and therefore, their motions are more efficiently sampled by the MD simulations. Correlation

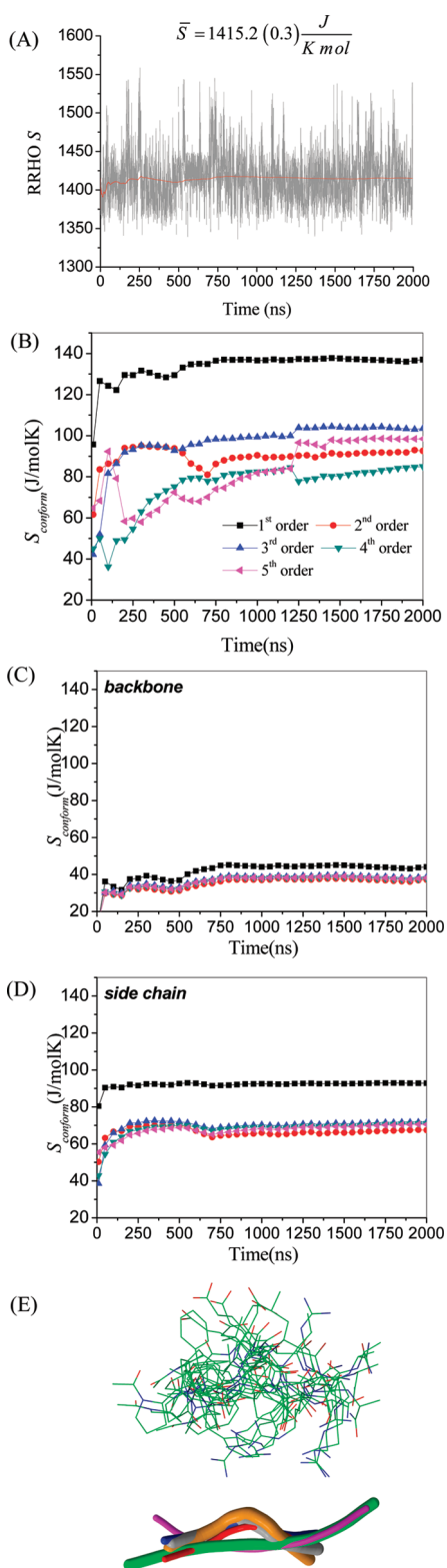


Figure 6. (A) RRHO entropy (in J/(K mol)) calculated for 10 000 snapshots extracted at 200 ps intervals from the 2.0 μ s MD simulations of PFG. The average value and its standard error in parentheses are also indicated. (B–D) Convergence plots of the PFG conformational entropy (in J/(K mol)). (E) Superposition of the most populated representative structures derived from clustering analyses both in wire-frame and ribbon model representation. Thickness of the ribbon models corresponds to the number of snapshots represented by each model.

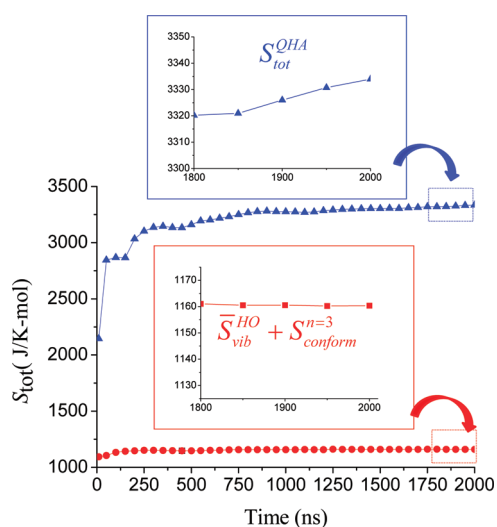


Figure 7. Convergence plots of the total entropy of PFG (excluding translation and rotation) estimated by the QHA method and the addition of the mean vibrational entropy and the conformational entropy at third order.

among the ψ and ϕ angles is only moderate (~ 6 J/(K mol)) and is largely captured by the second order approximation. Similarly, second- or third-order corrections to $S_{conform}$ are able to include most of the correlation effects due to the conformational motions of the side chains, but the magnitude of the concomitant entropy reduction is larger (~ 22 J/(K mol)). When comparing the plots in Figure 6B–D, it is clear that the source of the stronger correlation effects in the conformational entropy of the PFG peptide stem from the coupling between the backbone and side chain motions. We also see that, in terms of their convergence properties, the separate $S_{conform}$ values for the backbone and side-chain torsions are much more reliable than the global data as the segregated entropy plots reach stable plateaus and the free energy differences between the third-, fourth-, and fifth-order estimations are very small (~ 0.1 kJ/mol in terms of free energies).

Finally, we compare the additive entropy $\bar{S}_{vib}^{HO} + S_{conform}^{n=3}$ with the results provided by the quasi-harmonic approximation QHA, which constructs a pseudo-Hessian matrix directly from the covariance matrix of the Cartesian coordinates. To remove the overall translation of the center of mass and the overall rotation of the protein, the 2×10^6 MD snapshots employed in the QHA calculations were superposed on top of each other using a least-squares fit. As mentioned in the Introduction, the QHA method exhibits several disadvantages (e.g., neglecting supralinear correlations, approximating multimodal distributions to a unimodal one, etc.) that ultimately result in a large nonsystematic overestimation of S_{config} . This effect is clearly observed in Figure 7 as the QHA calculations lead to a very large entropy value (~ 3333 J/(K mol)). Although the entropy decomposition should give an upper limit to the true entropy, the $\bar{S}_{vib}^{HO} + S_{conform}^{n=3}$ value (~ 1160 J/(K mol)) is much lower than the QHA one even though the $-TS_{conform}$ contribution has a statistical uncertainty of several kJ/mol. Moreover, besides the larger overestimation of S_{config} the QHA calculations also exhibit worse convergence properties (see the two insets in Figure 7).

From the results of the test calculations on PFG, it can be concluded that the total conformational entropy has a considerable weight (10%) in its single-molecule entropy and that the effects of dynamic correlations among torsional angles are far

from being small. These calculations point out that both under-sampling and poor convergence with respect to the MIE order are two closely related problems of the S_{conform} calculations that need to be assessed in practical applications. However, although the uncertainty in the total S_{conform} of PFG can have an impact of several kJ/mol on free energy, we believe that meaningful results can be achieved from partially converged entropy curves like those shown in Figures 6B–D for PFG. For example, computation of entropy differences (e.g., upon peptide binding to a host molecule) could benefit from partial cancellation of errors and provide approximately constant ΔS values at different MIE orders.^{22,38} Alternatively, consideration of a subset of torsion angles (e.g., backbone ψ and ϕ angles) could be enough for capturing the relative change in entropy occurring in peptide folding or molecular association processes. In any case, an intensive MD sampling followed by the estimation of conformational entropy differences including correlation effects would be required to fully understand the thermodynamical forces controlling many biomolecular processes that alter the conformational dynamics of the involved molecules. In this respect, the present test calculations support the ability of our approach for including high order correlation effects that have been neglected so far in most of the previous studies using nonparametric methods for estimating the configurational entropy.

SUMMARY AND CONCLUSIONS

In this work, we have pursued the implementation of the partitioning of the intramolecular entropy into vibrational and conformational contributions as originally proposed by Karplus et al., in order to estimate the absolute entropy of single biomolecules from MD simulations. In our approach, a key element consists of the characterization of the conformational state of a given molecule by means of the discretization of the time evolution of its torsional motions about single bonds. This process leads naturally to the computation of the conformational entropy as Shannon entropy, which is subsequently added to the average translational–rotational–vibrational entropy computed by means of normal model calculations on a series of MD snapshots. To help overcome sampling limitations in the computation of the conformational probability mass function of large molecules, we use the mutual information expansion, which has also been employed in other entropy methods, to systematically include correlation effects among torsion angles. Although this protocol has been developed for treatment with relatively large systems, it must be emphasized that its core assumption, that is, the combination of the RRHO entropy with the conformational entropy, is formally equivalent to the “mixture of conformers” strategy that has been routinely used to predict absolute entropies of small molecules with good accuracy.

On the basis of the different test calculations that have been presented in this work, we can draw the following conclusions regarding the applicability and/or reliability of our approach: (a) The gross of the absolute entropy is computed with the RRHO approximation, which constitutes a straightforward computational protocol that avoids the need of discriminating between stiff or soft degrees of freedom. (b) From a quantitative point of view, the combination of the RRHO entropy with the conformational (or mixing) entropy yields results that are quite close to experimental data, as shown by our calculations on the alkane molecules or by other results previously reported in the literature. (c) For computing reliable conformational entropies of small- or

medium-sized molecules that exhibit a rich dynamical behavior, it is essential to capture correlation effects arising from the coupling of torsional motions through intramolecular interactions, and in this respect, the use of the MIE method together with the discretization of the torsional motions constitutes an interesting alternative capable of computing high order corrections quite efficiently. (d) Segregation of conformational entropy into different components (e.g., backbone and side chain terms) can be easily implemented, thus revealing the origin and relative importance of correlation effects.

Finally, it is also important to comment on some limitations of the RRHO-conformational entropy calculations on the basis of the results obtained for the PFG polypeptide. Thus, it is clear that the RRHO&MIE calculations for this kind of system may demand a considerable amount of computer time, particularly if energy minimizations and normal mode calculations have to be carried out on systems containing thousands of atoms, even though an MM method was employed. Similarly, the computation of conformational entropy corrections at high orders ($n > 5-8$) can be rather expensive too. In addition, application of the present approach to large molecules can suffer from convergence issues with respect to the simulation time needed to extract the probable mass functions of individual torsions and groups of torsions, and with reference to the MIE order that is required to capture correlation effects. We believe that the calculation of relative conformational entropies rather than absolute ones using low order approximations and/or for a subset of torsions may benefit from error cancellation. However, the problems of the RRHO&MIE technique should be mitigated by increasing the MD sampling, which in turn, is becoming more and more accessible thanks to the continuous improvement in the efficiency in computer hardware and simulation algorithms. Moreover, the fact that the current protocol has been shown to exhibit a much better convergence behavior than the QHA method in the case of the polypeptide entropy calculation is a promising result that should stimulate further methodological and computational experimentation aimed at overcoming the computational bottlenecks and/or convergence problems.

ASSOCIATED CONTENT

S Supporting Information. Derivation of the entropy partitioning (eq 8). Mathematical proof of the equivalence between eqs 9 and 10. Figure S1 showing convergence plots of conformational entropies for various alkane molecules. Figure S2 and Tables S1 and S2 summarizing the results of the “mixtures of conformers” entropy calculations on the alkane molecules. Figure S3 showing the time evolution of the PFG radius of gyration. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: +34-985103689. Fax: +34-985103125. E-mail: dimas@uniovi.es.

ACKNOWLEDGMENT

This research was supported by the following grants: FICyT (Asturias, Spain) IB05-076 and MEC (Spain) CTQ2007-63266. E.S. thanks MEC for his FPU contract.

