JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

# Combined Fragmentation Method: A Simple Method for Fragmentation of Large Molecules

Hai-Anh Le, Hwee-Jia Tan, John F. Ouyang, and Ryan P. A. Bettens*

Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543

Supporting Information

**ABSTRACT:** Here we present a new energy-based fragmentation method that is based on our previous work and combines the best elements of other energy-based fragmentation methods. Our new approach, termed "combined fragmentation method", is foremost simple to implement, robust, accurate, and produces small fragments, which are independent of conformation and size of the target molecule. Essentially small collections of bonded atoms in the target molecule are assigned to groups. Fragment molecules are formed by taking all bonded pairs of these groups. These fragments are then interacted with one another, and the interaction energy is simply added to the initial fragmentation energy. The method has been tested on numerous molecules of biological interest both in vacuum and in a continuum solvent.

## 1. INTRODUCTION

In recent years, theoretical studies of large molecular systems, such as proteins, have received increasing interest. With recent advancements in computer power, the size of the systems which can be studied has increased. Although ab initio electronic structure calculations can provide high accuracy, the steep scaling of required computational resources still severely limits their applicability to many systems of interest. Many attempts have been made to develop linear scaling techniques with respect to the size of the system, in which the key assumption is the local nature of the molecule's electronic environment.

One approach is to reduce the scaling of current electronic structure methods. In the early 1980's, it was pointed out by Pulay[1] that the steep computational scaling of traditional correlation methods is mainly the result of the use of delocalized canonical orbitals which destroy the locality of correlation effects and is therefore unphysical. In order to avoid this problem, several local correlation methods for perturbation theory up to fourth order were proposed by Pulay and Saebø.[2−5] The correlation energy is expressed in terms of localized quantities, like the localized oribtals, or directly in atomic orbitals.[6] The method was later developed further and generalized to coupled cluster theory by other groups, such as Werner and co-workers,[7−10] Head-Gordon et al.,[11] Schütz,[12] and Carter and co-workers.[13−15] Werner has recently introduced a series of F12 explicitly correlated methods[7] and basis sets[16] to overcome the slow convergence of the correlation energy with basis set size.

A different approach to large system problems is the hybrid QM/MM methods. For example, the effective potential (EFP) method[17] was developed as an explicit treatment of solvent molecules in the study of solvation of large systems. The solute is treated in a fully QM manner, while the solvent molecules are represented by a set of parameters obtained by means of ab initio calculations to describe the electrostatic, polarization, and exchange repulsion effects. This method was later generalized (EFP2)[18,19] to describe the nonbonded intermolecular interactions. Several damping formulas were included to take into account short-range and quantum effects. Other QM/MM methods include the explicit polarization potential (X-Pol) method,[20,21] the Truhlar multicoefficient correlation methods (MCMM),[22] and multilayer methods, such as ONIOM.[23]

The divide-and-conquer philosophy was first imported by Yang[24] in 1991 to enable calculations for large molecules. The system is spatially divided into disjoint subsystems called central regions; each is composed of a set of atomic orbitals. Atoms adjacent to the central regions are called the buffer regions and are included in constructing molecular orbitals (MOs) for the subsystems. The local density distribution corresponding to each subsystem region is calculated with these MOs, which is subsequently used in the calculation of the electronic density of the entire system. The density-based divide-and-conquer was originally limited to Kohn−Sham density functional theory (DFT)[25] but was later applied to the one-electron density matrix constructed by Hartree−Fock or semiempircal calculation by Yang and Lee.[26] Another method which was developed contemporarily with the divide-and-conquer method is the adjustable density matrix assembler approach (ADMA),[27−32] which applies the divide-and-conquer technique on the density matrix. Since then, many groups have proposed "divide-and-conquer-like" methodologies that divide the system of interest into smaller subsystems or fragments.[6] These are commonly known as fragment-based methods, or fragmentation methods. The wave function, energy, and properties of the fragments or subsystems can be evaluated simultaneously with high accuracy by performing conventional QM calculations and subsequently combined to predict the same properties for the whole system.

There are various ways to classify fragment-based methods. Li and co-workers classified fragmentation methods into density- and energy-based methods.[33] In density-based methods, the density of the whole system is calculated from the densities of the fragments followed by the energy, while the in the energy-based

**Received:** November 4, 2011
**Published:** December 19, 2011

Table 1. Hard (H) and Soft (S) Rules for Forming Sets of Atoms to Define Groups

| no. | rule |
|---|---|
| H1 | Singly valent (e.g., hydrogen and halogen) atoms are placed in the same group as the atom to which they are bonded. |
| H2 | Atoms bonded to each other by two or more bonds are placed in the same group. |
| H3 | Any atom in group, *i*, may not be bonded to more than one other atom outside of group *i*. |
| S1 | Atoms contained in an aromatic system are placed in the same group. |
| S2 | −OH and −SH atoms are placed in the same group as the atom to which they are bonded. |

methods, the energy of the system is calculated directly from the energies of the fragments. Suárez et al. grouped fragmentation methods into those using overlapping fragments and those using disjoint fragments.[34] A more elaborate classification is presented in a recent review by Gordon et al.[35]

The fragment molecular orbital (FMO) method[36] first introduced by Kitaura et al. in 1999 belongs to the density-based approach in the classification by Li. In this method, fragments are formed by detaching covalent bonds electrostatically, where two electrons from the broken bond are assigned to only one fragment. The Fock operators are then constructed, and the electronic energies of the fragment monomers are calculated self-consistently in the Coulomb bath of the rest of the system. In order to account for the two- and three-body interfragment interactions, the same procedure can be performed for the fragment dimers (FMO2) and fragment trimers (FMO3), though the energy calculations are only done once. Finally, the total energy of the system is calculated in terms of the monomer, dimer, and trimer energies. The FMO method has been considerably improved and applied to a variety of systems.[37−41] The electrostatically embedded many-body expansion (EEMB) method[42−45] is closely related to the FMO method, but the fragments in EEMB are embedded in a constant electrostatic potential, which is often represented using point charges. Recently, Jensen et al. introduced the effective fragment potential (EFMO) method,[46] which is a merger between the FMO and the EFP. The fragment energies in EFMO scheme are computed in vacuum, and the long-range and many-body interactions are included in Coulomb and polarization terms. These two distinct factors significantly reduce the computational cost of the EFMO in comparison with the FMO method.

