

# Understanding the many-body expansion for large systems. III. Critical role of four-body terms, counterpoise corrections, and cutoffs

Kuan-Yu Liu and John M. Herbert<sup>a)</sup>

*Department of Chemistry and Biochemistry, The Ohio State University, Columbus, Ohio 43210, USA*

(Received 1 June 2017; accepted 22 August 2017; published online 13 September 2017)

Papers I and II in this series [R. M. Richard *et al.*, *J. Chem. Phys.* **141**, 014108 (2014); K. U. Lao *et al.*, *ibid.* **144**, 164105 (2016)] have attempted to shed light on precision and accuracy issues affecting the many-body expansion (MBE), which only manifest in larger systems and thus have received scant attention in the literature. Many-body counterpoise (CP) corrections are shown to accelerate convergence of the MBE, which otherwise suffers from a mismatch between how basis-set superposition error affects subsystem versus supersystem calculations. In water clusters ranging in size up to (H<sub>2</sub>O)<sub>37</sub>, four-body terms prove necessary to achieve accurate results for both total interaction energies and relative isomer energies, but the sheer number of tetramers makes the use of cutoff schemes essential. To predict relative energies of (H<sub>2</sub>O)<sub>20</sub> isomers, two approximations based on a lower level of theory are introduced and an ONIOM-type procedure is found to be very well converged with respect to the appropriate MBE benchmark, namely, a CP-corrected supersystem calculation at the same level of theory. Results using an energy-based cutoff scheme suggest that if reasonable approximations to the subsystem energies are available (based on classical multipoles, say), then the number of requisite subsystem calculations can be reduced even more dramatically than when distance-based thresholds are employed. The end result is several accurate four-body methods that do not require charge embedding, and which are stable in large basis sets such as aug-cc-pVTZ that have sometimes proven problematic for fragment-based quantum chemistry methods. Even with aggressive thresholding, however, the four-body approach at the self-consistent field level still requires roughly ten times more processors to outmatch the performance of the corresponding supersystem calculation, in test cases involving 1500–1800 basis functions. *Published by AIP Publishing.* [<http://dx.doi.org/10.1063/1.4986110>]

## I. INTRODUCTION

Macromolecules, and also nanoscale molecular clusters and assemblies, serve as bridges between the quantum and classical limits and thus make interesting targets for quantum chemistry.<sup>1–22</sup> In contrast to semi-empirical, QM/MM, or force-field calculations, full electronic structure calculations on systems of this size usually require either massively parallel implementations of the underlying algorithms<sup>1</sup> (possibly in conjunction with linear-scaling versions of those algorithms<sup>2–4</sup>) or else implementations using graphical processing units.<sup>5</sup> An increasingly popular alternative, and one that is perhaps more easily amenable to large-scale parallelization,<sup>6</sup> is to adopt a fragment-based approach.<sup>7–22</sup> Fragment-based methods attempt to bypass the steep non-linear scaling of traditional quantum chemistry by decomposing a large system into (a potentially very large number of) small fragments. Insofar as calculations can be performed independently on subsystems composed of these fragments, the overall method is trivially parallelizable. Its utility depends upon the ability to reassemble the subsystem information

in a way that affords useful approximations to supersystem properties.

At some level, most fragment-based quantum chemistry methods rely on the many-body expansion (MBE) or generalizations thereof.<sup>23</sup> In such approaches the total energy, or any other property that can be expressed as a derivative of the energy,<sup>24</sup> is decomposed into a sum of contributions arising from monomers, dimers, trimers, . . . of fragments. High-order terms in the expansion are neglected in order to obtain a tractable approximation. Three-body terms sometimes contribute 15%–20% of the total inter-fragment interaction energy,<sup>25</sup> and can play a pivotal role in stabilizing, e.g.,  $\alpha$ -helix structures in peptides over long distances,<sup>21</sup> and are therefore usually retained. Four-body and higher-order terms are typically neglected, despite having been shown to be important in predicting relative conformational energies of proteins.<sup>14</sup> These terms are also definitely *not* negligible in water clusters,<sup>26–28</sup> where many-body polarization effects are significant.

It has been argued<sup>29–31</sup> that embedding the  $n$ -body subsystem quantum chemistry calculations in an environment of classical point charges, which serve to mimic the remaining fragments, will accelerate convergence of the MBE by replicating some portion of the many-body polarization effects that

<sup>a)</sup>herbert@chemistry.ohio-state.edu

are neglected when higher-order terms in the MBE are omitted. Papers I and II<sup>27,28</sup> strongly contest this idea, however, and suggest that much of the “conventional wisdom” regarding the MBE is either incorrect or at best does not generalize beyond the rather small systems (say,  $N \lesssim 10$  fragments) that have generally been used to benchmark truncated MBEs. Notably, small water clusters of this size were used as benchmarks in Refs. 29–31, but we obtained very different results upon examining water clusters up to  $N = 55$ .<sup>28,32</sup> At the three-body level, errors in the total interaction energies exceeded 15 kcal/mol by  $N = 30$ , and point-charge embedding did relatively little to reduce these errors, regardless of the details of how the charges were computed.<sup>28,32</sup> Only at the four-body level were errors reduced as low as a few kcal/mol.<sup>28</sup>

In general, however, one should not assume that inclusion of higher-order  $n$ -body terms in the MBE will necessarily afford better accuracy. This is perhaps counterintuitive, but the increasingly large number of subsystem calculations (each with error in the last digits) that are required as  $n$  increases engenders loss-of-precision issues that necessitate the use of far tighter convergence thresholds and drop tolerances than would ordinarily be required in a single electronic structure calculation.<sup>27,28</sup> In our experience, mainly with water clusters, these issues do not manifest in a significant way until the number of fragments reaches  $N \approx 30$ . Perhaps because supersystem calculations on large systems are required in order to notice this problem, it has largely been overlooked in most previous work on the MBE. The problem is especially acute when the software that runs the fragment-based calculation simply reads the output file of an electronic structure program, where quantities of interest are often truncated in their precision.<sup>27</sup> Especially in the presence of embedding charges, it is crucial to read binary scratch files or checkpoint files instead, in full machine precision.<sup>27</sup> (This fact has been mentioned in passing elsewhere<sup>34</sup> but without further analysis.)

The proper definition of “error” in the context of the MBE is debatable. Our group has long argued that the appropriate benchmark to assess the accuracy of a truncated  $n$ -body calculation is comparison to a supersystem calculation carried out at the same level of theory as that used for the subsystem calculations.<sup>23,27,28,32,33,35</sup> An alternative proposal is to compare to the best available benchmarks for a given system,<sup>36,37</sup> despite any disparities between levels of theory and basis sets. In our view, such a comparison makes the  $n$ -body expansion unsystematic and renders it essentially impossible to decipher how much of its success arises from error cancellation as opposed to capturing the true physics of the interactions.

A potentially more systematic and therefore more easily treatable source of error cancellation is basis-set superposition error (BSSE), whose effects were never discussed in the context of the MBE until recently.<sup>38,39</sup> BSSE can cause convergence of the MBE (with respect to  $n$ ) to become erratic because it may offset neglected many-body induction effects.<sup>33</sup> To address this problem, our group<sup>28,38</sup> and others<sup>40,41</sup> have developed many-body counterpoise (CP) corrections that are designed to approximate the supersystem Boys-Bernardi CP correction<sup>42</sup> (as generalized to an arbitrary number of monomers<sup>43,44</sup>), order-by-order in the MBE. It is

now clear that BSSE affects the supersystem calculation in a very different manner than it does the various subsystem calculations. In hindsight, this is unsurprising, insofar as BSSE stems from “borrowing thy neighbor’s basis functions” and there are simply fewer neighbors in the subsystem calculations. In the absence of CP corrections, it is therefore unclear whether  $n$ -body results should be compared, order-by-order, with a supersystem calculation. As such, our opinion of what constitutes an appropriate benchmark has evolved over time, and we now suggest that the most appropriate benchmark is to compare a CP-corrected supersystem calculation to a CP-corrected  $n$ -body calculation, each at the same level of theory.

The present work draws on Papers I and II in this series that documented precision problems<sup>27</sup> and accuracy problems<sup>28</sup> with the MBE. Here, we attempt to bring this discussion to a close by dealing with both issues. Accuracy is assessed in terms of CP-corrected calculations, and we extend the  $n$ -body approximation as far as required in order to obtain results of acceptable accuracy. Regarding what is “acceptable,” Ouyang and Bettens<sup>41</sup> note that for molecular dynamics applications at room temperature, each atom has  $(3/2)k_B T \approx 0.9$  kcal/mol of thermal energy, hence it makes little sense in that context to demand that single-point energies be orders-of-magnitude more accurate than this value. A “dynamic accuracy” criterion of  $0.1 \times (3/2)k_B \times (298 \text{ K}) = 0.09$  kcal/mol per fragment was suggested in Ref. 41, and we adopt this as our target accuracy. This level of accuracy will ultimately require four-body calculations, for which precision problems manifest for  $N \gtrsim 30$  unless thresholds are set tight enough to significantly slow down performance.<sup>28</sup>

To put this in perspective, a complete four-body calculation on the largest system considered here,  $(\text{H}_2\text{O})_{37}$ , consists of 74 518 distinct subsystems including 66 045 tetramers. At the  $\omega\text{B97X-V/aTZ}$  level that is used herein to examine relative energies of cluster isomers, the use of “tight” versus “loose” thresholds<sup>45</sup> (as defined in Ref. 28) increases the computation time for each water tetramer (368 basis functions) by a factor of two when running on a single processor. Precision problems can be circumvented, and the entire calculation significantly streamlined, by the introduction of thresholds for neglecting subsystem calculations that are unlikely to contribute significantly. This possibility, and the limits of its accuracy, is explored in the current work.

## II. THEORY AND METHODS

### A. Many-body expansion

The MBE expresses the total energy for a system of  $N$  fragments as

$$E = \sum_{I=1}^N E_I + \sum_{I=1}^N \sum_{J<I} \Delta E_{IJ} + \sum_{I=1}^N \sum_{J<I} \sum_{K<J} \Delta E_{IJK} + \dots \quad (1)$$

The two- and three-body corrections are

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

$$\begin{aligned} \Delta E_{IJK} = & E_{IJK} - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{JK} \\ & - E_I - E_J - E_K. \end{aligned} \quad (2b)$$

An  $n$ -body approximation, which we will denote as  $\text{MBE}(n)$ , truncates Eq. (1) at terms involving  $n$  fragments. If taken literally, however, Eq. (1) involves some redundant calculations because, e.g., the monomer energy  $E_I$  appears in  $\Delta E_{IJ}$ ,  $\Delta E_{IJK}$ , etc. Non-redundant formulas with appropriate combinatorial coefficients, for  $\text{MBE}(n)$  with arbitrary  $n$ , can be found in Ref. 27.