REFERENCES

- (1) Carlsson, J.; Aqvist, J. Calculations of Solute and Solvent Entropies from Molecular Dynamics Simulations. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5385.
- (2) Brady, G. P.; Sharp, K. E. Entropy in Protein Folding and in Protein-Protein Interactions. *Curr. Opin. Struct. Biol.* **1997**, *7*, 215.
- (3) Fitter, J. A. Measure of Conformational Entropy Change During Thermal Protein Unfolding Using Neutron Spectroscopy. *Biophys. J.* **2003**, *84*, 3924.
- (4) Bachmann, A.; Kiefhaber, T.; Boudko, S.; Engel, J.; Bächinger, H. P. Collagen Triple-Helix Formation in All-Trans Chains Proceeds by a Nucleation Growth Mechanism with a Purely Entropic Barrier. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13897.
- (5) Creamer, T. P.; Rose, G. D. Side-Chain Entropy Opposes α -Helix Formation but Rationalizes Experimentally Determined Helix-Forming Propensities. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 5937.
- (6) Chapagain, P. P.; Parra, J. L.; Gerstman, B. S.; Liu, Y. Sampling of States for Estimating the Folding Funnel Entropy and Energy Landscape of a Model α -Helical Hairpin Peptide. *J. Chem. Phys.* **2007**, *127*, 075103.
- (7) Choa, S. S.; Levya, Y.; Wolynesa, P. G. Quantitative Criteria for Native Energetic Heterogeneity Influences in the Prediction of Protein Folding Kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 434.
- (8) Schäfer, H.; Daura, X.; Mark, A. E.; van Gunsteren, W. F. Entropy Calculations on a Reversibly Folding Peptide: Changes in Solute Free Energy Cannot Explain Folding Behavior. *Proteins: Struct., Funct., Genet.* **2001**, *56*, 43.
- (9) Grünberg, R.; Nilges, M.; Leckner, J. Flexibility and Conformational Entropy in Protein-Protein Binding. *Structure* **2006**, *14*, 683.
- (10) Diehl, C.; Genheden, S.; Modig, K.; Ryde, U.; Akke, M. Conformational Entropy Changes Upon Lactose Binding to the Carbohydrate Recognition Domain of Galectin-3. *J. Biomol. NMR* **2009**, *45*, 157.
- (11) Stone, M. J. Nmr Relaxation Studies of the Role of Conformational Entropy in Protein Stability and Ligand Binding. *Acc. Chem. Res.* **2001**, *34*, 379.
- (12) Chang, C. A.; Chen, C.; Gilson, M. K. Ligand Configurational Entropy and Protein Binding. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534.
- (13) Baron, R.; McCammon, J. A. (Thermo)Dynamic Role of Receptor Flexibility, Entropy, and Motional Correlation in Protein-Ligand Binding. *ChemPhysChem* **2008**, *9*, 983.
- (14) Killian, B. J.; Yudenfreund-Kravitz, J.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K. Configurational Entropy in Protein-Peptide Binding: Computational Study of Tsg101 Ubiquitin E2 Variant Domain with an Hiv-Derived Ptap Nonapeptide. *J. Mol. Biol.* **2009**, *389*, 315.
- (15) Karplus, M.; Kushick, J. N. Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules* **1981**, *14*, 325.
- (16) Edholm, O.; Berendsen, H. J. C. Entropy Estimation from Simulations of Non-Diffusive Systems. *Mol. Phys.* **1984**, *51*, 1011.
- (17) Schlitter, J. Estimation of Absolute and Relative Entropies of Macromolecules Using the Covariance Matrix. *Chem. Phys. Lett.* **1993**, *215*, 617.
- (18) Andricioaei, I.; Karplus, M. On the Calculation of Entropy from Covariance Matrices of the Atomic Fluctuations. *J. Chem. Phys.* **2001**, *115*, 6289.
- (19) Baron, R.; Gunsteren, W. F. v.; Hünenberger, P. H. Estimating the Configurational Entropy from Molecular Dynamics Simulations: Anharmonicity and Correlation Corrections to the Quasi-Harmonic Approximation. *Trends Phys. Chem.* **2006**, *11*, 87.
- (20) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. Nearest-Neighbor Nonparametric Method for Estimating the Configurational Entropy of Complex Molecules. *J. Comput. Chem.* **2006**, *28*, 655.
- (21) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. Extraction of Configurational Entropy from Molecular Simulations Via an Expansion Approximation. *J. Chem. Phys.* **2007**, *127*.
- (22) Suárez, E.; Díaz, N.; Suárez, D. Entropic Control of the Relative Stability of Triple-Helical Collagen Peptide Models. *J. Phys. Chem. B* **2008**, *112*, 15248.
- (23) Li, D.-W.; Bruschiweiler, R. In Silico Relationship between Configurational Entropy and Soft Degrees of Freedom in Proteins and Peptides. *Phys. Rev. Lett.* **2009**, *102*, 118108.
- (24) Hensen, U.; Grubmüller, H.; Lange, O. F. Adaptive Anisotropic Kernels for Nonparametric Estimation of Absolute Configurational Entropies in High-Dimensional Configuration Spaces. *Phys. Rev. E* **2009**, *80*, 011913.
- (25) Hensen, U.; Lange, O. F.; Grubmüller, H. Estimating Absolute Configurational Entropies of Macromolecules: The Minimally Coupled Subspace Approach. *PLoS ONE* **2010**, *5*, e9179.
- (26) *Free Energy Calculations. Theory and Applications in Chemistry and Biology*; Chipot, C., Pohorille, A., Eds.; Springer-Verlag: Berlin, 2007.
- (27) Meirovitch, H. Methods for Calculating the Absolute Entropy and Free Energy of Biological Sys. *J. Mol. Recognit.* **2010**, *2*, 153.
- (28) Meirovitch, H.; Chelvaraja, S.; White, R. P. Methods for Calculating the Entropy and Free Energy and Their Application to Problems Involving Protein Flexibility and Ligand Binding. *Curr. Protein Pept. Sci* **2009**, *10*, 229.
- (29) Schäfer, H.; Mark, A. E.; Gunsteren, W. F. v. Absolute Entropies from Molecular Dynamics Simulation Trajectories. *J. Chem. Phys.* **2000**, *113*, 7809.
- (30) Chang, C.; Chen, W.; Gilson, M. K. Evaluating the Accuracy of the Quasiharmonic Approximation. *J. Chem. Theory Comput.* **2005**, *1*, 1017.
- (31) Cover, T. M.; Thomas, J. C. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, 2006.
- (32) Di Nola, A.; Berendsen, H. J. C.; Edholm, O. Free Energy Determination of Polypeptide Conformations Generated by Molecular Dynamics. *Macromolecules* **1984**, *17*, 2044.
- (33) Baron, R.; Hünenberger, P. H.; McCammon, J. A. Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties. *J. Chem. Theory Comput.* **2009**, *5*, 3150.
- (34) Numata, J.; Wan, M.; Knapp, E. Conformational Entropy of Biomolecules: Beyond the Quasi-Harmonic Approximation. *Genome Inform.* **2007**, *18*, 192.
- (35) Gorla, M. N.; Leonenko, N. N.; Mergel, V. V.; Novi-Inverardi, P. L. A New Class of Random Vector Entropy Estimators and Its Applications in Testing Statistical Hypotheses. *J. Nonparametr. Stat.* **2005**, *17*, 277.
- (36) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. Efficient Calculation of Configurational Entropy from Molecular Simulations by Combining the Mutual-Information Expansion and Nearest-Neighbor Methods. *J. Comput. Chem.* **2008**, *29*, 1605.
- (37) Ohkubo, Y. K.; Thorpe, I. F. Evaluating the Conformational Entropy of Macromolecules Using an Energy Decomposition Approach. *J. Chem. Phys.* **2006**, *124*, 024910.
- (38) Díaz, N.; Suarez, D.; Suarez, E. Kinetic and Binding Effects in Peptide Substrate Selectivity of Matrix Metalloproteinase-2: Molecular Dynamics and Qm/Mm Calculations. *Proteins* **2010**, *78*, 1.
- (39) Karplus, M.; Ichiye, T.; Pettit, B. M. Configurational Entropy of Native Proteins. *Biophys. J.* **1987**, *52*, 1083.
- (40) Zhou, H.-X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **2009**, *109*, 4092.
- (41) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Sausalito, CA, 2000.
- (42) Matsuda, H. Physical Nature of Higher-Order Mutual Information: Intrinsic Correlations and Frustration. *Phys. Rev. E* **2000**, *62*, 3098.
- (43) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; The University of Illinois Press: Urbana, IL, 1964.
- (44) DeTar, D. F. Theoretical Ab Initio Calculation of Entropy, Heat Capacity, and Heat Content. *J. Phys. Chem. A* **1998**, *102*, 5128.
- (45) DeTar, D. F. Calculation of Entropy and Heat Capacity of Organic Compounds in the Gas Phase. Evaluation of a Consistent Method without Adjustable Parameters. Applications to Hydrocarbons. *J. Phys. Chem. A* **2007**, *111*, 4464.
- (46) Block, D. A.; Armstrong, D. A.; Rauk, A. Gas Phase Free Energies of Formation and Free Energies of Solution of Rc-Centered Free Radicals from Alcohols: A Quantum Mechanical-Monte Carlo Study. *J. Phys. Chem. A* **1999**, *103*, 3562.