The energy-based approaches include the molecular tailoring approach (MTA),[47,48] the kernel energy method (KEM),[49−51] the molecular fractionation with conjugate caps (MFCC) method[52−56] and its generalized version the generalized energy-based fragmentation (GEBF) method,[33] the systematic fragmentation method (SFM),[57−59] and the isodesmic fragmentation method (IFM) developed in our group.[60−62] In these methods, the system is divided into overlapping fragments where the broken bonds are capped with hydrogen atoms or conjugate caps "concaps"[52] to form an initial set of fragments. The intersecting parts are subsequently eliminated, and *n*-body interactions are checked for double counting to form the final list of fragments with appropriate coefficients. The main difference between these methods is how the initial set of fragments are constructed. The MTA creates the initial fragment set by placing spheres of a certain dimension centering at each atom and assigning all atoms within each sphere to a fragment.[63] In the SFM,[57] the molecule is treated as a collection of mutually exclusive functional groups, and the initial fragments are formed by breaking bonds separated by *n* functional groups at fragmentation level-*n*. The IFM[60] follows the similar methodology as the SFM but differs in the way groups are combined to form fragments, resulting in smaller fragment sizes at higher levels of fragmentation. The MFCC method was designed specifically for protein−ligand systems, where the initial set of fragments are the constituent amino acids capped with concaps. In the GEBF method,[64] the initial fragments are formed in a similar fashion as the SFM at level 2, but hydrogen bonds are treated as "real" chemical bonds and a distance threshold is used to included nonbonded fragments. An additional feature of the GEBF method is that fragments are embedded in the field of point charges on all other atoms outside the fragments. The same idea has been incorporated into the electrostatic field-adapted MFCC method,[65] which also allows the addition of the surrounding environment using point charges. The recent SFM implementation[59] accounts for nonbonded interactions by including some fragment−fragment interactions, where distant fragments can be done electrostatically or by using EFP calculations.[19] A more comprehensive discussion of fragmentation methods can be found in the recent review given by Gordon et al.[35]

In the following section we will outline a simple energy-based fragmentation scheme, termed the combined fragmentation method (CFM), which resembles other energy-based methods but has overcome some of their limitations. The fragments of CFM are considerably smaller in size than those of MFCC and GEBF methods. Furthermore, the formation of groups following a number of simple rules prevents capping growth[60] and close contact capping hydrogens. Following this section, we give a simple example then go on to present our results and discussion for various molecules of biological interest.

## 2. THE COMBINED FRAGMENTATION METHOD

From the Lewis structure of the molecule, collect valence bonded atoms together into sets. Each set of valence bonded atoms is designated as a group. There are a few simple rules to follow in forming these groups, and those are provided in Table 1. In order to ensure the greatest possible computational efficiency, the number of atoms contained in each group should be kept as small as possible while satisfying these grouping rules.

Having placed all atoms in a molecule into groups, we next form a precursory set of fragments. The precursory fragments are generated in two steps: (a) form all possible pairs of valence bonded groups and (b) each group that appears more than once in (a) will itself generate a fragment. At this point the fragments are not yet valence completed molecules as they contain "dangling" or severed bonds. If valence completed fragment molecules are desired, then the severed bonds are capped with hydrogen atoms. A capping hydrogen atom is placed along the same direction as the bond that was severed at a distance of $r_{XH} = (r_{XY}/r^0_{XY})r^0_{XH}$. Atom X is in the fragment being capped that was bonded to atom Y, $r_{XY}$ is the bond length of the severed bond, and $r^0_{XY}$ and $r^0_{XH}$ are the standard bond lengths of the severed and

**Table 2. Standard Bond Lengths Used in this Work**

| bond | standard length (Å) |
|---|---|
| C−C | 1.54 |
| C−N | 1.47 |
| C−H | 1.09 |
| N−H | 1.01 |

X−H bonds, respectively. For the values of the standard bond lengths used in this work see Table 2.

Already the energies of these precursory fragment molecules could be linearly combined to produce a crude estimate of the total energy of the molecule.

$$E_{\text{precursory}} = \sum_{i}^{N_{\text{frag}}} c_i E(\hat{h}F_i) \tag{1}$$

where $\hat{h}$ represents capping the fragment directly to the right with hydrogen atoms and $c_i$ is the fragmentation coefficient for precursory fragment $F_i$. Note that the above description of the precursory fragmentation is the same as the level 1 fragmentation of refs 60 and 66. It is also of note that in order to obtain the energies of the fragment molecules and to precisely locate the capping hydrogens, obviously a three-dimensional structure would be needed, but only a Lewis structure or the atom connectivity matrix is necessary to derive the expression for fragmentation.

The coefficients to the energies, $c_i$, in eq 1 are all 1 for each of the fragments generated in step (a) above and −1 for each time the same fragment appears in step (b). No fractional coefficients can exist, and only in the case of branched molecules will coefficients less than −1 appear for the fragments generated in step (b). The main reasons for the crudeness of the total energy estimate from the above fragmentation are due to the small size of the fragment molecules (being either a pair of groups, or a single group) and the complete neglect of nonbonded interactions—interactions particularly important in biomolecules due to, at the very least, hydrogen bonding. This deficiency is to a great extent rectified in the following step of our fragmentation method.

An estimate of the nonbonded energy missing from eq 1 of our precursory fragmentation can be obtained by interacting pairwise the precursory fragments with each other. This treatment of the nonbonded interaction energy is similar to that described in the GEBF method[33] and the recent implementation of SFM,[59] although the original idea can be attributed to Zhang and Zhang in 2003.[52] In order to perform this interaction, we write the interaction energy between fragments, $F_i$ and $F_j$, as

$$\mathcal{O}(F_i, F_j) = E(\hat{h}\{F_i \cup F_j\}) - E(\hat{h}F_i) - E(\hat{h}F_j) + E(\hat{h}\{F_i \cap F_j\}) \tag{2}$$

where $\mathcal{O}$ is an interaction energy, $E$ the electronic energy of the molecule between the parentheses, and the symbols ∪ and ∩ represent union and intersection, respectively. This approach ensures that zero interaction energy is obtained if $F_i = F_j$ or if either fragment is a subfragment of the other. Note that under the assumption that $\mathcal{O}(F_i,F_j) = 0$, eq 2 is akin to that given by Ganesh et al. in eq 2 of ref 63. Thus the correction to the initial estimation of the total energy from precursory fragments is given by

$$\mathcal{O}_{\text{corr}} = \sum_{j>i}^{N_{\text{frag}}} c_i c_j \mathcal{O}(F_i, F_j) \tag{3}$$

and a new estimate of the total energy of the molecule, $E_{\text{CFM}}$, is given by adding eqs 3 to 1, i.e.,

$$E_{\text{CFM}} = E_{\text{precursory}} + \mathcal{O}_{\text{corr}} = \sum_{i}^{N_{\text{frag}}} c_i E(\hat{h}F_i) + \sum_{j>i}^{N_{\text{frag}}} c_i c_j \mathcal{O}(F_i, F_j) \tag{4}$$

Equation 4 accounts for all one- and two-body fragment interactions. In terms of groups, all the one-, two-, and significant three- and four-body interactions are accounted for. For noncyclic systems, one of the advantages of adding eqs 1 to 3 is that all of the terms that appear in eq 1 are canceled by terms of opposite sign in eq 3. The remaining expression on the right-hand side of eq 4 contains fragments with all of the groups bonded together (termed bonded fragment molecules) or precursory fragment pairs not directly bonded to each other (termed nonbonded fragment molecules). Thus eq 4 can be simply rewritten as

$$E_{\text{CFM}} = E_{\text{b}} + E_{\text{nb}} \tag{5}$$

Also of note is that capping hydrogen interactions are almost exactly canceled in the above treatment. Despite the improvement in the description of bonded and nonbonded interactions, up until this point in our discussion, the energies of all fragment molecules would be computed in the absence of any Coulomb field. As fragment molecules are meant to accurately represent part of the entire molecule in which they are embedded, a better description of the fragment is to place it in a Coulomb field that at least to some extent approximates what it would experience in the complete molecule. To this end we surround each fragment molecule with a set of point charges that are located only at the sites of nonfirst period atoms (i.e., heavy atoms) not present in the fragment molecule under consideration.

