JCTC Journal of Chemical Theory and Computation

Article

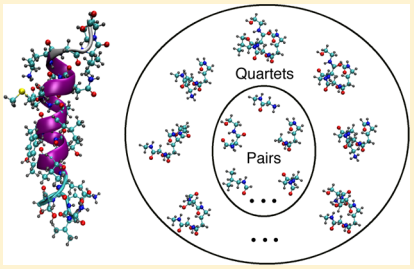
pubs.acs.org/JCTC

Pair−Pair Approximation to the Generalized Many-Body Expansion: An Alternative to the Four-Body Expansion for ab Initio Prediction of Protein Energetics via Molecular Fragmentation

Jie Liu and John M. Herbert*

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

ABSTRACT: We introduce a "pair−pair" approximation to the generalized many-body expansion (pp-GMBE) as an approximation to a traditional four-body expansion, the latter of which is accurate but quickly becomes numerically unstable and ultimately intractable as the number of "bodies" (fragments) increases. The pp-GMBE method achieves a good balance between accuracy and efficiency by defining significant fragment pairs and then fragment quartets. An efficient fragmentation scheme is introduced for proteins such that the largest subsystems contain about 60 atoms. Application of the pp-GMBE method to proteins with as many as 70 amino acids (1142 atoms) reveals that pp-GMBE energies are quite faithful to those obtained when the same level of density functional theory is applied to the entire macromolecule. When combined with embedding charges obtained from natural population analysis, the pp-GMBE approach affords absolute energies that differ by 1−3 kcal/mol from full supersystem results, but it yields conformational energy profiles that are practically indistinguishable from the supersystem calculation at the same level of theory.



1. INTRODUCTION

Ab initio quantum chemistry provides a general theoretical route to study molecular properties of medium-sized molecules containing ∼100 atoms. The computational cost of such methods scales as $O(N_A^a N_B^b)$, where N_A denotes the number of atoms and N_B is the number of basis functions per atom. The exponents a and b depend upon the ab initio method in question, but typically $a \geq 3$ and $b = 2$−4. Although a large number of algorithms that scale linearly with respect to N_A have been devised,[1−5] such methods tend to work best in small basis sets and therefore do not solve the "N_B problem". Molecular fragmentation approaches[4−6] are an alternative route to tractable macromolecular quantum chemistry calculations, but these methods have not been tested consistently or with regular success in the triple-ζ and larger basis sets that are necessary to obtain converged results, even for modern density functional theory (DFT).[7,8] When fragment-based methods do appear to work in large basis sets, this "success" is often due largely to error cancellation.[9,10]

The fragment-based approach is the one pursued here. Our approach is a chemically intuitive means to reduce the computational scaling with respect to N_A while preserving the ability to use reasonable basis sets. The N_A-atom system is partitioned into subsystems using a set of predefined rules. The total energy is computed by assembling the energies of smaller subsystems, each of which is assumed to be representative of its local electronic environment. For a fixed basis set, the computational scaling is reduced from $O(N_A^a)$ to $O(n_A^a)$, where n_A is the size of the largest subsystem (fragment). The subsystem calculations are completely independent of one another; hence, it is straightforward to combine the fragment-based approach with all levels of ab initio electronic structure theory and to parallelize the algorithm without extensive modification to existing quantum chemistry codes. The algorithm is embarrassingly parallelizable, so if a number of processors equal to the number of subsystem calculations is available, then the wall time becomes independent of system size.

A wide variety of fragment-based methods exist in the literature,[4,5] and Richard and Herbert[6] have established a conceptual framework that unifies many of them. The four elements proposed in ref 6 to define a particular fragment-based method are

- a fragmentation method,
- a capping method,
- an embedding method, and
- the number of layers.

In some fragment-based methods, the capping method is included as part of the fragmentation method. The fragmentation method together with the capping method determines how the subsystems are defined. Embedding and multilayer methods play an analogous role in the fragment-based method, namely, to recover the long-range interactions missing from the individual subsystems. The difference is that the embedding method includes only the Coulomb interactions while the multilayer method is able to introduce exchange and correlation interactions via some low-level method applied to the entire supersystem. In the latter case, one may significantly reduce the overall computational effort, as compared to

Received: October 7, 2015
Published: January 5, 2016

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

application of the high-level method to the entire supersystem, but linear scaling is sacrificed.

A good fragment-based method must achieve a delicate balance between accuracy and efficiency. In principle, the many-body expansion (MBE), in which the total energy is partitioned into a sum of monomer, dimer, trimer, ... energies is able to systematically improve the accuracy.[11] (Recent studies have questioned whether this is true in practice, however, at least for large systems.[9,10,12,13] For high-accuracy calculations, one must consider the effects of basis-set superposition error, which behaves quite differently in fragment-based approaches.[9,10,12,14]) A four-body expansion is usually sufficient to obtain an accurate total energy,[15] but the number of tetramers grows so rapidly with system size that four-body expansions are prone to serious loss-of-precision problems.[13] Two-body expansions are much more stable in large systems,[13] yet three- and four-body corrections are needed to obtain "chemical accuracy" of ∼1 kcal/mol.[16−20]

In this work, we will attempt to capture the most significant four-body interactions based on identifying nearby pairs of fragments. The pair−pair fragmentation method that we will introduce consists of two steps: (1) defining the monomers, which include significant fragment pairs, and (2) defining the *n*-mers (unions of the monomers), which include significant monomer pairs. This constitutes a "pair−pair" (pp) approximation to the generalized many-body expansion (GMBE) of Richard and Herbert[6,21,22] and *significantly* reduces the number of independent electronic structure calculations that are required, relative to a primitive four-body expansion or a two-body GMBE. (In the case of a 70-residue protein considered herein, for example, the number of individual subsystem calculations is reduced from more than 10^9 for a four-body expansion to fewer than 12 000 using the pp-GMBE, without sacrificing accuracy.) Charges are extracted from a low-level GMBE calculation in order to improve the accuracy of the method. Finally, we have parallelized our fragmentation code ("FRAGME∩T", developed in previous work[6,21]) in order to take advantage of very large numbers of processors, with subsystem calculations that are also relatively large. In the present work, we describe this methodology and test it for polypeptides and proteins, using a new fragmentation scheme that results in subsystems containing ≲60 atoms, which makes feasible the application of decent levels of electronic structure theory.

2. THEORY

2.1. Energy Expression. *2.1.1. General Energy Expression.* As with a few other fragment-based quantum chemistry methods,[5,23−25] the GMBE method is based on the principle of inclusion/exclusion (PIE), which is a theorem about the cardinality of sets. Let S_1, S_2, ..., S_N be subsets of some set S; these subsets need not be disjoint. There are two possible cases. First, the case where the aforementioned subsets form a complete partition of S, meaning that

$$|S| = \left| \bigcup_{n=1}^{N} S_n \right| \tag{1}$$

Alternatively, the partition might be incomplete, in which case

$$|S| = \left| \bigcup_{n=1}^{N} S_n \right| + \left| \bigcap_{n=1}^{N} \overline{S}_n \right| \tag{2}$$

Here, $\overline{S}_n = S \backslash S_n$ is the complement of S_n in S. The PIE states that the cardinality of the set $S_1 \cup S_2 \cup \cdots \cup S_N$ can be expressed as

$$\begin{aligned} \left| \bigcup_{n=1}^{N} S_n \right| &= \sum_{i=1}^{N} |S_i| - \sum_{i=1}^{N} \sum_{j>i}^{N} |S_i \cap S_j| \\ &\quad + \sum_{i=1}^{N} \sum_{j>i}^{N} \sum_{k>j}^{N} |S_i \cap S_j \cap S_k| \\ &\quad - \cdots + (-1)^{N-1} |S_1 \cap S_2 \cap \cdots \cap S_N| \\ &= \sum_{n=1}^{N} (-1)^{n+1} \sum_{\substack{i_1,i_2,\ldots,i_n \\ (i_1<i_2<\cdots<i_n)}}^{n} |S_{i_1} \cap S_{i_1} \cap \cdots \cap S_{i_n}| \end{aligned} \tag{3}$$

The GMBE is obtained by applying the PIE to a partition of the supersystem's Hamiltonian.[6,21] The latter is

$$\hat{H} = \sum_i \hat{h}_i + \sum_{i,I} \hat{V}_{iI} + \sum_{i<j} \hat{V}_{ij} + \sum_{I<J} \hat{V}_{IJ} \tag{4}$$

where lower-case indices are for electrons and upper-case indices denote nuclei. If the supersystem is divided into N primitive subsystems, then the total Hamiltonian can be expressed exactly as a linear combination of subsystem Hamiltonians, according to the PIE:[6,21]

$$\begin{aligned} \hat{H} &= \sum_{n=1}^{N} (-1)^{n+1} \sum_{\substack{i_1,i_2,\ldots,i_n \\ (i_1<i_2<\cdots<i_n)}}^{n} (\hat{H}_{i_1} \cap \hat{H}_{i_1} \cap \cdots \cap \hat{H}_{i_n}) \\ &\quad + \cap_{n=1}^{N} \hat{H}_n \end{aligned} \tag{5}$$

An energy expression follows as the expectation value of eq 5:

$$\begin{aligned} E &= \sum_{n=1}^{N} (-1)^{n+1} \sum_{\substack{i_1,i_2,\ldots,i_n \\ (i_1<i_2<\cdots<i_n)}}^{n} \langle \Phi | \hat{H}_{i_1} \cap \hat{H}_{i_1} \cap \cdots \cap \hat{H}_{i_n} | \Phi \rangle \\ &\quad + \langle \Phi | \cap_{n=1}^{N} \hat{H}_n | \Phi \rangle \\ &= \sum_{n=1}^{N} (-1)^{n+1} \sum_{\substack{i_1,i_2,\ldots,i_n \\ (i_1<i_2<\cdots<i_n)}}^{n} E_{i_1 \cap i_1 \cap \cdots \cap i_n} + E_{\overline{i}_1 \cap \overline{i}_2 \cap \cdots \overline{i}_N} \\ &= E_0 + \Delta E \end{aligned} \tag{6}$$

Here, Φ is the exact ground-state wave function for the entire supersystem, in which case E is the exact ground-state energy. The complementary energy

$$\Delta E = E_{\overline{i}_1 \cap \overline{i}_2 \cap \cdots \cap \overline{i}_N} \tag{7}$$

depends on the specific form of subsystem Hamiltonian.

