ACCOUNTS of chemical research

Article

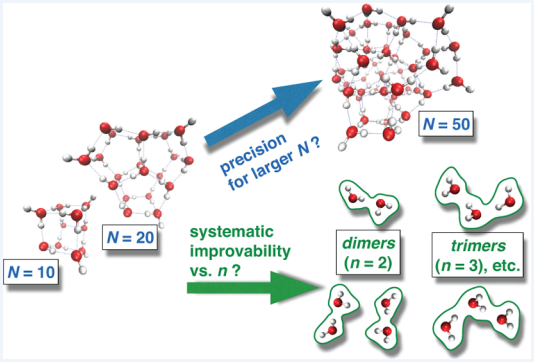
pubs.acs.org/accounts

Aiming for Benchmark Accuracy with the Many-Body Expansion

Ryan M. Richard,[‡] Ka Un Lao, and John M. Herbert*

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

CONSPECTUS: The past 15 years have witnessed an explosion of activity in the field of fragment-based quantum chemistry, whereby *ab initio* electronic structure calculations are performed on very large systems by decomposing them into a large number of relatively small subsystem calculations and then reassembling the subsystem data in order to approximate supersystem properties. Most of these methods are based, at some level, on the so-called many-body (or "*n*-body") expansion, which ultimately requires calculations on monomers, dimers, ..., *n*-mers of fragments. To the extent that a low-order *n*-body expansion can reproduce supersystem properties, such methods replace an intractable supersystem calculation with a large number of easily distributable subsystem calculations. This holds great promise for performing, for example, "gold standard" CCSD(T) calculations on large molecules, clusters, and condensed-phase systems.



The literature is awash in a litany of fragment-based methods, each with their own working equations and terminology, which presents a formidable language barrier to the uninitiated reader. We have sought to unify these methods under a common formalism, by means of a generalized many-body expansion that provides a universal energy formula encompassing not only traditional *n*-body cluster expansions but also methods designed for macromolecules, in which the supersystem is decomposed into overlapping fragments. This formalism allows various fragment-based methods to be systematically classified, primarily according to how the fragments are constructed and how higher-order *n*-body interactions are approximated. This classification furthermore suggests systematic ways to improve the accuracy.

Whereas *n*-body approaches have been thoroughly tested at low levels of theory in small noncovalent clusters, we have begun to explore the efficacy of these methods for large systems, with the goal of reproducing benchmark-quality calculations, ideally meaning complete-basis CCSD(T). For high accuracy, it is necessary to deal with basis-set superposition error, and this necessitates the use of many-body counterpoise corrections and electrostatic embedding methods that are stable in large basis sets. Tests on small noncovalent clusters suggest that total energies of complete-basis CCSD(T) quality can indeed be obtained, with dramatic reductions in aggregate computing time. On the other hand, naive applications of low-order *n*-body expansions may benefit from significant error cancellation, wherein basis-set superposition error partially offsets the effects of higher-order *n*-body terms, affording fortuitously good results in some cases. Basis sets that afford reasonable results in small clusters behave erratically in larger systems and when high-order *n*-body expansions are employed.

For large systems, and $(H_2O)_{N\gtrsim 30}$ is large enough, the combinatorial nature of the many-body expansion presents the possibility of serious loss-of-precision problems that are not widely appreciated. Tight thresholds are required in the subsystem calculations in order to stave off size-dependent errors, and high-order expansions may be inherently numerically ill-posed. Moreover, commonplace script- or driver-based implementations of the *n*-body expansion may be especially susceptible to loss-of-precision problems in large systems. These results suggest that the many-body expansion is not yet ready to be treated as a "black-box" quantum chemistry method.

1. INTRODUCTION: THE MANY-BODY EXPANSION

Fragment-based quantum chemistry methods have become a popular way to circumvent the highly nonlinear scaling (with respect to system size) of *ab initio* quantum chemistry calculations. The common theme among these methods is the decomposition of a large (super)system into smaller subsystems for distributed computing, followed by some attempt to reassemble this information to approximate some property of the supersystem, usually its energy or an energy derivative. Such methods appeal to what Kohn has called the near-sighted nature of electronic matter.[1] In addition to their pragmatic appeal for reasons of computational cost, there is an undeniable intuitive appeal insofar as chemists are accustomed to understanding molecules in terms of their functional groups.

A survey of the literature[2−4] reveals a plethora of fragment-based methods, each with its own terminology, motivation, and equations. Many of these methods are based at some level on the old idea of the many-body expansion (MBE), also called the *n*-body expansion, in which the supersystem energy is formally

Special Issue: Beyond QM/MM: Fragment Quantum Mechanical Methods

Received: March 12, 2014
Published: June 2, 2014

   dx.doi.org/10.1021/ar500119q | *Acc. Chem. Res.* 2014, 47, 2828−2836

expressed in terms of the energies of its constituent monomers, dimers, trimers, etc. The energy of a system of N fragments is thus expressed, formally but exactly, as

$$E = \sum_{I=1}^{N} E_I + \sum_{\substack{I,J \\ J>I}}^{N} \Delta E_{IJ} + \sum_{\substack{I,J,K \\ K>J>I}}^{N} \Delta E_{IJK} + \cdots \quad (1)$$

where E_I is the energy of the Ith fragment and

$$\Delta E_{IJ} = E_{IJ} - E_I - E_J \quad (2)$$

is a correction for two-body (pairwise-additive) interactions. Expressions for higher-order n-body corrections, ΔE_{IJK}, etc., can be found in ref 5.

If the n-body corrections become smaller as n increases, then eq 1 might sensibly be truncated at some finite n, affording an n-body approximation, $E^{(n)}$, to the total energy. This approximation can be written in closed form as[5]

$$E^{(n)} = \sum_{k=1}^{n} (-1)^{n-k} \binom{N-k-1}{n-k} \sum_{K=1}^{\binom{N}{k}} E_K^{(k)} \quad (3)$$

in which $E_K^{(k)}$ is the energy of the Kth k-mer of fragments, where K ranges over all $\binom{N}{k}$ unique k-mers.

Equation 3 suggests a sequence of approximations to the energy, becoming exact as $n \rightarrow N$, which require only subsystem calculations. Differentiation of eq 3 with respect to nuclear coordinates or external fields provides a corresponding n-body approximation to other properties such as multipole moments, static or dynamic polarizabilities, or magnetic resonance parameters.[6] In particular, forces on the nuclei (for use in geometry optimizations and molecular dynamics simulations) can be expressed as linear combinations of subsystem energy gradients.