This treatment is not new, as others[33,65] and ourselves[67] have also utilized a charge field in fragmentation calculations. Note that the effect of including a charge field is directly analogous to that described by Li et al.,[33] except that in the case where the total energy separates into a bonded and nonbonded energy, the sum of the bonded fragment coefficients is always 1 and the nonbonded always 0. For cyclic systems, the sum of all the coefficients is always 1. As a consequence, no electrostatic or induction interaction is double-counted. Furthermore, the bonded energy with the charge field present also includes the electrostatic monopole−multipole nonbonded interaction and the monopole-induced multipole induction interaction.

In this work the point charges were obtained from a contracted distributed multipole analysis.[68,69] We recognize the crudity of expressing the electrostatic field of the rest of the molecule not present in the fragment molecule with only distributed monopoles on heavy atoms. Indeed, we have shown (among others) that for accurate molecular electrostatic potentials at least distributed dipoles and quadrupoles should also be included.[62,70] Despite the crudity, a monopole field does appear to capture the essence of the Coulomb field generated by the remainder of the molecule. This is illustrated in Section 4 of this work, where we have for some systems also added distributed dipoles and quadrupoles. We intend in later work to perform a detailed analysis of the utility and accuracy of other Coulomb fields.

The charge field itself was obtained by first computing the wave function of the bonded fragment molecules appearing in the expression for $E_{\text{b}}$ in eq 5 and summing the evaluated charges contracted to each heavy atom using the same linear combination

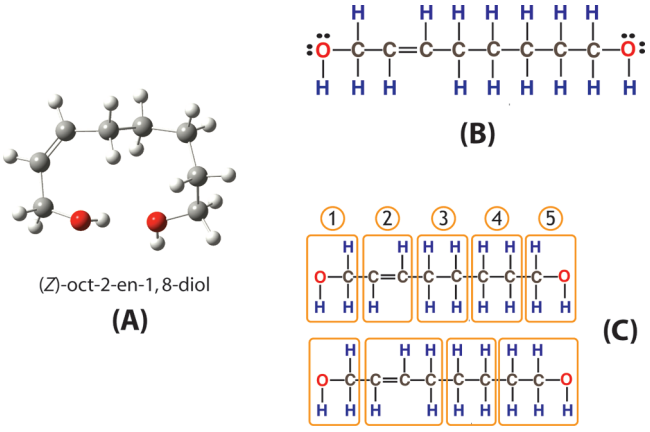


**Figure 1.** (A) (*Z*)-oct-2-ene-1,8-diol; (B) Lewis structure of (A); and (C) atoms sets to form groups.

as that used for $E_b$. These charges were then placed around the bonded fragment molecules, and their wave functions recomputed to obtain an updated set of charges. The process continued one more time before a converged set of charges was obtained. This final set of charges was then placed around all fragment molecules, both bonded and nonbonded, for the final fragmentation energy computation.

## 3. AN EXAMPLE OF THE CFM

Consider the relatively small molecule (*Z*)-oct-2-ene-1,8-diol given in Figure 1. The two −OH groups are H-bonded. Figure 1C shows the two possible sets of groups associated with this molecule following the grouping rules provided in Table 1. The first set with smaller groups is preferred. This five-group set forms the smallest fragments, according to the CFM algorithm, and will be labeled as 12345 in terms of its groups. The expression for $E_{\text{precursory}}$ is

$$E_{\text{precursory}}(12345) = E(12_H) + E({}_H23_H) + E({}_H34_H) + E({}_H45) - E({}_H2_H) - E({}_H3_H) - E({}_H4_H)$$

where each subscript "H" in the above expression represents a capping hydrogen. After interacting the precursory fragments with each other, we arrive at the following expression for $\mathcal{O}_{\text{corr}}$,

$$\begin{aligned}\mathcal{O}_{\text{corr}}(12345) = {} & E(1234_H) + E(12_{HH}45) + E({}_H2345) \\ & - E(12_{HH}4_H) - E({}_H234_H) - E({}_H2_{HH}45) \\ & - E(12_H) - E({}_H23_H) + E({}_H2_{HH}4_H) - E({}_H34_H) - E({}_H45) \\ & + E({}_H2_H) + E({}_H3_H) + E({}_H4_H)\end{aligned}$$

to which we add $E_{\text{precursory}}$ to finally obtain

$$\begin{aligned}E_{\text{CFM}}(12345) = {} & E(1234_H) + E({}_H2345) + E(12_{HH}45) \\ & - E(12_{HH}4_H) - E({}_H234_H) - E({}_H2_{HH}45) + E({}_H2_{HH}4_H)\end{aligned}$$

Here we note that

$$E_b(12345) = E(1234_H) + E({}_H2345) - E({}_H234_H) \quad (6)$$

and

$$E_{\text{nb}}(12345) = E(12_{HH}45) - E(12_{HH}4_H) - E({}_H2_{HH}45) + E({}_H2_{HH}4_H) \quad (7)$$

$E_b$ represents a much better estimate of the bonded energy compared to $E_{\text{precursory}}$, which in the case of "linear" molecules is the same expression as our level 3 fragmentation from our earlier work.[60] For this "linear" five-group molecule, the only two-body interaction not accounted for in eq 6 is the (1,5) interaction. Examination of eq 7 reveals that not only is this interaction accounted for but also the three-body interactions (1,2,5) and (1,4,5) as well as the four-body interaction (1,2,4,5). Thus eq 7 represents a significant improvement over our earlier attempts to account for nonbonded interactions. Finally, careful examination of eqs 6 and 7 shows that all the group-capping hydrogen and capping hydrogen−capping hydrogen interactions cancel almost exactly with only expected small terms remaining.

## 4. RESULTS AND DISCUSSION

In this section we will demonstrate the effectiveness of our CFM for various molecular systems both in vacuum and in a continuum solvent using PCM. All the conventional quantum chemistry calculations for the full systems and the fragments were performed using the Gaussian 09 package[71] at HF/6-311G(d).