Equation 6 is a universal energy expression that can serve as a starting point for making the necessary approximations (i.e., computing localized wave functions for the subsystems) that form the basis of a fragment-based approach. The results obviously depend on how the fragments are defined. In particular, the traditional MBE is simply an *n*-body truncation of the GMBE based on nonoverlapping fragments, and the "generalized energy-based fragmentation" (GEBF) method of Li and co-workers[24,26,27] is a one-body truncation but with overlapping fragments defined based on a distance threshold.

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

[For this reason, we have sometimes referred to the GEBF method as GMBE(1).[22]] The "many overlapping body expansion" (MOBE) of Mayhall and Raghavachari[28] and the systematic molecular fragmentation (SMF) method of Collins et al.[29−31] are truncations that in addition omit certain intersections from eq 6.[6,21] The "molecular fragmentation with conjugated caps" (MFCC) approach,[32−37] which predates the aforementioned methods, should be particularly mentioned in this context, as it has been widely applied to calculations on proteins.[37] The latest version[36] of MFCC bears much similarity to the method developed here, with some technical differences as explained herein, though our approach can be rigorously derived as an approximation to the GMBE. The traditional MBE as well as the GEBF/GMBE(1) approach will be considered further below.

2.1.2. Traditional Many-Body Expansion. For nonoverlapping fragments that do not cut across covalent bonds, it has long been recognized that the exact ground-state energy, E, can formally be expressed in terms of the energies of monomers, dimers, trimers, ... This is the traditional MBE, and the most conceptually straightforward way to write it is[11]

$$E = \sum_{I=1}^{N} E_I + \sum_{I=1}^{N}\sum_{J>I}^{N} \Delta E_{IJ} + \sum_{I=1}^{N}\sum_{J>I}^{N}\sum_{K>J}^{N} \Delta E_{IJK} + \sum_{I=1}^{N}\sum_{J>I}^{N}\sum_{K>J}^{N}\sum_{L>K}^{N} \Delta E_{IJKL} + \cdots \tag{8}$$

where the E_I are the monomer energies,

$$\Delta E_{IJ} = E_{IJ}^{(2)} - E_I^{(1)} - E_J^{(1)} \tag{9}$$

is a correction for dimers,

$$\Delta E_{IJK} = E_{IJK}^{(3)} - E_I^{(1)} - E_J^{(1)} - E_K^{(1)} - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{JK} \tag{10}$$

is a correction for timers, etc. An m-body approximation for the energy, $E^{(m)}$, simply consists of truncating eq 8 at the m-body level. As written in eq 8, this involves a lot of redundancy, which can be eliminated a priori to afford a compact expression for the m-body approximation to the total energy:[20]

$$E^{(m)} = \sum_{k=1}^{m}\sum_{\alpha=1}^{\binom{N}{k}} (-1)^{m-k}\binom{N-k-1}{m-k} E_\alpha^{(k)} \tag{11}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{12}$$

is a binomial coefficient and $E_\alpha^{(k)}$ is the energy of the αth unique k-mer of fragments. The coefficient of $E^{(k)}$ can be derived from eq 6:

$$C_k = \sum_{i=0}^{m-k} (-1)^i \binom{N-k}{i} = (-1)^{m-k}\left[\frac{(N-k-1)!}{(N-m-1)!(m-k)!}\right] \tag{13}$$

Higher-order terms that are neglected in truncating the MBE are sometimes included by performing a full supersystem calculation at a lower level of theory.[18,38−45]

2.1.3. Generalized Many-Body Expansion. In the GEBF method,[24,26,27] which we have also called GMBE(1) because it can be derived based on a one-body truncation of the GMBE,[6,22] the total energy expression is written as

$$E = \sum_{I}^{M} C_I E_I - \left(\sum_{I}^{M} C_I - 1\right)\sum_{A}\sum_{B>A} \frac{Q_A Q_B}{R_{AB}} \tag{14}$$

This energy expression can be derived from eq 6,[6] and the form of the second (charge-embedding) term will be discussed below. In the original paper introducing the GEBF method,[24] the coefficients C_I were derived for specific subsystems using a top-down procedure. The primitive subsystems correspond to the top subsets of S (i.e., the S_1, S_2, ..., S_N), and these get a coefficient $C_I = 1$. Next, the coefficient for the largest derivative subsystem is obtained by counting the number of times that this subsystem appears in any primitive subsystem. This process continues until all unique derivative subsystems have been determined. In the GMBE approach, the coefficient C_I for subsystem I is obtained instead by assembling all related coefficients in eq 6. In principle, the total energy expression can be written as in eq 6, but in practice we use eq 14, where each subsystem energy E_I is computed using a wave function for the localized fragment. The resulting energy expression, eq 14, is equivalent to the one used in the GEBF approach.[24]

2.1.4. Subsystem Energy Expressions. Next, let us turn to the specific energy expression for the subsystems and the resulting complementary energy expression, eq 7. In the gas phase, the subsystem Hamiltonian is independent of the supersystem:

$$\hat{H}_A^{\text{gas}} = \sum_{i\in A} \hat{h}_i + \sum_{\substack{i,I\\ i,I\in A}} \hat{V}_{iI} + \sum_{\substack{i<j\\ i,j\in A}} \hat{V}_{ij} + \sum_{\substack{I<J\\ I,J\in A}} \hat{V}_{IJ} \tag{15}$$

The subsystem energy is

$$E_A^{\text{gas}} = \langle\Phi|H_A|\Phi\rangle \approx E_{A,0} \tag{16}$$

where $E_{A,0}$ is the ground-state energy of isolated subsystem A. If $\bar{S}_1 \cap \bar{S}_2 \cap \cdots \cap \bar{S}_n \neq \varnothing$, then not all interaction terms are explicitly included in the subset $\{S_1, S_2, ..., S_N\}$ and the complementary energy ΔE is usually computed using an empirical method, such as a molecular mechanics (MM) force field. Furthermore, the residual interaction terms in ΔE might be ignored if the distance between interacting fragments is large.

A potentially more accurate approach is to embed the subsystem in some representation of the environment and thus write the subsystem Hamiltonian as

$$\hat{H}_A^{\text{embed}} = \sum_{i\in A} \hat{h}_i + \sum_{i\in A}\sum_{I} \hat{V}_{iI} + \sum_{i\in A}\sum_{j} \hat{V}_{ij} + \sum_{I\in A}\sum_{J} \hat{V}_{IJ} \tag{17}$$

Unfortunately, this expression does not satisfy the criterion $\hat{H}_{A\cap B} = \hat{H}_A \cap \hat{H}_B$. In order to build the total energy using the PIE, it is necessary to incorporate the self-energy ($\hat{H}_{\text{bath}}$) or two-particle interaction energy (V_{bath}) of the surrounding (bath) system into the subsystem Hamiltonian in order to satisfy this condition. Here, we use the interaction energy form, and the Hamiltonian of subsystem A is written as

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

$$\hat{H}_A^{\text{embed}} = \sum_{i\in A} \hat{h}_i + \sum_{i\in A}\sum_{I} \hat{V}_{iI} + \sum_{i\in A}\sum_{j} \hat{V}_{ij} + \sum_{I\in A}\sum_{J} \hat{V}_{IJ} + V_{\text{bath}} \quad (18)$$

In the electrostatic embedding case, the total energy of an empty subsystem (if $A = \varnothing$) is simply the self-interaction of all of the embedding charges. Otherwise, E_A^{embed} is equal to the sum of the energy of subsystem A, the interaction of A with the embedding charges, and the self-energy of the embedding charges. Since the sum of all coefficients in the PIE equals unity, the coefficient for $E_\varnothing$, the subsystem energy when $A = \varnothing$, is

$$C_\varnothing = 1 - \sum_{I}^{M} C_I \quad (19)$$

This is precisely the coefficient of the second term of eq 14. The total energy can, therefore, be rearranged as

$$E = \sum_{I}^{M} C_I(E_I - E_\varnothing) + E_\varnothing = \sum_{I}^{M} C_I(E_I^{\text{high}} - E_I^{\text{low}}) + E^{\text{low}} \quad (20)$$

Written in this way, it becomes clear that charge embedding is simply one special form of a multilayer method.[18,43,46,47]

2.2. GMBE Method. *2.2.1. Definition of Groups.* To apply the GMBE to polypeptides, we must decide upon a fragmentation scheme. It might seem natural to specify each amino acid residue as a "group" of atoms, as is done in the MFCC approach.[37] Such a choice corresponds to the cut labeled "X3" in Figure 1. However, two other groupings can be

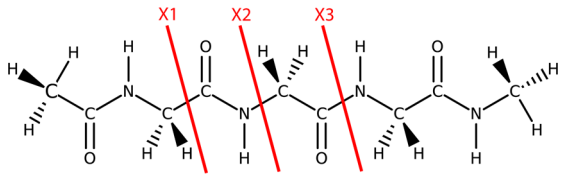


Figure 1. Definition of atom groups for polypeptides and proteins, using the peptide ACE−$(Gly)_3$−NME as an example.

envisaged,[48,49] severing either the bond between C═O and C_α (cut "X1" in Figure 1) or else the bond between C_α and N ("X2" in Figure 1). We find the second choice, severing the C−C bond between the carbonyl and C_α, to be the most natural, and we refer to this as the P1 partition scheme. This has several advantages over the other two cuts suggested in Figure 1, namely, (1) it keeps the peptide bonds intact, (2) it avoids problems arising from cutting the bond between C_α and N−H in proline, and (3) it severs a less polar bond.