In principle, the many-body expansion is applicable at any level of electronic structure theory, but our recent work has focused on obtaining benchmark-quality results for noncovalent clusters by applying traditional *ab initio* protocols but using eq 3 to dramatically reduce the cost of the calculations. Thus, our target benchmark is the complete-basis CCSD(T) energy, and the standard protocols include complete-basis extrapolation of counterpoise-corrected MP2/aug-cc-pVXZ energies, along with a triples correction

$$\delta E_{\mathrm{CCSD(T)}} = E_{\mathrm{CCSD(T)}} - E_{\mathrm{MP2}} \quad (4)$$

The n-body expansion reduces the computational scaling of these calculations from $O(N^p)$ to $O(n^p)$, where $p = 5$ for MP2 and $p = 7$ for CCSD(T), for example. Moreover, the subsystem calculations required by eq 3 are "embarrassingly parallelizable" in the sense that they are independent of one another and can easily be distributed across processors, leading to a dramatic reduction in both "wall time" and storage (memory and disk) requirements.

We have set our sights on a target accuracy of $\lesssim$1 kcal/mol. A large body of literature suggests that this level of accuracy is achievable or nearly achievable by considering at most $n = 3$ (see ref 2 and references therein). As such, fragment-based methods appear to offer a proverbial "free lunch" (or at worst, a heavily subsidized one), because even for a supersystem as small as $F^-(H_2O)_{10}$ with an $O(N^5)$ method such as MP2, the *total* aggregate computer time (including all subsystem calculations) required for a three-body expansion is significantly smaller than that required for a supersystem calculation.[7,8] The difference is even more pronounced for CCSD(T) calculations.

Despite this resumé of success, our recent work[5] has raised serious concerns about the generality of these high-accuracy results, which may not be indicative of the general performance of the n-body expansion, especially in large systems. In practice, we sometimes find that the sequence of n-body approximations is not convergent, owing to problems with finite precision that proliferate as n increases and to errors that grow rapidly as a function of system size.[5]

This Account is focused on (1) providing a unified formalism for discussing fragmentation methods based on the MBE, (2) understanding how such methods perform in large systems, and (3) assessing whether results are systematically improvable at all levels of electronic structure theory. Concerning the performance in large systems, we find that great care must be taken to address issues of finite precision, and regarding systematic improvability, we find that performance of the n-body expansion can be quite erratic, depending on the subsystem level of theory.

2. UNIFIED FORMALISM: THE *GENERALIZED* MANY-BODY EXPANSION

The discussion above, particularly eq 1, tacitly assumes that fragments I and J contain no nuclei in common, as would be appropriate (though not required) for applications involving noncovalent clusters. For macromolecular applications, fragmentation must sever covalent bonds and introduce caps on the severed valencies, similar to the link atoms or frozen orbital caps that are ubiquitous in QM/MM calculations. These caps represent potentially serious perturbations to the electronic structure, so to reduce their effects some investigators have taken to using overlapping fragments, which do contain nuclei in common. This choice, however, precludes straightforward use of the n-body expansion, because that would involve some double-counting, and an alternative to eq 3 must be developed. Work by others,[9−11] in the context of overlapping-fragment methods, provided a foundation for a *generalized* MBE (GMBE), which we ultimately derived in a rigorous and general way.[4,12] The GMBE provides a context for understanding the connections between a wide variety of seemingly disparate fragment-based methods.

2.1. Theory

The GMBE amounts to a sequence of n-body approximations

$$E \approx \mathcal{E}^{(1)} + \Delta\mathcal{E}^{(2)} + \cdots + \Delta\mathcal{E}^{(n)} \quad (5)$$

with each subsequent correction defined as $\Delta\mathcal{E}^{(n)} = \mathcal{E}^{(n)} - \mathcal{E}^{(n-1)}$. The quantity $\mathcal{E}^{(n)}$ is an overlap-corrected n-body energy that eliminates any double-counting:[4,12]

$$\mathcal{E}^{(n)} = \sum_{I=1}^{\binom{N}{n}} E_I^{(n)} - \sum_{I=1}^{\binom{N}{n}} \sum_{J>I}^{\binom{N}{n}} E_{I\cap J}^{(n)} + \cdots + (-1)^{\binom{N}{n}+1} E^{(n)}_{I\cap J\cap\cdots\cap\binom{N}{n}} \quad (6)$$

Here, $E_{I\cap J}^{(n)}$ is the energy of the subsystem formed from the intersection of the Ith and Jth n-mers of fragments, $F_I^{(n)} \cap F_J^{(n)}$. Equation 6 can be understood as follows.[13] The supersystem's Hamiltonian can be expressed exactly as a sum of the (overlapping) fragment Hamiltonians, plus additional terms involving all mutual intersections of these fragments, with

dx.doi.org/10.1021/ar500119q | Acc. Chem. Res. 2014, 47, 2828−2836

alternating signs as suggested by eq 6. (Fundamentally, this result comes from the set-theoretical inclusion/exclusion principle.[12]) Finally, one appeals to localized approximations of the form

$$\langle\Psi|\hat{H}(F_I^{(n)} \cap F_J^{(n)})|\Psi\rangle \approx \langle\Psi_{I\cap J}^{(n)}|\hat{H}(F_I^{(n)} \cap F_J^{(n)})|\Psi_{I\cap J}^{(n)}\rangle \quad (7)$$

where Ψ is the supersystem wave function and $\Psi_{I\cap J}^{(n)}$ is the localized wave function for $F_I^{(n)}\cap F_J^{(n)}$.