**4.1. Fifteen Small Proteins.** The X-ray crystal structures of fifteen small proteins were chosen from the Protein Data Bank (PDB).[72] The positions of the missing hydrogen atoms were uniquely determined from the coordinates of the heavy atoms using generalized valence shell electron pair repulsion theory. Missing carbon atoms were also added appropriately based on their hybridization state and the positions of adjacent atoms. For proteins which contain Gln, Asn, and His, all possible "flipped" states were considered to account for misassignment of element identities. Therefore, the presence of each Asn, Gln, or His gives two possible flipped states arising from the possible flipping of terminal groups. The presence of each unprotonated His results in two additional possible isomers for a total of four possible states for each His. After all possible flipped states were obtained, calculations at HF/STO-3G were performed, and the lowest energy structure was chosen. The final Cartesian coordinates can be found in the Supporting Information.

The CFM was carried out on these proteins, and the results are given in Table 3. One can see that CFM gives reasonably accurate results for most of the systems with errors generally less than 3 m-$E_h$ except for 3DS1. The precise reason for the relatively large error associated with this one protein remains unknown. In addition, the presence of a Coulomb field in the fragment calculations to describe the electronic environment in the complete molecule can improve the results, especially for highly polar (e.g., molecules containing helices) or charged molecules. The atomic charges or distributed monopoles used were obtained from Stone's distributed multipole analysis[69] using the wave function of the full system. The results using the "exact" partial charges would give a threshold for performance of CFM.

As discussed in Section 2, these point charges can be generated without prior knowledge of the whole system by first performing calculations on all the bonded fragments in the absence of the background charges to have an initial guess of the charges on heavy atoms in the whole molecule. These charges can then be incorporated into the fragment calculations to provide a better

**Table 3. CFM Results for 15 Small Proteins Obtained From the Protein Data Bank**

| | | | | | | CFM error (m-$E_h$) | | | |
|---|---|---|---|---|---|---|---|---|---|
| protein | $N_{atoms}$ | $N_+$ | $N_-$ | $N_{basis}$ of the largest frag. | total energy ($E_h$) | no field | charge | iterated charge | SCRF |
| 1D4T | 191 | 3 | 1 | 375 | −4344.94841 | 4.06 | −0.03 | 0.21 | - |
| 1OAI | 124 | 2 | 2 | 360 | −3254.33618 | 1.38 | −0.90 | 0.27 | −1.17 |
| 2ONW | 70 | 1 | 1 | 339 | −1889.79112 | −4.09 | 0.61 | 0.65 | 0.51 |
| 3DRF | 96 | 1 | 1 | 342 | −2513.51695 | 0.61 | 3.61 | 3.73 | −0.04 |
| 3DRJ | 72 | 2 | 1 | 315 | −1701.54758 | 0.11 | 4.34 | 3.45 | 0.13 |
| 3DS1 | 182 | 1 | 4 | 492 | −4645.66615 | −3.90 | 5.23 | 7.80 | −1.99 |
| 3FOD | 87 | 1 | 1 | 351 | −2049.15843 | −1.79 | 1.10 | 1.26 | 0.51 |
| 3FPO | 90 | 1 | 1 | 336 | −2489.95924 | −1.32 | −0.92 | 0.59 | 0.34 |
| 3FQP | 114 | 2 | 1 | 375 | −2508.42256 | 1.12 | −0.70 | −0.21 | −1.11 |
| 3FTK | 100 | 1 | 1 | 378 | −2665.07650 | 1.00 | 1.10 | 1.02 | −0.63 |
| 3FVA | 96 | 1 | 1 | 360 | −2605.10950 | −2.81 | −0.65 | −0.57 | 0.49 |
| 3G03 | 193 | 1 | 2 | 495 | −4793.93629 | 5.24 | 1.16 | 2.29 | −3.89 |
| 3HYD | 122 | 1 | 2 | 369 | −2757.19029 | 2.10 | 1.26 | 1.47 | 0.73 |
| 3NHC | 85 | 1 | 1 | 378 | −2445.24513 | −3.09 | 0.66 | −1.38 | 0.24 |
| 3NVG | 97 | 1 | 1 | 375 | −2726.09190 | 0.78 | −0.48 | 2.20 | 1.11 |

**Table 4. Dihedral Angles Used in This Work to Construct the Listed Secondary Structures**

| structure | $\phi$ (°) | $\psi$ (°) |
|---|---|---|
| $\alpha$-helix | −57 | −47 |
| $3_{10}$-helix | −49 | −26 |
| $\beta$-strand | −119 | 113 |

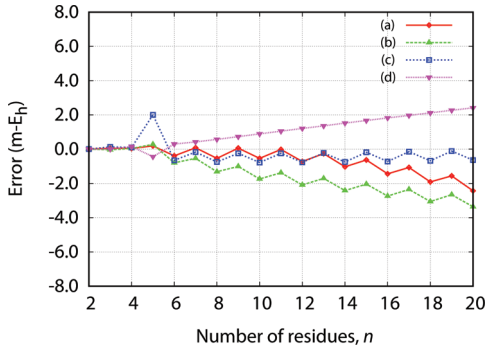


**Figure 2.** (a) CFM results for $\alpha$-helix $(Ala)_n$, $n \leq 20$, in the presence of Coulomb field described using distributed monopoles; (b) distributed monopoles and dipoles; (c) distributed monopoles, dipoles, and quadrupoles; and (d) finally the results when the peptide links are kept intact.

approximation of the atomic charges in the full system. This procedure is repeated until the set of charges on all the atoms or the fragmentation energy $E_{CFM}$ converges, which we found to take no more than three iterations.

From Table 3 it can be seen that the errors for 3DRF, 3DRJ, and 3DS1 are worsened when the charge field is present. The positive errors indicate that the fragmentation energies are lower than the energies obtained from ab initio calculations for the whole molecules. A possible explanation is that the point charges used "over polarize" the fragment molecules. This is not surprising as the electrostatic field of the rest of the molecule not present in the fragment molecules can only be crudely described using

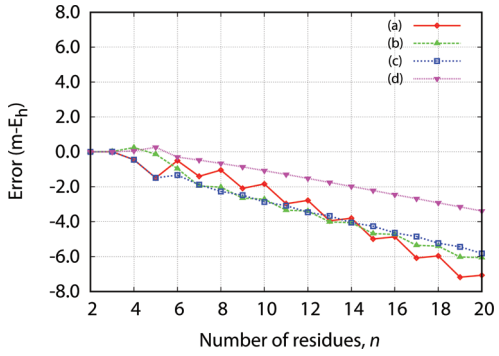


**Figure 3.** (a) CFM results for $3_{10}$-helix $(Ala)_n$, $n \leq 20$, in the presence of Coulomb field described using distributed monopoles; (b) distributed monopoles and dipoles; (c) distributed monopoles, dipoles, and quadrupoles; and (d) finally the results when the peptide links are kept intact.

point charges on heavy atoms (see Section 2 for more detailed discussion).

**4.2. Polyalanine Helices and $\beta$-Strands.** Due to the prevalence of $\alpha$- and $3_{10}$-helices and $\beta$-strands in proteins, these secondary structures were used to test the performance of the CFM. Canonical helices and $\beta$-strands consisting of alanine residues were constructed using the dihedral angles in Table 4. These polyalanines are terminated with amino and aldehyde groups at the N and C terminals, respectively.