## B. Counterpoise corrections

Define the interaction energy by removing the one-body contribution from the total energy,

$$E_{\text{int}} = E - \sum_{I=1}^N E_I. \quad (3)$$

The usual Boys-Bernardi CP correction for molecular dimers<sup>42</sup> resembles the two-body correction  $\Delta E_{IJ}$  performed in the dimer basis set. We might indicate this as

$$\Delta E_{IJ}^{\text{CP}} = E_{IJ}^{IJ} - E_I^{IJ} - E_J^{IJ}. \quad (4)$$

Following previous literature,<sup>28,38,46</sup> the subscripts denote real monomers (as above) whereas the superscripts denote where the basis functions are placed. Generalizing this to  $N$  monomers affords a generalization of the Boys-Bernardi idea,<sup>43,44</sup> and a CP-corrected interaction energy

$$E_{\text{int}}^{\text{CP}} = E_{IJK\dots N}^{IJK\dots N} - \sum_{I=1}^N E_I^{IJK\dots N}. \quad (5)$$

The quantity defined in this equation has been called the ‘‘site-site function counterpoise correction,’’<sup>43</sup> but we refer to it simply as the Boys-Bernardi CP correction, since it naturally generalizes the original dimer approach.<sup>42</sup>

The CP-corrected interaction energy in Eq. (5) can alternatively be expressed as

$$\begin{aligned} E_{\text{int}}^{\text{CP}} &= E_{IJK\dots N}^{IJK\dots N} - \sum_I E_I^I + \sum_I (E_I^I - E_I^{IJK\dots N}) \\ &= E_{\text{int}}^{\text{uncorr}} + \delta E^{\text{CP}}, \end{aligned} \quad (6)$$

where the ‘‘uncorrected’’ interaction energy is

$$E_{\text{int}}^{\text{uncorr}} = E_{IJK\dots N}^{IJK\dots N} - \sum_I E_I^I \quad (7)$$

and the CP correction is

$$\delta E^{\text{CP}} = \sum_I (E_I^I - E_I^{IJK\dots N}). \quad (8)$$

Equation (8) defines the  $N$ -body CP correction,<sup>43,44</sup> which has sometimes been criticized for its failure to account for ‘‘basis-set extension’’ effects,<sup>40,46,47</sup> although the good agreement between CP calculations and the alternative (and formally more complete) Valiron-Mayer function counterpoise corrections<sup>40,46</sup> suggests that any neglected effects are rather small.<sup>28,38</sup> The complete Valiron-Mayer approach also rapidly becomes intractable beyond just a few monomers. In view of this, we take Eq. (8) to define the counterpoise correction, in the spirit of Boys and Bernardi.<sup>43,44</sup> Even this procedure, however, requires  $N + 1$  calculations in the supersystem basis set, for a system of  $N$  fragments. Even more calculations are

required in the case of the generalized  $\text{MBE}$ ,<sup>23</sup> for which CP corrections have also been formulated.<sup>28</sup>

To circumvent this, and in view of Eq. (6), we approximate the CP-corrected total energy through a standard  $\text{MBE}(n)$  calculation applied to the supersystem energy  $E_{IJK\dots N}^{IJK\dots N}$  in Eq. (7) in conjunction with an  $n$ -body approximation to the summand in Eq. (8). We call this a many-body counterpoise (MBCP) correction,<sup>33,38</sup> truncated at order  $n$ , or  $\text{MBCP}(n)$  for short. Formulas for  $\delta E_I^{\text{MBCP}(n)}$ , which is the  $n$ -body approximation to the  $I$ th summand in Eq. (8), were derived previously through  $n = 4$ .<sup>33,38</sup> The two leading terms are

$$\delta E_I^{\text{MBCP}(2)} = (N - 1)E_I^I - \sum_{J \neq I}^N E_I^{IJ} \quad (9)$$

and

$$\begin{aligned} \delta E_I^{\text{MBCP}(3)} &= \delta E_I^{\text{MBCP}(2)} - \frac{1}{2}(N - 2)(N - 1)E_I^I \\ &\quad + (N - 2) \sum_{J \neq I}^N E_I^{IJ} - \sum_{J \neq I}^N \sum_{\substack{K > I \\ K \neq J}}^N E_I^{IJK}. \end{aligned} \quad (10)$$

Summing Eq. (9) and/or (10) over all monomers  $I$  affords the  $\text{MBCP}(n)$  approximation, for  $n = 2$  or 3.

Our original idea<sup>38</sup> was to combine the  $\text{MBE}(n)$  approximation for the supersystem energy  $E_{IJK\dots N}^{IJK\dots N}$  with the  $\text{MBCP}(n)$  approximation for the CP corrections  $E_I^{IJK\dots N}$ , for a consistent order-by-order truncation to the CP-corrected interaction energy. In practice, however, we find that the  $\text{MBCP}(n)$  corrections are quite small for  $n > 2$  (see Table S1 in the [supplementary material](#)). In the present work, we therefore include only the  $\text{MBCP}(2)$  correction.

An alternative way to interpret BSSE was introduced by Valiron and Mayer<sup>46</sup> and later adopted by others.<sup>39–41</sup> Within this formulation, one writes

$$E_{\text{int}}^{\text{CP}} = \sum_{IJ} \Delta E_{IJ}^{IJK\dots N} + \sum_{IJK} \Delta E_{IJK}^{IJK\dots N} + \dots \quad (11)$$

and then imagines that the total BSSE arises from two contributions: basis-set imbalance error (BSIE) and basis-set extension error (BSEE). This terminology, as well as arguments about whether BSEE is neglected by the Boys-Bernardi CP correction, has existed for a long time,<sup>47</sup> but in our opinion the distinction between the two effects is ambiguous and ill-defined. A recent attempt to distinguish the two effects, within the context of the  $\text{MBE}$ , can be found in Ref. 41, where it is stated that BSIE originates in the unbalanced comparison of  $n$ -body results, computed using subsystem basis sets, to supersystem results computed using the supersystem basis set. BSEE, according to this analysis, arises because subsystem calculations are stabilized by basis functions on nearby monomers. The latter ‘‘is important as these extension effects improve the quality of the total energy or binding energy by maximizing the flexibility of the wave function at the given basis set.’’<sup>41</sup> However, the quality of the subsystem calculations *also* improves if they are performed using the supersystem basis set, so it seems to us that BSIE and BSEE are inextricably entangled.

With Eq. (11) in mind, however, Ouyang and Bettens<sup>41</sup> introduced a CP scheme that is formally more general than our  $\text{MBCP}(n)$  approach, and in particular conforms more

closely to the Valiron-Mayer idea.<sup>40,46</sup> Nevertheless, our MBCP( $n$ ) approach is recovered as a low-order approximation to their “many-ghost, many-body expansion,” and it is found that MBCP(2) is sufficient to converge the CP correction,<sup>41</sup> as we have already suggested above. This provides further justification for the approximate CP correction employed here.

### C. Cost-reduction strategies

Reducing the number of subsystem calculations is crucial for obtaining good efficiency. One “dirty secret” of fragment-based approaches is that often quite large system sizes are required before the total computational time (measured in processor-hours) is actually less than the cost of the super-system calculation.<sup>14,28</sup> This is especially true when CP corrections are introduced, as these require a very large number of additional calculations.<sup>28</sup> It is true that the *wall time* (or time-to-solution) of the fragment-based calculation can be dramatically reduced via parallelization, although methods that rely on self-consistent updating of embedding charges will suffer some reduction in parallel scalability. Thresholds designed to eliminate unimportant subsystem calculations *a priori* not only reduce the cost, but by significantly reducing the number of subsystems they can also reduce finite-precision problems.

#### 1. Distance-based thresholds

We examine smooth distance-based cutoffs to discard some of the subsystems, based on a switching function

$$f(x) = \begin{cases} 1, & \text{if } x < 0 \\ 1 - x^3(10 - 15x + 6x^2), & \text{if } 0 \leq x \leq 1. \\ 0, & \text{if } x > 1 \end{cases} \quad (12)$$

Let  $R_{\max}$  denote the largest inter-fragment distance within a particular subsystem, measured in the present work in terms of the fragment centers of mass. (For fragments significantly larger than H<sub>2</sub>O, inter-fragment atom–atom distances are likely a better choice for  $R_{\max}$ , but the choice makes little difference here.) The cutoff procedure is characterized by two parameters:  $R_{\text{cut}1}$ , the distance for the onset of the threshold, and  $w$ , which indicates the width of the switching region or in other words how quickly  $f(x)$  switches between 0 and 1. Given these two parameters, we take

$$x = (R_{\max} - R_{\text{cut}1})/w \quad (13)$$

in Eq. (12). If  $R_{\max} \geq R_{\text{cut}1} + w$ , then  $f(x) = 0$  and the subsystem in question is neglected. [One could imagine adopting some small but non-zero drop tolerance for  $f(x)$ , say, on the order of the integral drop tolerance, but we have not done so here and do not expect that it would make much difference in clusters of this size.] For subsystems with  $R_{\max} < R_{\text{cut}1} + w$ , the energy is computed and then scaled by  $f(x)$  for use in the MBE. Each fragment in this work consists of a single H<sub>2</sub>O molecule and we will test various combinations of  $R_{\text{cut}1}$  and  $w$ . For brevity in the discussion that follows, we will use the notation  $(n_r, n_w)$  to indicate particular choices of the thresholds, where  $n_r$  and  $n_w$  are a pair of integers that specify the values of  $R_{\text{cut}1}$  and  $w$ , respectively, in Ångstroms.

Recent MBE calculations on alanine polypeptides have demonstrated that distance-based screening alone may artificially exclude certain important subsystems, namely, those characterized by a cooperative arrangement of dipole moments across length scales longer than the cutoff distance.<sup>21</sup> To account for this, we introduce a second distance parameter  $R_{\text{cut}2} < R_{\text{cut}1}$ , in the spirit of the connectivity-based analysis in Ref. 21. To understand the role of this second cutoff, consider the “connectivity diagrams” of trimers and tetramers that are illustrated in Fig. 1. In these diagrams, we connect with a line any pair of fragments that are separated by a distance less than  $R_{\text{cut}2}$ , whereas we imagine that disconnected fragments are separated by more than  $R_{\text{cut}1}$  and therefore these configurations are excluded by the  $R_{\text{cut}1}$  cutoff.