One can show that $E^{(n)} = \mathcal{E}^{(n)}$ in the special case of disjoint fragments,[4] so the GMBE is indeed a generalization of the MBE. However, eq 6 is valid for an arbitrary partition of the supersystem and in that sense is a generalization of the n-body expansion for arbitrary fragmentation methods. The $n = 1$ case of eq 6 had been suggested prior to our work,[10,11] as an energy formula for overlapping fragments, but the full form of the GMBE suggests a means for systematic improvement. A few other overlapping-fragment methods can be identified as including a proper subset of the terms in an n-body GMBE calculation,[9,14] but preliminary tests suggest the importance of these omitted terms in some cases.[12]

If the GMBE is simply a more complicated form of the MBE, in the sense that it has additional terms arising from intersections, then what is its utility? A hypothetical example is shown in Figure 1, where we have fragmented hexane on the

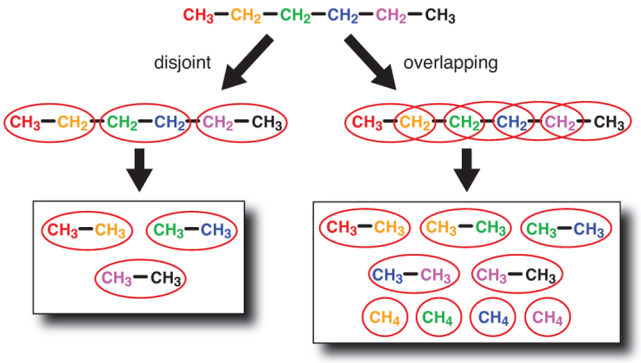


Figure 1. An illustration of disjoint versus overlapping fragments for a polyatomic molecule.

left into three disjoint fragments and on the right into five intersecting fragments. Even though the fragment sizes are the same in both cases, the disjoint fragmentation pattern neglects interactions between two pairs of covalently bonded carbon atoms, at the $n = 1$ level, and a two-body expansion is needed to include these interactions. For the cost of two additional ethane calculations and four methane calculations (the latter arising from intersections of fragments), these interactions can be included at the $n = 1$ level by using an overlapping fragmentation scheme. Thus, relaxing the restriction that fragments need to be disjoint allows us to incorporate more interactions at a given n. In practice, one might want to use fragments larger than a single carbon atom (although single-carbon fragments have been used in some previous studies[9]); nevertheless, Figure 1 can be viewed as a schematic example, and more generally the CH_2 and CH_3 groups in this figure could simply be some small pieces of a larger macromolecule, appropriately capped.

The GMBE provides the basis for a unified view of fragment-based methods, in that eq 6 provides a universal energy expression, regardless of how the fragments are formed, so that there is no need to appeal to different energy expressions for different methods. Various fragment-based methods are distinguished by the truncation order, n, by how the fragments are formed, and in the case that fragmentation severs covalent bonds by how the severed valencies are to be capped.[4]

2.2. Embedding

The n-body expansion is known to be slowly convergent, and a variety of multilevel approaches have been developed in which a two- or three-body expansion at a relatively high level of theory is supplemented with a lower-level treatment of higher-order terms.[15−19] In molecular systems, these higher-order terms are mostly many-body induction effects, and an alternative way to capture these is to perform a low-level calculation on the entire supersystem (at the Hartree−Fock level, say), which largely captures these effects, and then combine this with a low-order n-body treatment of the correlation energy.[20−22]

An alternative way to capture many-body polarization is to embed the low-order n-body calculations in some representation of the electrostatic potential of the rest of the system. In the fragment molecular orbital (FMO) method,[23,24] short-range electrostatic embedding is done using the actual fragment electron densities, then one switches to wave function-derived charges at longer range, whereas Dahlke and Truhlar[25] have pursued an approach in which only atom-centered point charges are used. In principle, these electrostatic embedding (EE) charges can be iteratively updated to mutual self-consistency alongside the fragment wave functions, although such a procedure complicates the formulation of analytic energy gradients,[4] a point that has not always been recognized in the literature.

2.3. Performance of the GMBE

Figure 2 compares the performance of the MBE versus the GMBE for a set of water and fluoride−water clusters.[4,13] We adopt the notation (G)MBE(n) for an n-body expansion, and EE-(G)MBE(n) to indicate the use of atom-centered embedding charges. The latter are noniterative Mulliken charges computed at the same level of theory that is used for

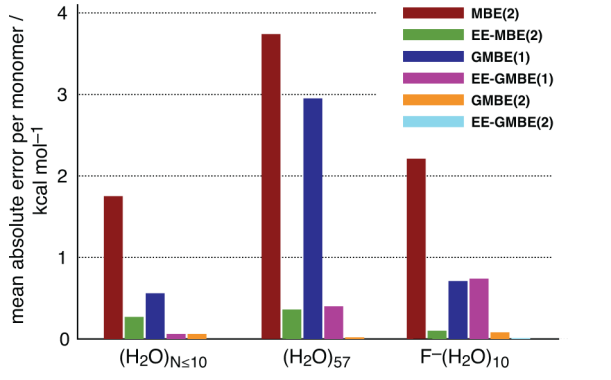


Figure 2. Mean accuracy of the (G)MBE for various sets of cluster isomers, with multiple isomers in each data set (data were obtained from ref 4). All n-body calculations were performed at the B3LYP/6-31+G(d,2p) level of theory, and accuracy is measured relative to a calculation of the entire cluster at the same level of theory. Calculations based on the GMBE employ three to four monomers per fragment, whereas MBE results use one monomer per fragment. The EE-GMBE(2) errors are so small as to be nearly invisible on this scale, except (barely) in the case of $F^-(H_2O)_{10}$.

dx.doi.org/10.1021/ar500119q | Acc. Chem. Res. 2014, 47, 2828−2836

the n-body calculations, an embedding procedure that has been shown to work remarkably well despite its simplicity.[26]

An important point to note is that we define "error" relative to a supersystem calculation performed at the same level of theory, which for the results in Figure 2 is the modest B3LYP/6-31+G(d,2p) level. This is a conscious choice, meant to explore the limitations of the (G)MBE itself and made with the recognition that eq 6 is formally exact for $n = N$. It is possible, by systematically varying the subsystem level of theory and basis set, that one might stumble upon a choice that accurately reproduces high-level supersystem benchmarks, but it is unclear what would be learned from such an exercise, and we are more interested in examining the convergence (or lack thereof) of the n-body expansion toward the supersystem result.

Results for water and fluoride−water clusters in Figure 2 show that MBE(2) is a poor approximation without electrostatic embedding, but EE-MBE(2) errors per monomer are quite small. The GMBE results in Figure 2 are based on placing three or four water molecules in each overlapping fragment, based on a distance criterion,[4] and a fairly large cluster ($N = 57$ in Figure 2) is required in order to discover that GMBE(1) is not a high-accuracy approach, in the absence of embedding. Notably, the GMBE(1) energy formula is the same one that has been written down in several previous studies,[9−11] based on the inclusion/exclusion principle.

It should also be noted that the errors in Figure 2 are plotted on a per-monomer basis, and upon multiplying the errors for $(H_2O)_{57}$ by a factor of 57, one discovers that the GMBE(2) and EE-GMBE(2) methods are the only ones that achieve an accuracy of <1 kcal/mol in the total energy. This is a notable victory for the intersecting-fragment approach.