The grouping rules described in Section 2 do not uniquely specify how atoms are assigned to groups. However, the group assignment does not significantly affect the accuracy of CFM as illustrated for $3_{10}$-helix $(Ala)_6$. For this molecule, there are a total of 34 possible different ways of assigning atoms to groups; all result in errors below 1.6 m-$E_h$ with standard deviation $\sigma \leq 0.9$ m-$E_h$. A systematic algorithm to determine all possible group assignments is described in the Appendix.

Results for all the canonical polyalanines $(Ala)_n$ $\alpha$- and $3_{10}$-helices, $n \leq 20$, are presented in Figures 2 and 3. Additional results including those for the $\beta$-strands are given in the Supporting Information. While accurate results are attained


dx.doi.org/10.1021/ct200783n | J. Chem. Theory Comput. 2012, 8, 469–478

Table 5. Comparison of CFM and GEBF Results for $(Ala)_{10}$ and $(Ala)_{18}$ in $3_{10}$- and $\alpha$-Helical and $\beta$-Strand Forms

| | GBEF | | CFM | |
|---|---|---|---|---|
| protein | $N_{basis}$ of largest frag. | error (m-$E_h$) | $N_{basis}$ of largest frag. | error (m-$E_h$) |
| $\beta$-strand acetyl$(ala)_{10}NH_2$ | 396 | 0.49 | 345 | −1.11 |
| $\alpha$-helix acetyl$(ala)_{10}NH_2$ | 628 | 0.27 | 345 | 0.33 |
| $3_{10}$-helix acetyl$(ala)_{10}NH_2$ | 616 | −0.32 | 345 | −0.19 |
| $\beta$-strand acetyl$(ala)_{18}NH_2$ | 387 | 1.07 | 330 | −2.95 |
| $\alpha$-helix acetyl$(ala)_{18}NH_2$ | 600 | −1.45 | 330 | −3.06 |
| $3_{10}$-helix acetyl$(ala)_{18}NH_2$ | 588 | −0.65 | 330 | −4.17 |

without the presence of Coulomb field in fragment calculations for $\beta$-strands (errors <0.1 m-$E_h$), the inclusion of charge field is important for accurate energy predictions of $\alpha$- and $3_{10}$-helices as they have huge dipole moments along the helical axes. Our results also suggest that the energies of nonpolar molecules, such as $\beta$-strands, can be reasonably predicted using the formula for the bonded energy, $E_b$ (errors <2.5 m-$E_h$).

It is apparent from Figures 2 and 3 that the errors for $\alpha$- and $3_{10}$-helices exhibit a small ($\approx$1 m-$E_h$) systematic odd−even variation. This zigzag pattern is attributed to the alternate breaking of peptide bonds of even residues and indeed disappears when the all peptide links are kept intact. However, preventing the breaking of peptide bonds can give rise to larger groups and therefore undesirably large fragments and is therefore unnecessary.

Figures 2 and 3 also show positive correlations between the absolute errors and the number of residues, $n$. The same linear trend is observed for the relative energies $\Delta_n = E_n - E_{n-1}$ of $\alpha$- and $3_{10}$-helices, where $E_n$ denotes the energy of $(Ala)_n$ and is often regarded as being caused by a so-called "cooperative" effect[73−76] between hydrogen bonds which increases with the chain length. This effect was shown to be simply a result of a large induction interaction generated by the parallel lying dipoles present in these helices, which is stronger for $\alpha$-helices than $3_{10}$-helices of the same length.[77] To account better for the through-space induction, distributed dipoles and quadrupoles were added in the fragment calculations. The errors with increasing number of residues become more flattened, and the results for $\alpha$-helices show a better improvement than $3_{10}$-helices as expected. Though $3_{10}$-helices still have considerable errors for $3_{10}$-helix $(Ala)_{20}$, it should be noted that the error at six residues, the longest chain length normally observed for $3_{10}$-helices in proteins,[77] is well within the acceptable range of $\pm$2 m-$E_h$.

The CFM was also applied to the polyalanines in the work of Li et al[33] (see Table 5). Although the GEBF method by Li and co-workers results in smaller errors, its fragments are notably larger. What should also be noted here is the different fragment sizes for different conformations in GEBF, which indicates the dependence of its energy expression on the three-dimensional structure of the molecule. On the contrary, CFM produces fragments solely based on the atom connectivity, and therefore the fragmentation of $\alpha$-helix, $3_{10}$-helix, and $\beta$-strand gives to the same sets of fragments, though the fragments can have different geometries. Furthermore, the formation of fragments from groups following the CFM scheme is independent of the molecular size. This can also be observed for the systems presented Table 3, where sizes of the largest fragments in various proteins are similar. These two features make CFM a black-box approach for the calculations of proteins, regardless of size, from their primary sequences. As with our previous helix series, the results here can also be improved by including distributed dipoles and quadrupoles in the fragment calculations (see the Supporting Information).

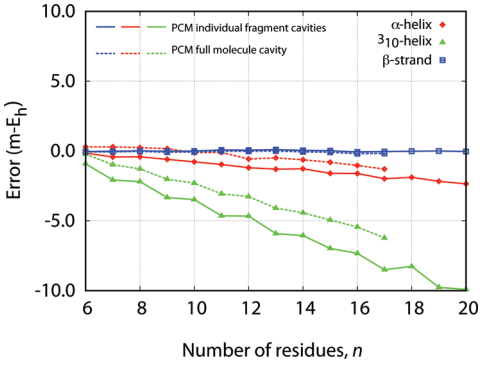


Figure 4. CFM-PCM errors for the canonical polyalanine series $\alpha$-helix, $3_{10}$-helix, and $\beta$-strand $(Ala)_n$, $6 \le n \le 20$, with the full molecule cavity (solid lines) and individual fragment cavities (dashed lines).

**4.3. Self-Consistent Reaction Field Results.** In this section the CFM was coupled with the polarizable continuum model (PCM),[78] termed CFM-PCM, to include the solvent effects in the calculation of the electronic energies. One important aspect of PCM is the definition of the cavity which hosts the solute molecule. The cavity surface is defined as the solute−solvent interface and used as a boundary condition to describe the electrostatic interaction between the solute and the continuum.[79] The choice of cavity is expected to affect the accuracy of PCM calculations. Indeed, as each fragment molecule is a part of the full system, the cavity surface of the whole molecule was used in the PCM fragment calculations (termed PCM full molecule cavity calculations) to compare with those results obtained by simply using each individual surface generated for each fragment molecule (termed PCM individual fragment cavity calculations). As pointed out by Mei et al.,[80] these two approaches represent two different solvation situations with the former only strictly correct. However, such calculations are expensive. Alternatively, because of the high degree of systematic overlap between our fragment molecules, there is a substantial degree of cancellation of error occurring in the PCM individual fragment cavity calculations, which we show shortly leads to only a small increase in error over the strictly correct PCM full molecule cavity calculations. It is also important to note that for our PCM-CFM calculations, we found it entirely unnecessary to add a charge field about each fragment molecule.