- (47) Guthrie, J. P. Use of Dft Methods for the Calculation of the Entropy of Gas Phase Organic Molecules: An Examination of the Quality of Results from a Simple Approach. *J. Phys. Chem. A* **2001**, *105*, 8495.
- (48) Bouchoux, G.; Bimbong, R. G.-B.; Nacer, F. Gas-Phase Protonation Thermochemistry of Glutamic Acid. *J. Phys. Chem. A* **2009**, *113*, 6666.
- (49) Bouchoux, G.; Bourcier, S.; Blanc, V.; Desaphy, S. Gas Phase Protonation Thermochemistry of Phenylalanine and Tyrosine. *J. Phys. Chem. B* **2009**, *113*, 5549.
- (50) East, A. L. L.; Radom, L. Ab Initio Statistical Thermodynamical Models for the Computation of Third-Law Entropies. *J. Chem. Phys.* **1997**, *106*, 6665.
- (51) Brown, R. A.; Case, D. A. Second Derivatives in Generalized Born Theory. *J. Comput. Chem.* **2006**, *27*, 1662.
- (52) Scott, A. P.; Radom, L. Harmonic Vibrational Frequencies: An Evaluation of Hartree–Fock, Møller–Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *J. Phys. Chem.* **1996**, *100*, 16502.
- (53) Johnson, R. D.; Irikura, K. K.; Kacker, R. N.; Kessel, R. Scaling Factors and Uncertainties for Ab Initio Anharmonic Vibrational Frequencies. *J. Chem. Theory Comput.* **2010**, *6*, 2822.
- (54) Gohlke, H.; Case, D. A. Converging Free Energy Estimates: Mm-Pb(Gb)Sa Studies on the Protein–Protein Complex Ras–Raf. *J. Comput. Chem.* **2003**, *25*, 238.
- (55) Lee, M. S.; Olson, M. A. Calculation of Absolute Protein–Ligand Binding Affinity Using Path and Endpoint Approaches. *Biophys. J.* **2009**, *90*, 864.
- (56) Taylor, C. C. Automatic Bandwidth Selection for Circular Density Estimation. *Comput. Stat. Data An.* **2008**, *52*, 3493.
- (57) Suárez, E.; Suárez, D. Manuscript in preparation.
- (58) Case, D. A.; Darden, T. A.; Cheatham, I. T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, K. F.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER*; 10th ed.; University of California: San Francisco, 2008.
- (59) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; R, L.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668.
- (60) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *14*, 1999.
- (61) 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*, C.02 ed.; Gaussian Inc.: Wallingford, CT, 2004.
- (62) Turk, B. E.; Huang, L. L.; Cantley, L. C. Determination of Protease Cleavage Site Motifs Using Mixture-Based Oriented Peptide Libraries. *Nat. Biotechnol.* **2001**, *19*, 661.
- (63) Kolossváry, I.; Guida, W. C. Low Mode Search. An Efficient, Automated Computational Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and Cyclic Peptides. *J. Am. Chem. Soc.* **1996**, *118*, 5011.
- (64) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate–DNA Helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401.
- (65) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648.
- (66) Dunning, T. H., Jr. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007.
- (67) *Modeling Unusual Nucleic Acid Structures*; Macke, T., Case, D. A., Eds.; American Chemical Society: Washington, DC, 1998.
- (68) Scott, D. W. Correlation of the Chemical Thermodynamic Properties of Alkane Hydrocarbons. *J. Chem. Phys.* **1974**, *60*, 3144.
- (69) Kilpatrick, J. E.; Pitzer, K. S. The Thermodynamics of 2,2-Dimethylbutane, Including the Heat Capacity, Heats of Transition, Fusion and Vaporization and the Entropy. *J. Am. Chem. Soc.* **1946**, *68*, 1066.
- (70) Finke, H. L.; Messerly, J. F. 3-Methylpentane and 3-Methylheptane: Low-Temperature Thermodynamic Properties. *J. Chem. Thermodyn.* **1973**, *5*, 247.
- (71) Huffman, H. M.; Gross, M. E.; Scott, D. W.; McCullough, J. P. Low Temperature Thermodynamic Properties of Six Isomeric Heptanes. *J. Phys. Chem.* **1961**, *65*, 495.

Automated Repulsive Parametrization for the DFTB Method

Zoltán Bodrog,* Bálint Aradi, and Thomas Frauenheim

Bremen Center for Computational Materials Science, University of Bremen, Am Fallturm 1, 28359 Bremen, Germany

ABSTRACT: The density-functional-based tight-binding method is an efficient scheme for quantum mechanical atomistic simulations. While the most relevant part of the chemical energies is calculated within a DFT-like scheme, a fitted correction function—the repulsive energy—is used to achieve results as close to *ab initio* counterparts as possible. We have developed an automatic parametrization scheme to ease the process of the repulsive energy fitting, offering a more systematic and much faster alternative to the traditional fitting process. The quality of the resulting repulsives can be tuned by selecting and weighting the fit systems and the important physical properties (energy, force, Hessian) of them. Besides driving DFT calculators in the fitting process automatically, the flexibility of our scheme also allows the usage of external data (e.g., molecular dynamics trajectories or experimental data) as a reference. Results with several elements show that our procedure is able to produce parameter sets comparable to handmade ones, yet requiring far less human effort and time.

1. INTRODUCTION

The density-functional-based tight-binding method (DFTB)¹ is an efficient quantum mechanical simulation method, which is an approximation to the density functional theory (DFT). While being typically orders of magnitude faster than its *ab initio* counterpart, it delivers results for many chemical problems with reasonable accuracy. The rigorously derived original DFTB method provides an excellent theoretical framework which can be systematically extended when higher accuracy is needed or some new chemical features should be described which could not be covered by previous schemes. The success of the framework can be judged by the huge amount of different systematic extensions which had been created over time (e.g., charge self-consistency for describing charge transfer,² inclusion of collinear and noncollinear spin,³ time-dependent⁴ and GW⁵ formalism to calculate excited state properties, Green-function technique to describe electron transport,⁶ etc.).

Common in all different DFTB extensions is the fact, that only the “most important” part of the total energy is calculated within an approximate quantum mechanical approach, while the rest, comprising the core–core repulsion and the double-counting terms, is taken into account as a fitted quasi-classical interaction energy (the so-called repulsive energy) between participating atoms, depending on the configuration of the atomic nuclei in the system. When carefully done, the fitted repulsive interaction can even compensate for parts of the error introduced with the approximations in the quantum-mechanically calculated parts of the DFTB energy. This division and the approximations in the quantum-mechanically calculated part allows calculations on chemical systems typically several orders of magnitude faster and using considerably less memory than *ab initio* calculations, while still maintaining a reasonable accuracy.

Due to its presence in all extensions and its effect on the accuracy of the total energies, the fitting of the repulsive interactions is a cardinal problem for the original DFTB scheme and all its extensions. However, the parametrization process for a broad range of chemical species is a rather tedious work, often taking months of valuable research time. Additionally, due to its

pairwise nature, the work necessary to extend an existing set with a new element increases with the set size, as the interaction of all elements with the new one has to be created.

To lower the barrier of extending DFTB to new chemical systems, several attempts have been made. First, Knaup et al. demonstrated their evolutionary algorithm work at fitting specific repulsives for the proton transfer in imidazole.⁷ Quite similar to our work and in time parallel to our early steps, an automated parametrization engine had been created by Gaus et al.⁸ that is able to fit repulsive energies in molecules. Its applicability has been demonstrated by fitting repulsives for carbohydrogen interactions, giving an accuracy comparable to the mio parameter set.⁹ Unfortunately, the fitting framework seems to have several limitations. First, it only considered molecular systems, making parametrizations for solids and surfaces rather difficult. Furthermore, it does not seem to have included any means of mass fit data production, i.e., does not seem to be capable of generating large energetic data sets with reasonably little human intervention. This is not an issue when the fit is done against experimental data, but it maintains a part of the “parametrization barrier” by making the usage of DFT calculator references and scanning off-equilibrium reference data much more difficult than necessary. Last, it is built around a fixed spline representation of the repulsives, which gives a limited flexibility.

In this paper, we describe a comprehensive automated fitting process for creating repulsive pair potentials. It was developed by trying to simplify the fitting procedure as far as possible while still keeping its applicability to almost any kind of chemical situation where one expects reasonable description by DFTB. We designed our algorithm to deal with a large variety of repulsive function shapes and to fit to energetic properties of not only molecules but also crystalline systems. Our algorithm contains interactive parts for not only the core fitting process but also defining and building fit and test systems and making large series of batch fits with various changing metaparameters (parameters

Received: May 13, 2011

Published: July 05, 2011

affecting the parametrization, e.g., preliminarily chosen cutoffs or polynomial degrees). It was designed to model the workflow and needs of an applied scientist creating new parameters, up to the point that it can be integrated into graphical chemistry and materials science program packages as a GUI-driven module. The applicability of the algorithm has been demonstrated by creating repulsive interactions for carbohydrates, titanium-organic chemistry, and crystalline zinc-oxide compounds. As it will be shown later, the obtained sets are of comparable quality to the current best handmade ones for those systems, while significantly reducing the human effort involved in their creation.

The paper is structured as follows: first, we give a short overview of the DFTB method and the parametrization problem. Then, we introduce our parametrizer automaton in detail. This is followed by a comparison of the automatically created sets mentioned above against their existing handmade counterparts.

2. METHODOLOGICAL BACKGROUND

2.1. The DFTB Method. The original, noniterative, perturbative DFTB¹ is a tight-binding DFT method used to calculate electronic structures of chemical systems. It solves the Kohn–Sham equations

$$\hat{H}_{\text{KS}}|\mu\rangle = [\hat{T} + \hat{V}_{\text{KS}}]|\mu\rangle = \varepsilon_{\mu}|\mu\rangle \quad (1)$$

which in an LCAO basis of Slater-type orbitals lead to

$$\sum_{\chi} H_{\phi\chi} c_{\mu,\chi} = \varepsilon_{\mu} \sum_{\omega} S_{\phi\omega} c_{\mu,\omega} \quad (2)$$

where

$$|\mu\rangle = \sum_{\phi} c_{\mu,\phi} |\phi\rangle, S_{\phi\omega} = \langle\phi|\omega\rangle, \\ H_{\phi\chi} = \langle\phi|\hat{T} + \hat{V}_{\text{KS}}|\chi\rangle \quad (3)$$

with the effective Kohn–Sham potential

$$V_{\text{KS}}(\mathbf{r}) = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' + \frac{\delta E_{\text{xc}}[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})} + V_{\text{ext}}(\mathbf{r}) \quad (4)$$

and V_{ext} being the potential of the atomic cores. The effective potential is decomposed into atom-centered contributions, and the integrals in the Hamiltonian matrix are calculated by taking only two-center terms into account. Since the two-center contributions can be tabulated *in advance*, the DFTB algorithm is extremely fast, but lacking any direct energy contributions involving more than two atoms (e.g., $\langle\phi|\hat{V}_{\text{KS}}|\chi\rangle$ -like terms having ϕ , V_{KS} , and χ from three different atoms).

Using compressed atomic orbitals, one can already get good results for many systems with the non-self-consistent application of the above scheme. In this perturbative approximation, the Kohn–Sham or electronic energy of the system, up to the pairwise approximation of the Hamiltonian, is

$$E_{\text{el}} = \sum_{\mu} \sum_{\phi,\chi}^{\text{occ}} \bar{c}_{\mu,\phi} H_{\phi\chi}^{(0)} c_{\mu,\chi} = T \\ + \int \left(\int \frac{\rho^{(0)}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' + \frac{\delta E_{\text{xc}}}{\delta\rho(\mathbf{r})} \Big|_{\rho^{(0)}} + V_{\text{ext}}(\mathbf{r}) \right) \rho(\mathbf{r}) d\mathbf{r} \quad (5)$$

with the first sum running over all occupied states (T is the total electronic kinetic energy). The total energy is

$$E = E_{\text{el}} - \frac{1}{2} \int \int \frac{\rho^{(0)}(\mathbf{r})\rho^{(0)}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}' \\ - \int \frac{\delta E_{\text{xc}}}{\delta\rho(\mathbf{r})} \Big|_{\rho^{(0)}} \rho^{(0)}(\mathbf{r}) d\mathbf{r} + E_{\text{xc}}^{(0)} + E_{\text{core}}^{(0)} \quad (6)$$

Here, E_{core} is the interaction energy of the atomic cores and the superscript (0) denotes quantities calculated from the nonperturbed superposition of starting atomic charge densities. The perturbative nature of the scheme ensures that the parts of the Hamiltonian and therefore the double-counting terms do not depend on the perturbatively calculated ρ charge density, but only on $\rho^{(0)}$.