A typical amino acid residue contains 7−24 atoms. The largest subsystem in the pair−pair fragmentation method will contain up to 96 atoms (fragment quartets) excluding the capping hydrogen atoms. These subsystems are too large to be calculated with high-level ab initio methods, such as MP2, and are computationally taxing even for some DFT methods with large basis sets. To reduce the cost, we could further divide residues with a long side chain into two parts (main chain part and side chain part), which we will call partition P2. As shown in Figure 2, all residues containing more than 15 atoms are divided into two parts. If one $-CH_2$ group connects to the C_α

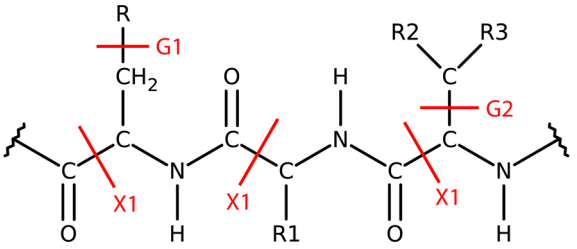


Figure 2. Definition of atom groups with partition P2, which subdivides amino acid residues into a main part and a side chain part, if the side chain R contains more than six atoms. R1 includes less than 10 atoms, and R2 + R3 includes more than 8 atoms. The cuts labeled "X1" correspond to the same label in Figure 1.

atom, then the main chain part contains the atoms in the main chain and the $-CH_2$ group (the group G1 in Figure 2). If one −CH group connects to the C_α atom, then the main chain part contains only the atoms in the main chain (G2 in Figure 2). The side chain part contains the remaining side chain atoms. As such, the largest group only contains 15 atoms, so the largest fragment quartet (and thus the largest individual electronic structure calculation) that is required in the P2 partition contains 60 atoms, excluding the capping hydrogen atoms. These largest subsystems include a fragment quartet, so assuming an average of 11 or 12 atoms per group, the largest fragments will consist of no more than about 48 atoms. Although this is somewhat larger than the typical fragment sizes used in the MOBE and SMF methods, it is comparable to the fragment size used to study proteins with the latest version of MFCC.[36] (See Table 1 of ref 5 for a comparison of typical fragment sizes for various methods.)

2.2.2. Fragmentation Method. The Hartree−Fock energy of a molecule in an atom-centered basis set can be expressed as

$$E = \sum_{\mu\nu} P_{\mu\nu}h_{\mu\nu} + \frac{1}{2}\sum_{\mu\nu\lambda\sigma} (\mu\nu\|\lambda\sigma)P_{\mu\nu}P_{\lambda\sigma} + \sum_{A}\sum_{B>A} \frac{Z_A Z_B}{R_{AB}} \quad (21)$$

with the usual notation, and antisymmetrized two-electron integrals. Although the Hamiltonian contains only one- and two-particle interactions, the total energy can be decomposed into the sum of the one-, two-, three-, and four-atom terms using eq 21, according to the four indices in $(\mu\nu\|\lambda\sigma)$. If a system is divided into N fragments, then the total energy can be decomposed using the MBE (eq 8), keeping up to four-body terms for high accuracy. The cost of such an approach will scale as $O(N^4)$, whereas with integral screening the traditional (supersystem) Hartree−Fock cost scales as $O(N^3)$ for a fixed basis set and for a system whose size grows like N. (We will see in our numerical results that there is a significant regime where fragment-based methods often require *more* computer time than the corresponding supersystem calculation, at least at the DFT level, and the fragment approach is only a "win" in terms of wall time because it parallelizes so well.)

The shell-pair data are central to the success of any modern integral program.[50] It is possible to construct all potentially important shell-quartets, which are the foundation of the two-electron integrals, by pairing the shell-pairs with one another. A similar strategy can be applied to fragment-based methods. The energy expression in eq 21 can be rewritten in a pair-interaction form

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

$$E = \sum_{IJ} \sum_{\mu\nu \in IJ} P_{\mu\nu} h_{\mu\nu} + \frac{1}{2} \sum_{IJ,KL} \sum_{\mu\nu \in IJ} \sum_{\lambda\sigma \in KL} (\mu\nu||\lambda\sigma) P_{\mu\nu} P_{\lambda\sigma} + \sum_{IJ} \sum_{\substack{A,B \in IJ \\ B>A}} \frac{Z_A Z_B}{R_{AB}} \quad (22)$$

where IJ and KL are pairs of atoms. In the formation of fragment pairs, all pairs of fragments in the supersystem are considered and categorized as either significant or negligible according to a distance criterion. Given the nearsightedness of electronic matter,[51,52] it is not surprising that most of the fragment pairs in a large molecule are negligible, that is, the number of significant fragment pairs increases linearly with the size of the molecule. In this work, a significant fragment pair is defined to mean two covalently bonded fragments or two hydrogen-bonded fragments. The criteria for a hydrogen bond are (1) the bond length between the donor atom (X = O, N, or F) is less than twice the sum of the van der Waals radii of X and H, and (2) the bond angle Y−H···X is larger than a specified cutoff ϕ_H, which we will ultimately take to be 130° based on numerical tests.

After defining the fragment pairs, the next step is to construct the important fragment quartets. As the distance between fragment pairs increases, the interaction is dominated by the Coulomb interaction and can be described by charge−charge interactions at long range. Thus, it is possible to set up a cutoff threshold to distinguish the significant pair−pair interactions. We use a simple distance cutoff, λ_c = 4.0 Å. We refer to this method, which constructs significant fragment quartets from the significant fragment pairs, as pp-GMBE.

2.2.3. Capping Method. The severed valencies introduced by fragmentation must be capped, and this is usually accomplished either using a simple hydrogen link atom or via some kind of frozen orbital cap. We use the former, and upon severing a covalent bond between atoms located at $\mathbf{r}_1$ and $\mathbf{r}_2$, a hydrogen atom is placed at a position $\mathbf{r}_{cap}$ defined by

$$\mathbf{r}_{cap} = \mathbf{r}_1 + \left(\frac{R_1 + R_H}{R_1 + R_2} \right) (\mathbf{r}_2 - \mathbf{r}_1) \quad (23)$$

where R_1, R_2, and R_H are the van der Waals radii of the indicated atoms. This is the approach taken in the SMF method of Collins et al.,[29−31] and in ref 6 we referred to this as the "SMF capping method". In contrast, MFCC calculations of proteins typically use functional group caps,[37] but results below will demonstrate that hydrogen caps are sufficient for use with the fragmentation schemes suggested herein.

An essential requirement for a valid fragmentation method is that the net number of capped hydrogen atoms must be zero. The GMBE method always satisfies this requirement since it complies with the PIE, as is easily shown. Let S_1, S_2, ..., S_N still be subsets of some set S, but let them now represent the covalent bonds between atoms. In the GMBE approach, $\overline{S}_1 \cap \overline{S}_2 \cap \cdots \cap \overline{S}_N = \varnothing$. Let $\sigma_{i\to j}$ denote the total number of the dangling bonds of atom i that point to atom j, which corresponds to subsystems that include atom i but not j. (We sever only single bonds, but the i−j bond may be severed in multiple subsystem calculations.) Let σ_{ij} be the total number of joint bonds between atoms i and j, which corresponds to the subsystems including both atoms i and j. The net number of bonds between atoms i and j is

$$\sigma_{ij} + \sigma_{i\to j} + \sigma_{j\to i} = 1 \quad (24)$$

and the net number of times that atom j appears in subsystem calculations is

$$\sigma_{i-j} + \sigma_{j\to i} = 1 \quad (25)$$

Inserting eq 25 into eq 24, one may deduce that $\sigma_{i\to j} = 0$. Because each dangling bond corresponds to a capped hydrogen atom, this result shows that the capped hydrogen atoms bonded to fragment i are canceled when all subsystems are considered. It is one way in which the GMBE, because it is based on a rigorous set-theoretical partition of the Hamiltonian, avoids double-counting (or undercounting) of interactions.

2.2.4. Embedding Charges. Dahlke and Truhlar[11] introduced two different versions of charge embedding in the context of what they call the "electrostatically embedded" MBE (EE-MBE). Embedding charges were either (a) obtained from a supersystem electronic structure calculation, e.g., as Mulliken atomic charges, or else (b) computed from subsystem calculations in the gas phase. The latter is more attractive for large systems as it avoids any supersystem calculation at all, and although at first glance it appears to be rather crude, EE-MB results are found to be surprisingly insensitive to the precise details of the embedding charges,[11,53] at least in systems such as small water clusters. As system size increases, this situation may change; even in larger water clusters, the difference between results obtained using various embedding charges is more pronounced.[13]

A very different charge embedding scheme was proposed in the context of GEBF.[24,27,54] Embedding charges on the central fragment were derived from natural population analysis[55] (NPA) in a primitive subsystem calculation, and an iterative method was introduced to calculate the natural charges. (This does complicate the formulation of analytic gradients, and only single-point energy calculations are considered here.) In the original GMBE method,[24] the embedding charge for a particular atom was defined as the average of its charge in all fragments. This extension to intersecting fragments is not physically straightforward because atoms contained within intersections might have different charges in different fragments. Although this crude averaging procedure worked surprisingly well in preliminary tests,[6] it is not without problems. Here, we propose a better way to extend charge embedding to intersecting fragments within the GMBE method.

Recalling the PIE, the total density of the supersystem can be expressed as a linear combination of the subsystem densities:[5,23,25]

$$\rho = \sum_{n=1}^{N} (-1)^{n+1} \sum_{\substack{i_1,i_2,\ldots,i_n \\ (i_1<i_2<\cdots<i_n)}}^{n} \rho_{i_1 \cap i_2 \cap \cdots \cap i_n} \quad (26)$$

This leads directly to an expression for the embedding charges

$$q_a = \sum_{n=1}^{N} (-1)^{n+1} \sum_{\substack{i_1,i_2,\ldots,i_n \\ (i_1<i_2<\cdots<i_n)}}^{n} \sum_{a \in i_1 \cap i_2 \cap \cdots \cap i_n} q_a' \quad (27)$$

where a is the atom index and q_a' is the charge on atom a in subsystem $i_1 \cap i_2 \cap \cdots \cap i_n$. In order to obtain more accurate embedding charges, the embedding charges are considered as the surrounding point charges to update the subsystem densities until the embedding charges converge. As in the GEBF approach, this updating procedure does not preserve the variational nature of the subsystem self-consistent field (SCF)

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

procedure, which will complicate the formulation of analytic energy gradients.