Another potentially significant advantage of the GMBE, which we have only begun to explore, is that this approach may be less sensitive to the details of precisely how the system is fragmented. In tests involving $F^-(H_2O)_6$ and $(^+H_3NCH_2CO_2^-)(H_2O)_{10}$ clusters, we consistently obtain high-accuracy results with the GMBE for a variety of fragmentation patterns, including cases in which the fragments were intentionally chosen in a nonintuitive way, placing spatially distant monomers together in the same fragment.[12] More effort is needed to test the generality of this conclusion, but if it proves to be robust this could be especially important for macromolecular applications, where the choice of fragments is less intuitive than it is for noncovalent clusters.

3. QUEST FOR HIGH ACCURACY

Too often, the reliability of EE-MBE and FMO methods has been judged based on comparison to supersystem calculations performed at modest levels of electronic structure theory (often Hartree−Fock or DFT, or occasionally MP2) in small basis sets. In this section, we consider high-accuracy calculations for noncovalent clusters, where the goal is to use n-body methods to reproduce the results of highly correlated approaches in large basis sets.

3.1. Counterpoise Corrections

For high-accuracy applications, one must confront the issue of basis-set superposition error (BSSE) or, in other words, the borrowing of a neighbor's basis functions that often leads to severe overestimation of interaction energies and disappears exceedingly slowly as the basis set approaches completeness. While the BSSE problem is well-known, as is its widely accepted solution via counterpoise (CP) correction,[27] it is useful to calibrate the reader as to the magnitude of this effect. For that purpose, Figure 3 illustrates the convergence of MP2/aug-cc-pVXZ calculations (X = D, T, Q) for an isomer of $(H_2O)_6$ whose MP2/CBS limit has been carefully established. Notice that a quadruple-ζ basis set is required to get within ±1 kcal/mol of this limit and that in the absence of CP corrections, even the CBS extrapolation misses the mark.

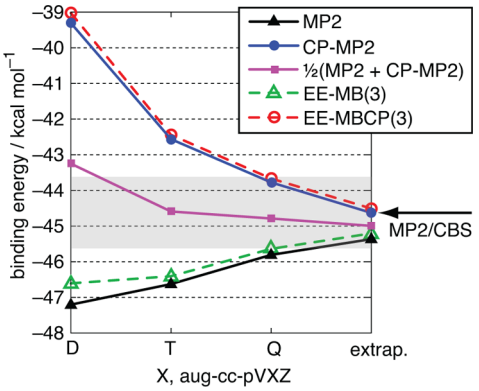


Figure 3. Convergence to the MP2/CBS limit for the "bag" isomer of $(H_2O)_6$. The gray region denotes ±1 kcal/mol from the benchmark MP2/CBS value. Reprinted from ref 7. Copyright 2013 American Chemical Society.

Insofar as the n-body expansion is designed to reproduce the results of a supersystem calculation performed at the same level of theory, one can expect slow basis-set convergence for EE-MBE(n) calculations as well, as shown in Figure 3 at the $n = 3$ level. What is needed is a CP correction that is valid order-by-order in the MBE. Once such correction had been introduced by Kamiya et al.,[28] but it requires an impractically large number of subsystem calculations. For $N = 20$ fragments, for example, 7340 distinct calculations are required at the $n = 3$ level (each in a trimer basis set), and at the $n = 4$ level, an additional 67 830 distinct calculations are required, each in a tetramer basis set. These calculations involve one to four actual monomers, with ghost atoms filling out the rest of the trimer or tetramer basis.

As an alternative, we have introduced a many-body CP correction that starts from the standard Boys−Bernardi correction and then applies a consistent n-body expansion to all terms. For an N-body cluster, the Boys−Bernardi CP correction is

$$\delta E^{\mathrm{CP}} = \sum_{I=1}^{N} (E_I^I - E_I^{IJK\ldots N}) \tag{8}$$

where superscripts indicate which monomers contribute basis functions and subscripts indicate which monomers contribute electrons and nuclei (the rest are ghost atoms). Application of an n-body approximation to δE^{CP} affords a correction that we have called MBCP(n), which is given by

$$\delta E^{\mathrm{MBCP}(2)} = \sum_{I=1}^{N} [(N-1)E_I^I - \sum_{J\neq I}^{N} E_I^{IJ}] \tag{9}$$

for $n = 2$ and

dx.doi.org/10.1021/ar500119q | Acc. Chem. Res. 2014, 47, 2828−2836

$$\delta E^{\mathrm{MBCP}(3)} = \sum_{I=1}^{N} [-\frac{1}{2}(N-4)(N-1)E_I^I + (N-3)\sum_{J\neq I}^{N} E_I^{IJ} - \sum_{J\neq I}^{N}\sum_{\substack{K=J+1 \\ (K\neq I)}}^{N} E_I^{IJK}] \quad (10)$$

for $n = 3$. It is then logically consistent to combine the MBCP(n) correction with an n-body approximation to the supersystem energy (eq 3). Figure 3 shows that the EE-MBCP(3) approach accurately reproduces the CP-corrected MP2 energy for $(H_2O)_6$ across the whole sequence of aug-cc-pVXZ basis sets and can be extrapolated to an accurate MP2/CBS binding energy.

For $N = 20$ fragments, the MBCP(3) correction requires 3800 calculations in a trimer basis, half as many as the CP correction of Kamiya et al.,[28] and the reduction is even greater at $n = 4$. While this still seems like (and is!) a substantial amount of computation, for a system with only $N = 11$ fragments, the cost of an MBCP(3) calculation at the RIMP2/aug-cc-pVQZ level is less than 30% (in *aggregate* computer time, considering all parallel processors) of what is required for a traditional, CP-corrected RIMP2 calculation in the same basis set.[7,8] For the MBCP(2) approach, the computer time is reduced by 98% and the accuracy is ∼1 kcal/mol,[7] which may be sufficient for rapid, rough screening of potential energy surfaces.

In the absence of CP corrections, one should be skeptical of ostensibly high-accuracy results obtained with n-body expansions, because BSSE (which overstabilizes the cluster) can partially compensate for neglect of higher-order n-body interactions in potentially unpredictable ways.[8] Examples are shown in Figure 4, which plots the convergence toward the MP2/CBS limit of MP2/aug-cc-pVXZ energies that are *not* CP-corrected. (The MP2/CBS limit is established using CP-corrected results.) In finite basis sets and in the absence of CP corrections, EE-MBE(2) and EE-MBE(3) approximations to the MP2 energy are actually *more* accurate (with respect to the MP2/CBS benchmark) than is MP2 itself! This is simply error cancellation, but there is no clear pattern to this cancellation, even in the relatively high-quality and systematic aug-cc-pVXZ sequence of basis sets. Results in less systematically defined basis sets are liable to behave even more erratically.