Figures 1(a)–1(c) exhaust all of the possible topologies for trimers, and Figs. 1(d)–1(i) show all possibilities for tetramers. Note that configurations (c) and (i), in which all inter-fragment distances are less than  $R_{\text{cut}2}$  (and therefore less than  $R_{\text{cut}1}$  as well), are always included and are shown only for completeness. In each of the remaining configurations, there is at least one inter-fragment distance greater than  $R_{\text{cut}1}$ , so each is excluded by the cutoff procedure of Eqs. (12) and (13).

Examining the trimer configurations, we wish to exclude configuration (a), which consists of a dimer of fragments plus a well-separated monomer, while retaining configuration (b), which might exhibit an energetically important chain-of-dipoles interaction but is excluded by the  $R_{\text{cut}1}$  cutoff procedure on the basis of its end-to-end distance. For the tetrameric cases, the configurations in Figs. 1(d) and 1(e) consist of strongly interacting dimers or trimers plus another

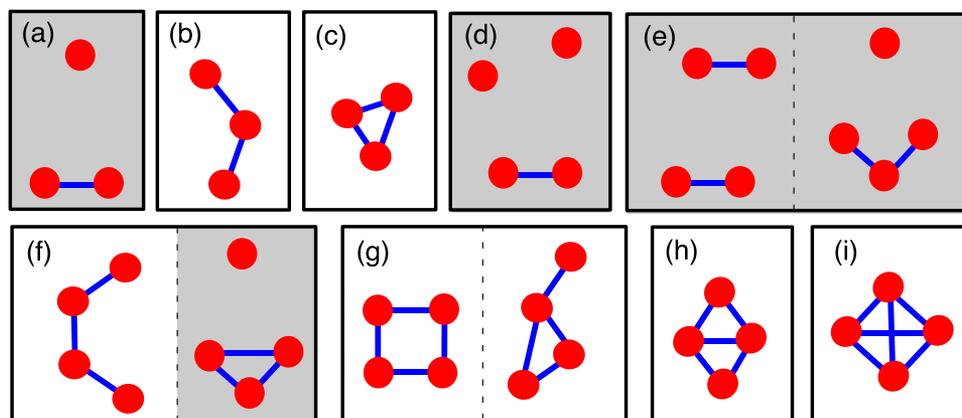


FIG. 1. Pictorial representations of all possible “connectivities” for trimers and tetramers of fragments (red circles). Solid blue lines indicate inter-fragment distances less than  $R_{\text{cut}2}$ , and the various cases are grouped according to how many of these there are. Fragments not connected by blue lines are farther apart than  $R_{\text{cut}1}$ . Trimers (a) and (b), and tetramers (d)–(h), are excluded by the  $R_{\text{cut}1}$  threshold [Eqs. (12) and (13)], but when we additionally employ the  $R_{\text{cut}2}$  criterion only configurations in the shaded boxes are excluded.

weakly interacting dimer or monomer(s). Since the strongly interacting dimers and trimers are already included in the two- and three-body calculations, respectively, we expect configurations (d) and (e) to make only minor contributions at the four-body level. As such, in these proof-of-concept calculations, we will use the  $R_{\text{cut}2}$  threshold to retain tetramers otherwise excluded by  $R_{\text{cut}1}$  only if they exhibit four or more inter-fragment distances less than  $R_{\text{cut}2}$ . This excludes the cases in Figs. 1(d) and 1(e), as well as one of the cases shown in Fig. 1(f). The  $R_{\text{cut}2}$  threshold, introduced in Ref. 21, has not yet been implemented in a smooth way nor will we attempt to do so now. Rather, we merely present results with  $R_{\text{cut}2}$  in order to compare the accuracy against those obtained using the smooth  $R_{\text{cut}1}$  threshold alone.

## 2. Energy-based thresholds

For systems where the individual fragments are substantially larger than  $\text{H}_2\text{O}$ , the distance-based thresholding discussed above may become less effective in reducing the number of subsystem calculations. Ouyang and Bettens<sup>21</sup> recently described an elegant, energy-based thresholding procedure in which classical multipole interactions are used as *a priori* estimates of the magnitude of higher-order terms in the MBE. (The monomer multipoles are available from the one-body calculations.) Trimers with classical interaction energies smaller than 0.25 kJ/mol, and tetramers with classical interactions <0.1 kJ/mol, were excluded from quantum calculations at the MBE(3) and MBE(4) levels, respectively. This procedure is quite new and has yet to be implemented in our code nor has it been implemented anywhere in conjunction with smoothing functions. Nevertheless, we can estimate its effectiveness after-the-fact by first computing all subsystem energies at the quantum level then using those results to discard certain subsystems according to the aforementioned energetic criteria.

## 3. Multi-level approaches

As compared to simply dropping well-separated subsystems outright, a more sophisticated approach might treat these small contributions to the MBE at a lower level of theory. We test two different approaches for doing so, taking the lower-level theory to be Hartree-Fock (HF) theory in either case. In the first scheme, we smoothly turn on a HF calculation using the switching function  $1 - f(x)$ , as the higher-level DFT method is turned off using the function  $f(x)$  [see Eq. (12)]. For any particular subsystem, the energy formula that is used is

$$E_{\text{subsys}} = f(x)E_{\text{subsys}}^{\text{DFT}} + [1 - f(x)]E_{\text{subsys}}^{\text{HF}}. \quad (14)$$

For subsystems that exist in the switching region, meaning that  $R_{\text{cut}1} \leq R_{\text{max}} \leq R_{\text{cut}1} + w$ , it is necessary to perform both the HF and the DFT calculations.

The second approach is an ONIOM-type formalism,<sup>48</sup> inspired by the fragment-based methods introduced by Raghavachari and co-workers,<sup>13,16,17,49,50</sup> who use a supersystem calculation performed at an inexpensive level of theory in order to capture long-range induction effects that would otherwise be omitted in a low-order  $n$ -body calculation. This is an alternative way to account for the cooperative, long-range

arrangements of fragment dipole moments. The subsystem energy formula used in this case is

$$E_{\text{subsys}} = (E_{\text{subsys}}^{\text{DFT}} - E_{\text{subsys}}^{\text{HF}})f(x) + E_{\text{supersys}}^{\text{HF}}. \quad (15)$$

Considering all subsystems, the terms  $E_{\text{subsys}}^{\text{DFT}}f(x)$  together constitute an  $n$ -body DFT calculation with smooth cutoffs, and subtracting  $E_{\text{subsys}}^{\text{HF}}f(x)$  prevents double-counting of the low-level calculations on the “model system” (to use ONIOM terminology<sup>48</sup>) in the presence of a low-level calculation  $E_{\text{supersys}}^{\text{HF}}$  on the “real system.” Note that the supersystem term in Eq. (15) is the same for each subsystem, so need only be computed once.

## III. RESULTS AND DISCUSSION

### A. Computational details

In Sec. III B, we examine how distance-based thresholds affect the accuracy of interaction energies computed for a sequence of water clusters,  $(\text{H}_2\text{O})_{N=6-37}$ . These structures were originally obtained from Ref. 51, where they were put forward as putative global minima (at each cluster size) on the TIP4P potential surface. They are used here without further optimization.

We use the affordable B3LYP/aug-cc-pVDZ (B3LYP/aDZ) level of theory, with an SCF convergence threshold  $\tau_{\text{SCF}} = 10^{-7}$  a.u. and a drop tolerance  $\tau_{\text{ints}} = 10^{-14}$  a.u. These are “tight” convergence thresholds, as defined in Paper II,<sup>28</sup> whereas looser thresholds may lead to precision problems in the MBE.<sup>27</sup> Both thresholds, especially  $\tau_{\text{ints}}$ , are significantly tighter than the default settings in common electronic structure programs.

In Sec. III B, we examine relative energies of four different structural motifs of  $(\text{H}_2\text{O})_{20}$ . These structures have also been considered in previous work on the MBE<sup>33</sup> and are taken from the extensive basin-hopping Monte Carlo search in Ref. 52. For these calculations, we employ a higher-quality level of theory, namely,  $\omega\text{B97X-V}^{53}/\text{aug-cc-pVTZ}$  ( $\omega\text{B97X-V/aTZ}$ ), with  $\tau_{\text{SCF}}$  and  $\tau_{\text{ints}}$  as above.

The SG-1 quadrature grid<sup>54</sup> is used for all calculations, as higher-quality grids have been examined and found to make little difference in the context of the MBE.<sup>27</sup> All calculations were performed using Q-CHEM, v. 4.2.<sup>55</sup>

### B. Interaction energies

Except where otherwise specified, in what follows we define the error in an  $n$ -body approximation to the interaction energy according to

$$\text{error} = E_{\text{int}}(n\text{-body}) - E_{\text{int}}(\text{supersystem}), \quad (16)$$

where one or both energies may be CP corrected, depending on the context. [For total interaction energies we report errors in size-intensive, per-monomer units, but Eq. (16) fixes the convention for the sign of the errors.] Errors will be compared to the dynamic accuracy threshold discussed above,<sup>41</sup> i.e., 10% of  $(3/2)k_B T$  per monomer at  $T = 298$  K, or in other words 0.09 kcal/mol/monomer.

Data comparing the full CP correction at the B3LYP/aDZ level versus its MPCP(2) approximation are shown for

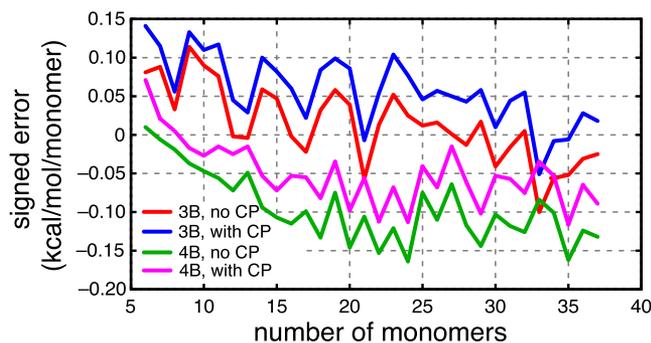


FIG. 2. Signed errors per monomer in three- and four-body total interaction energies for clusters  $(\text{H}_2\text{O})_{6-37}$ , at the B3LYP/aDZ level of theory. The  $n$ -body calculations labeled “no CP” are computed without MBCP corrections and compared to uncorrected supersystem energies, whereas those labeled “with CP” include MBCP(2) corrections and are compared to supersystem energies that include the full Boys-Bernardi CP correction.