In the GEBF method, the embedding charge are extracted from the central fragment, which means that the primitive subsystems should be large enough to describe environmental effects. This procedure is too expensive for high-level ab initio methods. In our charge embedding method, a much cheaper GMBE method (such as a truncated two-body expansion) is used to calculate the embedding charges, and this proves to be sufficient to capture the qualitative supersystem density and provide reasonable embedding charges.

One nontrivial (but seldom discussed) problem regarding embedding charges in macromolecular systems is the charge on "link" atoms. The total charge on the link (capping) atoms may not exactly cancel in the GMBE. In larger systems, or when the system is not charge-neutral, the cumulative charge error can be significant even while the charges on individual link atoms are small. In order to cancel these errors, we redistribute the charges on link atoms to the related bond-broken atoms. Generally, this operation leads to very small change on the bond-broken atoms while preserving the correct total charge.

3. RESULTS AND DISCUSSION

3.1. Embedding Charges. In this section, we consider α-helices and β-strands of polyalanine containing 10 residues capped with an acetyl (ACE) cap at the N-terminus and an *N*-methylacetamide (NME) cap at the C-terminus. We use these decapeptides, which we call α-$(Ala)_{10}$ and β-$(Ala)_{10}$, as test systems to compare various charge embedding methods. Both of these systems are charge-neutral, so we also use chignolin (PDB code: 1UAO) as an example of a charged polypeptide exhibiting a U-turn. Geometries of α-$(Ala)_{10}$ and β-$(Ala)_{10}$ were taken from ref 16, and the geometry of 1UAO was optimized with the Amber ff99SB force field,[56] using the AMBER program.[57] These structures are shown in Figure 3.

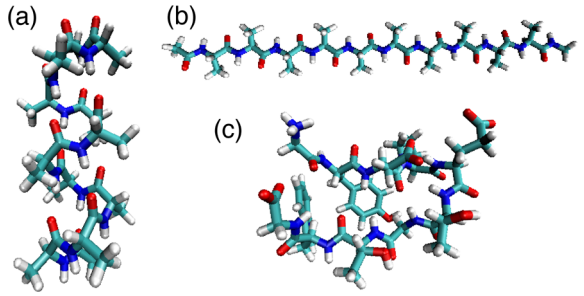


Figure 3. Macromolecular structures used to benchmark the embedding charges: (a) α-$(Ala)_{10}$, (b) β-$(Ala)_{10}$, and (c) the peptide 1UAO.

We first examine the effect of different point-charge embeddings on the total energy of these three polypeptides, using a supersystem calculation to obtain the embedding charges. As indicated above, this is not ideal for application to large molecules and will be replaced below by the subsystem-based embedding method suggested in eq 27, but for now, we use supersystem-derived embedding charges as a quick method to compare Mulliken charges, NPA charges,[55] and ChElPG charges.[58] As shown in Table 1, the differences among these three embedding schemes, and between the embedded methods and a pp-GMBE approach that uses no embedding at all, are less than 1 kcal/mol for the two neutral polyalanine systems. In the charged system 1UAO, however, the pp-GMBE without embedding charges leads to a large error, −5.2 kcal/mol, whereas the deviations remain <1 kcal/mol for the three embedded pp-GBME approaches. Since there is not much difference among these three charge schemes, we will settle on NPA charges due to their stability with respect to basis-set expansion[55] (unlike Mulliken charges) and because ChElPG charges can depend strongly on molecular conformation, especially in large molecules.[58]

Table 1. Energy Deviations (kcal/mol) between Supersystem and pp-GMBE Results at the HF/6-311G* Level Using Various Embedding Charges

charge scheme	α-$(Ala)_{10}$	β-$(Ala)_{10}$	1UAO
none	−0.16	0.44	−5.19
Mulliken	0.53	0.08	−0.42
ChElPG	−0.36	0.32	−1.23
NPA	0.96	0.24	−0.25

We next compare NPA embedding charges derived from two different GMBE-based methods to those obtained from a calculation on the supersystem in an effort to determine whether the supersystem calculation can be eliminated. Results are shown in Table 2, where the method labeled 2b-GMBE means that a cheaper GMBE method is used to approximate the supersystem density and thus compute the embedding charges using eq 27. Specifically, the 2b-GMBE method uses only the significant pairs (as defined above) and their intersections, but not the significant quartets, so that the largest subsystem consists of a fragment pair. The method labeled pp-GMBE is the one introduced in Section 2.2.2 that does use the fragment quartets. Environment charges for use in the next iteration are derived from those in the previous iteration, starting from a gas-phase calculation (no embedding), and the procedure is iterated to convergence, which takes no more than two iterations in these examples. In the two charge-neutral systems, embedding charges obtained from gas-phase calculations are accurate enough already without iteration, whereas in the charged 1UAO system, the method takes only an iteration or two to converge.

The performance of embedding charges derived from the 2b-GMBE method and the pp-GMBE method is very similar for α-$(Ala)_{10}$, β-$(Ala)_{10}$, and 1UAO except that it takes more iterations to converge the total energy using the 2b-GMBE charges. However, the computational cost for embedding charges derived from the pp-GMBE method is much greater than the cost to obtain 2b-GMBE charges, since the latter involves smaller and fewer subsystems. Hence, it is more efficient to use the 2b-GMBE approach to derive embedding charges. In case of larger systems, the convergence with embedding charges derived from the 2b-GMBE method is slightly slower, and we find that three iterations are required in some cases, such as 3T97-B, which contains 58 amino acids and has a net charge of +5.

3.2. Single-Point Energies of Proteins. We next consider the convergence of the total energy for the pp-GMBE using different distance cutoffs. Figure 4 shows the deviation of the total energy with respect to the full supersystem calculation, for distance cutoffs λ_c ranging from 2 to 5 Å. In addition, we compare two different criteria for the hydrogen-bond cutoff angle, $\phi_H = 130°$ versus $\phi_H = 150°$. In previous studies using the MFCC approach,[36,37] deviations with respect to supersystem results were found to be within 0.01 hartree when the

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

Table 2. Deviations between GMBE Charges Derived from Sub- versus Supersystem Calculations at the HF/6-311G* Level[a]

	α-$(Ala)_{10}$		β-$(Ala)_{10}$		1UAO	
iteration	2b-GMBE	pp-GMBE	2b-GMBE	pp-GMBE	2b-GMBE	pp-GMBE
1	0.008 (0.96)	0.001 (0.97)	0.002 (0.34)	0.000 (0.31)	0.013 (−0.77)	0.001 (−0.24)
2	0.002 (0.95)	0.000 (0.95)	0.002 (0.24)		0.005 (−0.16)	0.001 (−0.22)
3	0.002 (0.95)	0.000 (0.95)	0.002 (0.24)		0.005 (−0.17)	0.001 (−0.22)
4	0.002 (0.95)		0.002 (0.24)		0.005 (−0.16)	

[a]As a function of iterative-updating of the embedding charges, starting from gas-phase values. Energy deviations (in kcal/mol) are given in parentheses.

