Generalized Many-Body Expansion: A Fragment-Based Method for modeling Large Systems

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the Graduate

School of The Ohio State University

By

Kuan-Yu Liu, M.S.

Graduate Program In Chemistry

The Ohio State University

2019

Dissertation Committee:

John M. Herbert, Advisor

Sherwin J. Singer

Steffen Lindert

Abstract

In this work, we explore several promising aspects of generalized many-body expansion (GMBE) for computing energies and energy responses of large systems. The GMBE is recognized as one of the fragment-based methods and provides a unified view of those methods in electronic structure theory. The four elements shared by other fragment-based methods include a fragmentation method, a capping method, an embedding method, and the number of layer. We utilize the last two elements as well as the screening to approximate the non-additive many-body interactions. We found that energy-based thresholding affords linear scaling and high accuracy without embedding and multi-layer approximations. For the charge embedding scheme, the variational electrostatic embedded one-body approximation alongside symmetry-adapted perturbation theory proves promising for non-covalent interactions with the newly implemented modified Hirshfeld population analysis. We are able to reduce the cost for calculation the charge response term without compromising accuracy. The two prominent results of the GMBE show that we can apply this fragment-based method to relatively large systems due to its excellent parallel ability. Since the preliminary results of frequency calculations show the robustness of finite difference implementation, we are also looking forward to applying this methodology to energy responses such as spectroscopies.

Dedicated to Chin-Yu Sung and Chung-Jung Pan, my parents,
whose contributions have been immeasurable.

Acknowledgments

I must first acknowledge Professor Jen-Shiang Yu in Naitonal Chiao Tung University and Prefessor Chin-Hui Yu in National Tsing Hua University for introducing me to quantum chemistry as well as for guiding me through research. Then I thank The Ohio State University for providing an excellent Lab space and the Ohio Supercomputercenter for generous computation resources and support that facilitate my research. I must thank the members of the Herbert group, Zhi-Qiang You, Jie Liu, Ka Un Lao, and the rest for camaraderie, assistance, discussions, and unforgettable experiences. Finally, I have to indebtedly and humbly acknowledge my advisor John Herbert. I has been inspired by John's quantum mechanics lectures and I enjoy his handmade homework problem set. The detailed descriptions of each problems shows his passion and insightfulness which makes the routine work enjoyable. I must also thank his support, guidance, and friendship that companies with me in my pursuit of PhD studies. This document can not exist without him.

Vita

1985	Born, Puli, Taiwan
2005-2009	B.S. Biological Science and Technology, National Chiao Tung University, Taiwan
2009-2010	Military Service
2010-2011	Research Assistant, National Yang-Ming University, Taiwan
2011–2013	M.S. Biological Science and Technology, National Chiao Tung University, Taiwan
2013–2015	Graduate Teaching Associate, The Ohio State University
2015–2019	Graduate Research Associate, The Ohio State University

Publications

(5) J. Liu, B. Rana, K.-Y. Liu and J. M. Herbert. **Variational formulation of the generalized many-body expansion with self-consistent charge embedding: Simple and correct analytic energy gradient for fragment-based ab initio molecular dynamics.** *J. Phys. Chem. Lett.*, (in press)

(4) K. Carter-Fenk, K. U. Lao, K.-Y. Liu and J. M. Herbert. **Accurate and Efficient *ab Initio* Calculations for Supramolecular Complexes: Symmetry-Adapted Perturbation Theory with Many-Body Dispersion.** *J. Phys. Chem. Lett.*, **10**, 2706 (2019)

(3) K.-Y. Liu, and J. M. Herbert. **Understanding the many-body expansion for large systems. III. Critical role of four-body terms, counterpoise corrections, and cutoffs.** *J. Chem. Phys.,* **147**, 161729, (2017).

(2) K.-Y. Liu, J. Liu and J. M. Herbert. **Accuracy of finite-difference harmonic frequencies in density functional theory.** *J. Comput. Chem.,* **38**, 1678 (2017).

(1) K. U. Lao, K.-Y. Liu, R. M. Richard and J. M. Herbert. **Understanding the many-body expansion for large systems. II. Accuracy considerations.** *J. Chem. Phys.,* **144**, 164105 (2016).

Fields of Study

Major Field: Chemistry
Theoretical Physical Chemistry

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation

Characterization of "large" systems using electronic structure theory has attracted much attention thanks to the improvement of computing power and algorithm. However, the size of the systems is constrained by the fact that the computational time grows exponentially. Fragment-based methodologies, hence, have emerged to address this issue by treating supersystem problems as many-body problems involving relatively small fragments. In addition, such methodologies can benefit from embarrassingly parallelization. In this work, we employ generalized many-body expansion to approximate the supersystem by explicitly including indispensable non-additive interactions or implicitly including them via electrostatic embedding and multi-layer formalism. This method can apply to either non-bonded systems or covalent-bonded systems with the capability to obtain energies and energy responses.

1.2 Fundamentals

1.2.1 The Generalized Many-Body Expansion (GMBE)

The GMBE that employing overlapping fragments is derived from the principle of inclusion/exclusion; application of the GMBE requires calculations on subsystems that are formed from intersections of fragments. In an n-body GMBE, which we call GMBE(n), the approximate energy is

$$\varepsilon^{(n)} = \sum_{i=1}^{\binom{N_f}{n}} E_i^{(n)} - \sum_{i=1}^{\binom{N_f}{n}} \sum_{j>i}^{\binom{N_f}{n}} E_{i\cap j}^{(n)} + \cdots + (-1)^{\binom{N_f}{n}+1} E_{i\cap j\cap \ldots \cap \binom{N_f}{n}}^{(n)} . \tag{1.1}$$

Lower case indices $i, j, \cdots$ in Eq. (1.1) refer to n-mers of fragments, whose energies are $E_i^{(n)}, E_j^{(n)}, \ldots$, and $i \cap j$ is the subsystem formed from the intersection of n-mers i and j, with energy $E_{i\cap j}^{(n)}$. For general applications, construction of $i \cap j$ requires severing covalent bonds and capping the severed valencies. In addition to the GMBE, a wide variety of energy-based fragmentation schemes exist in the literature. They can be classified into groups according to four elements as follows: a fragmentation method, a capping method, an embedding method, and the number of layers. We will discuss the last two elements in Chapter 2 and 3.

1.2.2 The Traditional Many-Body Expansion (MBE)

If we only consider non-bonded clusters, the Eq. (1.1) is equivalent to the traditional many-body expansion. The total energy

$$E = \sum_{I=1}^{N_f} E_I + \sum_{I=1}^{N_f} \sum_{J<I} (E_{IJ} - E_I - E_J) + \cdots \tag{1.2}$$

is expressed as a sum of monomer energies (E_I), pairwise interaction energies ($E_{IJ} - E_I - E_J$), etc., becoming exact (by tautological definition) when $n = N_f$. We replace indices $i, j, \cdots$ by $I, J, \cdots$ to distinguish overlapping and non-overlapping monomers.

We are aware that the Eq. (1.2) will incur basis-set superposition error (BSSE), which is the result of unbalanced basis sets for n-body appoximation, where $n > 1$. Taking fragment E_{IJ} as an example, the naïve formula for the interaction energy is described as

$$E_{\text{int}} = E_{IJ} - E_I - E_J \tag{1.3}$$

and this results in overestimation of the interaction energy. Instead, the monomer energies should be computed using the supersystem basis set to avoid the BSSE. The procedure for the BSSE corrections called counterpoise (CP) is generalized as

$$\Delta E_{IJK\cdots} = E_{IJK\cdots} - \sum_{k=I,J,K,\cdots}^{N_f} E_k^{IJK\cdots} \tag{1.4}$$

where $E_i^{IJK\cdots}$ represents the energy of monomer k computed with basis functions on all monomers.

1.3 Electrostatic Embedding Scheme

1.3.1 Non-variational Embedding

In practice, we don't include all terms in Eq. (3.1). We truncate the expansion at lower order as the number of fragments increases combinatorially with order of truncation. Embedding subsystems in electric field allows the MBE to account for higher order

many-body effects. The energy expression can formulated as

$$E = \sum_{I=1}^{N_f} \widetilde{E_I} + \sum_{I=1}^{N_f} \sum_{J<I} (\widetilde{E_{IJ}} - \widetilde{E_I} - \widetilde{E_J}) + \cdots , \tag{1.5}$$

where

$$\widetilde{E_I} = E_I + \sum_{J \neq I, A \in J} \sum_{\mu\nu} P_{\mu\nu} \Big\langle \phi_\mu \Big| \frac{1}{\big\|\mathbf{r} - \mathbf{R}_A\big\|} \Big| \phi_\nu \Big\rangle q_A. \tag{1.6}$$

q_A represents the atomic point charge in fragment J. For brevity, we'll use $(I_A)_{\mu\nu}$ to indicate the electrostatic integral in Eq. (1.6). The $\widetilde{E_{IJ}}$ can be readily understood as

$$\widetilde{E_{IJ}} = E_{IJ} + \sum_{K \neq IJ, A \in K} \sum_{\mu\nu} P_{\mu\nu} (I_A)_{\mu\nu} q_A. \tag{1.7}$$

1.3.2 Variational Embedding

The aforementioned approach is non-variational because the embedded charge, q_A, is not varied with the molecular orbital coefficients of the subsystems. The variational version is also called "explicit polarization" (XPol) method which includes polarization effects through electrostatic embedding. For close-shell fragments, the energy of XPol is

$$E = \sum_A \left[2 \sum_a \mathbf{c}_a^\dagger \left(\mathbf{h}^A + \mathbf{J}^A - \frac{1}{2} \mathbf{K}^A \right) \mathbf{c}_a^\dagger + E_{\text{nuc}}^{\text{A}} \right] + E_{\text{embed}}. \tag{1.8}$$

Expression in the bracket of eq. 2.1 represents the Hatree Fock energy for each fragment expanded in absolutely localized molecular orbitals. E_{embed} is the sum of fragment embedded energies via electrostatic interactions. The corresponding Fock matrix can be written as

$$\mathbf{F}^{\text{A}} = f_{\mu\nu}^{\text{A}} - \frac{1}{2} \sum_{J \notin A} (I_J)_{\mu\nu} q_J + \sum_{I \in A} \frac{\partial E_{embed}}{\partial q_I} \frac{\partial q_I}{\partial P_{\mu\nu}}. \tag{1.9}$$

The $\frac{\partial q_I}{\partial P_{\mu\nu}}$ provides the charge response due to the variational of MO coefficient.

1.4 Multi-Layer appoximations

With the same goal of electrostatic embedding, multi-layer approaches try to include non-additive interactions by using lower level of theory, e.g. Hartree-Fock, on the supersystem. The most renowned example is the ONIOM-type formalism.

$$E_{\text{subsys}} = E^{\text{Low}}_{\text{supersys}} + (E^{\text{High}}_{\text{subsys}} - E^{\text{Low}}_{\text{subsys}}) \tag{1.10}$$

The $E^{\text{High}}_{\text{subsys}}$ is truncated Eq. (3.1) calculated with high level of theory while $E^{\text{Low}}_{\text{subsys}}$ is computed with low level of theory.

CHAPTER 2

Intermolecular energy decomposition analysis in large supramolecular complexes using symmetry-adapted perturbation theory

In our previous studies[8], charge embedding for two-body, three-body, and four-body expansion performs comparable to or worse than the non-embedding scheme. However, the variational charge embedded one-body expansion, the XPol, combining with the (anti)symmetry-adapted perturbation theory has become a promising tool to model non-covalent interactions. The "extended" version of symmetry-adapted perturbation theory (XSAPT), developed in our group over the past several years, extends traditional SAPT to noncovalent complexes larger than dimers, affording accurate interaction energies and a physically meaningful decomposition thereof. The original implementation of XSAPT is based on charges that are fit to reproduce molecular electrostatic potentials, which becomes a computational bottleneck in large systems. Charge embedding based on modified Hirshfeld atomic charges is reported here, which dramatically lowers the computational cost without compromising accuracy. This is especially beneficial in XSAPT calculations where the monomers are large, and calculations are presented on systems that include a DNA intercalation complex and the binding of a drug molecule to an enzyme.

2.1 Introduction

Quantum-based modeling of non-covalent interactions for large systems has become possible thanks to increases in computing power, but hardware improvements alone are insufficient to tackle the large supramolecular complexes of interest in drug discovery, which involve binding of ligands to proteins or DNA.[9,10] A plethora of fragment-based methodologies has emerged to address this issue by reducing the supersystem problem to a many-body problem involving relatively small fragments.[11–19] Along these lines, our group has been working on extended symmetry-adapted perturbation theory (XSAPT),[1,6,7,20–23] an accurate and efficient monomer-based method for computing intermolecular interaction energies that also generalizes traditional SAPT energy decomposition analysis (EDA) to the case of more than two monomers.