$(\text{H}_2\text{O})_{N=6-37}$  in Table S1 of the [supplementary material](#). Differences between  $\delta E^{\text{CP}}$  and its MBCP(2) approximation are smaller than 0.07 kcal/mol/monomer across the whole data set, with an average error of 0.04 kcal/mol/monomer. This is consistent with other results demonstrating that the higher-order MBCP( $n$ ) corrections are small.<sup>41</sup> As such, we will limit the CP corrections to MBCP(2) in what follows, despite our original intention of using a consistent MBE( $n$ )+MBCP( $n$ ) approximation to  $E_{\text{int}}^{\text{CP}}$ .

### 1. Role of CP correction

In Figs. 2 and 3, we examine size-dependent errors in MBE(3) and MBE(4) results and their MBCP(2)-corrected counterparts, in two different ways. In Fig. 2, the uncorrected MBE( $n$ ) results are compared to uncorrected supersystem interaction energies (i.e., none of the calculations includes any CP correction), whereas the MBCP(2)+MBE( $n$ ) results are compared to supersystem interaction energies that include the full Boys-Bernardi CP correction, i.e.,  $\delta E^{\text{CP}}$  in Eq. (8). Figure 3 compares both MBE( $n$ ) and MBCP(2)+MBE( $n$ ) results to supersystem energies that include  $\delta E^{\text{CP}}$ .

At the three-body level, errors are somewhat smaller when we ignore the issue of BSSE altogether, but grow larger when we attempt to account for it, whereas the opposite is true at

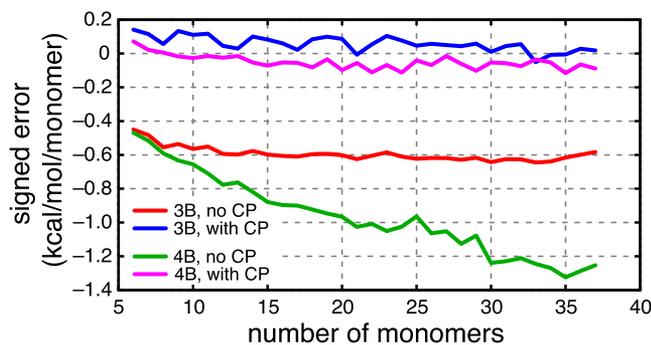


FIG. 3. Signed errors per monomer in three- and four-body total interaction energies for clusters  $(\text{H}_2\text{O})_{6-37}$ , at the B3LYP/aDZ level of theory. All of the  $n$ -body calculations, whether CP-corrected or not, are compared to supersystem calculations that include the full Boys-Bernardi CP correction.

the four-body level. These observations make sense in light of two facts: first, BSSE is always overstabilizing; and second, for water clusters, the non-pairwise terms often constitute stabilizing many-body induction effects. As such, the uncorrected MBE(3) results benefit from some error cancellation wherein neglected four-body terms are partially offset by BSSE, as observed in our previous work exploring extrapolations to the basis-set limit.<sup>33</sup> Note that the error in the CP-corrected interaction energy is

$$\text{error}(\text{CP}) = E_{\text{int}}^{\text{CP}} - \left( E_{\text{IJK}\dots\text{N}}^{\text{IJK}\dots\text{N}} - \sum_I E_I^{\text{IJK}\dots\text{N}} \right), \quad (17)$$

whereas the error in the uncorrected case is

$$\text{error}(\text{uncorr}) = E_{\text{int}}^{\text{uncorr}} - \left( E_{\text{IJK}\dots\text{N}}^{\text{IJK}\dots\text{N}} - \sum_I E_I^{\text{IJK}\dots\text{N}} \right). \quad (18)$$

Adding  $\delta E^{\text{CP}}$ , as defined in Eq. (8), to Eq. (17) results in precisely the right side of Eq. (18), which shows that the two definitions of error in Eqs. (17) and (18) are simply offset by the magnitude of the CP correction.

Examining Fig. 3, where all of the supersystem calculations include the full Boys-Bernardi correction and should therefore represent our best (or at least, most complete) benchmarks, we see that only the MBCP(2)-corrected  $n$ -body results are acceptable and lie essentially within our target accuracy of 0.09 kcal/mol/monomer for both  $n = 3$  and  $n = 4$ . Uncorrected MBE(3) results do not afford the target accuracy, and give rise to an error of  $\approx -0.6$  kcal/mol/monomer that is roughly constant as a function of cluster size. Errors in uncorrected MBE(4) results actually become larger as the cluster size increases.

### 2. Effects of cutoffs

To obtain a decent guess as to what might constitute a reasonable distance cutoff  $R_{\text{cut1}}$ , we examine the convergence of the total interaction energies at the MBE(4) level for  $(\text{H}_2\text{O})_{6-37}$ , in Fig. 4. These particular data do not apply any smoothing function but instead use a sharp drop criterion as a function of distance. A 6 Å cutoff recovers 97% of the total interaction energies, so in the interest of erring on the conservative side, we take this as our minimum value of  $R_{\text{cut1}}$  and also examine

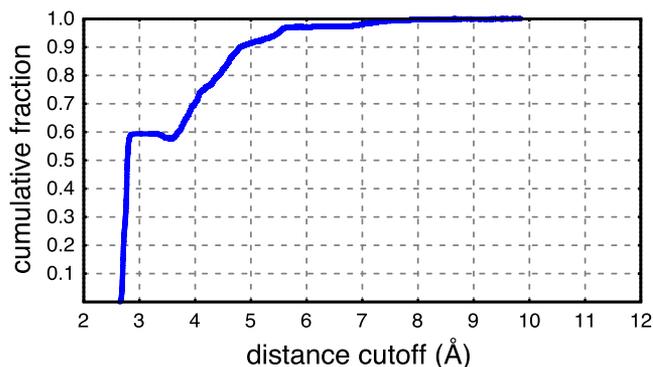


FIG. 4. Fraction of the cumulative interaction energy for water clusters (B3LYP/aDZ level) that is recovered by a four-body approximation, as a function of a sharp distance cutoff for the subsystem calculations. Interaction energies are not CP corrected, and the data point at each distance represents an average over cluster sizes from  $N = 6-37$ .

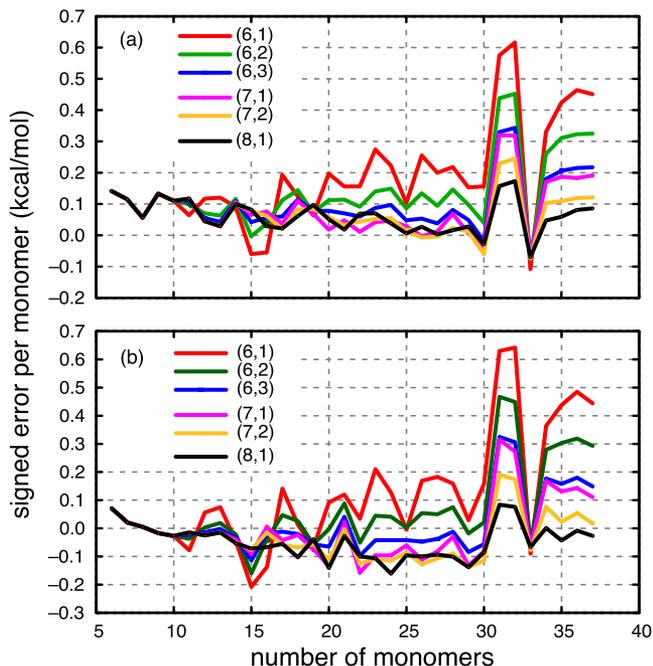


FIG. 5. Signed errors per monomer in CP-corrected (a) three-body and (b) four-body approximations to the total interaction energy for a sequence of water clusters, employing different values for the switching function parameters  $(n_r, n_w)$ . Subsystem calculations include the MBCP(2) counterpoise correction [Eq. (9)] but are compared to supersystem results including the full counterpoise correction [Eq. (5)].

$R_{\text{cut}1} = 7$  and  $8 \text{ \AA}$  along with  $w = 1, 2$ , and  $3 \text{ \AA}$ . [In the  $(n_r, n_w)$  notation introduced above, this means  $n_r = 6, 7$ , or  $8$  and  $n_w = 1, 2$ , or  $3$ .] Errors as a function of cluster size are plotted in Fig. 5, for both three- and four-body expansions. All calculations are CP-corrected.

It is obvious that neither MBE(3) nor MBE(4) has converged to the target accuracy until the cutoffs are pushed to  $(n_r, n_w) = (8, 1)$ , although  $(7, 2)$  comes close. Note that it is not easy to draw a direct connection between the choice of  $(n_r, n_w)$  and the number of subsystems that will be included in the calculation. For instance, in the present examples, the  $(6, 3)$ ,  $(7, 2)$ , and  $(8, 1)$  combinations involve the same subsystems but different values of  $f(x)$ , so the accuracy of each scheme is a bit different. Nevertheless, there is a clear trend in Fig. 5 that errors are reduced as we progress from  $(6, 3) \rightarrow (7, 2) \rightarrow (8, 1)$  thresholding, leading us to conclude that subsystems with inter-fragment distances in the range of  $6\text{--}9 \text{ \AA}$  are important in providing long-range stabilization.

Notice from Fig. 5 that errors are larger for the cluster sizes  $N = 31, 32$ , and  $34\text{--}37$ , anomalies that may result from a qualitative structural transition that occurs between  $N = 30$  and  $31$ , where the structures transition to large cages with cubic structures rather than pentaprismatic structures.<sup>56</sup> (Recall that our cluster structures are putative global minima at each value of  $N$ .<sup>51</sup>) In view of recent work by Ouyang and Bettens aimed at identifying important many-body interactions in polypeptides,<sup>21</sup> it may be the case that the sort of cooperative, chain-like interactions amongst fragment dipole moments that were identified in Ref. 21 are more important for the qualitatively different structures at  $N > 30$  than they are for the slightly smaller  $N \leq 30$  structures. To investigate this

TABLE I. Error statistics (maximum error and mean unsigned errors) for CP-corrected MBE(4) approximations to  $E_{\text{int}}^{\text{CP}}$ , using various thresholds  $(n_r, n_w)$ , in conjunction with the  $R_{\text{cut}2}$  threshold,  $R_{\text{cut}2} < R_{\text{cut}1}$ . Statistics include all  $(\text{H}_2\text{O})_N$  clusters,  $N = 6\text{--}37$ .