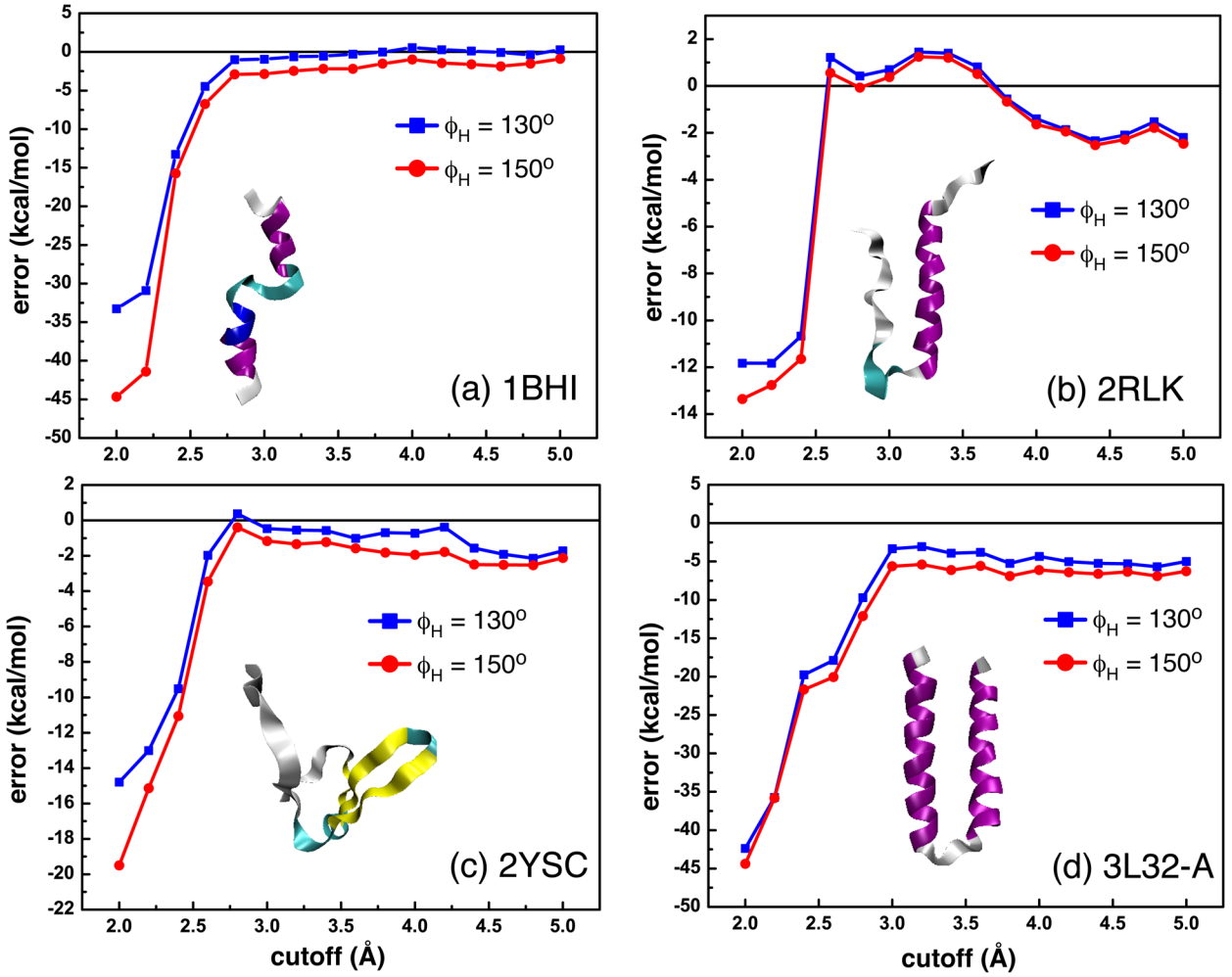


Figure 4. Errors in the pp-GMBE, with respect to the supersystem results, as a function of the distance cutoff λ_c for defining significant pair−pair interactions that are used to form fragment quartets and for two different values of the hydrogen-bond angle cutoff, ϕ_H. Results are shown for four different polypeptides, with PDB codes as indicated: (a) 1BHI, a 38-residue polypeptide containing both an antiparallel β-sheet and an α-helix, (b) 2RLK, a 37-residue β-hairpin, (c) 2YSC, a 39-residue polypeptide, and (d) 3L32-A, a 45-residue polypeptide consisting of two α-helices.

distance threshold $\lambda_c \geq 4.0$ Å. As shown in Figure 4, the relative energies with the pp-GMBE method converge when $\lambda_c \approx 3.5$−4.0 Å, which is very similar to the conclusions drawn in ref 36. We will use a cutoff $\lambda_c = 4$ Å, which for proteins includes most noncovalent interactions including C−H···π and π−π interactions. The two hydrogen-bond cutoffs in Figure 4 yield nearly identical results, although $\phi_H = 130°$ is slightly closer to the full supersystem results and will be used here.

We have performed pp-GMBE calculations on 18 different proteins at the HF/6-31G* level; see Figure 5 for the structures and PDB codes. Table 3 shows how pp-GMBE single-point energies for these proteins compare to full supersystem results at the same level of theory. We test both partitions P1 and P2 along with several sets of embedding charges. Protein structures were downloaded from the Protein Data Bank (PDB), and the data set includes various protein secondary structural motifs. The largest protein (2LH0) contains 70 amino acid residues and 1142 atoms.

Given the P1 partition, results for all three charge schemes agree very well with the full supersystem HF calculations, with mean absolute errors (MAEs) of just 1.6, 1.8, and 1.1 kcal/mol for Mulliken, ChElPG, and NPA charge embeddings, respectively. The NPA charge embedding scheme produces the smallest maximum error, at 2.5 kcal/mol. The P2 partition

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

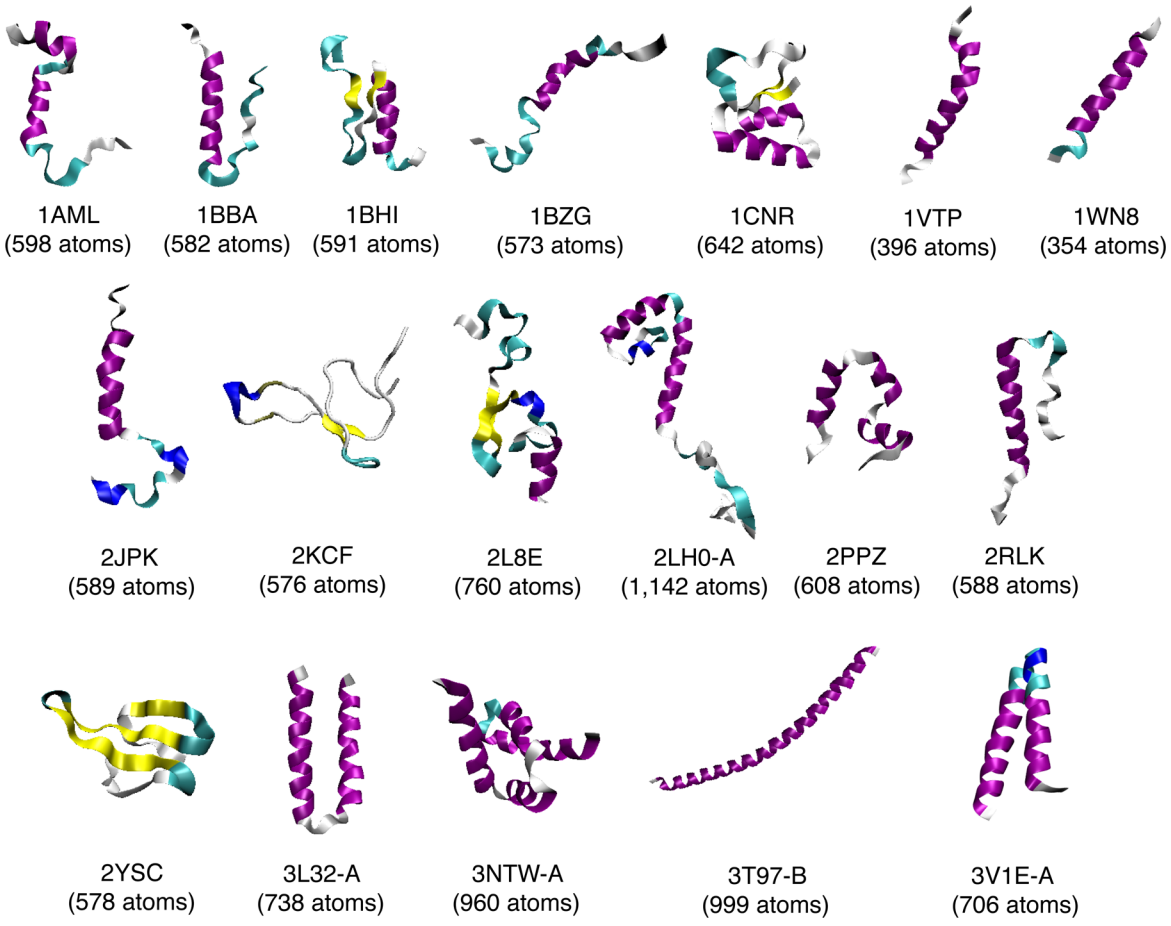


Figure 5. Structures of the 18 small proteins used in this study, with PDB codes as indicated.

Table 3. Deviations between the Full System Energy and Its pp-GMBE Approximation at the HF/6-31G* Level

			deviation (kcal/mol)					
			P1			P2		
system		full system (hartree)	Mulliken	ChElPG	NPA	Mulliken	ChElPG	NPA
1AML		−15142.596282	−0.31	−0.30	2.33	−2.01	−1.84	−0.18
1BBA		−15103.699913	2.13	−0.69	1.21	0.86	5.73	1.83
1BHI		−15990.191162	3.05	3.06	1.74	1.05	−0.75	0.52
1BZG		−13681.215668	1.84	0.75	1.80	1.36	1.46	1.95
1CNR		−18004.668080	−3.02	−3.56	−2.52	−3.08	−3.56	−0.92
1VTP		−10015.595416	0.46	0.60	0.65	−1.97	0.60	−0.52
1WN8		−8878.693155	−1.14	0.24	−0.17	−0.12	0.24	0.56
2JPK		−13855.875913	−0.91	−2.50	0.79	−0.50	−4.91	1.14
2KCF		−14599.940061	−0.53	−0.85	−1.19	1.67	−0.85	0.04
2L8E		−20270.601331	−1.16	−1.67	0.39	1.23	−2.47	3.13
2LH0	(Chain A)	−28500.910569	2.12	−0.42	1.02	1.57	−6.00	4.27
2PPZ		−14958.306351	−0.43	−1.58	−1.86	−0.37	−1.58	0.51
2RLK		−14590.340301	−0.23	−1.79	−0.04	−0.67	−10.01	−1.42
2YSC		−14634.918275	−0.37	−5.00	−0.86	−2.58	−5.00	−0.82
3L32	(Chain A)	−18055.090444	−4.27	−1.71	−1.96	−4.63	−1.71	−4.36
3NTW	(Chain A)	−23683.640326	−5.26	−1.24	−1.44	−4.57	−21.88	−0.27
3T97	(Chain B)	−24598.999033	−2.14	−5.03	−0.27	−2.01	−3.58	−0.28
3V1E	(Chain A)	−18015.416098	−0.34	−1.30	0.10	−2.23	−2.72	1.06
max			5.26	5.03	2.52	4.63	21.88	4.36
MAE			1.65	1.79	1.13	1.81	4.16	1.32
RMSE			2.18	2.31	1.36	2.20	6.45	1.85

produces results quite close to those obtained with P1, except in the case of ChElPG embedding charges, where the MAE is 4.2 kcal/mol and the maximum deviation is 21.9 kcal/mol. Much of this error arises from two protein structures, 2RLK

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

and 3NTW. It is unclear from the structures in Figure 5 why these particular examples should be problematic, but if they are removed from the data set, then the MAE for the P2/ChElPG approach is reduced to 2.4 kcal/mol and the RMSE is reduced to 3.1 kcal/mol. If only proteins with fewer than 700 atoms are considered, then the maximum P2/NPA error is 2.0 kcal/mol, for 1BZG.

Since most of the P2 error arises from the choice of embedding charges, it can be reasoned that the ChElPG charges suffer from dependence on conformation. Another important difference between P1 and P2 is subsystem size: the latter approach contains no more than 60 atoms per subsystem (excluding the capping hydrogen atoms), whereas the P1 partition may contain up to 96 atoms, although with an average of 60 atoms per subsystem since the average size of an amino acid is about 15 atoms. In view of the problems with ChElPG charges in the P2 case, it must be concluded that P2/NPA is an efficient yet stable approach for when high-level quantum chemistry calculations (albeit with moderate basis sets) are to be employed.