2.2 Theory

2.2.1 The XSAPT

Our XSAPT approach combines traditional dimer SAPT calculations with the variational explicit polarization or "XPol" method[24] to obtain the monomer wave functions. In this way, many-body polarization is included in the unperturbed monomer wave functions my means of self-consistent electrostatic embedding.[21] For closed-shell fragments, the XPol energy expression is

$$E = \sum_A \left[2 \sum_n (\mathbf{c}_n^A)^\dagger \Big(\mathbf{h}^A + \mathbf{J}^A - \tfrac{1}{2}\mathbf{K}^A \Big) \mathbf{c}_n^A + E_{\text{nuc}}^A \right] + E_{\text{embed}} \, . \tag{2.1}$$

The expression in square brackets represents the Hartree-Fock energy for monomer A, expressed in terms of "absolutely localized" molecular orbitals (MOs) $\mathbf{c}_n$.[25] The

final term, E_{embed}, is the sum of electrostatic embedding energies. The Fock matrix corresponding to Eq. (2.1) is

$$\mathbf{F}^A = f^A_{\mu\nu} - \frac{1}{2}\sum_{B\neq A}\sum_{b\in B} q_b(\mathbf{I}_B)_{\mu\nu} + \sum_{a\in A}\frac{\partial E_{\text{embed}}}{\partial q_a}\frac{\partial q_a}{\partial P_{\mu\nu}} \tag{2.2}$$

where $f^A_{\mu\nu}$ is the Fock matrix for isolated monomer A and

$$(\mathbf{I}_B)_{\mu\nu} = \left\langle \phi_\mu \middle| \frac{1}{\left\|\mathbf{r}-\mathbf{R}_B\right\|} \middle| \phi_\nu \right\rangle . \tag{2.3}$$

is a one-electron integral representation the electrostatic potential generated by the Gaussian function-pair $\phi_\mu\mathbf{r}\,\phi_\nu\mathbf{r}$ at the point $\mathbf{R}_B$.

We use the "SAPT0" energy formula,[26] which includes the intermolecular perturbation through second order:

$$\begin{aligned} E^{\text{SAPT0}}_{\text{int}} = {} & E^{(1)}_{\text{elst}} + E^{(1)}_{\text{exch}} + E^{(2)}_{\text{ind}} + E^{(2)}_{\text{exch-ind}} \\ & + E^{(2)}_{\text{disp}} + E^{(2)}_{\text{exch-disp}} . \end{aligned} \tag{2.4}$$

To include *intra*molecular electron correlation effects in an efficient fashion we adopt the SAPT(KS) variant of this theory,[27] where "KS" indicates that the MOs are obtained from Kohn-Sham density functional theory (DFT). SAPT(KS) dispersion energies are especially sensitive to problems with the asymptotic behavior of the exchange-correlation (XC) functional,[21,27] but by using range-separated hybrid functionals that are tuned for each monomer,[1,7,27] one can achieve dispersion energies that are no worse than Hartree–Fock-based SAPT0, while the other energy components are improved.[27]

XSAPT approximates the total interaction energy in a pairwise fashion based on Eq. (2.4), but non-pairwise-additive polarization effects are included from the XPol

wave functions.[23] The electrostatics, exchange, and (exchange-)induction contributions to Eq. (2.4) can be evaluated at $\mathcal{O}(N^3)$ cost but the dispersion and exchange-dispersion terms scale as $\mathcal{O}(N^4)$ and $\mathcal{O}(N^5)$, respectively. These are also the least accurate parts of a SAPT0 or SAPT(KS) calculation so we usually replace them, either with *ab initio* dispersion potentials ("+*ai*D"),[1,7,22,23] or with self-consistently-screened many-body dispersion (MBD).[6]

2.2.2 Charge Embedding Scheme

Construction of the XPol Fock matrix requires a prescription for how the embedding charges will be derived from the monomer wave functions, in order to evaluate the charge derivatives $\partial q_a / \partial P_{\mu\nu}$ that appear in Eq. (2.2). For this we have used "ChElPG" charges[28] that are fit to reproduced the molecular electrostatic potential, evaluated on a real-space grid outside of the van der Waals contact region. Although these charges are physically appealing, numerically stable,[29], and afford good accuracy for XSAPT calculations, the requisite equations for the charge derivatives are complicated,[21,29,30] and their implementation is costly.[1,29] Evaluation of the ChElPG charge derivatives quickly becomes the computational bottleneck for XSAPT calculations involving large monomers.[1] Figure 2.1(a) shows timing data for the buckycatcher/fullerene complex $C_{60}@C_{60}H_{28}$, demonstrating that fully one-third of the total XSAPT computation time is spent in evaluating ChElPG charge derivatives.

In view of this, we sought an alternative way to perform the charge embedding and settled on an approach known as "Charge Model 5" (CM5).[31] CM5 atomic charges are empirically-parameterized modifications of Hirshfeld charges,[32] the latter of which are

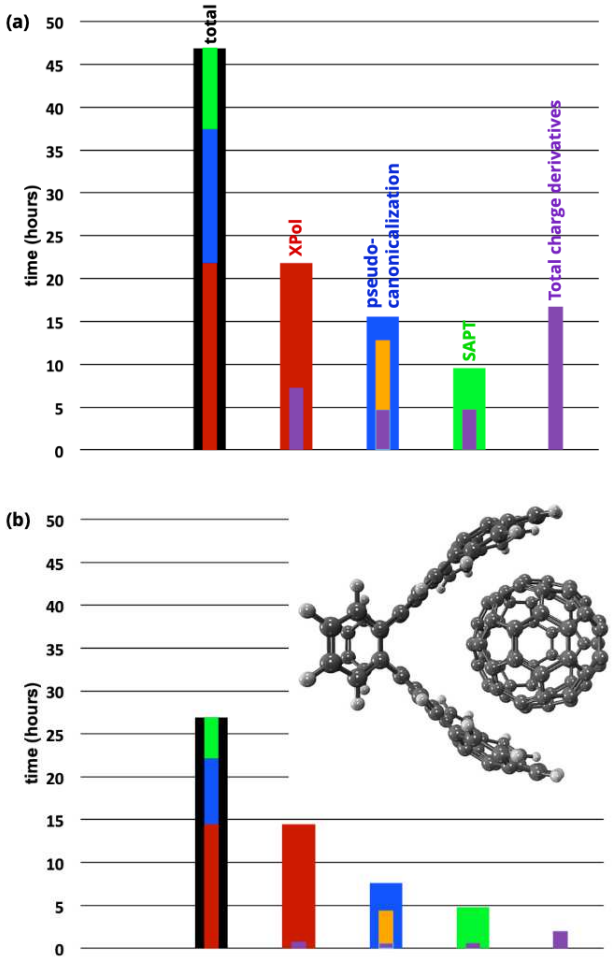

Figure 2.1: Timing data for XSAPT(KS)+*ai*D/hp-TZVPP calculations on C_{60}@$C_{60}H_{28}$ using (a) the original ChElPG implementation of XSAPT (data from Ref 1); and (b) the new CM5 implementation reported here, parallelized across all 28 cores of a single compute node. The black bar on the left represents the total wall time, broken down into red, blue, and green components representing the three major steps in the calculation. Charge derivatives are required in all three steps, and timing data for these are indicated in purple and summed on the right. The orange bar represents the Gram-Schmidt orthogonalization part of the pseudocanonicalization step, where the multithreading has been improved in the present implementation as compared to the one reported in Ref. 1.

derived from the molecular density $\rho(\mathbf{r})$ by using the superposition of isolated-atom densities $\widetilde{\rho}_A(\mathbf{r})$ to define a weight function

$$W_A(\mathbf{r}) = \frac{\widetilde{\rho}_A(\mathbf{r})}{\sum_B \widetilde{\rho}_B(\mathbf{r})} \tag{2.5}$$

that can be used to partition the molecular electron density into atomic contributions. Hirshfeld atomic charges are sometimes considered to be too small,[33] in the sense that the dipole moment obtained from them is smaller than the true dipole moment obtained from $\rho(\mathbf{r})$, and the CM5 parameterization corrects for this.

The basic formula for CM5 charges is

$$q_k^{\mathrm{CM5}} = Z_k - \int W_k(\mathbf{r})\ \rho(\mathbf{r})\ d\mathbf{r} + q_{\mathrm{param}} \tag{2.6}$$

where the empirical correction q_{param} depends on the Pauling bond order and other empirical parameters.[31] The requisite charge derivatives are simply

$$\frac{\partial q_k^{\mathrm{CM5}}}{\partial P_{\mu\nu}} = -\int W_k(\mathbf{r})\ \phi_\mu(\mathbf{r})\ \phi_\nu(\mathbf{r})\ d\mathbf{r}\ . \tag{2.7}$$

Integrals in Eq. (2.7) can be evaluated by quadrature in the same way that the DFT XC potential is evaluated, but a naïve implementation proves to be costly. Introducing a molecular quadrature grid consisting of points $\{\mathbf{r}_i\}$ and weights $\{w_i\}$, we have

$$\frac{\partial q_k^{\mathrm{CM5}}}{\partial P_{\mu\nu}} = -\sum_i w_i\ W_k(\mathbf{r}_i)\ \phi_\mu(\mathbf{r}_i)\ \phi_\nu(\mathbf{r}_i)\ . \tag{2.8}$$

The cost of this implementation scales with the number of atoms (N_{atoms}) and basis functions (N_{basis}) as $\mathcal{O}(N_{\mathrm{atoms}} \times N_{\mathrm{basis}}^2 \times N_{\text{mol-Leb}})$, where $N_{\text{mol-Leb}}$ represents number of Lebedev grid points that is required for accurate integration of the total molecular

density. For typical molecular quadrature grids this number ranges from ≈3,800 points per atom (low-quality, SG-1) to 15,000–19,000 points per atom (high quality, SG-3).[34] This formal scaling should be compared to that for the ChElPG charge derivatives, which is $\mathcal{O}(N_{\text{atoms}} \times N_{\text{basis}}^2 \times N_{\text{ESP-grid}})$,[29] where $N_{\text{ESP-grid}}$ represents the number of electrostatic potential grid points. Because the ChElPG procedure fits only to the long-range, slowly-varying parts of the electrostatic potential, and using an implementation of ChElPG charges based on atom-centered Lebedev grids,[29] it is possible to make $N_{\text{ESP-grid}} \ll N_{\text{mol-Leb}}$. In this case, no actual cost savings is realized by replacing ChElPG charges with CM5 charges.

That said, the cost to implement Eq. (2.8) can be dramatically reduced by recognizing that $W_A(\mathbf{r})$ vanishes far from $\mathbf{R}_A$, the position of nucleus A, because the free-atom density $\widetilde{\rho}_A(\mathbf{r})$ vanishes. As such, the integral required to compute $\partial q_k^{\text{CM5}}/\partial P_{\mu\nu}$ can be evaluated accurately and efficiently using just the atom-centered grid for atom k, not the entire molecular grid. In effect, we restrict the summation in Eq. (2.7) to just those grid points $i \in k$ contained within the atom-centered grid for atom k. This reduces the cost of the CM5 charge derivatives to $\mathcal{O}(N_{\text{atoms}} \times N_{\text{basis}}^2 \times N_{\text{atom-Leb}})$, which is the same cost as the XC quadrature step.

2.3 Results

The accuracy of the atomic-grid implementation of Eq. (2.8) has been tested by computing XSAPT interaction energies for the S22 data set.[35] The maximum deviation (with respect to an implementation that uses the full molecular quadrature grid) is

Table 2.1: Errors in XSAPT interaction energies as compared to benchmark values. The benchmarks are MP2/cc-pVTZ for the IHB data set and CCSD(T)/CBS for the rest. SAPT(KS) calculations for S22 and S66 are based LRC-ωPBE/hpTZVPP calculations for the monomers, whereas for the other two data sets we used ωB97X-V/ def2-TZVPPD.

Data Set	Error (kcal/mol)			
	maximum		MUE	
	CM5	ChElPG	CM5	ChElPG
S22	−1.1	−1.2	0.4	0.4
S66	−1.1	−1.1	0.3	0.3
IHB	−3.3	−5.3	1.1	1.7
ions	−3.8	−15.1	1.4	3.6

0.02 kcal/mol, with no systematic deviation. Figure 2.1(b) shows timings for the new XSAPT implementation as applied to $C_{60}@C_{60}H_{28}$. The time required to compute the charge derivatives has been reduced from 16.7 hours to 2.0 hours, with a secondary cost reduction coming from better parallelization of the repeated matrix multiplications required for the pseudocanonicalization step.[20] The speedup will be even greater for larger systems since $N_{\text{ESP-grid}}$ increases with molecular size but $N_{\text{atom-Leb}}$ does not.

The remainder of this work is dedicated to documenting the accuracy of the new CM5-based implementation of XSAPT. We first consider the standard S22[35] and S66[36] data sets consisting of dimers formed from charge-neutral molecules, along with the ionic hydrogen bonding (IHB) data set from Řezác and Hobza,[37] and an ion-pair data set from Lao and Herbert.[7] Error statistics for both CM5- and ChElPG-based implementations of XSAPT, as compared to the benchmark interaction energies for each data set, are listed in Table 2.1. Both charge schemes provide comparable results for S22 and S66 but significant differences are observed for ions. For the IHB data set

the largest absolute deviation between the CM5- and ChElPG-based XSAPT results occurs in the case of the imidazolium$\cdots$methylamine complex. Here, the charge difference $|q_{\mathrm{N}} - q_{\mathrm{C}}|$ between the heavy atoms in methylamine is an unrealistically large $3.8e$ in the case of ChElPG charges, versus only $0.6e$ for CM5 charges. For the ion-pair data set, the largest deviation is found in the complex of Cl^- with dimethyl ethyl amine, where the ChElPG atomic charges result in bond dipoles whose positive ends point toward the nitrogen atom whereas in the CM5 case they point away. A frequent criticism of ChElPG charges, at least when it comes to their use in force-field parameterization, is that the ChElPG procedure may sacrifice chemically-intuitive atomic partial charges in the interest of better fitting the molecular electrostatic potential,[28] a problem that becomes more severe for large molecules with "buried" atoms. We have previously considered that this criticism is not be relevant in the present context, since we have no interest in the atomic partial charges beyond their ability to reproduce the electrostatic potential, it appears that for monomers with net charge the CM5 charges produce both more intuitively-reasonable results *and* smaller errors in intermolecular interaction energies.