As the ρ dependence and all sophisticated electronic properties are included in the E_{el} electronic part, the rest of the DFTB total energy can be treated as an effective potential between atomic nuclei, the so-called repulsive energy. Due to its corrective nature, we handle it in a simplified approach and break it down to a sum of pairwise potentials between the atoms:¹

$$E_{\text{rep}} = E - E_{\text{el}} \approx \frac{1}{2} \sum_{i,j}^{\text{nuclei}} U_{\text{type}(ij)}(r_{ij}) \quad (7)$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between atoms i and j .

A systematic enhancement to the original DFTB scheme is the self-consistent-charges (SCC) extension,⁹ which makes the Hamiltonian depend on electronic density via a construction representing the charge fluctuations between atoms with point-like charges. This correction with respect to the total energy of the non-SCC DFTB is contained in the SCC total energy expression as

$$\Delta E = \frac{1}{2} \sum_{i,j} \gamma_{i,j} \Delta q_i \Delta q_j \quad (8)$$

where the γ 's are the effective interaction profiles of spherically symmetric diffuse charges, q_i is the Mulliken charge of atom i , and Δq_i is its change with respect to the neutral atomic population. These γ 's give back the Coulombic $1/r$ profile in large distances as well as the atomic chemical hardness at $r \rightarrow 0$. The above energy correction is realized in the Kohn–Sham Hamiltonian by

$$\Delta H_{\phi\chi} = \frac{1}{2} S_{\phi\chi} \sum_j (\gamma_{[\phi],j} + \gamma_{[\chi],j}) \Delta q_j \quad (9)$$

where $[\phi]$ and $[\chi]$ represent the atomic centers of orbitals ϕ and χ , respectively.

Having a Hamiltonian depending on molecular charge distribution makes self-consistent iterative calculations possible. SCC-DFTB needs a reparametrization with respect to the non-SCC one, however, as E_{el} and thus the difference between DFT total energy and E_{el} changes with the addition of self-consistency.

2.2. Parametrization with Pair Potentials. According to eq 7, the repulsive energy is broken down to pairwise potentials:

$$E_{\text{rep}}(\{\mathbf{r}_{\text{atoms}}\}) = \sum_{i < j} U_{AB}(r_{ij}) \quad (10)$$

where i and j both run over the atoms in the system and AB indicates the type of atom pair ij .

The parametrization process optimizes these $U_{AB}(r)$ pair potentials to cover the difference between the reference energies of certain fit systems and the corresponding electronic DFTB energy. The reference energies may be taken from experimental data or *ab initio* calculations. We prefer the latter, as it allows versatile reference data generation. The best parametrization can be viewed as the one where the set of pair potentials minimizes the error:

$$R = \sum (E_{\text{ref}} - E_{\text{DFTB}})^2 = \sum_{i < j} (U_{AB}(r_{ij}) - (E_{\text{ref}} - E_{\text{el}}))^2 = \min \quad (11)$$

Due to the approximative nature of DFTB, parametrizations lack universal transferability, but as the cases of successful parametrizations show, the validity of a good parameter set can extend to a wide range of problems.

2.3. Hand-Made Repulsive Potentials. In its usual course, parametrization for a bond type (e.g., the carbon–carbon bond) begins with stretching one bond of that kind in an appropriate molecule, as the simplest case, and creating a $U_{CC}(r)$ curve based on the energy difference between the DFT reference and DFTB:

$$U_{CC}(r) = E_{\text{DFT}}(r) - E_{\text{el}}(r) + \text{const} \quad (12)$$

with r being the length of the stretched bond. The constant term covers the limit of $E_{\text{DFT}}(r) - E_{\text{el}}(r)$ at $r \rightarrow \infty$ (this limit contains, e.g., the repulsive contributions from nonvarying bonds, that may not be known in detail at all) in order to ensure a zero limit for $U_{CC}(r)$, $r \rightarrow \infty$. Of course, one chooses stretched molecules so that stretching affects only one bond (or maybe several bonds, but in a totally equivalent way); all of the other pairs of atoms with changing distances remain outside the ranges of their respective repulsives.

One can construct a reasonable curve for a given interaction by merging curve sections created for different molecules which represent different chemical bonds between the considered elements. For example, a carbon–carbon pair potential can be constructed by taking the sections near 1.2 Å, 1.34 Å, and 1.54 Å from ethyne, ethene, and ethane, respectively, in order to take single, double, and triple carbon bonds into account. The resulting compound curves can then be heuristically improved by comparing DFTB results on some test systems to DFT data and fine-tuning them by hand. Unfortunately, the fine-tuning involves a tremendous amount of human work, making the fast extension of a given set or creating a new set from scratch rather difficult.

3. AUTOMATIC PARAMETRIZATION SCHEME

In order to reduce the work involved in creating repulsive potentials, we propose an automatic algorithm based on least-squares fitting of repulsive potentials to reference energy values. During our early automatic fitting attempts, we experimented also with genetic algorithms, but the simpler least-squares fits turned out to be easier to handle and far less resource-hungry while delivering results of the same or even better quality. The process to be described below is not limited to the bare fitting of the repulsive potentials $U_{AB}(r)$, but it also helps in selecting and producing fit systems and fit data, tuning the priorities of different systems or properties, etc., making the whole parametrization process largely automatic.

3.1. Least-Squares Fitting of Repulsive Potentials. In order to make a least-squares fitting for the pairwise repulsive

potentials possible, we express them in terms of some arbitrary basis functions as

$$U_{AB}(r_{ij}) = \sum_{\nu} \alpha_{AB,\nu} f_{AB,\nu}(r_{ij}) \quad (13)$$

where AB is the type of atomic pair ij . Substituting this into the pair potential structure of E_{rep} from eq 10, the total repulsive energy for a given system becomes a linear combination

$$E_{\text{rep}} = \sum_{AB,\nu} \alpha_{AB,\nu} X_{AB,\nu} \quad (14)$$

of the structure-dependent quantities

$$X_{AB,\nu} = \sum_{\substack{\text{type}(ij)=AB \\ i < j}} f_{AB,\nu}(r_{ij}) \quad (15)$$

The sum runs over all possible atom pairs where the pair ij belongs to pair type AB .

Using the above, the best $\alpha_{AB,\nu}$ coefficients may be easily approximated by a least-squares fit to energy values of several different distortions of a chemical system as a function of the changing X values. Due to the linearity of the energy as a function of $X_{AB,\nu}$'s, this fitting is a multidimensional linear regression.

Running over a sequence of distortions denoted by s of the same system (we will call this sequence a *fit path* and the distortions *fit steps*), the least-squares fit minimizes the overall error

$$R = \sum_s^{\text{all steps}} (E_{\text{rep}}^{(s)} - (E_{\text{ref}}^{(s)} - E_{\text{el}}^{(s)}))^2 \quad (16)$$

With expression 14 of the total repulsive energy, the stationary condition

$$\frac{\partial}{\partial \alpha_{AB,\nu}} R(\alpha_{AB,\nu}) = 0 \quad (17)$$

of the above error leads to a matrix expression of the coefficients $\alpha_{AB,\nu}$:

$$\mathbf{A} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{E} \quad (18)$$

The matrices \mathbf{E} , \mathbf{X} , and \mathbf{A} are constructed from the above energies, X structural constants, and α 's in the following way:

$$\mathbf{E} = \begin{pmatrix} E_{\text{ref}}^{(1)} - E_{\text{el}}^{(1)} \\ E_{\text{ref}}^{(2)} - E_{\text{el}}^{(2)} \\ \vdots \end{pmatrix} \quad (19a)$$

$$\mathbf{X} = \begin{pmatrix} X_{\text{HH},1}^{(1)} & X_{\text{HH},1}^{(2)} & \cdots \\ X_{\text{HH},2}^{(1)} & X_{\text{HH},2}^{(2)} & \cdots \\ \vdots & \vdots & \vdots \\ X_{\text{CH},1}^{(1)} & X_{\text{CH},1}^{(2)} & \cdots \\ X_{\text{CH},2}^{(1)} & \cdots & \cdots \\ \vdots & \vdots & \vdots \end{pmatrix} \quad (19b)$$

$$\mathbf{A} = \begin{pmatrix} \alpha_{\text{HH},1} \\ \alpha_{\text{HH},2} \\ \vdots \\ \alpha_{\text{CH},1} \\ \alpha_{\text{CH},2} \\ \vdots \end{pmatrix} \quad (19\text{c})$$

where, as an example, we assumed that the enumeration of the investigated atomic pairs begins with HH and contains CH.

As an example, a fit path could be built from a propane molecule with its middle carbon atom being shifted by 40 small random displacements around its equilibrium position. Each movement as well as the original configuration is a different fit step. The energy and structure data of these 41 steps would then give enough input to fit $U_{\text{CC}}(r)$ and $U_{\text{CH}}(r)$ ¹⁰ provided the number of independent fitting parameters $\alpha_{\text{AB},\nu}$ is well below 40, i.e., in this specific case, the number of basis functions used to describe one pairwise repulsive is well below 20. This criterion is normally fulfilled, but if not, increasing the amount of steps is always a straightforward remedy.

3.2. Fitting to Multiple Fit System Types and Objectives.

An important expectation toward repulsive potentials is their transferability to a broad range of different systems. Usually, this requires compromises; transferability can be reached via a tradeoff between individual systems. Our automatic parametrization scheme enables the optimization of this tradeoff by enabling the fit on multiple test systems (multiple fit paths) at the same time. Staying with the example of the C–C and C–H repulsive fitting, by taking several different carbohydrogen molecules and distorting them, one can generate several molecular fit paths for the fit. Additionally, taking bulk diamond (with various deformations) as an additional fit path, one can tune the transferability toward the description of crystalline systems as well.

The goal of the fit becomes the minimization of overall error along all fit paths, modifying eq 16 to

$$R = \sum_p^{\text{all paths}} \sum_{s \in p} (E_{\text{rep}}^{(ps)} - (E_{\text{ref}}^{(ps)} - E_{\text{el}}^{(ps)}))^2 \quad (20)$$

with p enumerating the paths and E_{rep} written as a function of α 's and X 's within each path in the same way as the one-path case. It should be noted here that the number of $\alpha_{\text{AB},\nu}$ parameters *does not depend* on the number of fit paths (nor on the number of steps in the individual fit paths). Its value is determined only by the choice of the basis functions to describe the repulsive interactions in question.