In proteins, not only covalent but also noncovalent interactions are important, with the latter playing a key role in determining secondary, tertiary, and quaternary structure. Correlated wave function approaches, scaling as $O(N_{basis}^5)$–$O(N_{basis}^7)$, have traditionally been necessary for the accurate description of noncovalent interactions, although a variety of DFT approaches to this problem have been put forth in recent years. These include dispersion corrected DFT (DFT+D),[59] along with nonlocal functionals such as ωB97X-V[8] and B97M-V.[60] Here, we will take the dispersion-corrected ωB97X-D functional[61] as an example to consider the accuracy of the pp-GMBE method when combined with DFT. Table 4 lists errors in pp-GMBE energies using the P2/NPA scheme. Although the errors are, in general, a bit larger at the ωB97X-D/6-31G* level than those observed at the HF/6-31G* level, the results are still in good agreement with full supersystem calculations.

Table 4. Errors in the pp-GMBE P2/NPA Method with Respect to Full Supersystem Calculations at the ωB97X-D/6-31G* Level

molecule		total energy (hartree)	deviation (kcal/mol)
1AML		−15228.320066	−0.21
1BBA		−15187.461036	−2.42
1BHI		−16076.053869	−2.19
1BZG		−13760.714345	−0.03
1CNR		−18098.695358	−6.51
1VTP		−10072.705786	−2.96
1WN8		−8927.072195	−1.30
2JPK		−13936.903344	−0.75
2KCF		−14682.518736	−2.34
2L8E		−20379.435805	−1.67
2LH0	(Chain A)	−28661.963961	1.79
2PPZ		−15042.290031	−0.60
2RLK		−14674.323078	−4.78
2YSC		−14719.120990	−1.53
3L32	(Chain A)	−18158.251179	−5.62
3NTW	(Chain A)	−23816.533876	1.46
3T97	(Chain B)	−24736.621922	−7.62
3V1E	(Chain A)	−18116.263876	−4.00
MAE			2.66
RMSE			3.41

3.3. Relative Energies of Proteins. More important than absolute energies for proteins are relative conformational energies. To test the latter, we employ a data set consisting of 20 conformations for each of two proteins: 1WN8 and 2KCF. Initial structures were obtained from the ensembles provided in the PDB and then optimized (in the absence of solvent) with the Amber ff99SB force field;[56] both the initial and the relaxed structures are shown in Figure 6. Whereas the initial structures (optimized subject to solution-phase NMR distance constraints) are quite similar save for the relatively mobile termini, gas-phase optimization leads to greater structural variety, although the primary features of secondary structure (an α-helix in 1WN8 and an antiparallel β-sheet in 2KCF) are preserved.

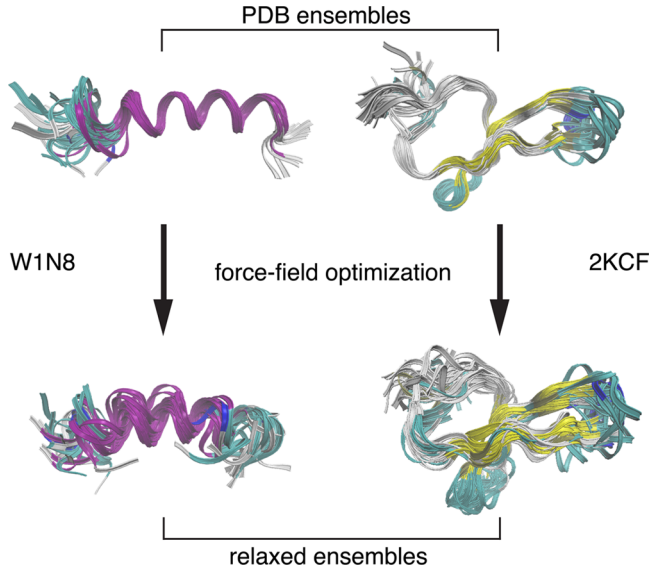


Figure 6. Top: Set of 20 solution-phase structures of 1WN8 and 2KCF obtained from the PDB. Bottom: The same set of structures, following gas-phase relaxation using the Amber ff99SB force field.

This structural variation affords an ensemble whose relative energies vary by nearly 100 kcal/mol, as shown in Figure 7, where we plot the relative energies for each of the 20 structures at both the HF/6-31G* and M06-2X/6-31G* levels of theory. While neither of these model chemistries is necessarily an appropriate level of theory for this problem—the former lacks any treatment of dispersion interactions, for example, and in the 6-31G* basis set, both methods are likely to be heavily beset by basis-set superposition error—the important observation is that the pp-GMBE approach faithfully reproduces the relative energies at both levels of theory, to the point that the pp-GMBE plots in Figure 7 are nearly indistinguishable from the full supersystem results. This is an important proof-of-principle result for when we attempt pp-GMBE calculations at higher levels of theory and in larger systems. Error statistics, averaged over 20 conformations (Table 5), show that pp-GMBE exhibits a maximum absolute deviation of only 1.8 kcal/mol for 1WN8 and 2.6 kcal/mol for 2KCF at the HF/6-31G* level. The corresponding values at the M06-2X/6-31G* level are 2.0 kcal/mol for 1WN8 and 4.5 kcal/mol for 2KCF.

One possible point of concern in these calculations is whether the ∼100 kcal/mol range of relative energies might

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

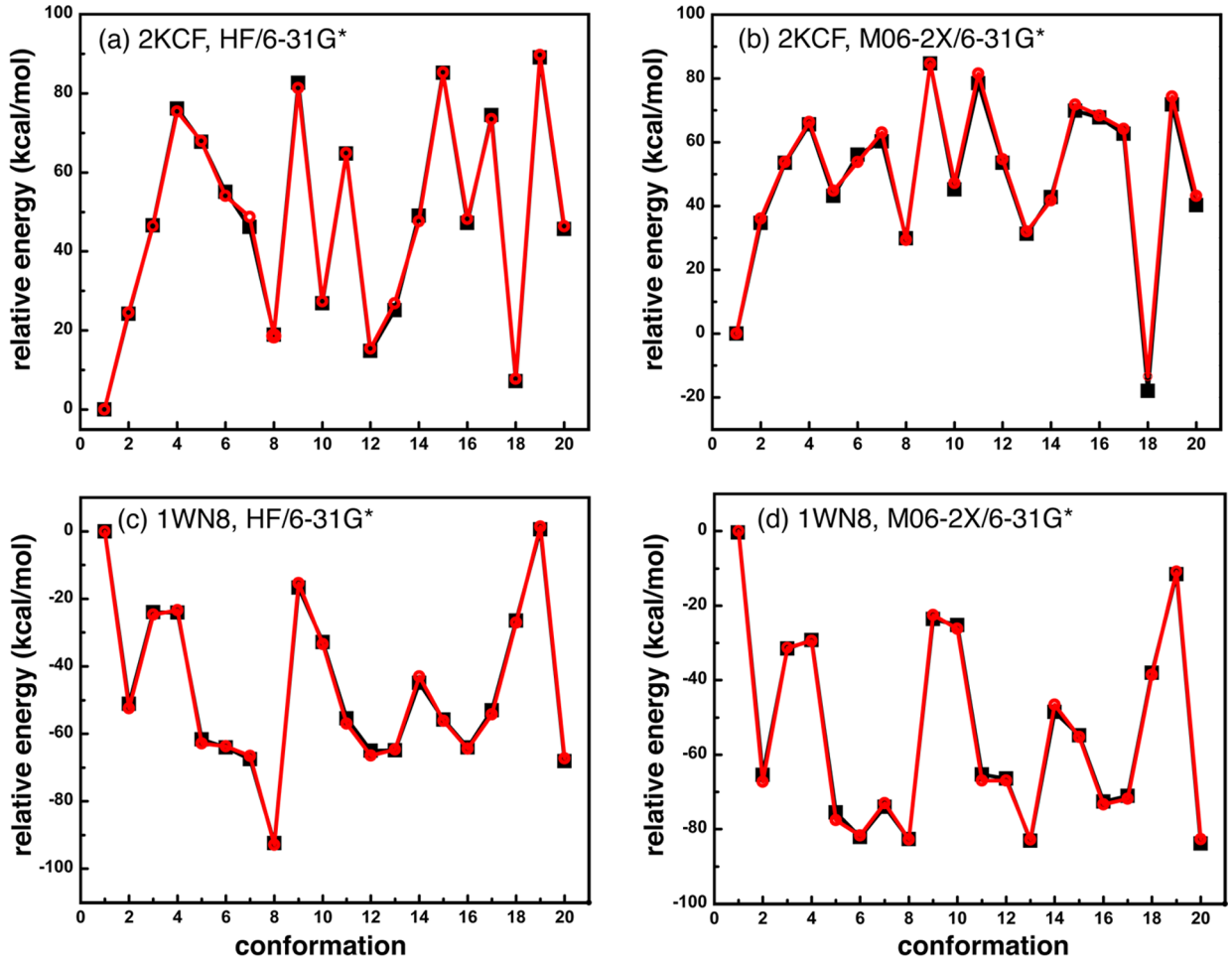


Figure 7. Relative energies for 20 different conformations of two proteins, computed at each of two levels of theory. Results in black are for the pp-GMBE method, whereas results in red represent a calculation on the entire protein at the same level of theory, but the two calculations are essentially indistinguishable on this scale.