We next examine the performance of CM5-based XSAPT in different basis sets. Table 2.2 shows mean unsigned errors (MUEs) for several different data sets containing ionic monomers, from Ref. 4. These include the AHB21 and CHB6 data sets in which one monomer is an anion or a cation, respectively, and also the IL16 data set consisting of ion-pairs taken from common ionic liquid constituent molecules. These systems are rather small, and perhaps for that reason the XSAPT results converge

Table 2.2: XSAPT error statistics for data sets containing ionic monomers,[4] using two different charge schemes and various basis sets.

Basis Set	MUE (kcal/mol)							
	AHB21		CHB6		IL16		overall	
	CM5	ChElPG	CM5	ChElPG	CM5	ChElPG	CM5	ChElPG
cc-pVDZ	8.3	9.0	2.4	2.7	10.7	11.6	8.7	9.4
jun-cc-pVDZ	1.2	1.3	–	–	1.4	2.9	1.3	2.0
aug-cc-pVDZ	0.9	2.9	1.2	0.8	3.0	7.2	1.8	4.3
cc-pVTZ	5.9	6.3	1.4	1.3	7.7	9.9	6.2	7.2
aug-cc-pVTZ	1.2	2.0	0.8	0.7	3.2	10.6	1.9	5.2
def2-TZVPP	3.5	3.0	0.5	0.5	1.3	2.3	2.3	2.4
def2-TZVPPD	1.2	1.1	1.0	1.1	1.9	7.7	1.4	3.6

already in the aug-cc-pVDZ basis set. A more detailed breakdown can be found in Tables A.5–A.7 of the Supplementary Material, and these data reveal that the difference between the CM5 and ChElPG charges is marginal for the AHB21 and the CHB6 data sets but quite pronounced for IL16, where both monomer units are ions. ChElPG charges have occasionally been used as a metric for intermolecular charge transfer, *e.g.*, for the ion pairs comprising ionic liquids.[38] This seems rather dubious in view of the problems documented here for charged monomers.

The S30L data set[39] consists of 30 large host/guest complexes, including the buckycatcher/C_{60} complex shown in Fig. 2.1. In Ref. 39, estimated gas-phase interaction energies for these complexes are estimated starting from experimental solution-phase binding free energies that are then back-corrected for vibrational entropy changes upon complexation, and for solvation contributions to the energy of complexation, resulting in estimated uncertainties of ∼2 kcal/mol in the benchmarks. We have previously used these complexes to test various versions of XSAPT,[1,6] and our older ChElPG-based implementation affords a MUE of 4.7 kcal/mol for these complexes

Table 2.3: Interaction Energies for ligand/macromolecule complexes shown in Fig. 2.2. The QMC result is from Ref. 5 and the counterpoise-corrected B97M-V/def2-TZVPPD result is from Ref. 6. XSAPT calculations use the partially-augmented def2-hp-TZVPP basis set defined in Ref. 7, which omits diffuse functions on hydrogen atoms.

Method	E_{int} (kcal/mol)	
	DNA/ ellipticine	HIV/ indinavir
QMC	−33.6	—
B97M-V (+counterpoise)	−41.3	—
XSAPT+*ai*D(CM5)	−36.7	−106.1
XSAPT+*ai*D(ChElPG)	−35.7	−103.9

that is competitive with the best-avaiable quantum chemistry approaches, at reduced cost even as compared to supramolecular DFT.[6] CM5-based XSAPT, however, affords a slightly lower MUE (4.1 kcal/mol), even while it accelerates the buckycatcher/ C_{60} calculation by more than a factor of 8.

Figure 2.2 shows a pair of model systems representing drug binding to a macromolecule, including a DNA/ellipticine intercalation complex[9] and a complex of the antiretroviral indinavir to HIV-2 protease.[10] Both the CM5- and ChElPG-based versions of XSAPT afford interaction energies in reasonable agreement with quantum Monte Carlo (QMC) calculations (see Table 2.3), and within 1.0 kcal/mol of one another. For the HIV/indinavir complex (323 atoms, or 10,626 basis functions using aug-cc-pVTZ), no reliable supersystem benchmark is available but the XSAPT results can serve as a good estimate.

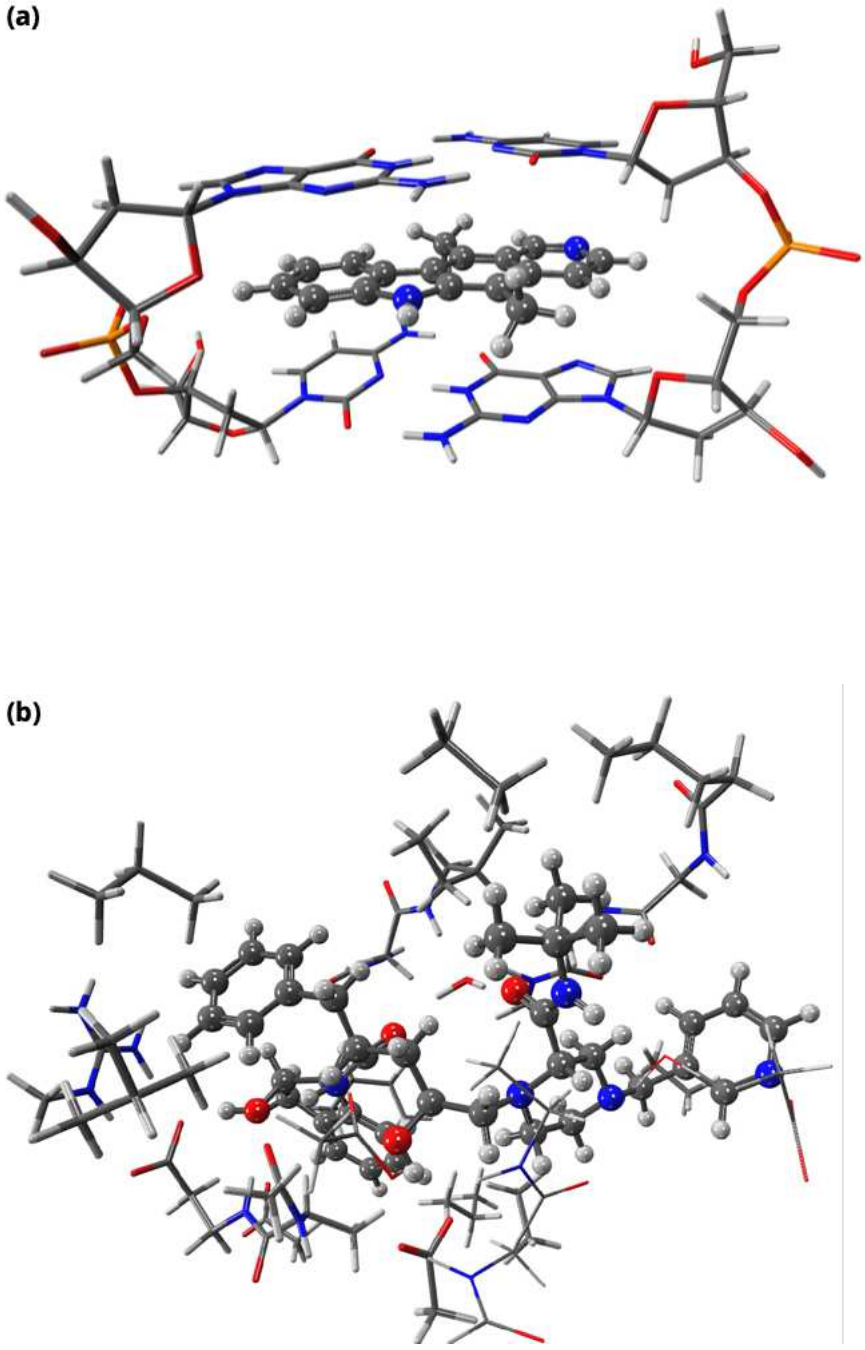

Figure 2.2: (a) Ellipticine intercalation complex with two base pairs (and backbone) of DNA (157 atoms), and (a) complex of the protease inhibitor indinavir with a model of the HIV-2 binding pocket (323 atoms). In both cases the drug molecule is shown in a ball-and-stick representation while the rest of the system is depicted with a tubular representation.

2.4 Conclusions

In summary, we have developed a CM5 charge-embedding scheme for use with the XSAPT methodology that improves both the accuracy and (especially) the efficiency of the method, as compared to our original ChElPG-based implementation. For benchmark data sets of non-covalent complexes, interaction energies computed with the CM5-based XSAPT procedure are consistently a bit more accurate than ChElPG-based results. For ion pairs, the CM5-based version considerably improves the accuracy, mainly by removing some outliers where the ChElPG embedding charges adopt counter-intuitive values. This improvement is coupled to a dramatic reduction in the cost of CM5-based XSAPT, which is 8.4× faster than the ChElPG version for a $C_{60}@C_{60}H_{28}$ complex. Note that ChElPG embedding charges are used also in one formulation of Ewald summation for quantum mechanics/molecular mechanics (QM/MM) calculations,[29,30] with charge derivatives as in Eq. (2.2) that prove to be a serious bottleneck in calculation the energy gradient, even for relatively small QM regions.[30] Work is underway in our group to implement a CM5-based version of the QM/MM-Ewald procedure.

CHAPTER 3

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

Our previous papers have attempted to shed light on precision and accuracy issues affecting the many-body expansion (MBE) but which only manifest in larger systems and 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, four-body terms prove necessary to achieve accurate results for both total interaction energies and relative isomer energies, and the sheer number of tetramers makes the use of cutoff schemes absolutely essential. To predict relative energies of water cluster isomers, two corrections based on a lower level of theory are introduced: either well-separated sub-clusters can be computed at a lower level of theory (as the higher-level calculation is subjected to a smooth, shorter-range cutoff); or else the entire supersystem can be computed at a low level of theory and combined with the MBE in an ONIOM-type paradigm. The latter results are 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 well below the number required when using distance-based thresholds. The result is an accurate four-body method that does not rely on charge embedding, is stable in large, diffuse basis sets, and shows excellent speedups relative to supersystem calculations.

3.1 Background

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.[40–61] 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[40] (possibly in conjunction with linear-scaling versions of those algorithms[41–43]), or else implementations using graphical processing units.[44] An increasing popular alternative, and one that is perhaps more easily amenable to large-scale parallelization,[45] is to adopt a fragment-based approach.[46–61] 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.[62] In such approaches the total energy, or any other property that can be expressed as a derivative of the energy,[63] 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,[64] and can play a pivotal role in stabilizing, *e.g.*, α-helix structures in peptides over long distances,[60] 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.[53] These terms are also definitely *not* negligible in water clusters,[8,65,66] where many-body polarization effects are significant.

It has been argued[67–69] 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 are neglected when higher-order terms in the MBE are omitted. Our previous work strongly contests this idea, however,[8,66] and suggests 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.67–69, but

we obtained very different results upon examining water clusters up to $N = 55$.[8,70] 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.[8,70] Only at the four-body level were errors reduced as low as a few kcal/mol.[8]

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 use of far tighter convergence thresholds and drop tolerances than would ordinarily be required in a single electronic structure calculation.[8,66] 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 previous work on the MBE. The problem is especially acute when the software than 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.[66] Especially in the presence of embedding charges, it is crucial to read binary scratch files or checkpoint files instead, in full machine precision.[66] (This fact has been mentioned in passing elsewhere,[71] but without further analysis.)

A related issue is how even to define "error" in the MBE. 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.[8,62,66,70,72,73] An alternative proposal is to compare to the best available benchmarks for a given system,[74,75] 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 cancelation is basis-set superposition error (BSSE), whose effects were never discussed in the context of the MBE until recently.[76,77] BSSE can cause convergence of the MBE (with respect to n) to become erratic, because it may offset neglected many-body induction effects.[72] To address this problem, our group[8,76] and others[78,79] have developed many-body counterpoise (CP) corrections that are designed to approximate the supersystem Boys-Bernardi CP correction[80] (as generalized to an arbitrary number of monomers[81,82]), 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 two previous papers in this series that documented precision problems[66] and accuracy problems[8] 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[79] note that for molecular dynamics applications at room temperature, each atom has $(3/2)k_BT \approx 0.9$ 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 Ref79, and we adopt this as our target accuracy per monomer. 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.[8]

To put this in perspective, a complete four-body calculation on the largest system considered here, $(H_2O)_{37}$, consists of 74,518 distinct subsystems including 66,045 tetramers. At the ωB97X-V/aTZ level that is used herein to examine relative energies of cluster isomers, the use of "tight" versus "'loose" thresholds? (as defined in Ref.8) 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 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.

3.2 Theory and Methods

3.2.1 Many-body expansion

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

$$E = \sum_{I=I}^{N} E_I + \sum_{I=1}^{N} \sum_{J<I} \Delta E_{IJ} + \sum_{I=1}^{N} \sum_{J<I} \sum_{K<J} \Delta E_{IJK} + \cdots \ . \tag{3.1}$$

The two- and three-body corrections are

$$\Delta E_{IJ} = E_{IJ} - E_I - E_J \tag{3.2a}$$

$$\Delta E_{IJK} = E_{IJK} - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{JK} \tag{3.2b}$$

$$- E_I - E_J - E_K \ .$$

An n-body approximation, which we will denote as MBE(n), truncates Eq. (3.1) at terms involving n fragments. If taken literally, however, Eq. (3.1) involves some redundant calculations because, *e.g.*, the monomer energy E_I appears in ΔE_{IJ}, ΔE_{IJK}, etc. Non-redundant formulas with appropriate combinatorial coefficients, for MBE(n) with arbitrary n, can be found in Ref.66.