The $E_{\text{rep}}^{(ps)}$ column vector of the repulsive energies for the multiple-path fitting is created by putting the $E^{(p)}$ vectors for each path on top of each other:

$$\mathbf{E} = \begin{pmatrix} \mathbf{E}^{(1)} \\ \mathbf{E}^{(2)} \\ \vdots \end{pmatrix} = \begin{pmatrix} E_{\text{ref}}^{(11)} - E_{\text{el}}^{(11)} \\ E_{\text{ref}}^{(12)} - E_{\text{el}}^{(12)} \\ \vdots \\ E_{\text{ref}}^{(21)} - E_{\text{el}}^{(21)} \\ E_{\text{ref}}^{(22)} - E_{\text{el}}^{(22)} \\ \vdots \end{pmatrix} \quad (21\text{a})$$

while the geometry matrix is created by putting single-path geometry matrices together in a similar way:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \\ \vdots \end{pmatrix} \quad (21\text{b})$$

For a multisystem fit, this all gives back the matrix eq 18 on \mathbf{A} .

The scheme proposed here is not restricted to obtaining repulsive potentials by fitting on energy differences between *ab initio* DFT and DFTB calculations. One can naturally extend it to interatomic forces or even Hessians as targets. This way, one gains the possibility of not just choosing the transferability range by selecting various systems for the fitting procedure but also of being able to select the properties which are required to be transferable to the maximum possible amount over those systems. Furthermore, by using energy differences between successive steps as a target instead of the absolute energies, fitting on molecular dynamics (MD) trajectories is also made efficient. Details for these three target extensions (force, Hessian, and energy difference) are given in the Appendix.

3.3. Weighting of Fit Targets. In the formalism described until now, every fit step contributes to the R overall error with the same weight. As this may not always be the desired behavior, we allow each step in each path to have an individual weight for its contribution to the total error. If the fit is done for multiple physical properties (e.g., energies and forces), each property can also be weighted differently.

The weighting issues come to play mainly in two areas. First, one typically would overweight near-equilibrium geometries to ensure a higher precision at near-equilibrium bond lengths at the cost of less precise description for strongly distorted geometries. Furthermore, weighting becomes a key issue when multiple physical properties are invoked into the fit, since the numerical values of the differences in the various properties (energy, force, etc.) must be converted to the same scale. This requires some experimenting, but it offers the possibility of balancing the performance of repulsives for various physical quantities. For example, heavy weights for forces are usually necessary when the fitted repulsives give poor results with geometry optimization otherwise.

3.4. Basis Function Shapes. We have experimented with several different $f_\nu(r)$ basis functions for the repulsive fitting. The splines used in most of the current DFTB implementations turned out to be inappropriate for a fitting procedure, as they tend to give very oscillatory behavior. As a straightforward alternative, we decided for the eq 22 cut-off polynomials first. They were used in the earliest DFTB-implementation¹¹ and still retain popularity with doing parametrization by hand, since it has been possible to do most of the parametrizations up to now with them. The zeroth and first-degree terms are omitted from such a polynomial to ensure a smooth decay at its cutoff distance r_0 :

$$f_\nu(r) = \begin{cases} (r - r_0)^\nu & \text{if } r < r_0 \text{ and } \nu \geq 2 \\ 0 & \text{otherwise} \end{cases} \quad (22)$$

This representation was shown to be successful at the relatively easy hydrocarbon parametrization.

To emulate spline-like behavior in our scheme, we also tested bases containing the above cut-off polynomials, but having no universal cutoff value (these bases can be regarded as sums of multiple single-cutoff bases). Bases with two cutoff values are

very efficient at improving polynomial repulsives, while more cutoff values bring up the oscillatory nature of splines. Less successful but still noteworthy examples of spline-like bases are wavelet bases, which we also probed.

Another important basis was the family of exponential functions. $e^{-a_\nu r^\nu}$ ($\nu = 1, 2, 3, \dots$) and their linear combinations, which seem to be a very natural choice for a repulsive function basis. These exponential functions proved to be a successful basis for our fittings with Ti and Zn. In these cases, a fairly tiny set of exponential basis functions (one to three of them) was quite enough to fit remarkably good parameter sets.

3.5. Further Automation in the Parametrization Workflow. Besides the automatized fitting process itself, there are three subprocesses of the parametrization workflow in which our program substantially lowers the human contribution.

- The path-building methods mentioned so far and some others are implemented to be executed automatically. They include bond stretchings, displacing atoms, uniform volume changes, linear interpolations between two configurations, and using predefined paths (e.g., MD trajectories or reaction paths).
- Instead of using fixed sets of metaparameters (input parameters determining the parametrization itself) for the fitting process, batch fits can use intervals of them. Scanning over all of these values in all of these intervals in every combination spares a lot of try-and-fail cycles for the user. At the end of the batch run, the set with the lowest total error on the targets (as defined in eq 20) is picked as the fittest solution.
- A module for defining test systems is built into our program too. It tests the energetical and geometrical performance of the fitted repulsives on the specified test systems. This way it can give a first-glance feedback about the performance of the fitted repulsive set on systems that were not necessarily fit systems.

4. RESULTS

4.1. Computational Tools. Since the molecular reference *ab initio* calculations in the handmade sets were mostly done using the Gaussian¹² code, we also used it as a reference for molecular systems. For the periodic systems, however, we have found the Siesta¹³ code far more stable (less prone to convergence failures) in our automatic fitting environment, where distorted systems far from the equilibrium must be calculated very often. Apart from stability issues, this choice is also a good cross-calculator and cross-methodology (e.g., between different xc functionals in DFT references) consistency check of our algorithm and in general for the DFTB parametrization philosophy. As will be seen from the results, this mixing of DFT references did not pose any problem. The DFTB calculations were carried out using the DFTB+ package.¹⁴

4.2. Carbohydrogen Systems. The carbohydrate case is a relatively easy case of parametrization in the sense that quite useful parameter sets can be fitted to it even with a small effort. Fitting to DFT references with the PBE exchange-correlation functional, the resulting parameter sets produce, according to our experiences, geometrical errors typically within a few 10^{-2} Å and atomization energy errors in the range of a few 10^{-2} au. This quality, which is almost comparable with the handmade mio set,⁹ is pretty easy to reach at an automatic fit with a nontrivial handful of fit systems and a couple of hours working with them.

Table 1. Molecular and Crystalline Data Calculated with the Three Parameter Sets (the mio Set⁹ and the Two Automatically Fitted Ones) Compared to Reference Values^a

property	reference	mio	hom	inhom
<i>methane</i>				
ΔE	0	7.3	-2.5	-0.1 (52.9)
C-H	1.093	1.089	1.094	1.080
<i>ethane</i>				
ΔE	0	17.7	-1.0	0.1 (94.8)
C-C	1.531	1.501	1.535	1.516
C-H	1.096	1.098	1.102	1.088
<i>ethene</i>				
ΔE	0	14.6	-2.4	-3.6 (68.6)
C=C	1.331	1.327	1.327	1.326
C-H	1.087	1.094	1.099	1.084
<i>ethyne</i>				
ΔE	0	21.7	10.2	-3.5 (53.1)
C≡C	1.205	1.203	1.200	1.204
C-H	1.067	1.075	1.080	1.066
<i>benzene</i>				
ΔE	0	52.9	-1.7	0.8 (170.8)
C-C	1.397	1.396	1.405	1.397
C-H	1.087	1.098	1.104	1.090
<i>butane</i>				
ΔE	0	38.7	2.8	1.6 (179.8)
(1,2) C-C	1.547	1.519	1.555	1.537
(2,3) C-C	1.536	1.518	1.552	1.534
(1) C-H	1.097	1.097	1.102	1.088
<i>isobutane</i>				
ΔE	0	38.0	1.9	1.5 (178.7)
C-C	1.535	1.518	1.552	1.534
C-H	1.097	1.098	1.102	1.088
<i>diamond</i>				
C-C	1.555	1.540	1.575	1.558
<i>cyclobutane</i>				
ΔE	0	40.0	7.3	13.4 (169.2)
C-C	1.557	1.539	1.569	1.534
C-H	1.095	1.102	1.107	1.094
<i>isobutene</i>				
ΔE	0	36.2	0.5	-2.7 (153.0)
C-C	1.509	1.493	1.524	1.505
C=C	1.337	1.341	1.34	1.339
C-H (in CH ₃)	1.099	1.100	1.104	1.090
C-H (in CH ₂)	1.087	1.093	1.099	1.084
<i>bicyclobutane</i>				
ΔE	0	26.9	-2.1	10.3 (143.4)
C-C (edge)	1.510	1.464	1.549	1.486
C-C (middle)	1.900	2.003	2.112	1.980
C-H (in CH ₂)	1.112	1.195	1.161	1.158
C-H (in CH)	1.095	1.066	1.021	1.065
<i>cyclobutene</i>				
ΔE	0	29.5	-1.8	7.4 (140.6)
C-C	1.573	1.569	1.597	1.538

Table 1. Continued

property	reference	mio	hom	inhom
C=C	1.519	1.524	1.548	1.493
C–H (in CH ₂)	1.097	1.104	1.109	1.097
C–H (in CH)	1.087	1.097	1.103	1.089
cyclopropane				
ΔE	0	18.9	–9.6	–0.2 (113.8)
C–C	1.509	1.489	1.523	1.502
C–H	1.087	1.096	1.100	1.087
propane				
ΔE	0	27.7	0.2	0.0 (136.6)
C–C	1.532	1.509	1.544	1.525
C–H (end)	1.097	1.098	1.102	1.088
C–H (middle)	1.099	1.107	1.110	1.097
cyclopropene				
ΔE	0	12.2	–13.8	–7.7 (83.7)
C–C	1.508	1.495	1.528	1.508
C=C	1.295	1.319	1.319	1.318
C–H (opposite to C=C)	1.095	1.107	1.109	1.096
C–H (neighbor to C=C)	1.080	1.090	1.095	1.081
spiropentane				
ΔE	0	29.4	–18.9	–2.0 (173)
C–C (“radial”)	1.485	1.479	1.508	1.488
C–C (outer)	1.530	1.508	1.547	1.524
C–H	1.088	1.097	1.102	1.088
methylene-cyclopropane				
ΔE	0	25.9	–10.6	–4.5 (128.7)
C=C	1.322	1.328	1.327	1.327
C–C (“radial”)	1.470	1.465	1.491	1.472
C–C (outer)	1.540	1.512	1.551	1.529
C–H (in CH ₂)	1.088	1.095	1.101	1.086
C–H (on ring)	1.089	1.098	1.102	1.089
propadiene				
ΔE	0	21.2	–2.6	–7.8 (83.6)
C=C	1.307	1.312	1.312	1.312
C–H	1.088	1.096	1.102	1.087
1,3-butadiene				
ΔE	0	49.9	16.2	10.0 (143.2)
C–C	1.439	1.436	1.457	1.441
C=C	1.392	1.372	1.373	1.370
C–H (middle)	1.089	1.098	1.103	1.089
C–H (end)	1.086	1.104	1.085	1.095
2-butyne				
ΔE	0	38.8	6.5	–1.2 (132.0)
C–C	1.462	1.455	1.477	1.461
C≡C	1.209	1.209	1.205	1.209
C–H	1.097	1.100	1.105	1.091
propyne				
ΔE	0	30.3	8.3	1.2 (92.6)
C–C	1.460	1.453	1.475	1.459
C≡C	1.207	1.206	1.203	1.207
C–H (in CH ₃)	1.097	1.100	1.104	1.090
C–H (in CH)	1.066	1.074	1.079	1.066

Table 1. Continued

property	reference	mio	hom	inhom
propene				
ΔE	0	24.9	–1.5	–3.7 (110.3)
C–C	1.502	1.485	1.517	1.497
C=C	1.333	1.334	1.334	1.333
C–H (in CH ₃)	1.098	1.100	1.105	1.091
C–H (in CH)	1.091	1.102	1.106	1.092
C–H (in CH ₂)	1.087	1.093	1.098	1.084

^a ΔE means atomization energy error relative to the reference in kcal/mol, and $A-B$ atom pairs denote distances of the appropriate neighboring atoms in Å. The column “hom” contains a fit without dissociation energy correction, “inhom” contains a fit with it. Values in parentheses indicate errors for the set with dissociation energy correction when used in a DFTB implementation without this correction scheme. Italicized names denote systems that were fit systems too; the other molecules are the rest of the carbohydrogen part of the G2¹⁸ test set.