CFM-PCM was applied to the canonical polyalanine series described in the previous section. As is evident from Figure 4, the

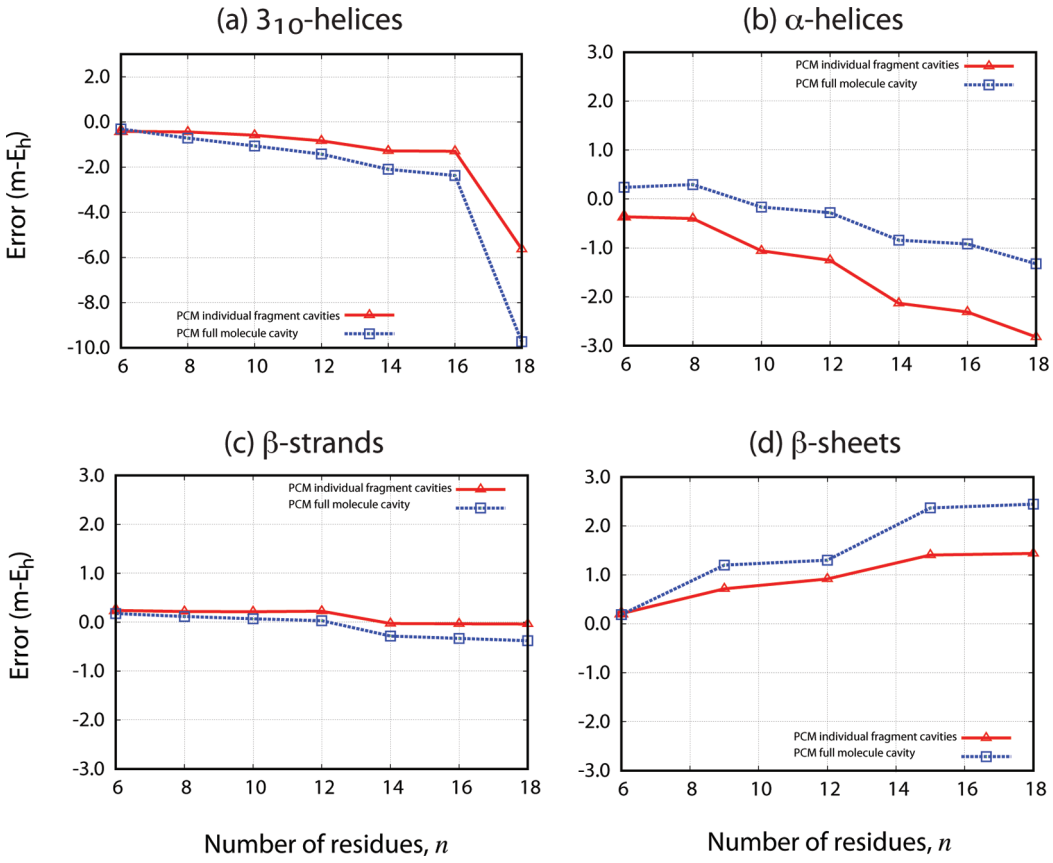


**Figure 5.** CFM-PCM results for a series of optimized polyalanines $(Ala)_n$ in the forms of $\alpha$-helix, $3_{10}$-helix, $\beta$-strand, and $\beta$-sheet.

inclusion of the full molecule cavity for every fragment does not give rise to significant improvement in the results and is therefore not necessary in CFM-PCM calculations. This is an important conclusion due to the extreme computational expense of the full molecule cavity caluclations.

CFM-PCM was also applied to a series of polyalanines $(Ala)_n$ from the work of Wieczorek and Dannenberg,[81] all of which are capped with the acetyl and amino groups at the N- and C-terminus, respectively, and the number of residues $n$ = 2, 4, 6, ..., 18. Note that these structures were optimized. The CFM-PCM results using the full molecule cavity and the individual fragment cavities are displayed in Figure 5. Except for an outlier $3_{10}$-helix $(Ala)_{18}$, which is likely due to the given molecular structure, the errors are all within $\pm 3$ m-$E_h$. The results also reflect good performance of CFM-PCM without the need of a Coulomb field to represent parts of the molecule which are not present in the fragments. A possible explanation for not requiring a Coulomb field is that the presence of a dielectric medium around the solute gives rise to stronger polarization than the long-range nonbonded interactions in different parts of the solute. It is noted here again that the computationally full molecule cavity can be ignored in CFM-PCM calculations.

In addition, the optimized polyalanine $3_{10}$-helices, as shown in Figure 5, exhibit significantly better results than the canonical helices. This is expected since the nonbonded interactions for the canonical peptides are stronger than for the optimized peptides as a result of the dipole moments being perfectly aligned.

As evident from Table 3, CFM-PCM also gives excellent results for the series of small proteins selected from the Protein Data Bank. The fragment calculations were all done in the absence of electrostatic field using the individual fragment cavities.

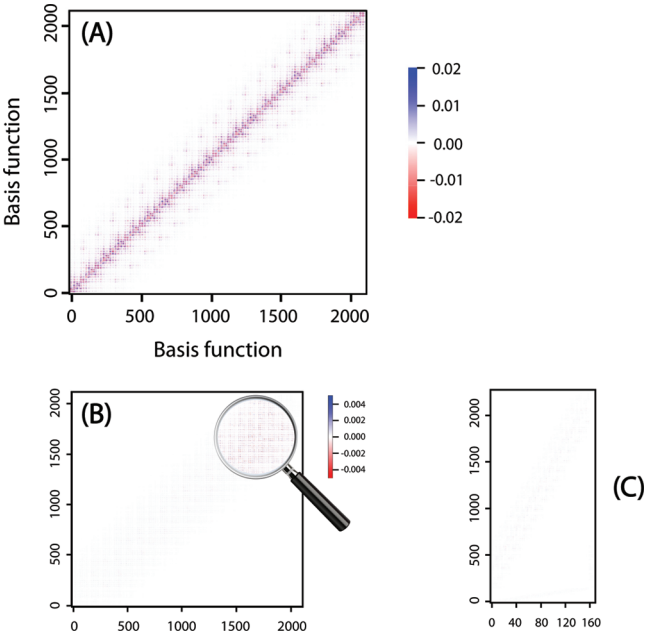


**Figure 6.** (A) Density matrix for $\alpha$-helix $(Ala)_{20}$ obtained from HF/6-311G(d); (B) the difference between (A) the resulting density matrix obtained from CFM; and (C) the difference density matrix for the capping hydrogens.


dx.doi.org/10.1021/ct200783n |J. Chem. Theory Comput. 2012, 8, 469–478

**4.4. Density Matrix.** In addition to the electronic energy and its derivatives, the density matrix of a molecule can also be predicted using CFM. This is done by expressing individual elements of the density matrix in terms of the appropriate superposition of matrix elements of the fragments. In other words, the fragment density matrices can be linearly combined to form the total density matrix.