3.2.2 Counterpoise corrections

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

$$E_{\text{int}} = E - \sum_{I=I}^{N} E_I \ . \tag{3.3}$$

The usual Boys-Bernardi CP correction for molecular dimers[80] resembles the two-body correction ΔE_{IJ} performed in the dimer basis set. We might indicate this as

$$\Delta E_{IJ}^{\mathrm{CP}} = E_{IJ}^{IJ} - E_{I}^{IJ} - E_{J}^{IJ} \; . \tag{3.4}$$

Following previous literature,[8,76,83] 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,[81,82] and a CP-corrected interaction energy

$$E_{\mathrm{int}}^{\mathrm{CP}} = E_{IJK\cdots N}^{IJK\cdots N} - \sum_{I=1}^{N} E_{I}^{IJK\cdots N} \; . \tag{3.5}$$

The quantity defined in this equation has been called the "site–site function counterpoise correction",[81] but we refer to it simply as the Boys-Bernardi CP correction, since it naturally generalizes the original dimer approach.[80]

The CP-corrected interaction energy in Eq. (3.5) can alternatively be expressed as

$$\begin{aligned} E_{\mathrm{int}}^{\mathrm{CP}} &= E_{IJK\cdots N}^{IJK\cdots N} - \sum_{I} E_{I}^{I} + \sum_{I} \Big(E_{I}^{I} - E_{I}^{IJK\cdots N} \Big) \\ &= E_{\mathrm{int}}^{\mathrm{uncorr}} + \delta E^{\mathrm{CP}} \end{aligned} \tag{3.6}$$

where the "uncorrected" interaction energy is

$$E_{\mathrm{int}}^{\mathrm{uncorr}} = E_{IJK\cdots N}^{IJK\cdots N} - \sum_{I} E_{I}^{I} \tag{3.7}$$

and the CP correction is

$$\delta E^{\mathrm{CP}} = \sum_{I} \Big(E_{I}^{I} - E_{I}^{IJK\cdots N} \Big) \; . \tag{3.8}$$

Equation (3.8) defines the N-body CP correction,[81,82] which has sometimes been criticized for its failure to account for "basis-set extension" effects,[78,83,84] although

the good agreement between CP calculations and the alternative (and formally more complete) Valiron-Mayer function counterpoise corrections[78,83] suggests that any neglected effects are rather small.[8,76] The complete Valiron-Mayer approach also rapidly becomes intractable beyond just a few monomers. In view of this, we take Eq. (3.8) to define the counterpoise correction, in the spirit of Boys and Bernardi.[81,82] 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 MBE,[62] for which CP corrections have also been formulated.[8]

To circumvent this, and in view of Eq. (3.6), we approximate the CP-corrected total energy through a standard MBE(n) calculation applied to the supersystem energy $E_{IJK\cdots N}^{IJK\cdots N}$ in Eq. (3.7) in conjunction with an n-body approximation to the summand in Eq. (3.8). We call this a many-body counterpoise (MBCP) correction,[72,76] truncated at order n, or MBCP(n) for short. Formulas for $\delta E_I^{\mathrm{MBCP}(n)}$, which is the n-body approximation to the Ith summand in Eq. (3.8), were derived previously through $n = 4$.[72,76] The two leading terms are

$$\delta E_I^{\mathrm{MBCP}(2)} = (N-1)E_I^I - \sum_{J\neq I}^{N} E_I^{IJ} \tag{3.9}$$

and

$$\begin{aligned}\delta E_I^{\mathrm{MBCP}(3)} = \delta E_I^{\mathrm{MBCP}(2)} &- \tfrac{1}{2}(N-2)(N-1)E_I^I \\ &+ (N-2)\sum_{J\neq I}^{N} E_I^{IJ} - \sum_{J\neq I}^{N}\sum_{\substack{K>J \\ K\neq I}}^{N} E_I^{IJK} \ .\end{aligned} \tag{3.10}$$

Summing Eqs. (3.9) and/or (3.10) over all monomers I affords the MBCP(n) approximation, for $n = 2$ or 3.

Our original idea[76] was to combine the MBE(n) approximation for the supersystem energy $E_{IJK\cdots N}^{IJK\cdots N}$ with the MBCP(n) approximation for the CP corrections $E_I^{IJK\cdots N}$, for a consistent order-by-order truncation to the CP-correction interaction energy. In practice, however, we find that the MBCP(n) corrections are quite small for $n > 2$. In the present work, we therefore include only the MBCP(2) correction.

An alternative way to interpret BSSE was introduced by Valiron and Mayer[83] and later adopted by others.[77–79] Within this formulation, one writes

$$E_{\text{int}}^{\text{CP}} = \sum_{IJ} \Delta E_{IJ}^{IJK\cdots N} + \sum_{IJK} \Delta E_{IJK}^{IJK\cdots N} + \cdots \tag{3.11}$$

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, have existed for a long time,[84] 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 MBE, can be found in Ref.79, 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".[79] 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. (3.11) in mind, however, Ouyang and Bettens[79] introduced a CP scheme that is formally more general than our MBCP(n) approach, and in particular conforms more closely to the Valiron-Mayer idea.[78,83] 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,[79] as we have already suggested above. This provides further justification for the approximate CP correction employed here.

3.2.3 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 supersystem calculation.[8,53] This is especially true when CP corrections are introduced, as these require a very large number of additional calculations.[8] 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 they number of subsystems they can also reduce finite-precision problems.

Distance-based thresholds

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

$$f(x) = \begin{cases} 1, & \text{if } x < 0 \\ 1 - x^3(10 - 15x + 6x^2), & \text{if } 0 \leq x \leq 1 \\ 0, & \text{if } x > 1 \end{cases} . \tag{3.12}$$

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_2O, 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_{cut1}, the distance for the onset of 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

$$x = (R_{\max} - R_{\text{cut1}})/w \tag{3.13}$$

in Eq. (3.12). If $R_{\max} \geq R_{\text{cut1}} + w$ 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 $R_{\max} < R_{\text{cut1}} + w$, 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_2O molecule and we will test various combinations of R_{cut1} and w. For brevity in the discussion that follows, we will use the notation (n_r, n_w) to indicate particular choices of the

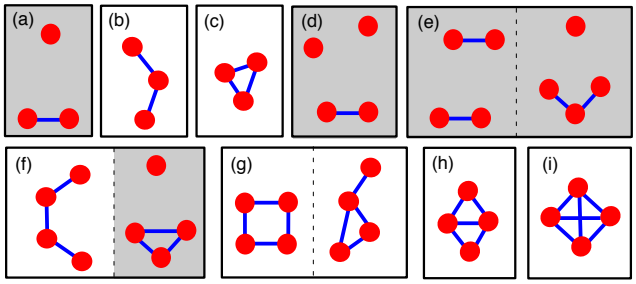

Figure 3.1: Pictorial representations of all possible "connectivities" for trimers and tetramers of fragments (red circles). Solid blue lines indicate inter-fragment distances less than R_{cut2}, and the various cases are grouped according to how many of these there are. Fragments not connected by blue lines are farther apart than R_{cut1}. Trimers (a) and (b), and tetramers (d)–(h), are excluded by the R_{cut1} threshold [Eqs. (3.12) and (3.13)], but when we additionally employ the R_{cut2} criterion only configurations in the shaded boxes are excluded.

thresholds, where n_r and n_w are a pair of integers that specify the values of R_{cut1} 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.[60] To account for this, we introduce a second distance parameter $R_{\text{cut2}} < R_{\text{cut1}}$, in the spirit of the connectivity-based analysis in Ref.60. To understand the role of this second cutoff, consider the "connectivity diagrams" of trimers and tetramers that are illustrated in Fig. 3.1. In these diagrams, we connect with a line any pair of fragments that are separated by a distance less than R_{cut2}, whereas we imagine that disconnected fragments are separated by more than R_{cut1} and therefore these configurations are excluded by the R_{cut1} cutoff.

Figures 3.1(a)–(c) exhaust all of the possible topologies for trimers, and Figs. 3.1(d)–(i) show all possibilities for tetramers. Note that configurations (c) and (i), in which all inter-fragment distances are less than R_{cut2} (and therefore less than R_{cut1} 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_{cut1}, so each is excluded by the cutoff procedure of Eqs. (3.12) and (3.13).

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_{cut1} cutoff procedure on the basis of its end-to-end distance. For the tetrameric cases, the configurations in Figs. 3.1(d) and 3.1(e) consist of strongly-interacting dimers or trimers plus another 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_{cut2} threshold to retain tetramers otherwise excluded by R_{cut1} only if they exhibit four or more inter-fragment distances less than R_{cut2}. This excludes the cases in Figs. 3.1(d) and 3.1(e), as well as one of the cases shown in Fig. 3.1(f). The R_{cut2} threshold, introduced in Ref.60, has not yet been implemented in a smooth way, nor will we attempt to do so now. Rather, we merely present results with R_{cut2} in order to compare the accuracy against those obtained using the smooth R_{cut1} threshold alone.

Energy-based thresholds

For systems where the individual fragments are substantially larger than H_2O, the distance-based thresholding discussed above may become less effective in reducing the number of subsystem calculations. Ouyang and Bettens[60] 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.

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. (3.12)]. For any particular

subsystem, the energy formula that is used is

$$E_{\text{subsys}} = f(x)E^{\text{DFT}}_{\text{subsys}} + [1 - f(x)]E^{\text{HF}}_{\text{subsys}} \, . \tag{3.14}$$

For subsystems that exist in the switching region, meaning that $R_{\text{cut1}} \leq R_{\text{max}} \leq R_{\text{cut1}} + w$, it is necessary to perform both the HF and the DFT calculation.

The second approach is an ONIOM-type formalism,[85] inspired by the fragment-based methods introduced by Raghavachari and co-workers,[52,55,56,86,87] 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

$$E_{\text{subsys}} = \big(E^{\text{DFT}}_{\text{subsys}} - E^{\text{HF}}_{\text{subsys}}\big) f(x) + E^{\text{HF}}_{\text{supersys}} \, . \tag{3.15}$$

Considering all subsystems, the terms $E^{\text{DFT}}_{\text{subsys}} f(x)$ together constitute an n-body DFT calculation with smooth cutoffs, and subtracting $E^{\text{HF}}_{\text{subsys}} f(x)$ prevents double-counting of the low-level calculations on the "model system" (to use ONIOM terminology[85]) in the presence of a low-level calculation $E^{\text{HF}}_{\text{supersys}}$ on the "real system". Note that the supersystem term in Eq. (3.15) is the same for each subsystem, so need only be computed once.

3.3 Results and Discussion

3.3.1 Computational details

In the first part of this work, we examine how distance-based thresholds affect the accuracy of interaction energies computed for a sequence of water clusters, $(H_2O)_{N=6-37}$. These structures were originally obtained from Ref.88, 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.

For these tests we use the affordable B3LYP/aug-cc-pVDZ (B3LYP/aTZ) level of theory, with an SCF convergence threshold $\tau_{\mathrm{SCF}} = 10^{-7}$ a.u. and a drop tolerance $\tau_{\mathrm{ints}} = 10^{-14}$ a.u. These are "tight" convergence thresholds, as defined in previous work,[8] whereas looser thresholds may lead to precision problems in the MBE.[66] Both thresholds, especially τ_{ints}, are significantly tighter than the default settings in common electronic structure programs.

In the second part of this work, we examine relative energies of four different structural motifs of $(H_2O)_{20}$. These structures have also been considered in previous work on the MBE,[72] and are taken from the extensive basin-hopping Monte Carlo search in Ref.89. For these calculations we employ a higher-quality level of theory, namely ωB97X-V[90]/aug-cc-pVTZ (ωB97X-V/aTZ), with τ_{SCF} and τ_{ints} as above.

The SG-1 quadrature grid[91] is used for all calculations, as higher-quality grids have been examined and found to make little difference in the context of the MBE.[66] All calculations were performed using Q-Chem, v. 4.2.[92]

3.3.2 Interaction energies

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

$$\text{error} = E_{\text{int}}(n\text{-body}) - E_{\text{int}}(\text{supersystem}) \,, \tag{3.16}$$

where one or both energies may be CP corrected, depending on context. For total interaction energies will report errors in size-intensive, per-monomer units, but Eq. (3.16) fixes the convention for the sign of the errors. Errors will be compared to the dynamic accuracy threshold discussed above,[79] *i.e.*, 10% of $(3/2)k_BT$ per monomer at $T = 298$ K, or in other words 0.09 kcal/mol/monomer.

Data comparing the full CP correction at the B3LYP/aug-cc-pVDZ level versus its MPCP(2) approximation are shown for $(H_2O)_{N=6-37}$ in Table B.1 of the Supplementary Material. Differences between δE^{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.[79] 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 $E_{\text{int}}^{\text{CP}}$.