Adding more configurations, the results can be further improved. In order to demonstrate the automatism in our procedure, we give the instructions used to generate those configurations:

- a methane molecule with its central carbon atom randomly displaced on five shells within a sphere¹⁵ of diameter 0.75 Å
- an ethane molecule with one carbon atom displaced on 10 equidistant shells within a 0.75 Å sphere
- a butane molecule with its 1-2 carbon–carbon bond stretched in 15 0.1 Å steps, from a shortening of 0.6 Å to a lengthening of 0.9 Å
- a benzene ring with one of its carbon atoms displaced on five equidistant shells within a 0.75 Å diameter sphere
- an ethene molecule with one carbon atom displaced on five shells in a 0.75 Å diameter sphere
- a series of random displacements similar to the above with an ethyne molecule
- a hydrogen molecule with its only bond shortened in eight and lengthened in 12 0.025 Å steps
- an isobutane molecule with its central carbon atom displaced in a 1 Å diameter sphere

As the mio set, the basis of comparison, was fitted to calculations with the B3LYP xc functional and the 6-31G* basis, we also used this as a reference. The force objective had a weight of three while energy had a weight of one, and each path had its near-equilibrium steps (at most three steps away from equilibrium) overweighted by five. For the diamond test system, we used the CRYSTAL2003 code¹⁶ (because of the problems with Gaussian mentioned above) with a 6-21G*¹⁷ basis set and a k -space mesh of an $8 \times 8 \times 8$ Monkhorst–Pack scheme.

During the fitting process, the automaton was allowed to sweep over the following metaparameters to search for the best fit:

- The cutoff of C–C: 2.0–2.3 Å
- The cutoff of C–H: 1.3–2.1 Å
- The cutoff of H–H: 1.3–2.1 Å
- The highest degree of polynomials: 10–12

The best fit was achieved with values of 2.3 Å, 2.1 Å, 1.3 Å, and 11 for the above metaparameters, respectively. For the sake of smoothness, the polynomials contained a minimal power of 4. Table 1 shows the performance of the resulting repulsive in comparison with the mio set (columns “mio” and “hom”) on the respective equilibrium structures. As our method aims at not only

Table 2. Reference Data and Its Comparison with Previous Handmade Parametrization¹⁹ (“znorg”) and the Automatically Created One (“auto”) for Zn and ZnO Crystals^a

property	reference	znorg	auto
<i>Zn hcp</i>			
ΔE (per Zn ₂)	0	115.5	94.1
Zn–Zn (#1)	2.523	2.796	2.433
Zn–Zn (#2)	2.886	2.864	2.931
Zn–Zn (#3)	3.831	4.051	3.788
Zn–Zn (#4)	4.591	4.872	4.524
<i>ZnO zinblend</i>			
ΔE (per ZnO)	0	22.3	–1.1
Zn–O	2.005	2.015	2.011
Zn–Zn	3.274	3.290	3.281
<i>ZnO wurtzite</i>			
ΔE (per Zn ₂ O ₂)	0	46.5	–0.7
Zn–O	2.017	2.015	2.018
Zn–O'	2.037	2.014	2.004

^aThe values given here refer to the equilibrium structure of each system. ΔE denotes the atomization energy difference with respect to the reference in kcal/mol. Atom pairs denote distances in Å.

describing equilibrium properties as close to *ab initio* results as possible, but also to provide a reasonable accuracy when dealing with structures out of equilibrium, we calculated also the energy errors over all nonequilibrium configurations in the fit paths. They remained generally within the error of 10^{-2} Hartree compared to the DFT reference except some of the extremely distorted geometries.

4.2.1. *Using One-Body Repulsive Terms.* With this carbohydrate fit, we also experimented with using one-body terms in the repulsive energy

$$E_{\text{rep}}(\{R_{\text{nuclei}}\}) = \sum_{i < j} U_{AB}(r_{ij}) + \sum_i U_A \quad (23)$$

One-body terms are a special case of inhomogeneous or dissociation energy terms: they represent a fixed, geometry-independent energy part as a sum of atomwise parts that do not come from the linear combinations of pairwise basis functions and that maintain the asymptotic value of E_{rep} at the dissociation limit. One-body energies are the only mathematically correct means of putting any correction to dissociation energy because only a sum of one-atomic dissociation energy terms behaves like an extensive quantity, i.e., is an additive function of stoichiometry. This fact strongly encourages investigating their use.

As the results in Table 1 illustrate (column “inhom”), one-body terms can slightly improve geometry results via eliminating the need of trying to set absolute atomization energy levels using the pair potential profiles. The resulting one-body terms were $U_{\text{C}} = 0.030633\text{H}$ and $U_{\text{H}} = 0.017967\text{H}$ for C and H atoms, respectively. The optimal cutoff distances were (determined by a similar batch) equal to those of the homogeneous case.

To maintain compatibility with the current DFTB implementations lacking one-body repulsive parts, we also took another way of improving results by one-body terms into account. Using them only at the fitting process but dropping them after it retains improved geometries and reaction energies calculated with the produced set yet leaves the pair potential structure of the repulsive energy built

Table 3. Titanium–Oxygen Compound Reference Data and Its Comparison with Previous Handmade Parametrization²⁰ (“tiorg”) and the Automatically Created One (“auto”)^a

property	reference	tiorg	auto
<i>TiO</i>			
ΔE	0	55.0	40.2
Ti–O	1.586	1.592	1.586
<i>Ti₂O₂ planar</i>			
ΔE	0	87.7	69.4
Ti–Ti	2.198	2.355	2.092
Ti–O	1.857	1.891	1.866
<i>Ti₂O₂ nonplanar</i>			
ΔE	0	68.0	57.6
Ti–Ti	2.127	2.249	2.133
Ti–O	1.838	1.888	1.826
<i>Ti₂O₄ #1 (dibridged with end O atoms in cis position)</i>			
ΔE	0	49.4	81.3
Ti–Ti	2.716	2.800	2.635
bridging Ti–O	1.848	1.887	1.812
end Ti–O	1.622	1.606	1.589
O–Ti–Ti (ending O)	126.1	124.7	123.7
<i>Ti₂O₄ #2 (dibridged with end O atoms in trans position)</i>			
ΔE	0	145.7	82.8
Ti–Ti	2.709	2.726	2.709
bridging Ti–O	1.840	1.831	1.806
end Ti–O	1.625	1.608	1.590
O–Ti–Ti (ending O)	123.7	122.3	122.2
<i>Ti₂O₄ #3 (tribridged with an O atom at one end)</i>			
ΔE	0	123.3	66.4
Ti–Ti	2.394	2.540	2.399
bridging Ti–O (opposite to end O)	1.763	1.801	1.742
end Ti–O	1.628	1.606	1.586
<i>Ti hcp</i>			
ΔE (per Ti)	0	–40.6	5.4
Ti–Ti	2.900	2.993	2.915
<i>TiO₂ anatase</i>			
ΔE (per TiO ₂)	0	22.5	–4.5
shortest Ti–Ti	3.028	2.996	3.082
shortest Ti–O	1.933	1.921	1.957
		1.995	1.958
<i>TiO₂ rutile</i>			
ΔE (per TiO ₂)	0	19.4	1.1
shortest Ti–Ti	3.559	3.613	3.605
shortest Ti–O	1.974	1.914	1.976
		1.992	1.995

^aThe values given here refer to the equilibrium structure of the various systems. ΔE means atomization energy difference with respect to the reference in kcal/mol. Atom pairs denote neighbor distances in Å; triples denote angles in degrees (double distance values show artificially broken symmetries of DFTB-optimized lattices). Italicized system names denote fit systems.

in DFTB intact (deteriorated bare atomization energy values are shown in parentheses in the appropriate column of Table 1).

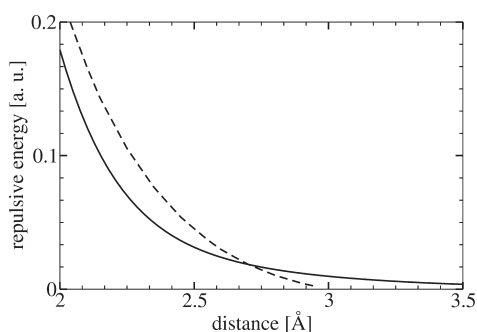


Figure 1. Comparison of the tiorg (dashed) and the automatically generated (solid line) Ti–Ti repulsives in the area of the sharp cutoff of the former.

4.3. Zinc–Oxygen Compounds. As a further demonstration for our fitting procedure, we attempted to create a parametrization for the Zn–O interaction. A high-quality and well-tested parameter set had been recently created manually for the zinc–organic chemistry by Moreira et al.,¹⁹ which should serve as an etalon for our Zn–O repulsive. For the DFT references, the same settings had been used as for the handmade parametrization (PBE functional, double- ζ polarized basis, norm-conserving Troullier–Martins pseudopotentials, $8 \times 8 \times 8$ Monkhorst–Pack scheme for k sampling). The fit paths were made with distortions applied to the test systems (see Table 2) in addition to Zn–Zn and Zn–O dimers with very low weights. The distortions applied to crystalline paths were uniform volume scaling and moving a Zn atom around. We show a comparison between the performance of the two Zn–O sets in Table 2. As fit targets, we used the two energy targets (energy and energy differences between steps weighted by 1:10); step weighting was by 10 and 2 in the immediate and in a wider neighborhood of equilibria. Here, the basis of repulsives consisted of exponential functions of type $e^{-a_2 r^2}$ and $e^{-a_3 r^3}$, as these shapes offered good results quickly in situations where absolute energy targets were not heavily weighted. As can be seen in Table 2, the resulting set is superior in the crystalline properties to the handmade one.