$R_{\text{cut}2}$ ( $\text{\AA}$ )	Error (kcal/mol/monomer)					
	(6,1)		(7,1)		(8,1)	
	Maximum	MUE	Maximum	MUE	Maximum	MUE
4	0.64	0.17	0.31	0.08	0.16	0.06
5	0.54	0.12	0.30	0.09	0.16	0.06
6	0.40	0.10	0.24	0.08	0.17	0.06
7	...	...	0.17	0.07	0.15	0.06
8	...	...	...	...	0.14	0.06

possibility, we introduce the second threshold parameter  $R_{\text{cut}2}$ , as discussed in Sec. II C 1. Error statistics employing both  $R_{\text{cut}1}$  and  $R_{\text{cut}2}$  are summarized in Table I. For the  $(n_r, n_w) = (7, 1)$  and  $(8, 1)$  schemes, errors converge by  $R_{\text{cut}2} = 7 \text{ \AA}$ , and they converge to values not worse than what we encountered prior to introducing  $R_{\text{cut}2}$  (see Table I).

Figure 6 plots the signed errors for three- and four-body approximations using the  $(n_r, n_w) = (6, 1)$ ,  $(7, 1)$ , and  $(8, 1)$  schemes but this time with  $R_{\text{cut}2} = 7 \text{ \AA}$ . At the three-body level, the errors are reduced for the  $(6, 1)$  and  $(7, 1)$  schemes as compared to results where the  $R_{\text{cut}2}$  threshold is absent. At the four-body level,  $(7, 1)$  results with  $R_{\text{cut}2} = 7 \text{ \AA}$  are close to the target accuracy of  $0.09 \text{ kcal/mol/monomer}$ . Therefore, in Fig. 7, we examine  $(7, 1)$  results with  $R_{\text{cut}2} = 7 \text{ \AA}$  more closely, plotting them alongside results obtained with no cutoffs whatsoever, as a reference, and also results in which energy rather than distance cutoffs are employed. Following the recommendation in Ref. 21, for the energy-based scheme, we discard all trimers whose interaction energies are  $< 0.25 \text{ kJ/mol}$  and all tetramers whose interaction energies are  $< 0.10 \text{ kJ/mol}$ . (The energy-based scheme retains all dimers, whereas the distance-based scheme discards sufficiently distant dimers.) Results demonstrate that both cutoff strategies faithfully track the reference calculations, at both the three- and four-body levels. Absolute errors, with respect to a CP-corrected supersystem calculation, are not much larger than  $0.2 \text{ kcal/mol/monomer}$  for any of the clusters examined here.

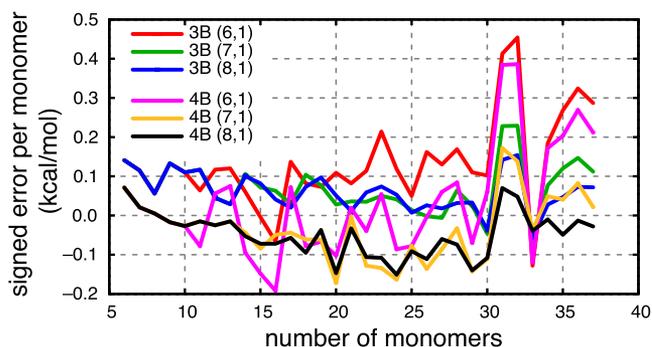


FIG. 6. Signed errors per monomer for three- and four-body approximations to the total interaction energies, in conjunction with MBCP(2) counterpoise corrections, for  $(\text{H}_2\text{O})_N$  clusters. Various  $(n_r, n_w)$  combinations are used, with  $R_{\text{cut}2} = 7 \text{ \AA}$  in each case.

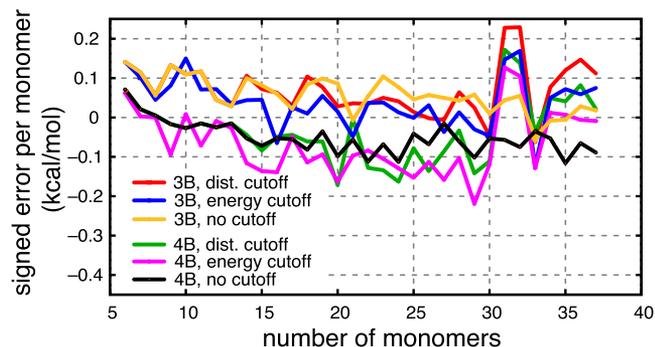


FIG. 7. Signed errors per monomer in three- and four-body approximations to the total interaction for  $(\text{H}_2\text{O})_{6-37}$ , including MBCP(2) corrections. The “distance cutoff” results use the thresholds  $(n_r, n_w) = (7, 1)$  along with  $R_{\text{cut}2} = 7 \text{ \AA}$ . The “energy cutoff” results do not employ any distance-based thresholding but discard all trimers whose interaction energies are  $< 0.25 \text{ kJ/mol}$  and all tetramers whose interaction energies are  $< 0.10 \text{ kJ/mol}$ .

### C. Relative energies

We next examine three- and four-body expansions as applied to predicting relative energies of  $(\text{H}_2\text{O})_{20}$  isomers. Cluster geometries, consisting of twenty low-energy isomers each from the four families of isomers on the  $(\text{H}_2\text{O})_{20}$  potential surface, were taken from Ref. 52 without further optimization. These structures have been used by us in previous work,<sup>28,33</sup> and examples of the four classes of isomers are depicted in Fig. 8. Benchmark energies were computed at the CP-corrected  $\omega\text{B97X-V/aTZ}$  level, and error with respect to these benchmarks is defined as

$$\text{error} = E_{\text{rel}}^{n\text{-body}} - E_{\text{rel}}^{\text{supersys}}. \quad (19)$$

Both energies in Eq. (19) are CP-corrected, using MBCP(2) in the  $n$ -body case and a full Boys-Bernardi correction in the supersystem case. Our target accuracy for these calculations is “chemical accuracy” of 1 kcal/mol with respect to a supersystem calculation performed using the same density functional and basis set.

Errors in the relative isomer energies are plotted in Fig. 9, using  $(n_r, n_w)$  cutoffs but not the  $R_{\text{cut}2}$  threshold, as the

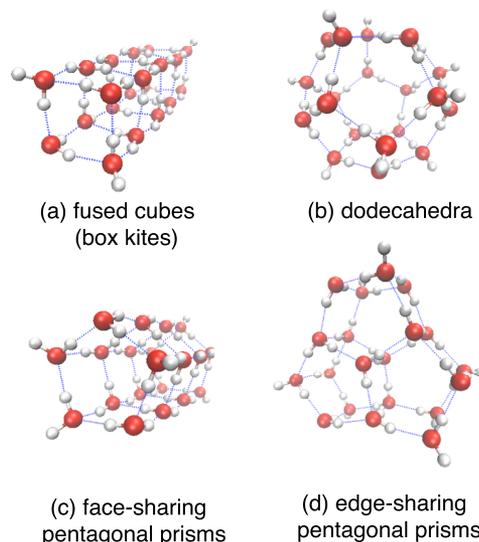


FIG. 8. Examples of the four families of  $(\text{H}_2\text{O})_{20}$  isomers.

latter only becomes important in larger clusters. To achieve the target accuracy of 1 kcal/mol requires the use of our most conservative thresholding strategy,  $(n_r, n_w) = (8, 1)$ , in which case there are only three isomers out of 80 where the error exceeds 1 kcal/mol. A detailed examination (see Table S2 in the supplementary material) reveals that, within the isomers belonging to a given family, these three outliers exhibit the largest stabilization energies arising from sub-clusters separated by 8–9 Å. The contrast is especially apparent for isomers 10 and 18 of the edge-sharing pentagonal prism motif, where the 8–9 Å sub-clusters contribute  $-4.42$  and  $-4.11 \text{ kcal/mol}$ , respectively, to the total interaction energy, whereas this value does not exceed  $-0.45 \text{ kcal/mol}$  for any other isomer in this family, and in a few cases, it is actually repulsive. (The difference lies primarily in the arrangement of monomer dipole moments, which in the case of isomers 10 and 18 makes all of the two-body interactions attractive, whereas for other isomers about half of the two-body interactions in the 8–9 Å range are repulsive.) The contrast is not quite as stark in the case of fused-cube isomer

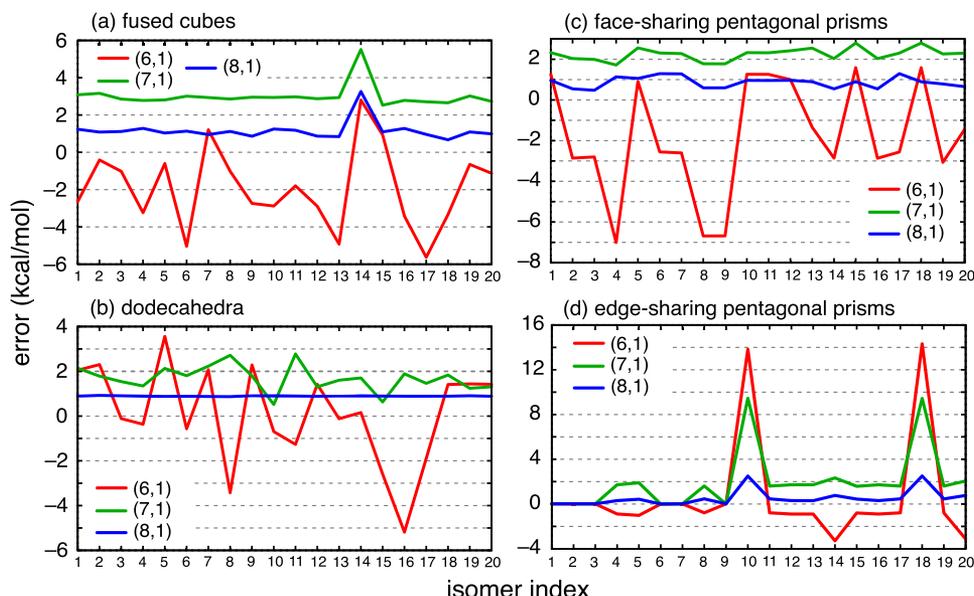
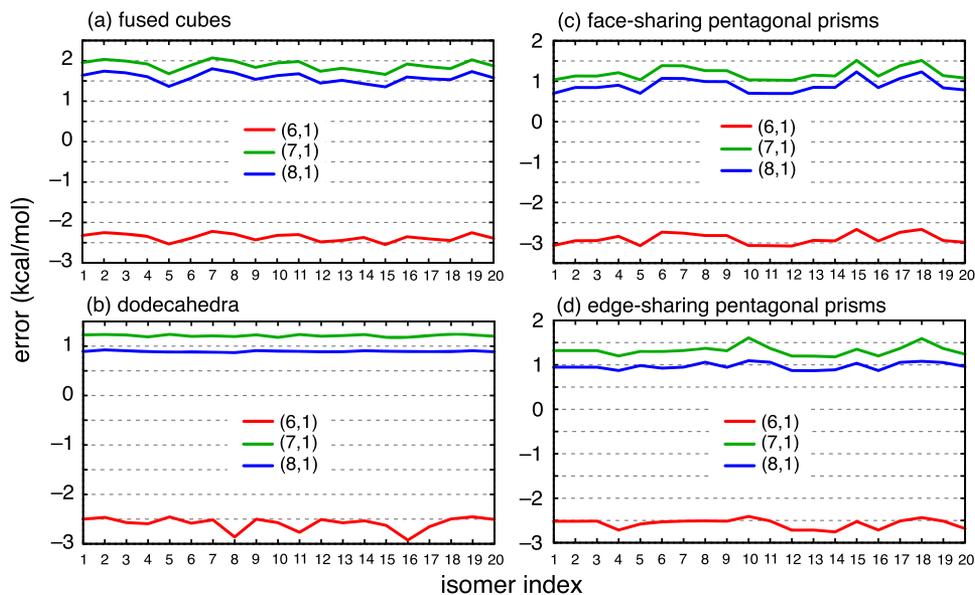