Table 5. Statistical Deviations of the Absolute Energies (in kcal/mol) for 20 Conformation of 1WN8 and 2KCF with Respect to Full System Calculation

	1WN8		2KCF	
	HF[a]	M06-2X[a]	HF[a]	M06-2X[a]
max	2.00	4.41	4.18	9.32
MAE	0.96	1.53	1.72	6.01
RMSE	1.13	2.56	1.98	6.20

[a]Basis set is 6-31G*.

somehow be an artifact of, say, minimum-energy bond lengths and angles that are slightly different in the force field than in the ab initio calculations. (Thus, certain structures might be pushed slightly up the repulsive wall of the HF or M06-2X potential, exacerbating the energy differences between them.) Even if it is true, this is not an indictment of the pp-GMBE approximation per se, but to examine this issue, we took seven structures of 1UAO (Figure 3c) from the PDB and then relaxed them in the gas phase at two different levels of theory: the ff99SB force field, as in the calculations above, and using the HF-3c method,[62] as implemented in ORCA software.[63] HF-3c is a semiempirical method in which a minimal-basis Hartree−Fock calculation is combined with three empirical corrections: one for basis-set superposition error, one for basis-set incompleteness, and one for dispersion. This approach has been shown to perform well for large-molecule geometry optimizations and is affordable enough to be used for such. Figure 8 plots the resulting relative energies. For this particular polypeptide, the relative energies span a range of about 50 kcal/mol, but, importantly, they do so at both the force field level and at the HF-3c level. This result, in conjunction with the results for 1WN8 and 2KCF above, suggests that relative energies of 50−100 kcal/mol are simply what one can expect when structures from a solution-phase ensemble are pulled out of solution and subjected to gas-phase calculations. This is an interesting observation in the context of using such calculations to benchmark force fields, suggesting that comparisons between force field and ab initio energetics are meaningless in the absence of solvent.

3.4. Parallelization. The serial efficiency of traditional supersystem methods is highly optimized, but the potential for parallelization is fundamentally limited by their iterative nature; each step requires the results of the previous iteration before it can proceed. The work done in each iteration can be parallelized, and this can be made to scale fairly well across a single node, but in attempts to scale beyond a single node, the efficiency often suffers greatly due to latencies in communication. Multithreading offers only a partial solution, as two-electron integral calculations can become easily bottlenecked by passage of information through memory in such a calculation,

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

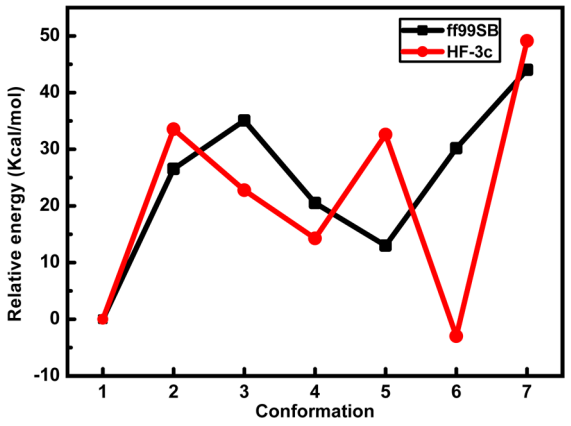


Figure 8. Relative energies for seven isomers of 1UAO, computed using either the ff99SB force field or the HF-3c semiempirical method. In each case, the same set of starting structures was obtained from the PDB, but these structures were then relaxed in the gas phase at either the ff99SB or HF-3c level of theory so that the geometries are computed at the same level of theory that is subsequently used to compute the relative energies.

e.g., to evaluate the vertical recurrence relations required for higher angular momentum functions.[50]

Fragment-based methods are designed to overcome these limitations. Iterations are limited to fragments only, and the subsystem energy calculations are entirely independent of one another and thus trivial to distribute across any arrangement of processors. This internode parallelism could be further augmented by shared memory intranode parallelism, such as open multi-processing (OpenMP) multithreading. We have implemented the pp-GMBE method with (a) parallelism using the message-passing interface (MPI) in order to parallelize subsystem energy computations across cores within a distributed memory model and (b) a MPI + OpenMP parallel version that parallelizes subsystem energy computations across nodes and parallelizes each subsystem calculation with OpenMP. (This is highly efficient for more expensive levels of theory.) Given N cores for N subsystems, the total wall time is essentially reduced to the time required for the longest subsystem calculation.

Figure 9 shows both the total computer time and the wall time required for single-point energy calculations on the set of 18 proteins given in Table 3. Even for the largest of these proteins (2LH0, at 1142 atoms), the total computer time for the pp-GMBE calculations is significantly greater than that required for the supersystem calculations. However, using just 120 processors, the wall time for the pp-GMBE calculations can be significantly less than that required for the full system calculations. On 120 processors, the pp-GMBE calculations take 1−2 h for most of our protein data set (at the HF/6-31G* level), and only about 7 h for 2LH0.

In addition, the pp-GMBE method significantly reduces the number of independent electronic structure calculations that are required, relative to a traditional four-body expansion or a two-body GMBE. For example, there are 445 groups in the P2 partition of 2LH0-A, and for a traditional four-body expansion, this means ${}_{445}C_4 \approx 1.6 \times 10^9$ individual tetramer calculations, plus ${}_{445}C_3 \approx 1.5 \times 10^6$ trimers and some comparatively trivial number of dimers and monomers. These are intractable numbers. On the other hand, these 445 groups generate 152 monomers if we set the hydrogen-bond angle threshold ϕ_H to 130°, so for the two-body GMBE, the total number of the largest subsystems (four groups) is ${}_{152}C_2 = 11\,552$. (Smaller subsystems, namely, intersections, are also required, but clearly this is a significant reduction relative to the four-body expansion.) For the pp-GMBE approximation to the two-body GMBE, the number of largest subsystems is a mere 1266.

4. SUMMARY

We have introduced a pp-GMBE method for fragment-based quantum chemistry, which is an approximation to the two-body GMBE and an alternative to the traditional four-body expansion, with significant fragment pairs and quartets treated quantum mechanically in the presence of embedding charges representing the rest of the system. Relative to a four-body expansion, this approach significantly reduces the number of subsystem calculations and thus the computational time. An efficient and accurate fragmentation scheme is introduced for proteins, reducing the size of the largest subsystem to about 60 atoms.

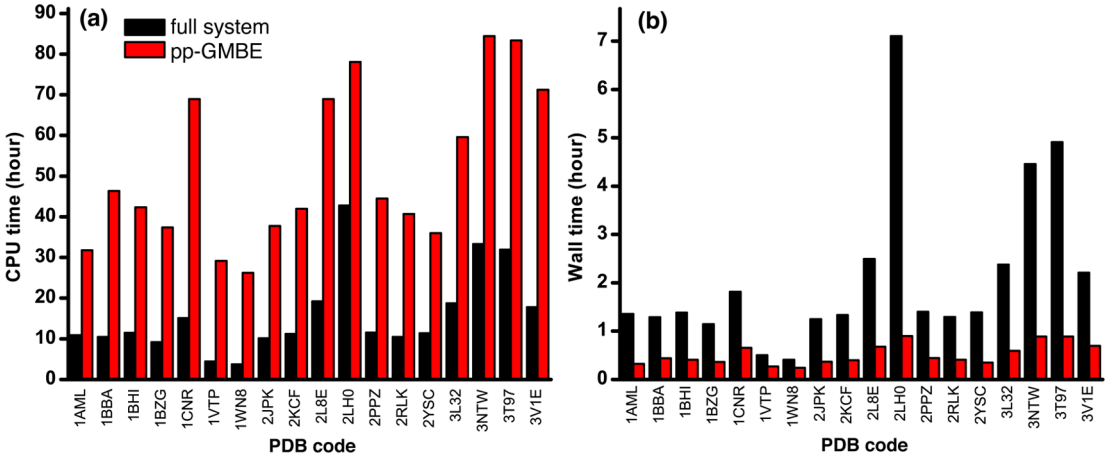


Figure 9. Total (a) computer time and (b) wall time (i.e., time-to-solution) required for single-point energy calculations for 18 different proteins at the HF/6-31G* level. The pp-GMBE calculations were performed on 10 nodes (with 12 processors per node) with a MPI + OpenMP parallelized version of our FRAGMENT code,[6,21] as a driver for Q-CHEM.[64] The supersystem calculations were performed on a single node, multithreaded across all 12 processors using Q-CHEM.

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

A comparison of the pp-GMBE method with full system calculations for a set of proteins has been carried out at the Hartree−Fock and DFT levels of theory. Mean absolute errors approaching 1 kcal/mol are achievable with the pp-GMBE and our P1/NPA partitioning and charge-embedding scheme. Other charge embeddings work well in some cases, but ChElPG charges turn out to be surprisingly unstable, with maximum errors in excess of 21 kcal/mol in some cases. For relative conformational energies of a given protein, the pp-GMBE P1/NPA approach provides results that are nearly indistinguishable from calculations at the same level of theory applied to the entire protein.

In contrast to the four-body expansion, it is necessary to set up two parameters for pp-GMBE calculations. One is the monomer (fragment dimer) criterion, and the other is the distance cutoff. In principle, the result is more accurate when more monomers are involved and the distance cutoff is larger, but the number of subsystems increases rapidly as does the computational time. It is necessary to find a balance between accuracy and efficiency. With more accurate quantum chemistry methods, the criterion of the significant fragment pairs will change. In this work, only hydrogen bonds are included, and for more general applications, it will be necessary to find a general way to include all significant pairs. The MPI + OpenMP version of the pp-GMBE within the FRAGME∩T code[6] allows all individual subsystem calculations to be multithreaded using OpenMP. This provides a powerful tool to carry out pp-GMBE calculations using correlated wave function levels of theory, with appropriate basis sets, or using the latest nonlocal density functionals, which are more expensive to evaluate than the ones used here for testing purposes. We hope to report on such calculations in the future.

AUTHOR INFORMATION

Corresponding Author

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

Funding

This work was supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences under Award No. DE-SC0008550. Calculations were performed at the Ohio Supercomputer Center under project PAA-0003.[65]

Notes

The authors declare the following competing financial interest(s): J.M.H. serves on the Board of Directors of Q-Chem, Inc.

ACKNOWLEDGMENTS

J.M.H. is a Camille Dreyfus Teacher−Scholar.