Figure 6 displays the total density matrix of $3_{10}$-helix $(Ala)_{20}$ and the errors resulting from CFM calculations for the whole molecule as well as the capping hydrogens. The results suggest that not only can CFM provide a good prediction for the total electronic energy but also can describe satisfactorily the electronic environment around the molecule. Furthermore, the effect induced by the presence of capping hydrogens is indeed negligible and therefore does not contribute significantly to the fragmentation error.

## 5. CONCLUSION

In this paper we introduced a combined fragmentation method (CFM), which forms fragments from a Lewis structure of target molecule following a simple algorithm. The method was tested extensively both in vacuum and in the continuum solvent using polarizable continuum model for biological molecules. Test systems include 15 small proteins selected from the Protein Data Bank, a series of canonical $\alpha$-helix, $3_{10}$-helix, and $\beta$-strand polyalanines, and a series of optimized polyalanines from the work of Wieczorek and Dannenberg.[81]

The results for these molecular systems prove that the CFM is a reliable method for energy prediction of large molecules. For nonpolar systems (e.g., $\beta$-strand), the bonded energy gave a reasonably good approximation of the total energy but including nonbonded interactions could further improve the results. In addition, the errors for these systems are negligible even without the presence of a Coulomb field. However, the presence of such a Coulomb field is important for accurate energy predictions of highly polar (e.g., $\alpha$- and $3_{10}$-helices) or charged systems. Distributed dipoles and quadrupoles could also be included in the field description to better account for the through-space induction for helices.

We also compared our CFM with the GEBF method by Li and co-workers.[33] Although our method exhibits slightly larger errors for the six polyalanines, it should be noted that CFM produces smaller fragments and fragments solely based on the atom connectivity matrix. Results for a set of 15 small proteins show that the fragment size does not depend on the size of the whole system. Therefore, CFM could be utilized as a black-box approach for energy calculations of proteins, regardless of size, from their primary sequences. In addition, the density matrix can also be computed in a similar fashion as for the energy.

Lastly, CFM exhibits excellent results for molecular energy predictions in a solvent continuum. This can be achieved without the presence of Coulomb field, regardless of the polarity of the parent system, in CFM-PCM calculations and by using cavities generated for individual fragments.

## APPENDIX

The algorithm used to generate all possible ways to assign atoms to disjoint groups for a given molecule is presented here. A group assignment is termed a graph. Each atom of the molecule is represented by a vertex $v_i$ and an edge between two vertices represents a chemical bond (Table 6).

The *makeGroups* and *makeGraphs* algorithms are presented in Tables 7 and 8, in which $|S|$ is used to denote the cardinality of set $S$. Note that if there are user-defined groups, the procedure can be stopped at $i > 1$ in step 3 of 7 as the first group being predefined must always be present in all graphs, and therefore the consideration of alternate starting point is unnecessary.

**Table 6**

| | |
|---|---|
| step 1 | Read the Cartesian coordinates, atom connectivity matrix, and user-defined groups. |
| step 2 | Perform an edge contraction by identifying singly valent atoms and atoms to which they are bonded; denote the new set of vertices as $V = \{v_1, v_2, ...\}$. |
| step 3 | Run *makeGroups* to form all possible groups $G = \{g_1, g_2, ...\}$ satisfying the grouping rules and identify group connectivity. |
| step 4 | Run *makeGraphs* to form all possible graphs $F = \{f_1, f_2, ...\}$, then compute their scores. The best graph is one with smallest groups of comparable sizes. |

**Table 7. The *makeGroups* Algorithm**

| | |
|---|---|
| step 1 | Set $i = 0$, $n = 0$, and $G = 0$. |
| step 2 | While $i \leq \|V\|$, let $i = i + 1$. Create a new group $g = \{v_i\}$ and add it to the set of groups $G = G \cup \{g\}$. |
| step 3 | While $n \leq \|G\|$, set $n = n + 1$. |
| step 4a | Check the grouping rules if $g_n$ forms a group. |
| | If true, then go to step 3, else $G = G - \{g_n\}$ and form set $N$ containing groups $g_j = g_n \cup \{v_j\}$ such that $v_j$ is connected to $v_k \in g_n$ and $g_j \notin g_n \ \forall j$. |
| step 4b | If $N = 0$, then go to step 3, else $G = G \cup N$ and go to step 4a. |

**Table 8. The *makeGraphs* algorithm**

| | |
|---|---|
| step 1 | Set $i = 0$, $n = 0$, and $F = 0$. |
| step 2 | While $i \leq \|G\|$, let $i = i + 1$. Create a new graph $f = \{g_i\}$ and add it to the set of graphs $F = F \cup \{f\}$. |
| step 3 | While $n \leq \|F\|$, set $n = n + 1$. |
| step 4a | Check if $f_n$ forms a graph, i.e., if $\cup f_n = V$. |
| | If true, then go to step 3, else $F = F - \{f_n\}$ and form set $M$ containing "incomplete" graphs $f_j = f_n \cup \{g_j\}$ such that $g_j$ is connected to $g_k \in f_n$, $g_j \cap g_k = 0$ and $f_j \notin f_n \ \forall j$. |
| step 4b | If $M = 0$, then go to step 3, else $G = G \cup M$ and go to step 4a. |

## ■ ASSOCIATED CONTENT

**Supporting Information.** The algorithm used to generate all possible vertex contractions is given in section S1. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: chmbrpa@nus.edu.sg. Telephone: +65 6516 2846.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENT

The authors thanks the National University of Singapore's support from the Academic Research Fund, grant number R-143-000-402-112. The authors also thank the Centre for Computational Science and Engineering for the use of their computers.