4.4. Titanium–Organic Repulsives. After the hydrocarbon fits, our next test of the fitting automaton was producing a titanium–oxygen set and extending it to a titanium–organic set. For this parametrization, a good-quality handmade set (tiorg) had been recently created by Dolgonos et al.²⁰ We used the same reference structures and *ab initio* reference data (various molecular systems calculated with the B3LYP functional and with mixed SDD+ basis set) augmented with crystalline reference systems. For the reference calculations of the periodic systems, the PBE functional, double- ζ plus polarized basis functions, and norm-conserving Troullier–Martins pseudopotentials had been used. K-point sampling was set to an $8 \times 8 \times 8$ Monkhorst–Pack scheme with both Siesta and DFTB in this fit session.

In order to fit repulsive functions for the Ti–Ti and Ti–O interactions, we used a fit set including a titanium dimer (with a very low weight), a TiO₂ molecule, a planar Ti₂O₂ molecule, a tribridged Ti₂O₄, the bulk hcp titanium, and the bulk anatase and rutile forms of TiO₂. The molecular fit paths were created by stretching bonds and displacing titanium atoms while the crystalline paths were created by uniformly changing the volume of the crystal lattices and by using crystals with displaced titanium atoms. We used both energy and force targets (generally weighted 1:2) in the fit.

Table 4. Reference Data and Its Comparison with Previous Hand Made Parametrization²⁰ (“tiorg”) and the Automatically Created One (“auto”) for Various Titanium Compounds^a

property	reference	tiorg	auto
<i>Ti(CH₃)₄</i>			
ΔE_b	0	64.6	180.2
Ti–C	2.072	2.096	2.025
<i>Ti(CH₃)₂</i>			
ΔE_b	0	–38.8	93.4
Ti–C	2.038	2.096	2.025
C–Ti–C	113.7	110.2	109.9
<i>crystalline TiC</i>			
ΔE	0	111	91.7
Ti–C	2.141	2.159	2.170
Ti–i	3.024	3.047	3.067
<i>Ti(NH₂)₄</i>			
ΔE_b	0	30.6	287.4
Ti–N	1.899	1.902	1.853
<i>H₃Ti(NH₂)</i>			
ΔE_b	0	12.0	76.2
Ti–N	1.846	1.898	1.837
<i>HN=Ti=NH</i>			
ΔE_b	0	15.5	156.1
Ti–N	1.707	1.703	1.671
N–Ti–N	114.8	114.7	113.7
<i>crystalline TiN</i>			
ΔE	0	196.6	192.2
Ti–N	2.094	2.159	2.115
Ti–Ti	2.958	3.043	2.982
<i>Ti₂H₂ (dibridged planar)</i>			
ΔE	0	123.5	131
Ti–Ti	1.985	1.967	2.011
Ti–H	1.868	1.827	1.899
Ti–H–Ti	64.2	65.2	63.9

^a The values given here refer to the equilibrium structure of each system. ΔE denotes the atomization energy difference with respect to the reference in kcal/mol. ΔE_b indicates the binding energy between the central Ti atom and the ligands compared to the reference value in kcal/mol. Atom pairs denote neighbor distances in Å; triples denote angles in degrees. Italicized system names refer to fit systems.

In a fit session of a few days, we were able to produce a set of Ti–Ti and Ti–O repulsive potentials which reproduce energy and geometrical data in the same quality as the reference handmade set. A detailed comparison is given in Table 3. These results were obtained using $e^{-a_1 r}$ - and $e^{-a_2 r^2}$ -type exponential functions as basis functions for the fit because this analytical basis gave very good results quickly with the Ti and Ti–O chemistry.

After creating the repulsives for the Ti–Ti and Ti–O interactions, we extended the set to a complete Ti–organic set, still using exponential basis functions for expressing the repulsive potentials. The extension turned out to be more difficult than expected, mainly due to the sudden cutoff in the handmade Ti–Ti repulsive giving a very stable 3 Å Ti–Ti distance in hcp

titanium, titanium nitride, and titanium carbide. This feature (shown in Figure 1) can hardly be reconstructed with analytical sets. Although this peculiar shape is the numerically most convenient way to confine the range of the Ti–Ti repulsive well below the second-neighbor distance and gives good results for various systems, it may be an interesting question for future investigations whether it is a precise representation of the underlying physics.

Similar to the case of the titanium–oxygen fit, we used the same molecules (TiH₄, Ti(CH₃)₄, Ti(NH₂)₄) as during the handmade parametrization²⁰ extended by crystalline fit systems (TiN and TiC). The fit paths were similar constructions to the titanium–oxygen case. Every molecule had two fit paths with the relevant bonds stretched and the titanium atom displaced, while the crystalline systems had a volume change path and a path with a titanium atom displaced. We handled the relative difficulty of Ti–C and Ti–N fitting compared to Ti–Ti and Ti–O by lowering the relative weights of the energy targets in the Ti–C and Ti–N case. As can be seen from the results (Table 4), this resulted in fairly good geometries at the expense of accuracy in energies (one-body terms as a tool for resolving the conflict between energy and geometry accuracy was not used here).

5. CONCLUSION

We are suggesting a new fitting mechanism to create repulsive potentials for the DFTB method in an automatic way using least-squares fitting on automatically generated reference data. Using the proposed scheme, we fitted new repulsives for carbohydrogen systems, zinc and zinc oxide crystal structures, and titanium-containing organic compounds. Due to its efficiency and high degree of automatization, the fitting took in each case at most a couple of days' work of a researcher with the new fitting scheme. Comparing the new fits against existing handmade fits showed that we were able to create a general-purpose parametrization engine for the DFTB method. The engine enables us to optimize new parameters from scratch for any group of systems where the DFTB formalism with the pairwise repulsive potentials gives a reasonable description of the underlying physics. While these new parameter sets are very close in accuracy to handmade sets, they require considerably less time and human effort to be created.

The fitting procedure was planned to be as easy to handle, as comprehensive and as interactive as possible. As a demonstration of its easy integrability into current quantum chemical tools, it had been included into the Material Studio program package, using a graphical user interface to control the parametrization process.

APPENDIX

In the following sections, we give a detailed derivation about how the energy fitting procedure described above can be extended to objectives other than the basic energy objective (of them, the first two are fully implemented and tested in our program). This enables the extension of the fitting procedure to further physical properties (forces and frequencies) and the effective use of energy data from existing MD trajectories in the fitting process.

Fitting to Forces. The force objective from the repulsive interaction is the repulsive force F_i acting on atom i projected

onto a unit vector (a direction) \mathbf{u}

$$F_{i,\mathbf{u}} = \mathbf{F}_i \cdot \mathbf{u} = \sum_{j \neq i, \nu} \alpha_{\text{type}(ij), \nu} f'_{\text{type}(ij), \nu}(r_{ij}) \frac{\mathbf{r}_{ij}}{r_{ij}} \cdot \mathbf{u} \quad (24)$$

This can be, similar to the energy expression 14, decomposed into a linear combination

$$F_{i,\mathbf{u}} = \sum_{AB, \nu} \alpha_{AB, \nu} X_{AB, \nu}^{(i, \mathbf{u})} \quad (25)$$

with coefficients $\alpha_{AB, \nu}$, built of geometry-dependent factors

$$X_{AB, \nu}^{(i, \mathbf{u})} = \sum_{\substack{\text{type}(ij) = AB \\ j \neq i}} f'_{AB, \nu}(r_{ij}) \frac{\mathbf{r}_{ij}}{r_{ij}} \cdot \mathbf{u} \quad (26)$$

containing the first derivatives of the basis functions $f_{AB, \nu}(r)$. Because the α coefficients in these force components are the same as the ones used for the energy fitting, fitting to energies and forces can be unified when both are required. If $(F_{\text{ref}} - F_{\text{el}})_{i, \mathbf{u}}$ takes the place of $E_{\text{ref}} - E_{\text{el}}$ and the above new X 's are used as independent variables, fitting to force components can be simply regarded as additional new fit paths. The matrices \mathbf{E} and \mathbf{X} can then be extended in the same way as in eqs 21a and 21b.

Fitting to MD Trajectory Energies. A problem that often compromises fitting to MD trajectories (or to large molecules where only a tiny part is distorted) is the fact that equilibrium bond lengths are heavily overweighted by their overwhelming presence in the sample fit paths. This can make efficient fitting to ranges of bond lengths other than the covalent equilibrium impossible with the original energy target described above.

A remedy of this problem can be found by fitting to *energy differences between subsequent fit steps* instead of energies of each fit step. As

$$\begin{aligned} \Delta E_{\text{rep}}^{(s)} &= E_{\text{rep}}^{(s+1)} - E_{\text{rep}}^{(s)} \\ &= \sum_{AB, \nu} \alpha_{AB, \nu} (X_{AB, \nu}^{(s+1)} - X_{AB, \nu}^{(s)}) \end{aligned} \quad (27)$$

is a linear combination of structural quantities of type $X^{(s+1)} - X^{(s)}$, it is a valid target in our least-squares fit scheme. This modified energy target, however, contains virtually nothing arising from those bonds which do not change over the fit path; thus the overweighting of unchanged bonds is avoided. Of course, if fitting to absolute energy values at molecular equilibrium bond lengths is required, it can be brought back by an appropriate weighting between the original energy objective and the current one, or by defining additional molecular fit paths.

As an alternative use, the fit target based on the energy differences can also be used in cases where retrieving force data from a DFT reference is for some reason problematic or meaningless (e.g., with symmetric distortions of symmetric systems, atomwise total forces are constant zero). Using small distortion steps and the energy difference fit target, one automatically obtains a fit mimicking the fit on certain force or stress tensor components.