Role of CP correction

In Figs. 3.2 and 3.3 we examine size-dependent errors in MBE(3) and MBE(4) results and their MBCP(2)-corrected counterparts, in two different ways. In Fig. 3.2,

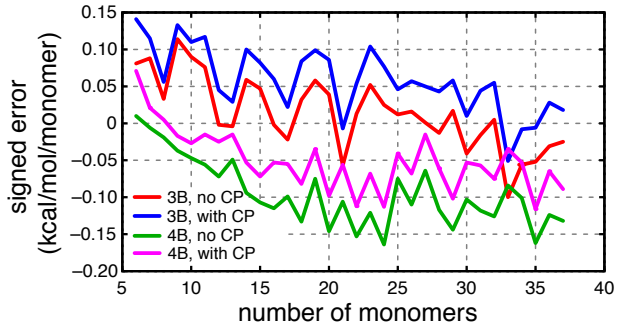

Figure 3.2: Signed errors per monomer in three- and four-body total interaction energies for clusters $(H_2O)_{6-37}$, at the B3LYP/aDZ level of theory. The n-body calculations labeled "no CP" are computed without MBCP corrections and compared to uncorrected supersystem energies, whereas those labeled "with CP" include MBCP(2) corrections and are compared to supersystem energies that include the full Boys-Bernardi CP correction.

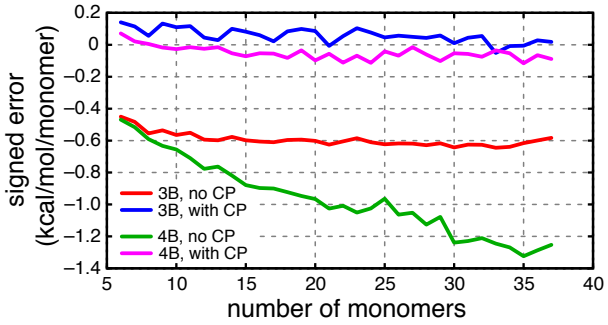

Figure 3.3: Signed errors per monomer in three- and four-body total interaction energies for clusters $(H_2O)_{6-37}$, at the B3LYP/aDZ level of theory. All of the n-body calculations, whether CP-corrected or not, are compared to supersystem calculations that include the full Boys-Bernardi CP correction.

the 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.*, δE^{CP} in Eq. (3.8). Figure 3.3 compares both MBE(n) and MBCP(2)+MBE(n) results to supersystem energies that include δE^{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 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 stabilizing four-body terms are partially offset by BSSE, as observed in our previous work exploring extrapolations to the basis-set limit.[72] Note that the error in the CP-corrected interaction energy is

$$\mathrm{error(CP)} = E_{\mathrm{int}}^{\mathrm{CP}} - \left(E_{IJK\cdots N}^{IJK\cdots N} - \sum_{I} E_{I}^{IJK\cdots N} \right) \tag{3.17}$$

whereas the error in the uncorrected case is

$$\mathrm{error(uncorr)} = E_{\mathrm{int}}^{\mathrm{uncorr}} - \left(E_{IJK\cdots N}^{IJK\cdots N} - \sum_{I} E_{I}^{I} \right) . \tag{3.18}$$

Adding δE^{CP} as defined in Eq. (3.8) to Eq. (3.17) results in precisely the right side of Eq. (3.18), which shows that the two definitions of error in Eqs. (3.17) and (3.18) are simply offset by the magnitude of the CP correction.

Examining Fig. 3.3, 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, and give rise to an error of ≈ -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 cluster size increases.

Effects of cutoffs

To obtain a decent guess as to what might constitute a reasonable distance cutoff R_{cut1}, we examine the convergence of the total interaction energies at the MBE(4) level for $(H_2O)_{6-37}$, in Fig. 3.4. 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_{cut1}, and also examine $R_{\mathrm{cut1}} = 7$ and 8 Å along with $w = 1$, 2, and 3 Å. [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. 3.5, 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

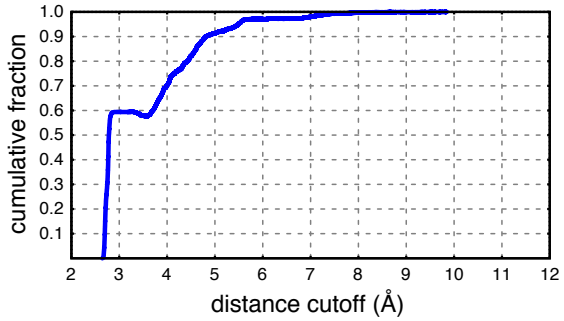

Figure 3.4: Fraction of the cumulative interaction energy for water clusters (B3LYP/ aDZ level) that is recovered by a four-body approximation, as a function of a sharp distance cutoff for the subsystem calculations. Interaction energies are not CP corrected, and the data point at each distance represents an average over cluster sizes from $N = 6$–37.

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. 3.5 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 6–9 Å are important in providing long-range stabilization.

Notice from Fig. 3.5 that errors are larger for the cluster sizes $N = 31$, 32, and 34–37, 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 pentaprismic structures.[93] (Recall that our cluster structures are putative global minima at each value of N.[88]) In view of recent work by Ouyang and

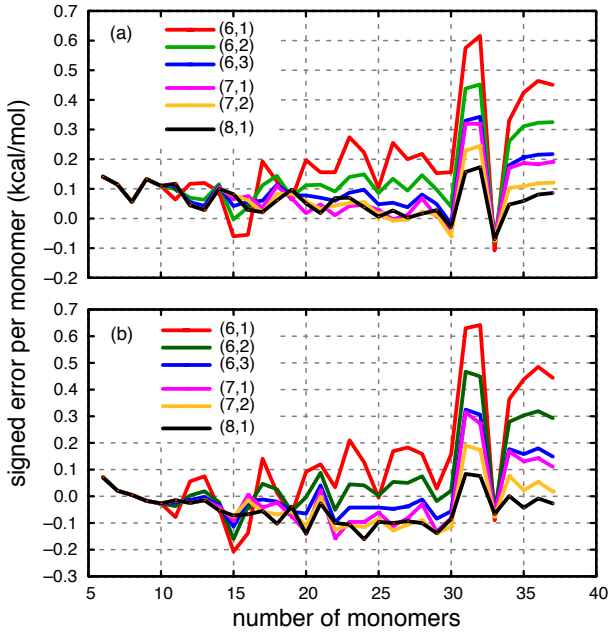

Figure 3.5: Signed errors per monomer in CP-corrected (a) three-body and (b) four-body approximations to the total interaction energy for a sequence of water clusters, employing different values for the switching function parameters (n_r, n_w). Subsystem calculations include the MBCP(2) counterpoise correction [Eq. (3.9)] but are compared to supersystem results including the full counterpoise correction [Eq. (3.5)].

Table 3.1: Error statistics (maximum error and mean unsigned errors) for CP-corrected MBE(4) approximations to $E_{\rm int}^{\rm CP}$, using various thresholds (n_r, n_w), in conjunction with the $R_{\rm cut2}$ threshold, $R_{\rm cut2} < R_{\rm cut1}$. Statistics include all $(H_2O)_N$ clusters, $N = 6$–37.

$R_{\rm cut2}$ (Å)	error (kcal/mol/monomer)					
	(6,1)		(7,1)		(8,1)	
	max	MUE	max	MUE	max	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

Bettens aimed at identifying important many-body interactions in polypeptides,[60] it may be the case that the sort of cooperative, chain-like interactions amongst fragment dipole moments that were identified in Ref60 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 possibility, we introduce the second threshold parameter $R_{\rm cut2}$, as discussed in Section 3.2.3. Error statistics employing both $R_{\rm cut1}$ and $R_{\rm cut2}$ are summarized in Table 3.1. For the $(n_r, n_w) = (7, 1)$ and $(8, 1)$ schemes, errors converge by $R_{\rm cut2} = 7$ Å, and they converge to values not worse than what we encountered prior to introducing $R_{\rm cut2}$ (see Table 3.1).

Figure 3.6 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 $R_{\rm cut2} = 7$ Å. At the three-body level, the errors are reduced for the (6,1) and (7,1) schemes as compared to results where the $R_{\rm cut2}$ threshold is absent. At the four-body level, (7,1)

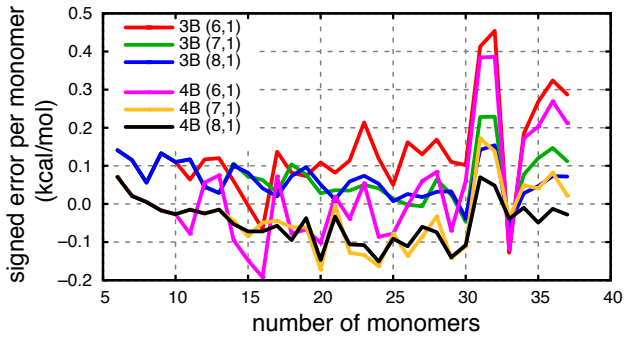

Figure 3.6: Signed errors per monomer for three- and four-body approximations to the total interaction energies, in conjunction with MBCP(2) counterpoise corrections, for $(H_2O)_N$ clusters. Various (n_r, n_w) combinations are used, with $R_{\text{cut2}} = 7$ Å in each case.

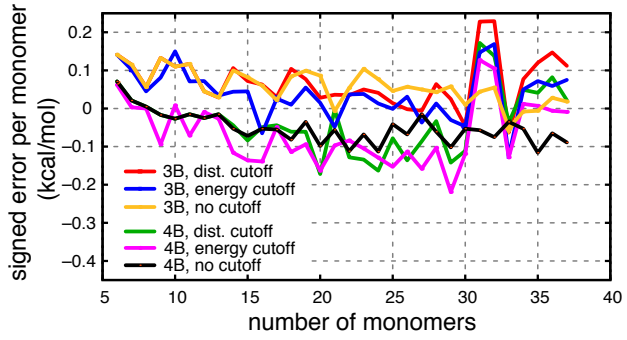

Figure 3.7: Signed errors per monomer in three- and four-body approximations to the total interaction for $(H_2O)_{6-37}$, including MBCP(2) corrections. The "distance cutoff" results use the thresholds $(n_r, n_w) = (7, 1)$ along with $R_{\text{cut2}} = 7$ Å. The "energy cutoff" results do not employ any distance-based thresholding, but discard all trimers whose interaction energies are < 0.25 kJ/mol and all tetramers whose interaction energies are < 0.10 kJ/mol.

results with $R_{\text{cut2}} = 7$ Å are close to the target accuracy of 0.09 kcal/mol/monomer. Therefore, in Fig. 3.7 we examine (7,1) results with $R_{\text{cut2}} = 7$ Å 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.60, for the energy-based scheme we discard all trimers whose interaction energies are < 0.25 kJ/mol and all tetramers whose interaction energies are < 0.10 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 counterpoise-corrected supersystem calculation, are not much larger than 0.2 kcal/mol/monomer for any of the clusters examined here.

Energy cutoffs

Previous results of energy cutoffs is reverse-engineered meaning that we've already calculated all possible interactions. Realistically, we need a model to predict the significant interactions. Therefore, we adopt effect fragment potential[94] (EFP) which is a computationally inexpensive of modeling interaction energies in non-bonded systems. To understand why the energy thresholding works, we examine four different clusters in Figure 3.8. They share the same trend that when we gradually discard non-additive interactions with small contributions, further stabilization in interaction energies is observed and the dynamical accuracy can be obtained except hydrofluoric acid clusters where higher order interactions play an important role in stabilization.

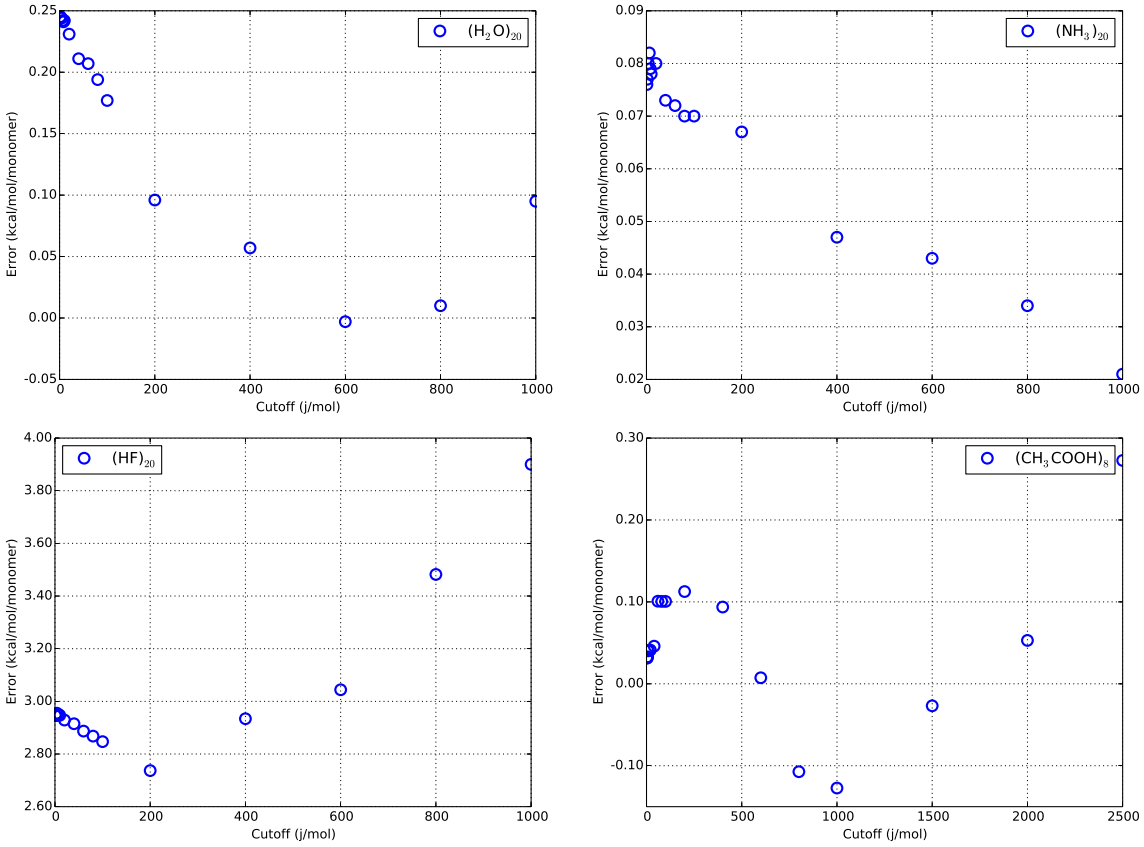

Figure 3.8: Errors in interaction energies with respect to supersystem benchmark at the level of MP2/aug-cc-pVDZ for selected noncovalent clusters.