3.2. CCSD(T) Corrections

For best accuracy, "gold standard" CCSD(T) CBS results are desirable. With the MP2/CBS limit in hand, these are obtained by adding a correction, $\delta E_{\mathrm{CCSD(T)}}$, as defined in eq 4, evaluated in a triple-ζ basis set.[29] In the spirit of this work, we apply an n-body approximation to both the MP2 and CCSD(T) energies in this equation, to obtain a (relatively!) low-cost CCSD(T) correction.

Figure 5 shows the errors engendered by two- and three-body approximations to $\delta E_{\mathrm{CCSD(T)}}$, which are vanishingly small at the three-body level and <0.2 kcal/mol at the two-body level. In conjunction with MP2/CBS results from Figure 3, this suggests that total energies within ∼0.2 kcal/mol of CCSD(T)/CBS benchmarks can be obtained for $(H_2O)_6$, with a computational strategy whose bottleneck steps are MP2/aug-cc-pVQZ calculations on trimers and CCSD(T)/heavy-aug-cc-pVTZ calculations on dimers.

Both aspects of this strategy need to be investigated in larger clusters, however. In the more strongly interacting $F^-(H_2O)_{10}$ cluster, for example, the two-body triples correction is ∼2.4 kcal/mol (averaged over several isomers), whereas the three-body approximation to $\delta E_{\mathrm{CCSD(T)}}$ is ∼1.8 kcal/mol.[30] One might therefore guess that the four-body approximation would differ by an even smaller amount (<0.6 kcal/mol) from the three-body approximation, insofar as the n-body expansion converges monotonically, but this may not be the case and warrants further exploration.

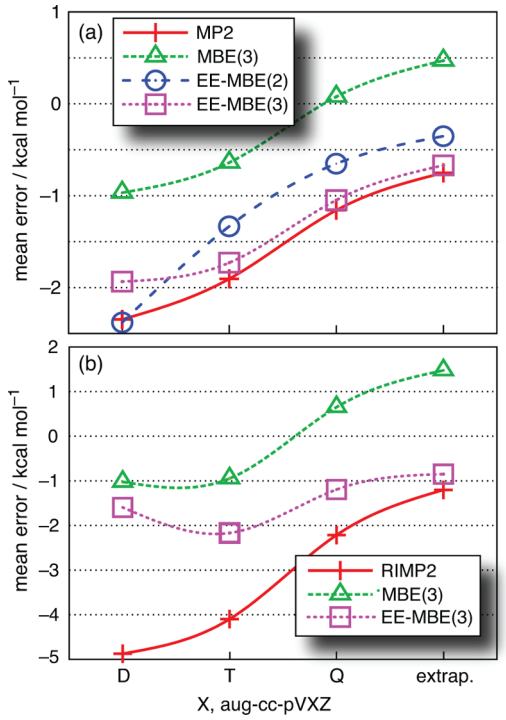


Figure 4. Mean errors in (RI)MP2/aug-cc-pVXZ energies for (a) $(H_2O)_6$ and (b) $F^-(H_2O)_{10}$, computed with respect to (RI)MP2/CBS benchmarks and averaged over a set of isomers. The (RI)MP2 errors remain nonzero upon extrapolation in this figure because the benchmark extrapolation uses counterpoise-corrected (RI)MP2 results. Reprinted from ref 8. Copyright 2013 American Institute of Physics.

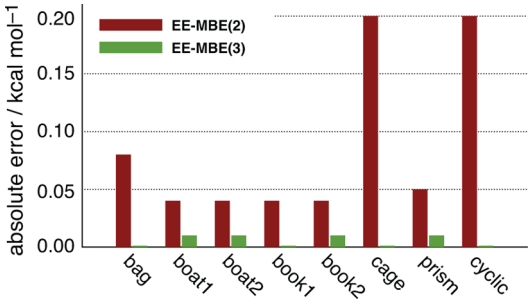


Figure 5. Absolute errors in two- and three-body approximations to $\delta E_{\mathrm{CCSD(T)}}$, relative to full-cluster CCSD(T) benchmarks. Data are taken from ref 7.

3.3. Embedding Methods

Although a few results in the literature suggest that EE-MBE(n) results are fairly insensitive to the nature of the point charges used in the embedding,[26] differences are more pronounced in our hands and become even more so in large clusters.[5] Preliminary results at a modest level of theory (Figure 6) suggest that varying the nature of the electrostatic embedding

dx.doi.org/10.1021/ar500119q | Acc. Chem. Res. 2014, 47, 2828−2836

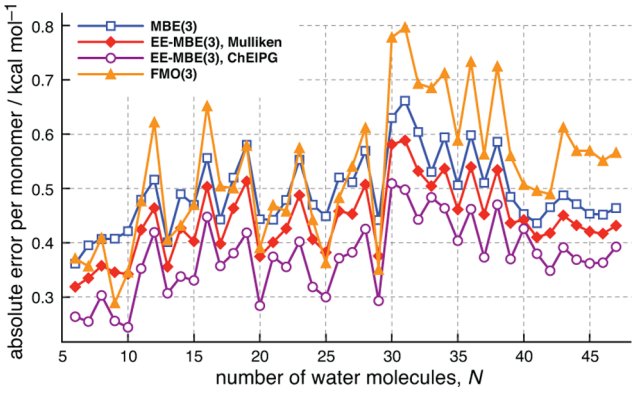


Figure 6. Absolute error per monomer in the binding energies of $(H_2O)_N$ clusters (representing putative global minimum structures from ref 31) for various three-body expansions. Errors are measured relative to supersystem calculations at the same level of theory (B3LYP/cc-pVDZ).

may be a route to higher accuracy. In the figure, we compare EE-MBE(3) results using two different point-charge embeddings: gas-phase Mulliken charges and iteratively updated "ChElPG" charges. The latter are fit in order to reproduce the electrostatic potential outside of the atomic van der Waals radii and thus represent the best electrostatic embedding charges, in a well-defined sense. For the sequence of water clusters in Figure 6, these charges do indeed outperform Mulliken embedding for EE-MBE(3) calculations, and Mulliken embedding in turn outperforms the nonembedded MBE(3) approximation.