FIG. 9. Signed errors for relative energies of  $(\text{H}_2\text{O})_{20}$  cluster isomers, employing MBE(4)+MBCP(2) and various  $(n_r, n_w)$  thresholds. Energies were computed at the  $\omega\text{B97X-V/aTZ}$  level. Each panel presents data for a different family of isomers (see Fig. 8), but all 80 isomers are plotted on a common energy scale even though the vertical axes differ between panels.



14, although the 8–9 Å interactions are still  $\approx 1$  kcal/mol more stabilizing than for any of the other fused-cube isomers. For the other two families of isomers there are no such outliers, and as such the results with (8,1) thresholds are more consistent in these cases.

These  $(\text{H}_2\text{O})_{20}$  clusters are too small to benefit from the alternative  $R_{\text{cut}2}$  threshold introduced above, so to improve the results we turn to two other *ad hoc* strategies described in Sec. II C 3. The first approach gradually turns on a HF/aTZ calculation at long range, as the switching function is turning off the DFT calculation; see Eq. (14). Results in Fig. 10 using (8,1) thresholds show that the relative energies are more consistent across isomers than when the long-range interactions are simply neglected, although for the fused-cube isomers the errors are  $\approx 0.5$  kcal/mol greater, even while the aforementioned outlier is eliminated. Nevertheless, this hybrid scheme comes close to achieving the desired accuracy of 1 kcal/mol, at least with (8,1) thresholds. For (7,1) thresholds, the errors

remain fairly consistent across isomers but are increased to  $\sim 1.5$  kcal/mol. Errors for the (6,1) scheme are clearly unacceptable.

As an alternative to low-level calculations of just the long-range subsystems, we also examine an ONIOM-type approach [Eq. (15)] using a DFT-based MBE as the high-level calculation ( $\omega\text{B97X-V/aTZ}$ ) and HF/aTZ as a low-level supersystem calculation. Results are shown in Fig. 11 and are extremely accurate in comparison to either of the previous two approaches. (Note the much smaller energy scale in Fig. 11 versus either of Fig. 9 or 10.) In this case, errors in relative energies do not exceed 0.5 kcal/mol, *even when (6,1) thresholding is employed*. This eliminates a great many subsystems as compared to (8,1) thresholds. For example, at the (6,1) level, we must retain 144, 536, and 1160 subsystems for  $n = 2, 3,$  and  $4$ , respectively, as compared to 190, 1,140, and 4845 subsystems when no thresholds are employed. For the (8,1) scheme, very few subsystems can be neglected in  $(\text{H}_2\text{O})_{20}$ . Granted, this

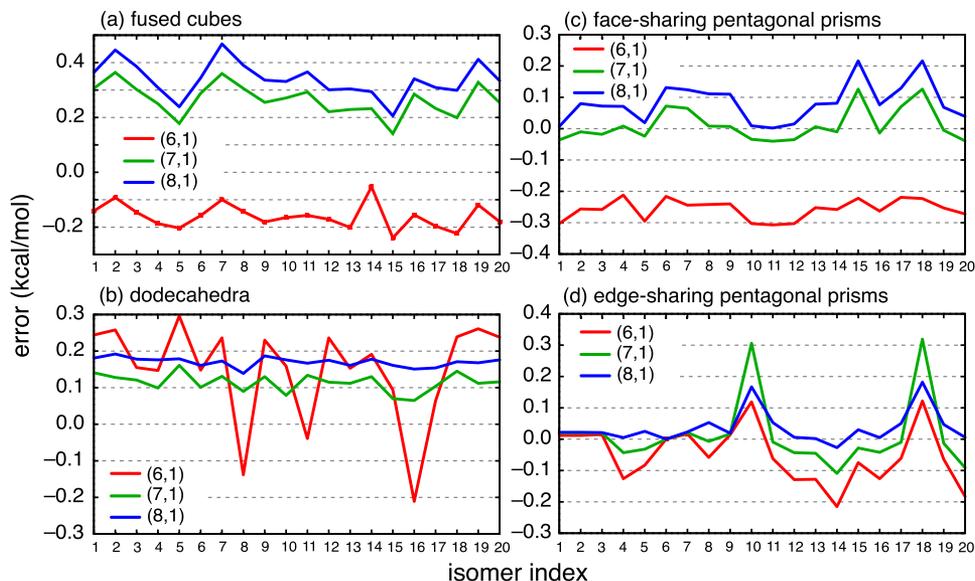


FIG. 11. Signed errors for relative energies of  $(\text{H}_2\text{O})_{20}$  cluster isomers, employing MBE(4)+MBCP(2) and various  $(n_r, n_w)$  thresholds. Energies up to the  $R_{\text{cut}1}$  cutoff were computed at the  $\omega\text{B97X-V/aTZ}$  level and corrected using a HF/aTZ calculation for the entire supersystem, using the ONIOM-style correction in Eq. (15). Each panel presents data for a different family of isomers (see Fig. 8), but all 80 isomers are plotted on a common energy scale even though the vertical axes differ between panels.

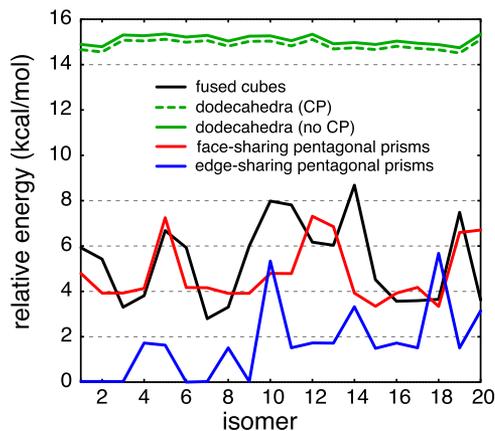


FIG. 12. Relative energies of twenty isomers from each of four motifs of  $(\text{H}_2\text{O})_{20}$ , computed at the  $\omega\text{B97X-V/aTZ}$  level using the (8,1) cutoff scheme. Except for the dodecahedral isomers, the difference between CP-corrected and uncorrected results is indistinguishable within the thickness of the lines.

reduction comes at the expense of introducing a single super-system calculation at the HF level, though as the high-level method becomes even more expensive—a correlated wave function calculation, for example, rather than DFT—the cost of the low-level supersystem calculation may not be so egregious. As such, this composite approach may have a useful domain of applicability, even if it becomes intractable as  $N \rightarrow \infty$ . (We return to this issue below, with timings.)

Finally, we revisit the relative energies of the  $(\text{H}_2\text{O})_{20}$  isomers examined in Ref. 33. New data at the  $\omega\text{B97X-V/aTZ}$  level are plotted in Fig. 12, using an (8,1) cutoff scheme. Although  $\delta E^{\text{CP}}$  is around 2.80 kcal/mol for the edge-sharing-pentagonal-prism, face-sharing-pentagonal-prism, and fused-cube isomers, this sizable correction is about the same for all isomers and the CP-corrected relative energies for these three families cannot be distinguished from the uncorrected energies. On the other hand,  $\delta E^{\text{CP}} \approx 2.55$  kcal/mol for dodecahedral isomers, so this correction matters at the level of  $\approx 0.25$  kcal/mol when trying to establish the energies of the dodecahedra relative to those of the other isomers. We observed the same phenomenon at the MP2 level in previous work,<sup>33</sup> namely, that CP correction matters only for predicting the energies of the dodecahedral isomers relative to those of the other three families. (It is also true that  $\delta E^{\text{CP}}$  was a bit larger than 1 kcal/mol in those previous calculations,<sup>33</sup> consistent with the observation that BSSE is typically larger in post-Hartree-Fock calculations as compared to DFT calculations.)

#### D. Computational cost

Our analysis suggests that fragments separated by 6–9 Å are indispensable in obtaining accurate total interaction energies. For distance-based thresholding, this places a fairly strong limit on the number of subsystems that can be discarded while maintaining faithful accuracy with respect to the supersystem calculation. For example, the number of subsystems that must be retained for  $(\text{H}_2\text{O})_{37}$ , using (7,1) thresholds with or without  $R_{\text{cut}2} = 7$  Å, is listed in Table II. The reduction is quite dramatic when  $R_{\text{cut}2}$  is *not* considered, but only moderate when it is. Similar trends are reflected in

TABLE II. Number of subsystems required for an MBE(4) calculation on the  $(\text{H}_2\text{O})_{37}$  cluster considered here, using the (7,1) thresholding scheme for  $R_{\text{cut}1}$  with and without  $R_{\text{cut}2} = 7$  Å. The number of subsystems required for energy-based thresholding ( $E_{\text{cut}}$ ) is also shown.

Subsystem	Full	$R_{\text{cut}1}$	$R_{\text{cut}1} + R_{\text{cut}2}$	$E_{\text{cut}}^a$
Monomers	37	37	37	37
Dimers	666	504	504	666
Trimers	7770	3751	5141	908
Tetramers	66045	17856	38278	999
Total	74518	22310	43923	2573

<sup>a</sup>The energy-based scheme does not cull monomers or dimers.