## ■ REFERENCES

(1) Pulay, P. *Chem. Phys. Lett.* **1983**, *100*, 151.
(2) Saebø, S.; Pulay, P. *J. Chem. Phys.* **1988**, *88*, 1884.
(3) Saebø, S.; Pulay, P. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213.
(4) Pulay, P.; Saebø, S.; Wolinski, K. *Chem. Phys. Lett.* **2001**, *344*, 543.
(5) Saebø, S.; Pulay, P. *J. Chem. Phys.* **2001**, *115*, 3975.
(6) *Linear-scaling Techniques in Computational Chemistry and Physics: Methods and Applications; Challenges and Advances in Computational Chemistry and Physics*; Zalesny, R Papadopoulos, M. G. Mezey, P. G. Leszczynski, J, Ed.; Springer: Hoboken, NJ, 2011; Vol. *13*.
(7) Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 241101.
(8) Schutz, M.; Werner, H. *J. Chem. Phys.* **2001**, *114*, 661.
(9) Schutz, M.; Werner, H. *Chem. Phys. Lett.* **2000**, *318*, 370.
(10) Schutz, M.; Hetzer, G.; Werner, H. *J. Chem. Phys.* **1999**, *111*, 5691.
(11) Subotnik, J. E.; Sodt, A.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 034103.
(12) Schutz, M. *Phys. Chem. Chem. Phys.* **2002**, *4*, 3941.
(13) Walter, D.; Venkatnathan, A.; Carter, E. *J. Chem. Phys.* **2003**, *118*, 8127.
(14) Venkatnathan, A.; Szilva, A.; Walter, D.; Gdanitz, R.; Carter, E. *J. Chem. Phys.* **2004**, *120*, 1693.
(15) Chwee, T. S.; Szilva, A. B.; Lindh, R.; Carter, E. A. *J. Chem. Phys.* **2008**, *128*, 224106.
(16) Peterson, K. A.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 084102.
(17) Day, P.; Jensen, J.; Gordon, M.; Webb, S.; Stevens, W.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968.
(18) Gordon, M.; Freitag, M.; Bandyopadhyay, P.; Jensen, J.; Kairys, V.; Stevens, W. *J. Phys. Chem. A* **2001**, *105*, 293.
(19) Gordon, M. S.; Mullin, J. M.; Pruitt, S. R.; Roskop, L. B.; Slipchenko, L. V.; Boatz, J. A. *J. Phys. Chem. B* **2009**, *113*, 9646.
(20) Gao, J. *J. Phys. Chem. B* **1997**, *101*, 657.
(21) Xie, W.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1890.
(22) Lynch, B.; Truhlar, D. *J. Phys. Chem. A* **2003**, *107*, 3898.
(23) Svensson, M.; Humbel, S.; Froese, R.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357.
(24) Yang, W. *Phys. Chem. Lett.* **1991**, *66*, 1438.
(25) Parr, R.; Yang, W. *Density-functional theory of atoms and molecules; International series of monographs on chemistry*; Oxford University Press: Oxford, U.K., 1994.
(26) Yang, W. T.; Lee, T. S. *J. Chem. Phys.* **1995**, *103*, 5674.
(27) Szekeres, Z.; Mezey, P. *Mol. Phys.* **2005**, *103*, 1013.
(28) Exner, T. E.; Mezey, P. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 4061.
(29) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2004**, *108*, 4301.
(30) Exner, T. E.; Mezey, P. G. *J. Comput. Chem.* **2003**, *24*, 1980.
(31) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2002**, *106*, 11791.
(32) Mezey, P. *J. Math. Chem.* **1995**, *18*, 141.
(33) Li, W.; Li, S.; Jiang, Y. *J. Phys. Chem. A* **2007**, *111*, 2193.
(34) Suarez, E.; Diaz, N.; Suarez, D. *J. Chem. Theory Comput.* **2009**, *5*, 1667.
(35) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. *Chem. Rev.* **2012**, *112*, 632–672.
(36) Kitaura, K.; Sawai, T.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *312*, 319.
(37) Fedorov, D. G.; Ishida, T.; Uebayasi, M.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 2722.
(38) Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904.
(39) Fedorov, D. G.; Jensen, J. H.; Deka, R. C.; Kitaura, K. *J. Phys. Chem. A* **2008**, *112*, 11808.
(40) Fedorov, D. G.; Slipchenko, L. V.; Kitaura, K. *J. Phys. Chem. A* **2010**, *114*, 8742.
(41) Fedorov, D.; Kitaura, K. *The fragment molecular orbital method: practical applications to large molecular systems*; CRC Press, Taylor & Francis:Boca Raton, FL, 2009.
(42) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1.
(43) Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 33.
(44) Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1573.
(45) Sorkin, A.; Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 683.
(46) Steinmann, C.; Fedorov, D. G.; Jensen, J. H. *J. Phys. Chem. A* **2010**, *114*, 8705.
(47) Babu, K.; Gadre, S. R. *J. Comput. Chem.* **2003**, *24*, 484.
(48) Babu, K.; Ganesh, V.; Gadre, S. R.; Ghermani, N. E. *Theor. Chem. Acc.* **2004**, *111*, 255.
(49) Huang, L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2005**, *103*, 808.
(50) Huang, L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2006**, *106*, 772.
(51) Huang, L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2006**, *106*, 447.
(52) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599.
(53) Zhang, D. W.; Chen, X. H.; Zhang, J. Z. H. *J. Comput. Chem.* **2003**, *24*, 1846.
(54) Zhang, D. W.; Xiang, Y.; Gao, A. M.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 1145.
(55) Zhang, D.; Zhang, J. *J. Chem. Theory Comput.* **2004**, *3*, 43.
(56) Chen, X. H.; Zhang, J. Z. H. *J. Theor. Comput. Chem.* **2004**, *3*, 277.
(57) Collins, M. A.; Deev, V. A. *J. Chem. Phys.* **2006**, *125*, 104104.
(58) Collins, M. A. *J. Chem. Phys.* **2007**, *127*, 024104.
(59) Addicoat, M. A.; Collins, M. A. *J. Chem. Phys.* **2009**, *131*, 104103.
(60) Bettens, R. P. A; Lee, A. M. *J. Phys. Chem. A* **2006**, *110*, 8777.
(61) Lee, A. M.; Bettens, R. P. A *J. Phys. Chem. A* **2007**, *111*, 5111.
(62) Le, H.-A.; Lee, A. M.; Bettens, R. P. A *J. Phys. Chem. A* **2009**, *113*, 10527.
(63) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. *J. Chem. Phys.* **2006**, *125*, 104109.
(64) Hua, S.; Hua, W.; Li, S. *J. Phys. Chem. A* **2010**, *114*, 8126.
(65) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.
(66) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.
(67) Bettens, R. P. A; Lee, A. M. *Chem. Phys. Lett.* **2007**, *449*, 341.
(68) Stone, A. J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047.
(69) Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128.
(70) Le, H.-A.; Bettens, R. P. A *J. Chem. Theory Comput.* **2011**, *7*, 921.
(71) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.;

Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; 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, .; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09 Revision A.1. Gaussian Inc.: Wallingford CT, 2009.

(72) 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.

(73) Wieczorek, R.; Dannenberg, J. *J. Am. Chem. Soc.* **2003**, *125*, 8124.

(74) Wieczorek, R.; Dannenberg, J. *J. Am. Chem. Soc.* **2003**, *125*, 14065.

(75) Ireta, J.; Neugebauer, J.; Scheffler, M.; Rojo, A.; Galvan, M. *J. Phys. Chem. B* **2003**, *107*, 1432.

(76) Tkatchenko, A.; Rossi, M.; Blum, V.; Ireta, J.; Scheffler, M. *Phys. Chem. Lett.* **2011**, *106*, 118102.

(77) Wu, Y.; Zhao, Y. *J. Am. Chem. Soc.* **2001**, *123*, 5313.

(78) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117.

(79) Scalmani, G.; Frisch, M. J. *J. Phys. Chem. A* **2010**, *132*, 114110.

(80) Mei, Y.; Ji, C.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 094906.

(81) Wieczorek, R.; Dannenberg, J. *J. Am. Chem. Soc.* **2004**, *126*, 14198.