One might anticipate even better results if the actual monomer electron densities are used to compute the intermolecular Coulomb interactions, which is the approach taken (at least for nearby monomers) in the FMO method.[23,24] FMO(3) results are also shown in Figure 6, but in fact these errors are significantly larger than those obtained from MBE(3) calculations with no embedding at all! FMO calculations are sometimes prone to large errors for large basis sets, especially where diffuse functions are involved, and point-charge embedding often performs better in such cases.[32] However, the calculations in Figure 6 use only the cc-pVDZ basis set, yet errors remain large. The reasons for this are unclear.

4. SYSTEMATIC TESTS FOR LARGER SYSTEMS

Although numerous studies have documented the performance of the MBE in small noncovalent clusters, relatively little attention has been paid to how these methods perform as a function of increasing system size. We have recently begun to explore the performance of n-body expansions in larger clusters,[5] with some provocative results that are summarized in this section.

Given a function f that depends on a set of independent variables $\{x_i\}$, propagation-of-errors (PoE) analysis suggests that the uncertainty in the value of f is

$$\delta f = \left[\sum_i \left(\frac{\partial f}{\partial x_i}\right)^2 (\delta x_i)^2\right]^{1/2} \tag{11}$$

where δx_i is the uncertainty in x_i. In the present case, $f = E^{(n)}$ is the approximate energy obtained from an n-body expansion, and the x_i are the independent subsystem energies. For a self-consistent field (SCF) convergence threshold of $10^{-\alpha}$ hartree, we assume that the $(\alpha + 1)$st decimal digit of the energy is a random number and take $10^{-(\alpha+1)}$ hartree as the uncertainty in the subsystem energies. A PoE analysis of eq 3 is then used to estimate how these uncertainties manifest in $E^{(n)}$.[5]

The combinatorial nature of the MBE causes the uncertainty in $E^{(n)}$ to grow highly nonlinearly as a function of system size, N.[5] Error estimates obtained from PoE analysis are plotted as a function of N in Figure 7, and for MBE(3), the uncertainty

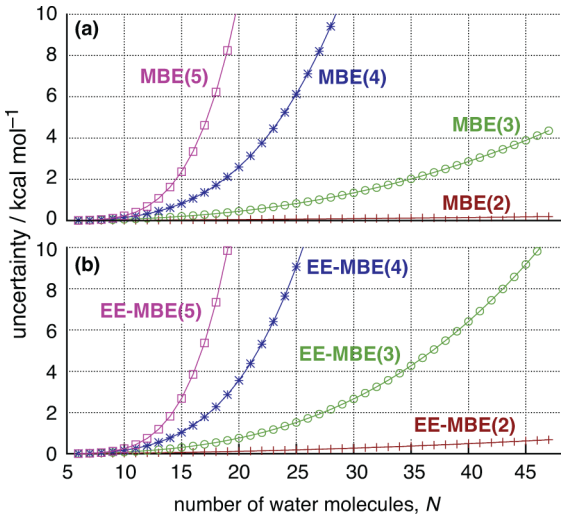


Figure 7. Total uncertainty (as estimated by PoE analysis) in an n-body expansion (a) without embedding, assuming an uncertainty of 10^{-6} hartree in each subsystem calculation, and (b) with embedding, assuming an additional uncertainty of 10^{-6} au in the embedding charges.

crosses the 1 kcal/mol mark around $N \approx 30$. The situation is markedly worse when embedding charges are added, and for EE-MBE(3) calculations, the PoE uncertainty crosses the 1 kcal/mol threshold for $N \approx 20$. (The latter analysis assumes an uncertainty of 10^{-6} au in the embedding charges, consistent with the six decimal digits to which these charges are often output by electronic structure programs.) Although the data in Figure 7 are generic estimates, in actual practice we have noted that for $n \geq 3$ and for systems of size $(H_2O)_{N\gtrsim 30}$, discrepancies on the order of several kilocalories per mole can appear between implementations of EE-MBE(n) that use a script or a driver program (which simply reads the text output from an electronic structure program), compared with an implementation that reads binary scratch files in full double precision.[5]

These results underscore the need for high precision in the subsystem energy calculations. The use of embedding charges exacerbates this need but seems preferable to including higher-order n-body terms whose number proliferates rapidly. We have systematically studied the precision of EE-MBE(n) as a function of the SCF convergence threshold (τ_{SCF}) and the integral screening threshold (τ_{ints}),[5] and some sample results are shown in Table 1 for "loose" versus "tight" thresholds. Through at least $N = 40$ water molecules, results with the two sets of thresholds are unchanged for an EE-MBE(2) calculation but differ by ≈1.5 kcal/mol for EE-MBE(3) calculations on $(H_2O)_{40}$ and by even more for EE-MBE(4) calculations. The dramatic variation in accuracy as a function of τ_{ints} (which is primarily associated with shell-pair screening[5]) is disheartening, because the "tight" value $\tau_{ints} = 10^{-14}$ au that is used in Table 1

dx.doi.org/10.1021/ar500119q | Acc. Chem. Res. 2014, 47, 2828−2836

Table 1. Total Errors (in kcal/mol) in EE-MBE(n) Calculations[a] of $(H_2O)_N$ Clusters Using Two Sets of Thresholds

	EE-MBE(2)		EE-MBE(3)		EE-MBE(4)		EE-MBE(5)	
N	loose[b]	tight[c]	loose[b]	tight[c]	loose[b]	tight[c]	loose[b]	tight[c]
10	−8.1	−8.1	2.4	2.4	−0.3	−0.3	−0.0	−0.0
20	−24.2	−24.3	5.6	6.0	0.1	−0.6	−2.8	−0.2
30	−55.2	−55.2	15.1	16.0	1.3	−1.1	−16.2	−2.6
40	−74.6	−74.6	16.8	18.4	4.6	−0.5	−54.1	−8.1

[a]Subsystem calculations performed at the B3LYP/cc-pVDZ level with TIP3P embedding charges and compared with a supersystem B3LYP/cc-pVDZ calculation. [b]$\tau_{SCF} = 10^{-5}$ au and $\tau_{ints} = 10^{-9}$ au. [c]$\tau_{SCF} = 10^{-6}$ au and $\tau_{ints} = 10^{-14}$ au.

will be extremely costly for correlated wave function calculations.