The investigation of the correlation plots for the EFP and QM calculations in Figure 3.9 shows that the EFP is a good model to predict interaction energies. Further analysis of sum of three-body interactions, Figure 3.10, indicates that further stabilization occurs when small contributions are neglected.

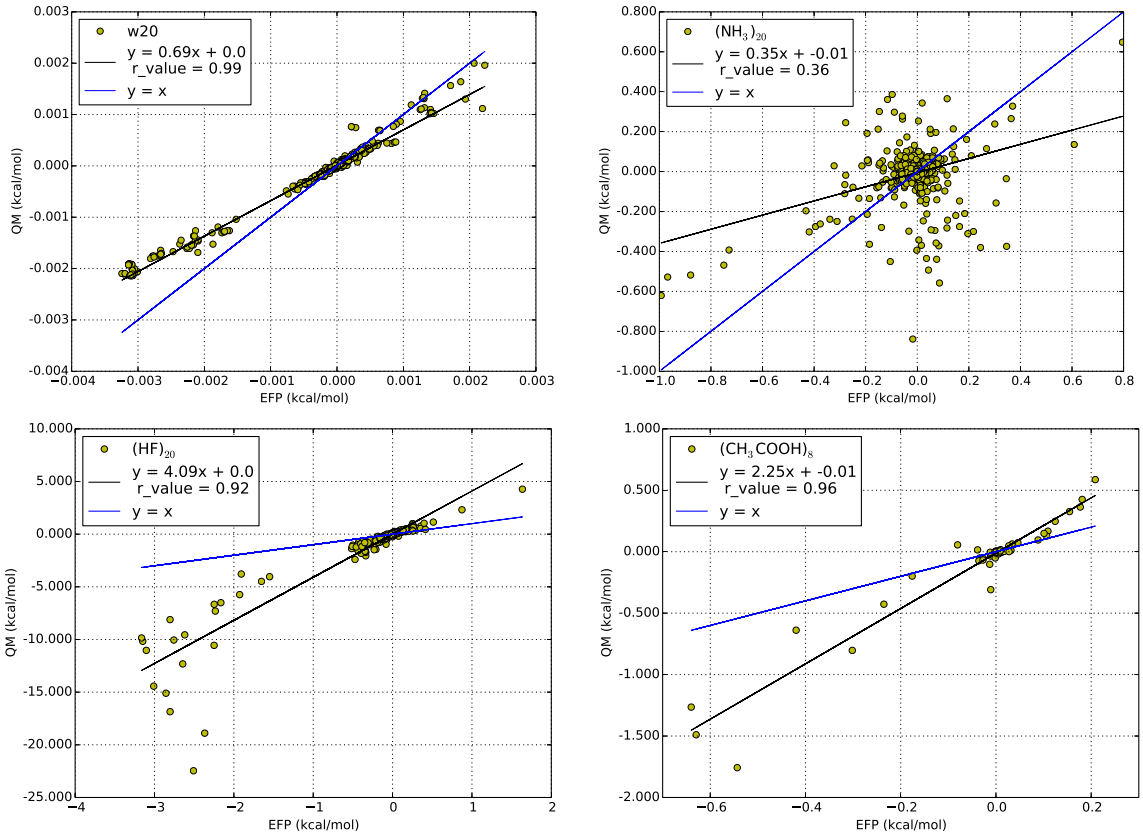

Figure 3.9: Correlation plots of three body interactions calculated by MP2 and the EFP.

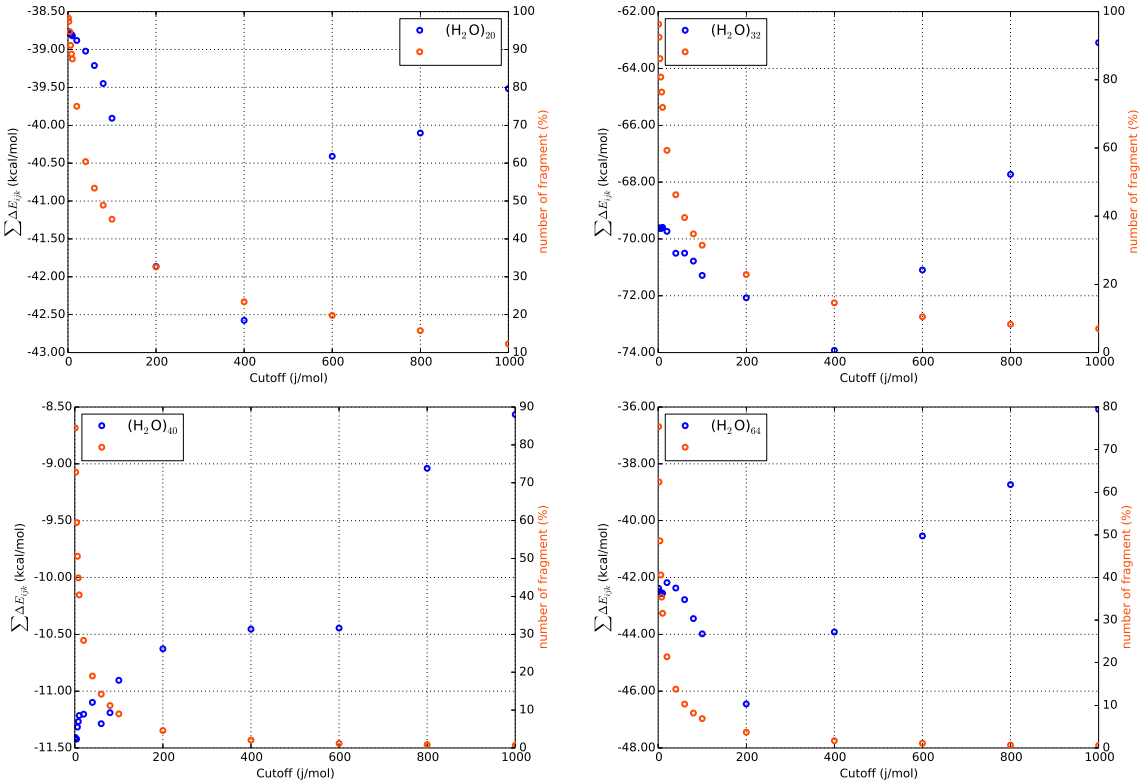

Figure 3.10: Sum of three body interactions using different cutoffs for selected water clusters. The second y-axis represent the number of fragment required after screening.

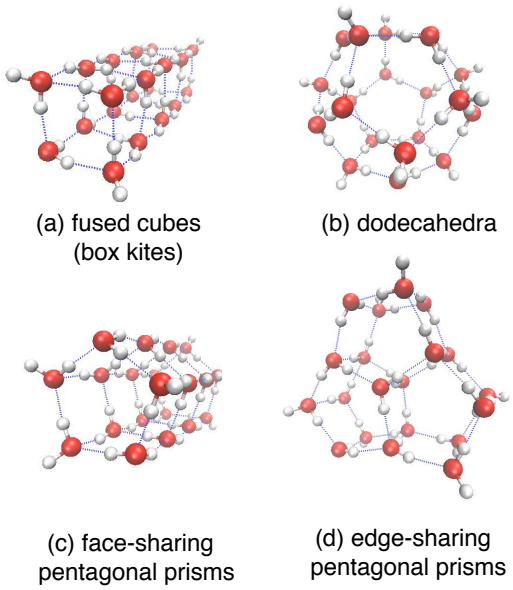

Figure 3.11: Examples of the four families of $(H_2O)_{20}$ isomers.

3.3.3 Relative energies

We next examine three- and four-body expansions as applied to predicting relative energies of $(H_2O)_{20}$ isomers. Cluster geometries, consisting of twenty low-energy isomers each from the four families of isomers on the $(H_2O)_{20}$ potential surface, were taken from Ref.89 without further optimization. These structures have been used by us in previous work,[8,72] and examples of the four classes of isomers are depicted in Fig. 3.11. Benchmark energies were computed at the CP-corrected ωB97X-V/aTZ level, and error with respect to these benchmarks is defined as

$$\text{error} = E_{\text{rel}}^{n\text{-body}} - E_{\text{rel}}^{\text{supersys}} \ . \tag{3.19}$$

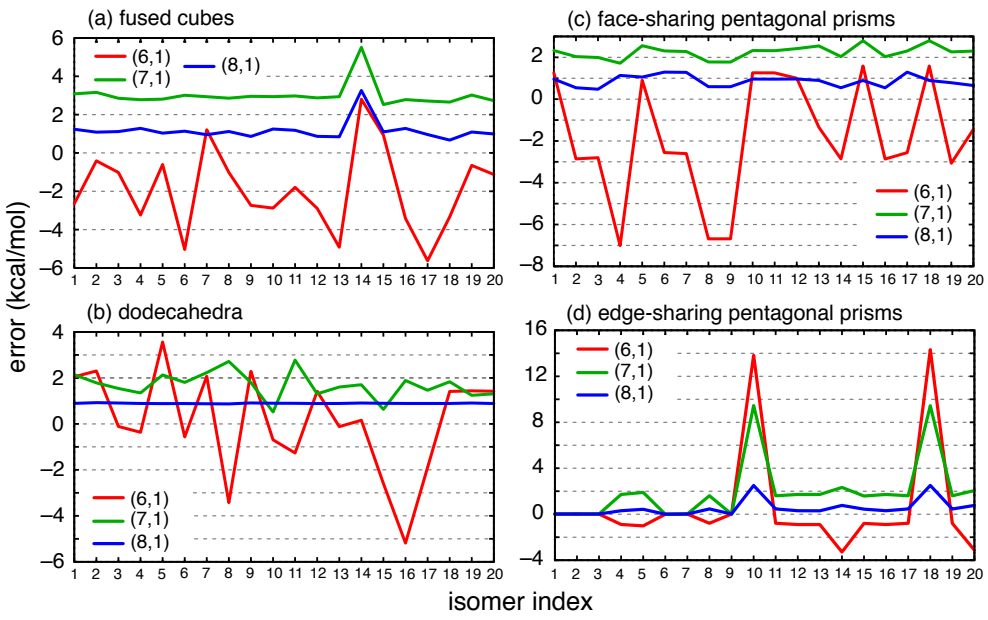

Figure 3.12: Signed errors for relative energies of $(H_2O)_{20}$ cluster isomers, employing MBE(4)+MBCP(2) and various (n_r, n_w) thresholds. Energies were computed at ωB97X-V/aTZ level. Each panel presents data for a different family of isomers (see Fig. 3.11), but all 80 isomers are plotted on a common energy scale even thought the vertical axes differ between panels.

Both energies in Eq. (3.19) 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. 3.12, using (n_r, n_w) cutoffs but not the R_{cut2} threshold, as the 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 B.2 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 prisms motif, where the 8–9 Å sub-clusters contribute -4.42 and -4.11 kcal/mol, respectively, to the total interaction energy, whereas this value does not exceed -0.45 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 14, although the 8–9 Å interactions are still ≈ 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 $(H_2O)_{20}$ clusters are too small to benefit from the alternative R_{cut2} threshold introduced above, so to improve the results we turn to two other *ad hoc* strategies described in Section 3.2.3. 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. (3.14). Results in Fig. 3.13 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 ≈ 0.5 kcal/mol

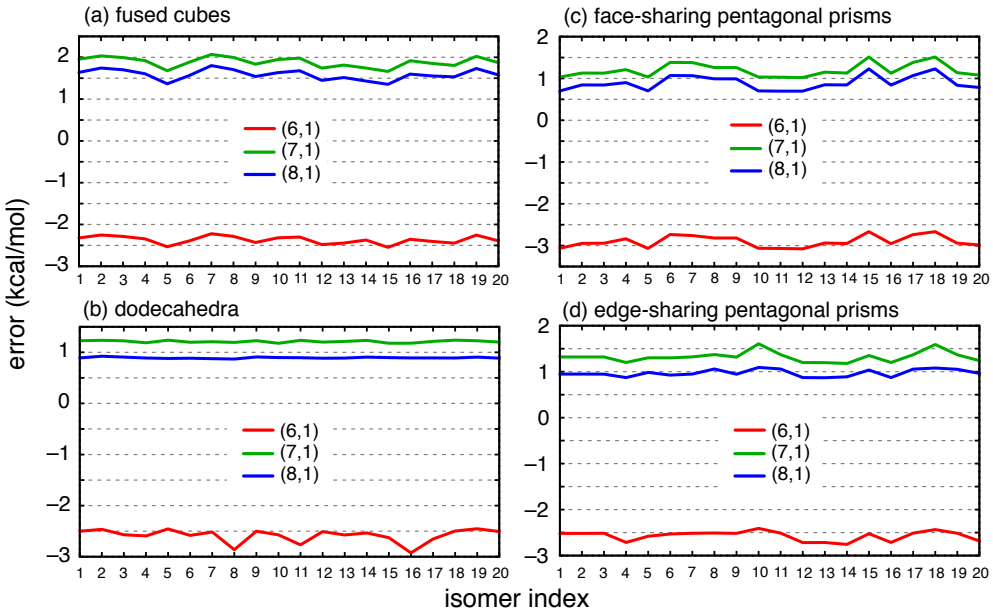

Figure 3.13: Signed errors for relative energies of $(H_2O)_{20}$ cluster isomers, employing MBE(4)+MBCP(2) and various (n_r, n_w) thresholding schemes. Energies up to the R_{cut1} cutoff were computed at ωB97X-V/aTZ level and supplemented with HF/aTZ for the long-range interactions, according to Eq. (3.14). Each panel presents data for a different family of isomers (see Fig. 3.11), but all 80 isomers are plotted on a common energy scale even thought the vertical axes differ between panels.