Fig. 13, which plots the fraction of the subsystems that are retained in the MBE(3) and MBE(4) approximations using various cutoffs, where the data are averaged over all water clusters  $(\text{H}_2\text{O})_{6-37}$ . For MBE(4) with (7, 1) thresholds, which was sufficient to obtain high-accuracy interaction energies for clusters with  $N \leq 30$ , more than 60% of the subsystems can be discarded, although this fraction drops to about 25% upon inclusion of the  $R_{\text{cut}2} = 7$  Å criterion that was necessary in larger clusters. Note that the fraction of subsystems that can be discarded will increase as system size grows.

The energy-based cutoff scheme is far more successful, essentially by construction, and eliminates 96.5% of the subsystem calculations as compared to an MBE(4) calculation with no cutoffs whatsoever. As compared to the (7,1) distance-based cutoff scheme, the energy-based scheme requires only 11.5% as many sub-cluster calculations. At present, our implementation of this approach is “cheating,” given that we have computed all of the sub-cluster energies *a priori* at the QM level and then thrown out the ones with sufficiently small interaction energies, *a posteriori*, but this suffices to demonstrate the promise of the energy-based approach. In Ref. 21, the energy-based scheme was introduced by Ouyang and Bettens based on classical multipole approximations to the sub-cluster energies, and it remains to implement a proper energy-based thresholding scheme using smooth cutoffs. Such efforts are underway in our group.

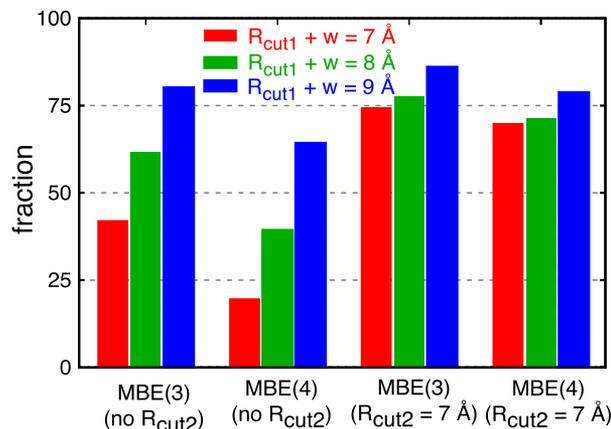


FIG. 13. Fraction of subsystem calculations required for various MBE( $n$ ) approximations and thresholding schemes, averaged over  $(\text{H}_2\text{O})_{N=6-37}$ . Note that any  $(n_r, n_w)$  combination with the same value of  $n_r + n_w$  results in the same subsystems, so we label the cutoffs in terms of  $R_{\text{cut}1} + w$ .

TABLE III. Timing data for MBE(4) calculations of  $(\text{H}_2\text{O})_{37}$  (without CP corrections) at the B3LYP/aDZ level using the (7,1) thresholding scheme for  $R_{\text{cut}1}$  with and without  $R_{\text{cut}2} = 7 \text{ \AA}$ . These are the same calculations as used to count the number of subsystems in Table II. Wall times reflect the cost to run on a single 28-core node,<sup>57</sup> so except for the supersystem calculation the wall time should decrease linearly with the number of nodes.

Method	Time (h)	
	CPU	Wall
MBE(4), no cutoffs	1601.7	58.4
MBE(4), $R_{\text{cut}1}$	496.3	18.1
MBE(4), $R_{\text{cut}1} + R_{\text{cut}2}$	951.4	34.6
MBE(4), $E_{\text{cut}}$	252.1	9.4
Supersystem	15.4	0.9

Actual timing data for a supersystem and various MBE(4) calculations on  $(\text{H}_2\text{O})_{37}$ , at the B3LYP/aDZ level and without CP corrections, are presented in Table III. These calculations reflect the subsystem counts that appear in Table II. All calculations were threaded across 28 processors within a single node,<sup>57</sup> and we note that the ratio of CPU time to wall time is  $\approx 27$  for each of the MBE(4) calculations, indicating near-perfect parallel scalability across a single node. (The parallel speedup is only about  $17\times$  for the supersystem calculation.) Note also that the wall times reported in Table III for the MBE(4) calculations reflect what would be required *if only a single node were used*. As such, the time-to-solution should decrease linearly as the number of nodes is increased, up to a very large number of nodes given the very large number of subsystems. At the same time, it is worth mentioning that for this particular calculation where the supersystem includes 1517 basis functions, ten times as many processors are required to make the MBE(4) wall time competitive with that of the supersystem calculation, even with our most aggressive thresholding scheme.

Lastly, one might object to our use of a supersystem HF calculation in the ONIOM-style procedure as this destroys the linear-scaling nature of the MBE. It bears note, however, that the prefactor on the  $\mathcal{O}(N)$  scaling of MBE( $n$ ) is extremely large *unless* the number of processors available amounts to a significant fraction of the number of subsystems. Batches of processors numbering in the thousands or tens of thousands may be unavailable on commodity clusters, and where they are available at supercomputer centers the queue times may be quite long for such requests. Furthermore, Raghavachari and co-workers have shown that there is a useful mid-size regime where a low-level supersystem remains tractable but a high-level calculation would not be. For systems in this size range, the combination of high-level fragment calculation with a low-level supersystem calculation can afford useful results.<sup>13,16,17,50</sup>

To put this in perspective, Table IV shows timing data for calculations on one isomer of  $(\text{H}_2\text{O})_{20}$  using a variety of thresholds and also lists the time required for a HF/aTZ supersystem calculation. As above, all calculations are multithreaded across all 28 cores of one node.<sup>57</sup> As in the  $(\text{H}_2\text{O})_{37}$  example, wall times for the MBE(4) calculations should decrease linearly with the number of nodes. The supersystem HF/aTZ calculation takes 2.0 h on a single node, as compared to 17.4 h for

TABLE IV. Timing data (in hours) for  $(\text{H}_2\text{O})_{20}$ , edge-sharing pentagonal prism isomer 10, with all calculations multithreaded across a single 28-core node.<sup>57</sup> For the short-range DFT + long-range HF method of Eq. (14), the total time is the sum of the two MBE(4) timings (HF + DFT), whereas the ONIOM-style method in Eq. (15) also includes the supersystem HF time.

Method	(6,1)		(7,1)		No thresholds	
	CPU	Wall	CPU	Wall	CPU	Wall
MBE(4), HF <sup>a</sup>	280.4	12.4	630.7	27.2	738.4	35.4
MBE(4), DFT <sup>b</sup>	424.6	17.4	970.1	39.0	1291.5	51.5
Supersystem, HF <sup>a</sup>	...	...	...	...	39.0	2.0

<sup>a</sup>Hartree-Fock/aTZ.

<sup>b</sup> $\omega\text{B97X-V/aTZ}$ .

a MBE(4) calculation at the  $\omega\text{B97X-V/aTZ}$  level, even with relatively loose (6,1) thresholds, which afford acceptable accuracy within the ONIOM-style paradigm. Thus, the lower-level supersystem calculation is cheaper (in terms of wall time) than the higher-level MBE(4) calculation until the latter is run on 9 nodes or 252 processors. This example demonstrates that a supersystem calculation of this size (1840 basis functions) need not be an overwhelming bottleneck, in many hardware configurations.

#### IV. CONCLUSIONS

We have demonstrated the performance of distance-based, connectivity-based, and energy-based cutoffs in the context of the many-body expansion, for total interaction energies of water clusters  $(\text{H}_2\text{O})_{6-37}$  and for relative energies of many different isomers of  $(\text{H}_2\text{O})_{20}$ . To achieve an accuracy better than 0.1 kcal/mol/monomer, or in other words 10% of  $k_B T$  (at  $T = 300 \text{ K}$ ) in the total interaction energy, without simply relying on error cancellation, requires a four-body expansion with counterpoise corrections. The latter can be approximated at the two-body level. This alone is a significant conclusion, given the paucity of fragment-based calculations that include four-body terms, or the even smaller number that include counterpoise corrections. Only fairly conservative distance-based thresholds achieve this level of accuracy, resulting in only about a 30% reduction in the number of subsystem calculations required, for the systems considered here, but the energy-based thresholds suggested in Ref. 21 seem much more promising in this respect.

Regardless of the details of the thresholding scheme, this work demonstrates that four-body calculations in large systems are both feasible and accurate, without resort to charge embedding. The intuitive appeal of embedding notwithstanding, there are reasons to avoid it because it hinders parallelization and also significantly complicates the formulation of analytic energy derivatives.<sup>58-60</sup> Absent embedding, the MBE is perfectly scalable up to a very large number of processors; derivatives are straightforward; and the method is stable even in large, augmented basis sets such as aug-cc-pVDZ and aug-cc-pVTZ, for which some fragment-based methods that employ embedding may experience problems.<sup>61,62</sup> Nevertheless, it is worth acknowledging that a large number of processors is required in order for the four-body approximation to outperform an efficient SCF implementation, even for systems

involving 1500–1800 basis functions and small fragments. This remains true even in light of preliminary results using a very promising energy-based thresholding scheme<sup>21</sup> that appears to achieve a far more dramatic reduction in the number of subsystem calculations, as compared to distance-based thresholding.

## SUPPLEMENTARY MATERIAL

See [supplementary material](#) for additional data for counterpoise corrections and (H<sub>2</sub>O)<sub>20</sub> interaction energies.

## ACKNOWLEDGMENTS

This work was supported primarily by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences under Award No. DE-SC0008550, with additional support from the National Science Foundation (No. CHE-1300603). Calculations were performed at the Ohio Supercomputer Center under Project No. PAA-0003.<sup>63</sup> J.M.H. is a fellow of the Alexander von Humboldt Foundation.