The threshold-dependent data in Table 1 bear out the PoE prediction that precision problems will grow worse as a function of n for a given system size. Together, these results call into question the assumption that the n-body expansion represents a systematically improvable method, as a function of n. In support of this assertion, we plot size-dependent errors in EE-MBE(n) results in Figure 8. There are several take-home

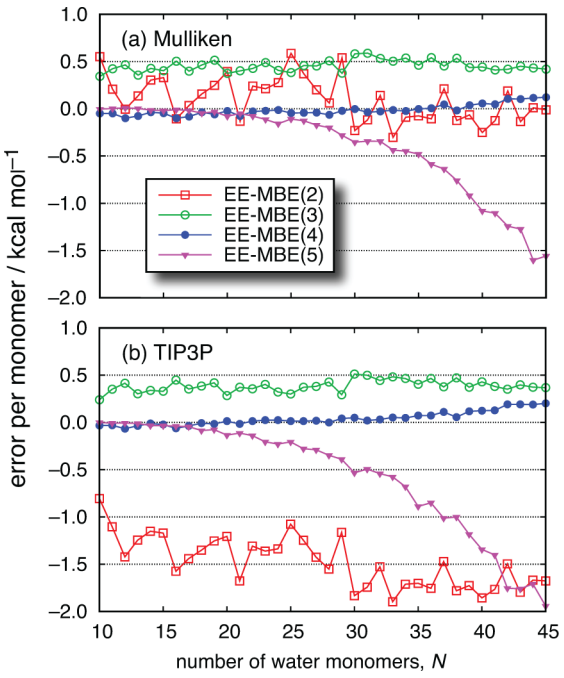


Figure 8. Errors per monomer, with respect to a supersystem calculation at the same level of theory (B3LYP/cc-pVDZ), in EE-MBE(n) results using (a) Mulliken and (b) TIP3P embedding charges. Thresholds were set at $\tau_{SCF} = 10^{-5}$ au and $\tau_{ints} = 10^{-9}$ au.

points from this plot. First, notice the rapid divergence (with respect to system size) of the EE-MBE(5) results, even when the error is measured on a per-monomer basis. Although the individual five-body energy corrections may be small, they proliferate rapidly in number as system size increases. Second, no attempt has been made to correct these results for BSSE, and therefore the accuracy of the supersystem benchmarks is questionable, but in any case the extent to which uncorrected n-body results can approximate the uncorrected supersystem energy is interesting. The data in Figure 8 suggest that $n = 4$ is required to achieve accuracy better than ∼0.5 kcal/(mol·monomer) for large systems.

5. FUTURE OUTLOOK

Our generalized many-body expansion demonstrates how a variety of fragment-based quantum chemistry methods, including those based on overlapping fragments, can be unified under a common theoretical framework. At the moment, however, our numerical results regarding the properties of this expansion are somewhat conflicting. We have demonstrated very high accuracy with an overlapping two-body expansion, where "accuracy" is defined with respect to a supersystem calculation computed at the same level of theory that is used in the subsystem calculations. Low-level DFT benchmarks can be reproduced with extraordinarily high accuracy in systems as large as $(H_2O)_{57}$, and high-accuracy CCSD(T)/CBS benchmarks for systems such as $F^-(H_2O)_{10}$ are reproduced for the "right reasons", via three-body expansions of MP2/aug-cc-pVXZ energies combined with three-body counterpoise corrections and extrapolated to the MP2/CBS limit, followed by a two-body CCSD(T) correction.

Results based on the traditional (nonoverlapping) n-body expansion, however, are less promising when extended to large systems. The combinatorial growth in the number of subsystem calculations (with respect to both system size, N, and truncation order, n) leads to precision problems that amplify uncertainties in the subsystem energies, necessitating the use of tighter-than-normal numerical thresholds and suggesting that, as a practical matter, only three- or possibly four-body expansions are numerically feasible. As such, the n-body expansion does not, in practice, afford a systematically improvable sequence of approximations and cannot naively be assumed to converge to the correct result as n increases.

An alternative to high-order expansions is to sum the high-order terms in closed-form by means of a supersystem calculation performed at a low level of theory,[20−22] which can be done in an affordable way even for large systems if the supersystem calculation uses a classical force field.[16] Even with tight numerical thresholds, however, we find that total error grows rapidly as a function of N even at the two-body level, although the error per fragment may still be acceptable at the three-body level. As such, it is imperative to test the accuracy of even low-order expansions in a systematic way for large systems.

Despite results in small clusters suggesting that the details of electrostatic embedding matter little,[26] results presented here for both point-charge and density embeddings suggest that these details matter quite significantly. We also find that the accuracy varies in unpredictable ways as the basis set is changed,[5,8] at least some of which is a result of BSSE that has often gone uncorrected in n-body methods, which can partially offset the neglect of higher-order terms in the expansion.[8] Given all of these caveats, at present we must conclude that it remains to be seen whether the n-body expansion, or a

dx.doi.org/10.1021/ar500119q | Acc. Chem. Res. 2014, 47, 2828−2836

generalized *n*-body expansion,[4] can be turned into a "black box" quantum chemistry method where one can reliably anticipate the error as a function of both *n* and *N*.

AUTHOR INFORMATION

Corresponding Author

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

Present Address

‡R.M.R.: School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA.

Funding

This research 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 no. PAA-0003. J.M.H. is an Arthur P. Sloan Foundation Fellow and a Camille Dreyfus Teacher-Scholar.

Notes

The authors declare no competing financial interest.

Biographies

Ryan Richard attended Cleveland State University and did undergraduate research with Prof. David Ball. He received a Ph.D. from The Ohio State University in 2013 under the supervision of Prof. John Herbert, where his thesis focused on fragment-based quantum chemistry. He is currently a postdoctoral researcher with Prof. David Sherrill at the Georgia Institute of Technology.

Ka Un Lao was born in Macau and received B.S. and M.S. degrees in chemistry from National Tsing Hua University in Hsinchu, Taiwan, where he worked with Prof. Chin-Hui Yu. He is currently a graduate student under Prof. John Herbert at The Ohio State University, where his research focuses on intermolecular interactions using fragment-based methods and symmetry-adapted perturbation theory.

John Herbert obtained his Ph.D. from the University of Wisconsin—Madison in 2003, where he studied under Prof. John Harriman. Following postdoctoral work at The Ohio State University and at the University of California—Berkeley, he joined The Ohio State University as an Assistant Professor in 2006, where he was promoted to the rank of Associate Professor in 2011 and to Professor in 2014. His research interests include intermolecular interactions, solvation phenomena, and electronic processes in systems that are big, wet, messy, and warm. His research group is a major contributor to the Q-Chem[33] electronic structure software.

REFERENCES

(1) Prodan, E.; Kohn, W. Nearsightedness of electronic matter. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 11635−11638.