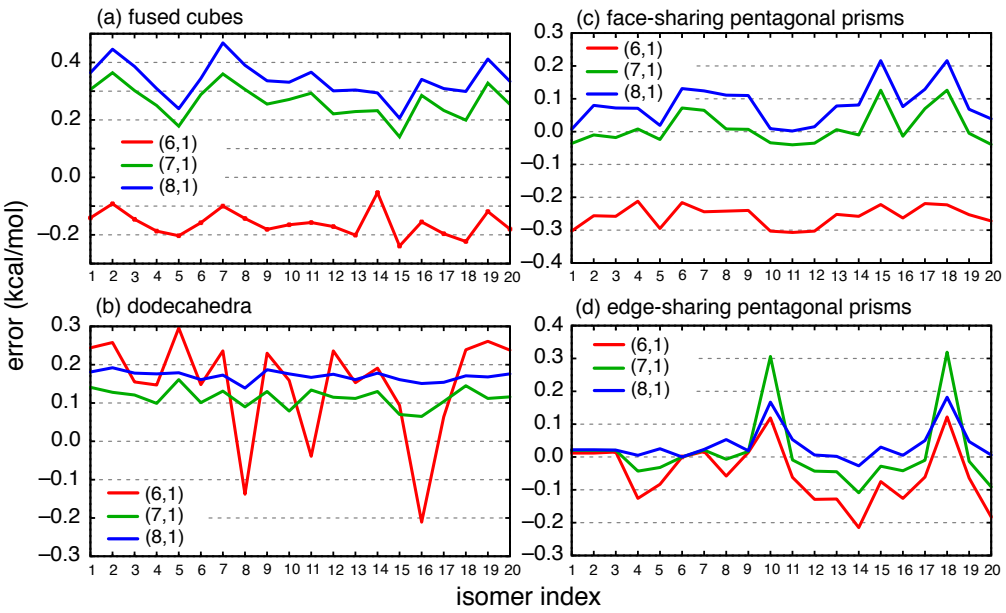

Figure 3.14: Signed errors for relative energies of $(H_2O)_{20}$ cluster isomers, employing MBE(4)+MBCP(2) and various (n_r, n_w) thresholds. Energies up to the R_{cut1} cutoff were computed at ωB97X-V/aTZ level and corrected using a HF/aTZ calculation for the entire supersystem, using the ONIOM-style correction in Eq. (3.15). Each panel presents data for a different family of isomers (see Fig. 3.11), but all 80 isomers are plotted on a common energy scale even thought the vertical axes differ between panels.

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 ∼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. (3.15)] using a DFT-based MBE as the high-level calculation (ωB97X-V/aTZ) and HF/aTZ as a low-level supersystem

calculation. Results are shown in Fig. 3.14 and are extremely accurate in comparison to either of the previous two approaches. (Note the much smaller energy scale in Fig. 3.14 versus either of Figs. 3.12 or 3.13.) 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 1,160 subsystems for $n = 2$, 3, and 4, respectively, as compared to 190, 1,140, and 4,845 subsystems when no thresholds are employed. For the (8,1) scheme, very few subsystems can be neglected in $(H_2O)_{20}$. Granted, this reduction comes at the expense of introducing a single supersystem 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 $N \to \infty$. (We return to this issue, with timings, in Section 3.3.4.)

Finally, we revisit the relative energies of the $(H_2O)_{20}$ isomers examined in Ref.72. New data at the ωB97X-V/aTZ level are plotted in Fig. 3.15, using an (8,1) cutoff scheme. Although δE^{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, $\delta E^{\mathrm{CP}} \approx 2.55$ kcal/mol for dodecahedral isomers, so this correction matters at the level of ≈ 0.25 kcal/mol when trying to establish the energies of the dodecahedra

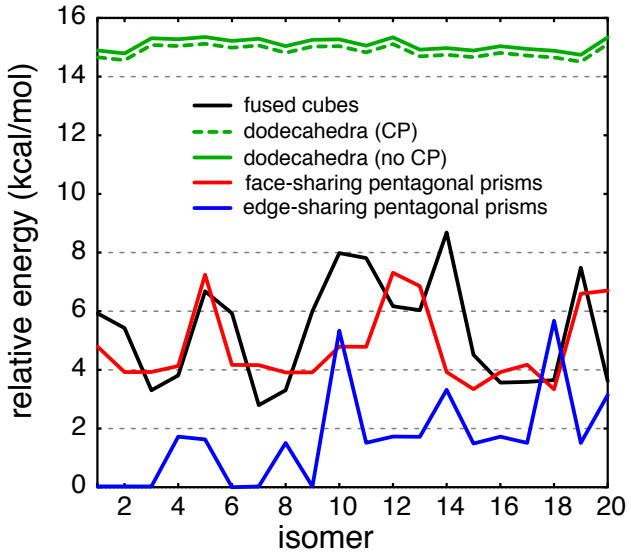

Figure 3.15: Relative energies of twenty isomers from each of four motifs of $(H_2O)_{20}$, computed at the ωB97X-V/aTZ level using the (8,1) cutoff scheme. Except for the dodecahedral isomers, the difference between CP-corrected and uncorrected results is indistinguishable within the thickness of the lines.

Table 3.2: Number of subsystem required for an MBE(4) calculation on the $(H_2O)_{37}$ cluster considered here, using the (7,1) thresholding scheme for R_{cut1} with and without $R_{\text{cut2}} = 7$ Å. The number of subsystems required for energy-based thresholding (E_{cut}) is also shown.

subsystem	full	R_{cut1}	$R_{\text{cut1}}+$ R_{cut2}	$E_{\text{cut}}{}^{a}$
monomers	37	37	37	37
dimers	666	504	504	666
trimers	7,770	3,751	5,141	908
tetramers	66,045	17,856	38,278	999
total	74,518	22,310	43,923	2,573

[a]The energy-based scheme does not cull monomers or dimers

relative to those of the other isomers. We observed the same phenomenon at the MP2 level in previous work,[72] 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 δE^{CP} was a bit larger than 1 kcal/mol in those previous calculations,[72] consistent with the observation that BSSE is typically larger in post-Hartree–Fock calculations as compared to DFT calculations.)

3.3.4 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 $(H_2O)_{37}$, using (7,1) thresholds with

or without $R_{\text{cut2}} = 7$ Å, are listed in Table 3.2. The reduction is quite dramatic when R_{cut2} is *not* considered, but only moderate when it is. Similar trends are reflected in Fig. 3.16, 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 $(H_2O)_{6-37}$. 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 $R_{\text{cut2}} = 7$ 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. 60, 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.

Actual timing data for a supersystem and various MBE(4) calculations on $(H_2O)_{37}$,

Table 3.3: Timing data for MBE(4) calculations of $(H_2O)_{37}$ (without CP corrections) at the B3LYP/aDZ level using the (7,1) thresholding scheme for R_{cut1} with and without $R_{\text{cut2}} = 7$ Å. These are the same calculations as used to count the number of subsystems in Table 3.2. Wall times reflect the cost to run on a single 28-core node,[?] so except for the supersystem calculation the wall time should decrease linearly with the number of nodes.

Method	time (hours)	
	CPU	wall
MBE(4), no cutoffs	1,601.7	58.4
MBE(4), R_{cut1}	496.3	18.1
MBE(4), $R_{\text{cut1}} + R_{\text{cut2}}$	951.4	34.6
MBE(4), E_{cut}	252.1	9.4
supersystem	15.4	0.9

at the B3LYP/aDZ level and without CP corrections, are presented in Table 3.3. These calculations reflect the subsystem counts that appear in Table 3.2. All calculations were threaded across 28 processors within a single node, and we note that the ratio of CPU time to wall time is ≈ 27 for each of the MBE(4) calculations, indicating near-perfect parallel scalability across a single node. (The parallel speedup is only about 17× for the supersystem calculation.) Note also that the wall times reported in Table 3.3 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 1,517 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

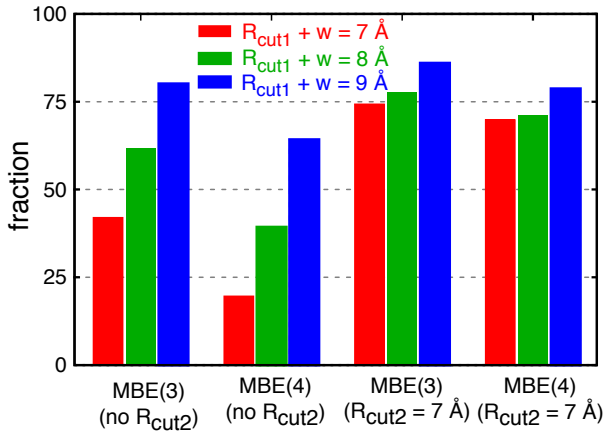

Figure 3.16: Fraction of subsystem calculations required for various MBE(n) approximations and thresholding schemes, averaged over $(H_2O)_{N=6-37}$. Note that any (n_r, n_w) combination with the same value of $n_r + n_w$ results in the same subsystems, so we label the cutoffs in terms of $R_{\text{cut1}} + w$.

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

Table 3.4: Timing data (in hours) for $(H_2O)_{20}$, edge-sharing pentagonal prism isomer 10, with all calculations multithreaded across a single 28-core node.[?] For the short-range DFT + long-range HF method of Eq. (3.14), the total time is the sum of the two MBE(4) timings (HF + DFT), whereas the ONIOM-style method in Eq. (3.15) also includes the supersystem HF time.

Method	(6,1)		(7,1)		no thresholds	
	CPU	wall	CPU	wall	CPU	wall
MBE(4), HF[a]	280.4	12.4	630.7	27.2	738.4	35.4
MBE(4), DFT[b]	424.6	17.4	970.1	39.0	1291.5	51.5
supersystem, HF[a]	—	—	—	—	39.0	2.0

[a]Hartree-Fock/aTZ

[b]ωB97X-V/aTZ

calculation with a low-level supersystem calculation can afford useful results.[52,55,56,87]

To put this in perspective, Table 3.4 shows timing data for calculations one one isomer of $(H_2O)_{20}$ 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.[?] As in the $(H_2O)_{37}$ example, wall times for the MBE(4) calculations should decrease linearly with the number of nodes. The supersystem HF/aTZ calculation takes 2.0 hours on a single node, as compared to 17.4 hours for a MBE(4) calculation at the ωB97X-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. The latter is not an outrageous number, but shows that for calculations of this size (1,840 basis functions), the supersystem calculation need not be an overwhelming bottleneck.

3.4 Conclusion

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 $(H_2O)_{6-37}$ and for relative energies of many different isomers of $(H_2O)_{20}$. To achieve an accuracy better than 0.1 kcal/mol/monomer (*i.e.*, 10% of k_BT at $T = 300$ K) in the total interaction energy, without simply relying on error cancellation, requires a four-body expansion with counterpoise corrections, although 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 counterpose corrections. However, only fairly conservative distance-based thresholds suffice to achieve this level of accuracy, resulting in only about a 30% reduction in the number of subsystem calculations required.

Nevertheless, this work demonstrates that routine four-body calculations *are* feasible, and also accurate, without resort to charge embedding that both hinders the parallelism and also complicates the formulation of analytic energy derivatives. We document significant decreases in both the total CPU time and the time-to-solution (wall time), relative to supersystem calculations, using as few as 28 processors, with a method that is essentially perfectly scalable due to the lack of any self-consistent embedding charges. This approach is stable even in large, augmented basis sets such as aug-cc-pVTZ, for which some fragment-based methods that employ embedding can experience problems.[95,96] Preliminary results using an energy-based thresholding

scheme suggest that this approach may achieve a far more dramatic reduction in the number of calculations, if the subsystem energies can be estimated *a priori* by means of classical multipole approximations.[60]

CHAPTER 4

Accuracy of finite-difference harmonic frequencies in density functional theory

In this chapter, we jump few steps ahead to pave the way for the GMBE energy responses. As analytical gradient is not always available, we test the robustness of finite-difference approach here. Analytic Hessians are often viewed as essential for the calculation of accurate harmonic frequencies, but the implementation of analytic second derivatives is non-trivial and solution of the requisite coupled-perturbed equations engenders a sizable memory footprint for large systems, given that these equations are not required for energy and gradient calculations in density functional theory. Here, we benchmark the alternative approach to harmonic frequencies based on finite differences of analytic first derivatives, a procedure that is amenable to large-scale parallelization. Not only for absolute frequencies but also for isotopic and conformer-dependent frequency shifts in flexible molecules, we find that the finite-difference approach exhibits mean errors < 0.1 cm^{-1} as compared to results based on an analytic Hessian. For very small frequencies corresponding to non-bonded vibrations in non-covalent complexes (for which the harmonic approximation is questionable anyway), the finite-difference error can be larger, but even in these cases the errors can

be reduced below 0.1 cm^{-1} by judicious choice of the displacement step size and a higher-order finite-difference approach. The surprising accuracy and robustness of the finite-difference results suggests that availability of the analytic Hessian is not so important in today's era of commodity processors that are readily available in large numbers.