REFERENCES

(1) Yang, W. *J. Mol. Struct.: THEOCHEM* **1992**, *255*, 461.
(2) Li, S.; Shen, J.; Li, W.; Jiang, Y. *J. Chem. Phys.* **2006**, *125*, 074109.
(3) Wu, F.; Liu, W.; Zhang, Y.; Li, Z. *J. Chem. Theory Comput.* **2011**, *7*, 3643.
(4) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. *Chem. Rev.* **2012**, *112*, 632.
(5) Collins, M. A.; Bettens, R. P. *Chem. Rev.* **2015**, *115*, 5607.
(6) Richard, R. M.; Herbert, J. M. *J. Chem. Phys.* **2012**, *137*, 064113.
(7) Mardirossian, N.; Head-Gordon, M. *J. Chem. Theory Comput.* **2013**, *9*, 4453.
(8) Mardirossian, N.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904.
(9) Richard, R. M.; Lao, K. U.; Herbert, J. M. *J. Chem. Phys.* **2013**, *139*, 224102.
(10) Ouyang, J. F.; Cvitkovic, M. W.; Bettens, R. P. A. *J. Chem. Theory Comput.* **2014**, *10*, 3699.
(11) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 46.
(12) Richard, R. M.; Lao, K. U.; Herbert, J. M. *J. Phys. Chem. Lett.* **2013**, *4*, 2674.
(13) Richard, R. M.; Lao, K. U.; Herbert, J. M. *Acc. Chem. Res.* **2014**, *47*, 2828.
(14) Ouyang, J. F.; Bettens, R. P. A. *J. Chem. Theory Comput.* **2015**, *11*, 5132.
(15) Li, S.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215.
(16) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *120*, 6832.
(17) Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2004**, *389*, 129.
(18) Bates, D. M.; Smith, J. R.; Janowski, T.; Tschumper, G. S. *J. Chem. Phys.* **2011**, *135*, 044123.
(19) Qi, H. W.; Leverentz, H. R.; Truhlar, D. G. *J. Phys. Chem. A* **2013**, *117*, 4486.
(20) Richard, R. M.; Lao, K. U.; Herbert, J. M. *J. Chem. Phys.* **2014**, *141*, 014108.
(21) Richard, R. M.; Herbert, J. M. *J. Chem. Theory Comput.* **2013**, *9*, 1408.
(22) Jacobson, L. D.; Richard, R. M.; Lao, K. U.; Herbert, J. M. *Annu. Rep. Comput. Chem.* **2013**, *9*, 25.
(23) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. *J. Chem. Phys.* **2006**, *125*, 104109.
(24) Li, W.; Li, S.; Jiang, Y. *J. Phys. Chem. A* **2007**, *111*, 2193.
(25) Reid, D. M.; Collins, M. A. *J. Chem. Phys.* **2013**, *139*, 184117.
(26) Hua, W.; Fang, T.; Li, W.; Yu, J.-G.; Li, S. *J. Phys. Chem. A* **2008**, *112*, 10864.
(27) Hua, S.; Hua, W.; Li, S. *J. Phys. Chem. A* **2010**, *114*, 8126.
(28) Mayhall, N. J.; Raghavachari, K. *J. Chem. Theory Comput.* **2012**, *8*, 2669.
(29) Collins, M. A.; Deev, V. A. *J. Chem. Phys.* **2006**, *125*, 104104.
(30) Addicoat, M. A.; Collins, M. A. *J. Chem. Phys.* **2009**, *131*, 104103.
(31) Collins, M. A. *J. Chem. Phys.* **2014**, *141*, 094108.
(32) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599.
(33) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 031103.
(34) Chen, X.; Zhang, Y.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 184105.
(35) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 044903.
(36) Wang, X.; Liu, J.; Zhang, J. Z. H.; He, X. *J. Phys. Chem. A* **2013**, *117*, 7149.
(37) He, X.; Zhu, T.; Wang, X.; Liu, J.; Zhang, J. Z. H. *Acc. Chem. Res.* **2014**, *47*, 2748.
(38) Tschumper, G. S. *Chem. Phys. Lett.* **2006**, *427*, 185.
(39) Elsohly, A. M.; Shaw, C. L.; Guice, M. E.; Smith, B. D.; Tschumper, G. S. *Mol. Phys.* **2007**, *105*, 2777.
(40) Bates, D. M.; Smith, J. R.; Tschumper, G. S. *J. Chem. Theory Comput.* **2011**, *7*, 2753.
(41) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 1342.
(42) Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 33.
(43) Beran, G. J. O. *J. Chem. Phys.* **2009**, *130*, 164115.
(44) Beran, G. J. O.; Nanda, K. *J. Phys. Chem. Lett.* **2010**, *1*, 3480.
(45) Sebetci, A.; Beran, G. J. O. *J. Chem. Theory Comput.* **2010**, *6*, 155.
(46) Howard, J. C.; Tschumper, G. S. *J. Chem. Phys.* **2013**, *139*, 184113.
(47) Mayhall, N. J.; Raghavachari, K. *J. Chem. Theory Comput.* **2011**, *7*, 1336.
(48) Zhang, D. W.; Chen, X. H.; Zhang, J. Z. H. *J. Comput. Chem.* **2003**, *24*, 1846.
(49) Isegawa, M.; Wang, B.; Truhlar, D. G. *J. Chem. Theory Comput.* **2013**, *9*, 1381.
(50) Gill, P. M. W. *Adv. Quantum Chem.* **1994**, *25*, 141.

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584

(51) Kohn, W. *Phys. Rev. Lett.* **1996**, *76*, 3168.
(52) Prodan, E.; Kohn, W. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 11635.
(53) Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1573.
(54) Li, W.; Hua, W.; Fang, T.; Li, S. In *Computational Methods for Large Systems: Electronic Structure Approaches for Biotechnology and Nanotechnology*; Reimers, J. R., Ed.; Wiley: Hoboken, NJ, 2011; pp 227−258.
(55) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735−746.
(56) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 712.
(57) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. *WIREs Comput. Mol. Sci.* **2013**, *3*, 198.
(58) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.
(59) Grimme, S. *WIREs Comput. Mol. Sci.* **2011**, *1*, 211.
(60) Mardirossian, N.; Head-Gordon, M. *J. Chem. Phys.* **2015**, *142*, 074111.
(61) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.
(62) Sure, R.; Grimme, S. *J. Comput. Chem.* **2013**, *34*, 1672.
(63) Neese, F. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73.
(64) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; Ghosh, D.; Goldey, M.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Khaliullin, R. Z.; Kús, T.; Landau, A.; Liu, J.; Proynov, E. I.; Rhee, Y. M.; Richard, R. M.; Rohrdanz, M. A.; Steele, R. P.; Sundstrom, E. J.; Woodcock, H. L., III; Zimmerman, P. M.; Zuev, D.; Albrecht, B.; Alguire, E.; Austin, B.; Beran, G. J. O.; Bernard, Y. A.; Berquist, E.; Brandhorst, K.; Bravaya, K. B.; Brown, S. T.; Casanova, D.; Chang, C.-M.; Chen, Y.; Chien, S. H.; Closser, K. D.; Crittenden, D. L.; Diedenhofen, M.; DiStasio, R. A., Jr.; Do, H.; Dutoi, A. D.; Edgar, R. G.; Fatehi, S.; Fusti-Molnar, L.; Ghysels, A.; Golubeva-Zadorozhnaya, A.; Gomes, J.; Hanson-Heine, M. W. D.; Harbach, P. H. P.; Hauser, A. W.; Hohenstein, E. G.; Holden, Z. C.; Jagau, T.-C.; Ji, H.; Kaduk, B.; Khistyaev, K.; Kim, J.; Kim, J.; King, R. A.; Klunzinger, P.; Kosenkov, D.; Kowalczyk, T.; Krauter, C. M.; Lao, K. U.; Laurent, A.; Lawler, K. V.; Levchenko, S. V.; Lin, C. Y.; Liu, F.; Livshits, E.; Lochan, R. C.; Luenser, A.; Manohar, P.; Manzer, S. F.; Mao, S.-P.; Mardirossian, N.; Marenich, A. V.; Maurer, S. A.; Mayhall, N. J.; Oana, C. M.; Olivares-Amaya, R.; O'Neill, D. P.; Parkhill, J. A.; Perrine, T. M.; Peverati, R.; Pieniazek, P. A.; Prociuk, A.; Rehn, D. R.; Rosta, E.; Russ, N. J.; Sergueev, N.; Sharada, S. M.; Sharma, S.; Small, D. W.; Sodt, A.; Stein, T.; Stück, D.; Su, Y.-C.; Thom, A. J. W.; Tsuchimochi, T.; Vogt, L.; Vydrov, O.; Wang, T.; Watson, M. A.; Wenzel, J.; White, A.; Williams, C. F.; Vanovschi, V.; Yeganeh, S.; Yost, S. R.; You, Z.-Q.; Zhang, I. Y.; Zhang, X.; Zhao, Y.; Brooks, B. R.; Chan, G. K. L.; Chipman, D. M.; Cramer, C. J.; Goddard, W. A., III; Gordon, M. S.; Hehre, W. J.; Klamt, A.; Schaefer, H. F., III; Schmidt, M. W.; Sherrill, C. D.; Truhlar, D. G.; Warshel, A.; Xu, X.; Aspuru-Guzik, A.; Baer, R.; Bell, A. T.; Besley, N. A.; Chai, J.-D.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Gwaltney, S. R.; Hsu, C.-P.; Jung, Y.; Kong, J.; Lambrecht, D. S.; Liang, W.; Ochsenfeld, C.; Rassolov, V. A.; Slipchenko, L. V.; Subotnik, J. E.; Van Voorhis, T.; Herbert, J. M.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Mol. Phys.* **2015**, *113*, 184.
(65) *Ohio Supercomputer Center*. http://osc.edu/ark:/19495/f5s1ph73 (accessed Dec. 23,2015).

DOI: 10.1021/acs.jctc.5b00955
J. Chem. Theory Comput. 2016, 12, 572−584