- <sup>1</sup>M. Katouda, A. Naruse, Y. Hirano, and T. Nakajima, *J. Comput. Chem.* **37**, 2623 (2016).
- <sup>2</sup>D. J. Cole, C.-K. Skylaris, E. Rajendra, A. R. Venkitaraman, and M. C. Payne, *Europhys. Lett.* **91**, 37004 (2010).
- <sup>3</sup>S. J. Fox, J. Dziedzic, T. Fox, C. S. Tautermann, and C.-K. Skylaris, *Proteins* **82**, 3335 (2014).
- <sup>4</sup>W. Li, Z. Ni, and S. Li, *Mol. Phys.* **114**, 1447 (2016).
- <sup>5</sup>H. J. Kulik, N. Luehr, I. S. Ufimtsev, and T. J. Martinez, *J. Phys. Chem. B* **116**, 12501 (2012).
- <sup>6</sup>G. D. Fletcher, D. G. Fedorov, S. R. Pruitt, T. L. Windus, and M. S. Gordon, *J. Chem. Theory Comput.* **8**, 75 (2012).
- <sup>7</sup>D. G. Fedorov, T. Ishida, M. Uebayasi, and K. Kitaura, *J. Phys. Chem. A* **111**, 2722 (2007).
- <sup>8</sup>D. G. Fedorov, Y. Alexeev, and K. Kitaura, *J. Phys. Chem. Lett.* **2**, 282 (2011).
- <sup>9</sup>N. Sahu, S. D. Yeole, and S. R. Gadre, *J. Chem. Phys.* **138**, 104101 (2013).
- <sup>10</sup>N. Sahu and S. R. Gadre, *Acc. Chem. Res.* **47**, 2739 (2014).
- <sup>11</sup>S. Li, W. Li, and J. Ma, *Acc. Chem. Res.* **47**, 2712 (2014).
- <sup>12</sup>X. He, T. Zhu, X. W. Wang, J. F. Liu, and J. Z. H. Zhang, *Acc. Chem. Res.* **47**, 2748 (2014).
- <sup>13</sup>A. Saha and K. Raghavachari, *J. Chem. Theory Comput.* **11**, 1212 (2015).
- <sup>14</sup>J. Liu and J. M. Herbert, *J. Chem. Theory Comput.* **12**, 572 (2016).
- <sup>15</sup>H. Nakata and D. G. Fedorov, *J. Phys. Chem. A* **120**, 9794 (2016).
- <sup>16</sup>K. V. J. Jose and K. Raghavachari, *Mol. Phys.* **113**, 3057 (2015).
- <sup>17</sup>K. V. J. Jose and K. Raghavachari, *J. Chem. Theory Comput.* **13**, 1147 (2017).
- <sup>18</sup>J. D. Hartman and G. J. O. Beran, *J. Chem. Theory Comput.* **10**, 4862 (2014).
- <sup>19</sup>J. D. Hartman, G. M. Day, and G. J. O. Beran, *Cryst. Growth Des.* **16**, 6479 (2016).
- <sup>20</sup>G. J. O. Beran, J. D. Hartman, and Y. N. Heit, *Acc. Chem. Res.* **49**, 2501 (2016).
- <sup>21</sup>J. F. Ouyang and R. P. A. Bettens, *J. Chem. Theory Comput.* **12**, 5860 (2016).
- <sup>22</sup>J. Liu, L. Qi, J. Z. H. Zhang, and X. He, *J. Chem. Theory Comput.* **13**, 2021 (2017).
- <sup>23</sup>R. M. Richard and J. M. Herbert, *J. Chem. Phys.* **137**, 064113 (2012).
- <sup>24</sup>J. Gauss, in *Modern Methods and Algorithms of Quantum Chemistry*, NIC Series, 2nd ed., edited by J. Grotenndorst (John von Neumann Institute for Computing, Jülich, 2000), Vol. 3, pp. 541–592.
- <sup>25</sup>G. A. Cisneros, K. T. Wikfeldt, L. Ojamäe, J. Lu, Y. Xu, H. Torabifard, A. P. Bartók, G. Csányi, V. Molinero, and F. Paesani, *Chem. Rev.* **116**, 7501 (2016).
- <sup>26</sup>J. Cui, H. Liu, and K. D. Jordan, *J. Phys. Chem. B* **110**, 18872 (2006).
- <sup>27</sup>R. M. Richard, K. U. Lao, and J. M. Herbert, *J. Chem. Phys.* **141**, 014108 (2014).
- <sup>28</sup>K. U. Lao, K.-Y. Liu, R. M. Richard, and J. M. Herbert, *J. Chem. Phys.* **144**, 164105 (2016).
- <sup>29</sup>E. E. Dahlke and D. G. Truhlar, *J. Chem. Theory Comput.* **3**, 46 (2007).
- <sup>30</sup>E. E. Dahlke and D. G. Truhlar, *J. Chem. Theory Comput.* **3**, 1342 (2007).
- <sup>31</sup>E. E. Dahlke, H. R. Leverentz, and D. G. Truhlar, *J. Chem. Theory Comput.* **4**, 33 (2008).
- <sup>32</sup>R. M. Richard, K. U. Lao, and J. M. Herbert, *Acc. Chem. Res.* **47**, 2828 (2014).
- <sup>33</sup>R. M. Richard, K. U. Lao, and J. M. Herbert, *J. Chem. Phys.* **139**, 224102 (2013).
- <sup>34</sup>G. D. Chen, J. Weng, G. Song, and Z. H. Li, *J. Chem. Theory Comput.* **13**, 2010 (2017).
- <sup>35</sup>R. M. Richard and J. M. Herbert, *J. Chem. Theory Comput.* **9**, 1408 (2013).
- <sup>36</sup>J. Friedrich, H. Yu, H. R. Leverentz, P. Bai, J. I. Siepmann, and D. G. Truhlar, *J. Phys. Chem. Lett.* **5**, 666 (2014).
- <sup>37</sup>D. Yuan, Y. Li, Z. Ni, P. Pulay, W. Li, and S. Li, *J. Chem. Theory Comput.* **13**, 2696 (2017).
- <sup>38</sup>R. M. Richard, K. U. Lao, and J. M. Herbert, *J. Phys. Chem. Lett.* **4**, 2674 (2013).
- <sup>39</sup>J. F. Ouyang, M. W. Cvitkovic, and R. P. A. Bettens, *J. Chem. Theory Comput.* **10**, 3699 (2014).
- <sup>40</sup>M. Kamiya, S. Hirata, and M. Valiev, *J. Chem. Phys.* **128**, 074103 (2008).
- <sup>41</sup>J. F. Ouyang and R. P. A. Bettens, *J. Chem. Theory Comput.* **11**, 5132 (2015).
- <sup>42</sup>S. F. Boys and F. Bernardi, *Mol. Phys.* **19**, 553 (1970).
- <sup>43</sup>B. H. Wells and S. Wilson, *Chem. Phys. Lett.* **101**, 429 (1983).
- <sup>44</sup>G. S. Tschumper, in *Reviews in Computational Chemistry*, edited by K. B. Lipkowitz and T. R. Cundari (Wiley-VCH, 2009), Vol. 26, Chap. 2, pp. 39–90.
- <sup>45</sup>In the notation to be introduced in Sec. III A, “loose” thresholds correspond to  $\tau_{\text{SCF}} = 10^{-5}$  a.u. and  $\tau_{\text{ints}} = 10^{-9}$  a.u., whereas “tight” thresholds employ  $\tau_{\text{SCF}} = 10^{-7}$  a.u. and  $\tau_{\text{ints}} = 10^{-14}$  a.u. This nomenclature is consistent with that used in our previous work.<sup>28</sup>
- <sup>46</sup>P. Valiron and I. Mayer, *Chem. Phys. Lett.* **275**, 46 (1997).
- <sup>47</sup>F. B. van Duijneveldt, J. G. C. M. van Duijneveldt-van de Rijdt, and J. H. van Lenthe, *Chem. Rev.* **94**, 1873 (1994).
- <sup>48</sup>L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. F. Ke, F. Y. Liu, H. B. Li, L. N. Ding, and K. Morokuma, *Chem. Rev.* **115**, 5678 (2015).
- <sup>49</sup>N. J. Mayhall and K. Raghavachari, *J. Chem. Theory Comput.* **8**, 2669 (2012).
- <sup>50</sup>K. V. J. Jose, D. Beckett, and K. Raghavachari, *J. Chem. Theory Comput.* **11**, 4238 (2015).
- <sup>51</sup>S. Kazachenko and A. J. Thakkar, *J. Chem. Phys.* **138**, 194302 (2013).
- <sup>52</sup>D. J. Wales and M. P. Hodges, *Chem. Phys. Lett.* **286**, 65 (1998).
- <sup>53</sup>N. Mardirossian and M. Head-Gordon, *Phys. Chem. Chem. Phys.* **16**, 9904 (2014).
- <sup>54</sup>P. M. W. Gill, B. G. Johnson, and J. A. Pople, *Chem. Phys. Lett.* **209**, 506 (1993).
- <sup>55</sup>Y. Shao, Z. Gan, E. Epifanovsky, A. T. B. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kúš, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. Woodcock III, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Didenhofen, R. A. DiStasio, Jr., H. Do, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. D. Hanson-Heine, P. H. P. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyayev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, C. M. Oana, R. Olivares-Amaya, D. P. O’Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, P. A. Pieniazek, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, N. Sergueev, S. M. Sharada, S. Sharma, D. W. Small, A. Sodt, T. Stein, D. Stück, Y.-C. Su, A. J. W. Thom, T. Tsuchimochi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, V. Vanovschi, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhao, B. R. Brooks, G. K. L. Chan, D. M. Chipman, C. J. Cramer, W. A. Goddard III, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. Schaefer III, M. W. Schmidt,

- C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xu, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. Van Voorhis, J. M. Herbert, A. I. Krylov, P. M. W. Gill, and M. Head-Gordon, *Mol. Phys.* **113**, 184 (2015).
- <sup>56</sup>S. Kazachenko and A. J. Thakkar, *Chem. Phys. Lett.* **476**, 120 (2009).
- <sup>57</sup>Dell PowerEdge C6320 two-socket servers with Intel Xeon E5-2680 v4 (Broadwell, 14 cores, 2.40 GHz) processors, 128 GB memory.
- <sup>58</sup>N. J. Mayhall, K. Raghavachari, and H. P. Hratchian, *J. Chem. Phys.* **132**, 114107 (2010).
- <sup>59</sup>T. Nagata, D. G. Fedorov, and K. Kitaura, in *Linear-Scaling Techniques in Computational Chemistry and Physics*, Challenges and Advances in Computational Chemistry and Physics, edited by R. Zalesny, M. G. Papadopoulos, P. G. Mezey, and J. Leszczynski (Springer, New York, 2011), Vol. 13, Chap. 2, pp. 17–64.
- <sup>60</sup>T. Nagata, K. Brorsen, D. G. Fedorov, K. Kitaura, and M. S. Gordon, *J. Chem. Phys.* **134**, 124115 (2011).
- <sup>61</sup>D. G. Fedorov, L. V. Slipchenko, and K. Kitaura, *J. Phys. Chem. A* **114**, 8742 (2010).
- <sup>62</sup>D. G. Fedorov and K. Kitaura, *Chem. Phys. Lett.* **597**, 99 (2014).
- <sup>63</sup>See <http://osc.edu/ark:/19495/f5s1ph73> for Ohio Supercomputer Center.