(2) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation methods: A route to accurate calculations on large systems. *Chem. Rev.* **2012**, *112*, 632−672.

(3) Beran, G. J. O.; Hirata, S. Fragment and localized orbital methods in electronic structure theory. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7559−7561.

(4) Richard, R. M.; Herbert, J. M. A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. *J. Chem. Phys.* **2012**, *137*, No. 064113.

(5) Richard, R. M.; Lao, K. U.; Herbert, J. M. Understanding the many-body expansion in large systems. I. Precision considerations. Manuscript submitted for publication.

(6) Gauss, J. In *Modern Methods and Algorithms of Quantum Chemistry*, 2nd ed.; Grotendorst, J., Ed.; NIC Series; John von Neumann Institute for Computing: Jülich, Germany, 2000; Vol. 3; pp 541−592.

(7) Richard, R. M.; Lao, K. U.; Herbert, J. M. Achieving the CCSD(T) basis-set limit in sizable molecular clusters: Counterpoise corrections for the many-body expansion. *J. Phys. Chem. Lett.* **2013**, *4*, 2674−2680.

(8) Richard, R. M.; Lao, K. U.; Herbert, J. M. Approaching the complete-basis limit with a truncated many-body expansion. *J. Chem. Phys.* **2013**, *139*, No. 224102.

(9) Deev, V.; Collins, M. A. Approximate *ab initio* energies by systematic molecular fragmentation. *J. Chem. Phys.* **2005**, *122*, No. 154102.

(10) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. Molecular tailoring approach for geometry optimization of large molecules: Energy evaluation and parallelization strategies. *J. Chem. Phys.* **2006**, *125*, No. 104109.

(11) Li, W.; Li, S.; Jiang, Y. Generalized energy-based fragmentation approach for computing the ground-state energies and properties of large molecules. *J. Phys. Chem. A* **2007**, *111*, 2193−2199.

(12) Richard, R. M.; Herbert, J. M. The many-body expansion with overlapping fragments: Analysis of two approaches. *J. Chem. Theory Comput.* **2013**, *9*, 1408−1416.

(13) Jacobson, L. D.; Richard, R. M.; Lao, K. U.; Herbert, J. M. Efficient monomer-based quantum chemistry methods for molecular and ionic clusters. *Annu. Rep. Comput. Chem.* **2013**, *9*, 25−56.

(14) Mayhall, N. J.; Raghavachari, K. Many-overlapping-body (MOB) expansion: A generalized many body expansion for non-disjoint monomers in molecular fragmentation calculations of covalent molecules. *J. Chem. Theory Comput.* **2012**, *8*, 2669−2675.

(15) Fedorov, D. G.; Ishida, T.; Kitaura, K. Multilayer formulation of the fragment molecular orbital method (FMO). *J. Phys. Chem. A* **2005**, *109*, 2638−2646.

(16) Beran, G. J. O. Approximating quantum many-body intermolecular interactions in molecular clusters using classical polarizable force fields. *J. Chem. Phys.* **2009**, *130*, No. 164115.

(17) Řezáč, J.; Salahub, D. R. Multilevel fragment-based approach (MFBA): A novel hybrid computational method for the study of large molecules. *J. Chem. Theory Comput.* **2010**, *6*, 91−99.

(18) Góra, U.; Podeszwa, R.; Cencek, W.; Szalewicz, K. Interaction energies of large clusters from many-body expansion. *J. Chem. Phys.* **2011**, *135*, No. 224102.

(19) Mayhall, N. J.; Raghavachari, K. Molecules-in-molecules: An extrapolated fragment-based approach for accurate calculations on large molecules and materials. *J. Chem. Theory Comput.* **2011**, *7*, 1336−1343.

(20) Stoll, H. Correlation energy of diamond. *Phys. Rev. B* **1992**, *46*, 6700−6704.

(21) Tschumper, G. S. Multicentered integrated QM:QM methods for weakly bound clusters: An efficient and accurate 2-body:many-body treatment of hydrogen bonding and van der Waals interactions. *Chem. Phys. Lett.* **2006**, *427*, 185−191.

(22) Dahlke, E. E.; Truhlar, D. G. Electrostatically embedded many-body correlation energy, with applications to the calculation of accurate second-order Møller−Plesset perturbation theory energies for large water clusters. *J. Chem. Theory Comput.* **2007**, *3*, 1342−1348.

(23) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment molecular orbital method: An approximate computational method for large molecules. *Chem. Phys. Lett.* **1999**, *313*, 701−706.

(24) Fedorov, D. G.; Kitaura, K. Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. *J. Phys. Chem. A* **2007**, *111*, 6904−6914.

(25) Dahlke, E. E.; Truhlar, D. G. Electrostatically embedded many-body expansion for large systems, with applications to water clusters. *J. Chem. Theory Comput.* **2007**, *3*, 46−53.

(26) Leverentz, H. R.; Truhlar, D. G. Electrostatically embedded many-body approximation for systems of water, ammonia, and sulfuric

acid and the dependence of its performance on embedding charges. *J. Chem. Theory Comput.* **2009**, *5*, 1573−1584.
(27) van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Lenthe, J. H. State of the art in counterpoise theory. *Chem. Rev.* **1994**, *94*, 1873−1885.
(28) Kamiya, M.; Hirata, S.; Valiev, M. Fast electron correlation methods for molecular clusters without basis set superposition errors. *J. Chem. Phys.* **2008**, *128*, No. 074103.
(29) Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis set convergence of the coupled-cluster correction, $\delta_{MP2}^{CCSD(T)}$: Best practices for benchmarking non-covalent interactions and the attendant revision of the S22, NBC10, and HSG databases. *J. Chem. Phys.* **2011**, *135*, No. 194102.
(30) Lao, K. U.; Herbert, J. M. An improved treatment of empirical dispersion and a many-body energy decomposition scheme for the explicit polarization plus symmetry-adapted perturbation theory (XSAPT) method. *J. Chem. Phys.* **2013**, *139*, No. 034107.
(31) Kazachenko, S.; Thakkar, A. J. Water nanodroplets: Predictions of five model potentials. *J. Chem. Phys.* **2013**, *138*, No. 194302.
(32) Fedorov, D. G.; Slipchenko, L. V.; Kitaura, K. Systematic study of the embedding potential description in the fragment molecular orbital method. *J. Phys. Chem. A* **2010**, *114*, 8742−8753.
(33) Krylov, A. I.; Gill, P. M. W. Q-Chem: An engine for innovation. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3*, 317−326.