Fitting to Hessians. Similar to the forces, the repulsive contribution to the Hessian matrix of a chemical system can also be projected onto unit vectors \mathbf{u} and \mathbf{v} (these unit vectors can be regarded as virtual displacements of atoms). When both \mathbf{u} and \mathbf{v}

are on the same i th atom (i.e., we examine the i th “on-site” 3×3 hyperdiagonal block of the $3N \times 3N$ collective molecular Hessian),

$$\begin{aligned} H_{i,uv} &= \mathbf{uHv} = \sum_{j \neq i, mn} u_m \frac{\partial^2 U_{\text{type}(ij)}(r_{ij})}{\partial x_{i,m} \partial x_{i,n}} v_n \\ &= \sum_{j \neq i, mn, v} \alpha_{\text{type}(ij),v} u_m \frac{\partial^2 f_{\text{type}(ij),v}(r_{ij})}{\partial x_{i,m} \partial x_{i,n}} v_n \\ &= \sum_{AB} \sum_{\substack{\text{type}(ij)=AB \\ j \neq i, v}} \alpha_{\text{type}(ij),v} \left(\frac{1}{r^2} \frac{\partial^2 f_v}{\partial r^2} - \frac{1}{r^3} \frac{\partial f_v}{\partial r} \right) (\mathbf{u} \cdot \mathbf{r})(\mathbf{v} \cdot \mathbf{r}) \\ &\quad + \frac{1}{r} \frac{\partial f_v}{\partial r} (\mathbf{u} \cdot \mathbf{v}) \end{aligned} \quad (28)$$

(with $\mathbf{r}_{ij} = (x_{j,1} - x_{i,1}, x_{j,2} - x_{i,2}, x_{j,3} - x_{i,3})$ at the beginning and $f_v = f_{\text{type}(ij),v}(r_{ij})$, $r = r_{ij}$ and $\mathbf{r} = \mathbf{r}_{ij}$ in the last step). So, with

$$\begin{aligned} X_{AB,v}^{(i,uv)} &= \sum_{\substack{\text{type}(ij)=AB \\ j \neq i}} \left(\frac{1}{r^2} \frac{\partial^2 f_v}{\partial r^2} - \frac{1}{r^3} \frac{\partial f_v}{\partial r} \right) (\mathbf{u} \cdot \mathbf{r})(\mathbf{v} \cdot \mathbf{r}) \\ &\quad + \frac{1}{r} \frac{\partial f_v}{\partial r} (\mathbf{u} \cdot \mathbf{v}) \end{aligned} \quad (29)$$

the usual linear combinations can be written again

$$H_{i,uv} = \sum_{AB,v} \alpha_{AB,v} X_{AB,v}^{(i,uv)} \quad (30)$$

With the \mathbf{E} vector composed of $(H_{\text{ref}} - H_{\text{el}})_{i,uv}$'s and the \mathbf{X} matrix composed of the above X 's, the fitting of the Hessian can be included as an additional path into the fitting scheme.

When \mathbf{u} and \mathbf{v} are on the i th and j th atoms, respectively,

$$\begin{aligned} H_{ij,uv} &= \mathbf{uHv} = u_m \frac{\partial^2 U_{\text{type}(ij)}(r_{ij})}{\partial x_{i,m} \partial x_{j,n}} v_n \\ &= -u_m \frac{\partial^2 U_{\text{type}(ij)}(r_{ij})}{\partial x_{i,m} \partial x_{i,n}} v_n \end{aligned} \quad (31)$$

Therefore, a similar construction applies to “off-site” Hessian parts, but with the opposite sign and without the summation over j .

As a linear combination of the above \mathbf{u} and \mathbf{v} atomic virtual displacements, every collective distortion of a molecule can be constructed. This knowledge can be used to fit to Hessians of DFT reference algorithms that give no detailed Hessian matrix but only vibrational modes and frequencies in their output. If \mathbf{e} is a ($3N$ -component) collective eigenmode of the molecular Hessian with ω frequency,

$$\omega^2 \mathbf{M} \mathbf{e} = \mathbf{H} \mathbf{e} = (\mathbf{H}_{\text{el}} + \mathbf{H}_{\text{rep}}) \mathbf{e} \quad (32)$$

where \mathbf{M} is the diagonal mass matrix. The vector of equations contained in

$$\omega^2 \mathbf{M} \mathbf{e} - \mathbf{H}_{\text{el}} \mathbf{e} = \mathbf{H}_{\text{rep}} \mathbf{e} \quad (33)$$

can then be used as a new fit path with the left hand side as a vector of \mathbf{E} values and the right hand side as the usual linear combinations coming from the repulsives and using α 's as coefficients.²¹ Note that the last equation contains explicit Hessian data from DFTB only.

Fitting to the Stress Tensor. The repulsive part of the stress tensor in periodical systems is calculated as

$$\begin{aligned} \sigma_{mn} &= -\frac{1}{\tilde{V}} \frac{\partial \tilde{E}_{\text{rep}}}{\partial \epsilon_{mn}} = \frac{1}{\tilde{V}} \sum_{\substack{i \in \text{a cell} \\ j}} F_{ij,m} r_{ij,n} \\ &= -\frac{1}{\tilde{V}} \sum_{i \in \text{a cell}} \frac{1}{r_{ij}} \frac{\partial U(r_{ij})}{\partial r_{ij}} r_{ij,m} r_{ij,n} \\ &= -\frac{1}{\tilde{V}} \sum_{\substack{i \in \text{a cell} \\ j}} \alpha_{\text{type}(ij),v} f'(r_{ij}) \frac{r_{ij,m} r_{ij,n}}{r_{ij}} \end{aligned} \quad (34)$$

where ϵ_{mn} is the strain tensor, \tilde{V} is the unit cell volume, \tilde{E}_{rep} is the cellwise repulsive energy, $r_{ij,m}$ is a component of the relative position vector \mathbf{r}_{ij} from the i th atom to the j th, and r_{ij} is the length of it. A double projection of σ_{mn} onto unit vectors \mathbf{u} and \mathbf{v} can be written

$$\sigma_{uv} = \sum_{mn} \sigma_{mn} u_m v_n = \sum_{AB,v} \alpha_{AB,v} X_{AB,v}^{(uv)} \quad (35)$$

if our structural quantities are

$$X_{AB,v}^{(uv)} = -\frac{1}{\tilde{V}} \sum_{\substack{i \in \text{a cell} \\ j}} f'(r_{ij}) \frac{(\mathbf{r}_{ij} \mathbf{u})(\mathbf{r}_{ij} \mathbf{v})}{r_{ij}} \quad (36)$$

So with the above X 's, σ_{uv} can be another valid target of our repulsive fitting algorithm.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bodrog.zoltan@bccms.uni-bremen.de.

ACKNOWLEDGMENT

The authors thank Gotthard Seifert, Martin Persson, and Grygoriy Dolgonos for fruitful discussions. Z.B. acknowledges support from the *Scientific Computing in Engineering* doctoral school at the University of Bremen.

REFERENCES

- (1) Seifert, G. J. *Phys. Chem. A* **2007**, *111*, 5609–5613.
- (2) Elstner, M.; Porezag, D.; Jungnickel, G.; Elstner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.
- (3) Köhler, C.; Seifert, G.; Frauenheim, T. *Chem. Phys.* **2005**, *309*, 23–31.
- (4) Niehaus, T. *THEOCHEM* **2009**, *914*, 38–49.
- (5) Niehaus, T.; Rohlfing, M.; Sala, F. D.; Carlo, A. D.; Frauenheim, T. *Phys. Rev. A* **2005**, *71*, 022508.
- (6) Pecchia, A.; Carlo, A. D. *Rep. Prog. Phys.* **2004**, *67*, 1497.

(7) Knap, J. M.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5637–5641.

(8) Gaus, M.; Chou, C.-P.; Witek, H.; Elstner, M. *J. Phys. Chem. A* **2009**, *113*, 11866–11881.

(9) Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 5614–5621.

(10) Displacements of a carbon atom in a carbohydrogen molecule alter the C–C and C–H bonds (but not the H–H bonds) and thus enable fitting to these repulsive potentials only.

(11) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947.

(12) Frisch, M. J. et al. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2003.

(13) Soler, J. M.; Artacho, E.; Gale, J. D.; Garcia, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. *J. Phys. (Paris)* **2002**, *14*, 2745.

(14) Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

(15) “Displaced on n shells in a sphere” means that the atom is dislocated with a random vector on n spherical shells around its original position; the n equidistant shells are defined within the largest sphere, from radius 0 up to the largest radius. The random vectors are generated isotropically, one with length zero and at least four on each nontrivial shell. This way, a path with an atom jumping around n times contains at least $4n$ steps, plus one for the original configuration.

(16) Saunders, V.; Dovesi, R.; Roetti, C.; Orlando, R.; Zicovich-Wilson, C. M.; Harrison, N.; Doll, K.; Civalleri, B.; Bush, I.; D’Arco, P.; Llunell, M. *CRYSTAL2003 User’s Manual*; University of Torino: Torino, Italy, 2003.

(17) Dovesi, R.; Causa, M.; Orlando, R.; Roetti, C.; Saunders, V. *J. Chem. Phys.* **1990**, *92*, 7402–7411.

(18) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(19) Moreira, N. H.; Dolgonos, G.; Aradi, B.; da Rosa, A. L.; Frauenheim, T. *J. Chem. Theory Comput.* **2009**, *5*, 605–614.

(20) Dolgonos, G.; Aradi, B.; Moreira, N. H.; Frauenheim, T. *J. Chem. Theory Comput.* **2010**, *6*, 266–278.

(21) An important issue is whether we must compare DFT-equilibrium Hessians to DFTB-equilibrium Hessians or we must compare Hessians of the very same geometry (practically, the DFT equilibrium geometry, as DFT calculators tend to compute (real) eigenvalues and eigenmodes instead of outputting raw Hessian matrices). From a theoretical point of view, the latter comparison is more valid, while from a semiempirical point of view, the former comparison is much more justified.

Fully Numerical All-Electron Solutions of the Optimized Effective Potential Equation for Diatomic Molecules [*Journal of Chemical Theory and Computation* **2009**, *5*, 1731–1740. DOI: 10.1021/ct800485v]. Adi Makmal, Stephan Kümmel, and Leeor Kronik*

With this Erratum, we provide corrections for minor technical errors we found in our original publication.¹ These corrections do not alter any of the conclusions drawn in the paper:

- (1) In Table 9, the xOEP 1σ eigenvalue of LiH should be $-2.06(6)$ Hartree.
- (2) In appendix B, eq (B4) should be

$$\varepsilon_{i\sigma}^1 = \langle \varphi_{i\sigma} | \Delta v_{i\sigma} | \varphi_{i\sigma} \rangle + \langle \psi_{i\sigma} | H_{KS} | \varphi_{i\sigma} \rangle^* - \langle \varphi_{i\sigma} | H_{KS} | \psi_{i\sigma} \rangle$$

where the Kohn–Sham eigenvalues are assumed to be real. This differs from the form given in the original article by complex conjugation of the second term on the right-hand side of the equation and by the sign of the third term on the right-hand side of the equation.

- (3) In appendix B, eq (B5) should be

$$\underbrace{[H_{KS} - \varepsilon_{i\sigma}^0 + \varepsilon_{i\sigma}^0 | \varphi_{i\sigma} \rangle \langle \varphi_{i\sigma} | - | \varphi_{i\sigma} \rangle \langle \varphi_{i\sigma} | H_{KS}]}_A \underbrace{|\psi_{i\sigma} \rangle}_x = \underbrace{-\langle \overline{\Delta v_{i\sigma}}(\mathbf{r}) - \Delta v_{i\sigma}(\mathbf{r}) | \varphi_{i\sigma} \rangle}_b$$

This differs from the form given in the original article by sign of the third term on the left-hand side of the equation.

- (4) Our mention of previous optimized effective potential work for single atoms unfortunately omitted the work of Cinal and Holas.²

REFERENCES

- (1) Makmal, A.; Kümmel, S.; Kronik, L. *J. Chem. Theory Comput.* **2009**, *5*, 1731.
- (2) Cinal, M.; Holas, A. *Phys. Rev. A* **2007**, *76*, 042510.

DOI: 10.1021/ct200381z

Published on Web 06/30/2011