4.1 Introduction

In quantum chemistry, analysis of harmonic vibrational frequencies provides important information about the stability of structures located via geometry optimization and serves as a first point of contact with vibrational spectroscopy. The conventional wisdom has long held that the "proper" (and most accurate) way to compute harmonic frequencies is to derive and implement analytic second derivatives of the energy with respect to displacements of the nuclei, *i.e.*, the analytic Hessian. This exercise is non-trivial, however, even at the level of density functional theory (DFT), to which we limit the following discussion. Calculation of the analytic Hessian requires second functional derivatives $\delta^2 E_{\mathrm{xc}}/\delta\rho^2$ whereas energy and gradient calculations require only first derivatives, $\delta E_{\mathrm{xc}}/\delta\rho$. Solution of so-called coupled-perturbed equations is also required,[97] engendering a memory footprint of $\mathcal{O}(N_{\mathrm{basis}}^2 N_{\mathrm{atoms}})$. Although this footprint can be split into segments across batches of atoms,[98] reducing the memory requirement by a factor of $N_{\mathrm{atoms}}/N_{\mathrm{segments}}$, two-electron integrals must be recomputed for each segment. Derivation and implementation of analytic Hessians for correlated wave function methods is even more involved.

The finite-difference (FD) approach, in contrast, is simple and parallelizes trivially, with different displacements performed on different processors and without the need (at the DFT level) to solve memory-intensive coupled-perturbed equations. This is potentially important in situations where a large number of processors are available but come with severe limits on wall time, a configuration that is often encountered at supercomputer centers. In addition, to the best of our knowledge the analytic Hessian of the VV10 non-local correlation functional[99] has yet to be implemented in any quantum chemistry software, meaning that analytic Hessians are unavailable for several very promising new functionals such as ωB97X-V,[90] B97M-V,[100] and ωB97M-V.[101]

Historically, FD results have been viewed as inferior in quality to analytic Hessian results, and in some quantum chemistry applications this may indeed be the case. In this study, we set out to quantify the extent to which the FD approach can be trusted for harmonic vibrational frequencies computed using DFT. Not only are the absolute vibrational frequencies of interest, but also isotope- and conformer-dependent frequency shifts, as these are often the relevant observables in experimental vibrational spectroscopy.

4.2 Computational Details

Calculations were performed using the B3LYP, B3LYP-D3,[102] and ωB97X-D functionals,[103,104] as indicated below, for which analytic Hessians are available for comparison to FD results. The SG-1 quadrature grid[91] was used for all calculations. Geometries were optimized subject to convergence thresholds (in atomic units) of

1.0×10^{-6}, 1.2×10^{-3}, and 3.0×10^{-4} on the stepwise energy difference, the stepwise atomic displacement, and maximum component of the gradient, respectively. FD calculations were performed using a home-built driver code, FRAGME∩T, originally developed for fragment-based quantum chemistry calculations,[8,53,62,66,73] but which is readily adapted to the present purpose. The FRAGME∩T code is currently interfaced with several quantum chemistry programs including Q-CHEM,[92] GAMESS,[105], PSI4,[106] and NWCHEM;[107] Q-CHEM is used for all of the electronic structure calculations presented here.

Unless stated otherwise, the FD calculations presented herein use the traditional three-point stencil,

$$f''(x_0) = \frac{f'(x_0 + h) - f'(x_0 - h)}{2h} + \mathcal{O}(h^2) \ , \tag{4.1}$$

with a step size $h = 0.001$ Å. Here, $f'(x) = \partial E / \partial x$ represents the analytic energy gradient. For non-bonded modes, we also explore the use of a five-point stencil,

$$\begin{aligned} f''(x_0) = \frac{1}{12h}\Big[& -f'(x_0 + 2h) + 8f'(x_0 + h) \\ & - 8f'(x_0 - h) + f'(x_0 - 2h)\Big] + \mathcal{O}(h^4) \ . \end{aligned} \tag{4.2}$$

4.3 Numerical Results

Benchmark Data Sets

We first ask the simple question of how well the FD approach reproduces the vibrational frequencies themselves. We examine this question first using the F38 database of small-molecule vibrational frequencies,[108] which was designed to include a broad

range of vibrational frequencies for small molecules. Individual FD frequencies in Table 4.1, computed at two different levels of theory, exhibit excellent agreement with analytical frequencies, with mean unsigned errors (MUEs) of only 0.01 cm^{-1} and a maximum error of 0.03 cm^{-1}. Statistical results shown in Fig. 4.1 demonstrate that similar accuracy is obtained in various basis sets and at various levels of theory, with nearly all of the errors being < 0.1 cm^{-1}.

We also examine six of the large non-covalent complexes in the L7 data set,[2] whose structures are shown in Fig. 4.2 and which are bound primarily by dispersion. (Coordinates for the optimized structures are provided in the Supporting Information.) For such complexes, we expect low-frequency vibrations along intermolecular coordinates, which might be problematic for the FD approach. We compute FD frequencies using the B3LYP/6-311G** and B3LYP-D3/6-31+G* levels of theory, with overall error statistics for each complex listed in Table 4.2. Although the MUEs in the FD frequencies, when averaged over all vibrational modes, are small (1–2 cm^{-1}), such averaging hides the larger errors in the low-frequency modes. Maximum errors at the B3LYP-D3/6-31+G* level range up to 32 cm^{-1} for the C3A and C3GC complexes, corresponding in both cases to a wobbling mode of circumcoronene whose frequency is $\nu = 913$ cm^{-1} (C3A) and $\nu = 1078$ cm^{-1} (G3GC). The distribution of FD errors for two of these complexes can be found in the Supporting Information.

The frequencies quoted above are not *extremely* low, especially for complexes having numerous frequencies below 100 cm^{-1}, and indeed we obtained very accurate FD results for frequencies on the order of $\sim$1000 cm^{-1} in the F38 data set. Therefore

Table 4.1: Analytical frequencies and errors (ΔFD) in the FD result, in cm^{-1}, for the F38 data set.

Molecule	B3LYP/ 6-311G**		ωB97X-D aug-cc-pVTZ	
	analytical	ΔFD	analytical	ΔFD
C_2H_2	642.81	−0.01	764.12	−0.02
	642.81	−0.01	764.12	0.01
	773.55	−0.03	855.70	0.02
	773.55	−0.01	855.70	0.04
	2070.11	0.00	2085.36	0.00
	3420.37	−0.01	3421.20	0.02
	3523.31	−0.01	3529.42	0.01
CH_4	1340.04	0.00	1349.56	−0.01
	1341.36	0.00	1376.83	−0.01
	1341.85	0.00	1392.07	0.00
	1559.63	0.00	1585.72	0.00
	1560.26	0.00	1589.53	0.00
	3026.91	0.00	3034.65	−0.02
	3132.41	−0.01	3150.21	0.02
	3132.91	−0.01	3151.16	0.01
	3133.00	−0.01	3159.19	0.01
Cl_2	501.11	0.00	589.28	−0.01
CO_2	666.45	0.00	690.00	−0.01
	666.45	−0.01	690.00	0.00
	1376.38	0.00	1393.89	0.01
	2437.53	0.00	2435.03	0.01
N_2	2448.01	−0.01	2494.88	0.01
N_2O	607.38	−0.01	634.91	−0.01
	607.38	−0.01	634.91	−0.01
	1335.97	0.00	1359.73	0.00
	2356.09	0.00	2397.62	0.01
OH	3700.02	−0.03	3769.58	0.01
CO	2222.51	0.00	2245.50	0.01
F_2	984.85	−0.01	1094.45	0.01
H_2CO	1199.21	0.00	1199.61	0.02
	1270.24	0.01	1242.44	0.01
	1538.71	0.00	1504.97	0.00
	1825.92	−0.01	1841.80	0.01
	2869.79	0.00	2903.22	−0.02
	2919.26	0.00	2968.16	0.00
H_2	4418.58	−0.02	4431.91	0.00
H_2O	1636.14	0.01	1634.53	−0.01
	3814.26	−0.01	3888.69	0.01
	3910.36	−0.01	3996.34	0.01
HCN	787.26	0.00	843.42	−0.03
	787.26	0.00	843.42	0.00
	2201.69	0.01	2228.30	0.00
	3454.67	0.01	3455.38	0.02
HF	4125.36	−0.01	4158.40	0.01
NH_3	1073.31	0.01	1034.35	0.00
	1682.04	0.00	1668.94	−0.01
	1682.74	0.00	1688.55	−0.01
	3457.65	−0.01	3507.31	0.02
	3575.98	−0.01	3632.64	0.02
	3576.60	0.00	3634.94	0.01
Average		0.01		0.01

Table 4.2: Error statistics for finite-difference vibrational frequencies for complexes in the L7 data set.

Complex[a]	max. error (cm^{-1})		MUE[b] (cm^{-1})	
	B3LYP/ 6-311G*	B3LYP-D3/ 6-31+G*	B3LYP/ 6-311G*	B3LYP-D3/ 6-31+G*
C3A	4.00	32.43	0.59	2.62
C3GC	7.50	32.26	0.74	2.16
GCGC	−1.07	−4.52	0.07	1.19
GGG	−0.55	4.84	0.03	1.06
CBH	−2.60	−23.97	0.39	1.46
PHE	−2.40	−6.37	0.23	0.33
Average			0.38	1.47

[a]See Fig. 4.2. [b]Averaged over all vibrational modes.

the large FD errors in these L7 cases must reflect the flatness of the potential energy surface along the non-bonded vibrational modes, and one can reasonably argue that it is inappropriate to apply the harmonic approximation to these sorts of vibrations. This, combined with the accuracy of the FD approach for medium- to high-frequency modes, and its computational advantages in terms of low memory and ease of parallelization, lead us to conclude that the FD approach can be useful even in non-covalent complexes such as these.

In view of the larger FD errors for non-bonded modes, however, we have performed a systematic study of the effects of the FD displacement step size, h, in a more computationally-tractable non-bonded system, namely, the parallel-displaced, π-stacked isomer of the benzene dimer. In addition, we test both the three-point and five-point stencil algorithms, Eqs. (4.1) and (4.2). Results are shown in Table 4.3. For

Table 4.3: Error statistics in finite-difference vibrational frequencies for the parallel-displaced isomer of $(C_6H_6)_2$ for various finite-difference schemes.

Step size,	max. error (cm^{-1})		MUE[a] (cm^{-1})	
h (Å)	3-point[b]	5-point[c]	3-point[b]	5-point[c]
0.0100	3.30	−0.03	0.23	0.01
0.0050	0.90	−0.04	0.05	0.01
0.0010	−1.04	−1.22	0.13	0.17
0.0005	2.17	2.87	0.19	0.20
0.0001	6.63	6.63	0.59	0.59

[a]Averaged over all vibrational modes. [b]Eq. (4.1).
[c]Eq. (4.2).

step sizes $h \leq 0.001$ Å (*i.e.*, equal to or smaller than our default value), the five-point algorithm leaves the maximum FD error unchanged or even slightly increased. Due to the very flat nature of the potential energy surface along the mode in question, however, larger step sizes can be more successful, especially when used with the five-point algorithm. For $h = 0.01$ Å the five-point algorithm reduces the errors to the level obtained for F38, namely, < 0.1 cm^{-1}. Hence, even given the aforementioned caveat regarding the appropriateness of the harmonic approximation for non-bonded modes, it is possible to use the FD approach to reproduce even very small harmonic frequencies.

Conformation-Dependent Frequency Shifts

An important aspect of making contact between *ab initio* frequency calculations and experimental vibrational spectroscopy is the ability to capture the vibrational frequency shifts engendered by conformational changes in a molecule. We examine these here, for water clusters and for conformational isomers of several hydrocarbons.

Isomers of a flexible (tryptamine)$\cdots$(H_2O) complex are also considered below in the context of isotopic frequency shifts.

Vibrational frequencies in clusters $(H_2O)_{2-6}$ have been benchmarked in a previous study using CCSD(T) calculations.[109] The red-shifted hydrogen-bonded O–H stretching vibrations are found to be sensitive to the level of theory, with errors compared to CCSD(T) results that range from nearly zero to more than 100 cm^{-1}. In the present work, we wish to establish whether the FD approach can capture differences in the O–H frequencies for water molecules in different hydrogen-bonding environments. Average errors in *absolute* vibrational frequencies for water clusters are provided in Table 4.4.

At the level of B3LYP/6-311G**, maximum errors for $(H_2O)_{2-5}$ are 0.03, 0.01, 0.04, and 0.05 cm^{-1} for n = 2, 3, 4, and 5, respectively, and at the ωB97X-D/aug-cc-pVTZ level these maximum errors are 0.02, 0.03, 0.08, and 0.04 cm^{-1}. We also examine four different conformers of $(H_2O)_6$, for which we find no FD error larger than 1 cm^{-1}, and in that particular case, the outlier corresponds to the lowest vibrational frequency (ν = 37.71 cm^{-1}) rather than an O–H stretching mode. Average FD errors (Table 4.4) are 0.02 and 0.01 cm^{-1} for these water clusters.

Given the results for the small-molecule F38 database, it is safe to assume that the FD frequencies for a single H_2O molecule are quite accurate. As such, the FD errors in vibrational frequencies can be taken to be equivalent to the errors in the vibrational red shifts associated with hydrogen bond. These errors, for the O–H stretching modes, are listed in Table 4.5 and are < 0.1 cm^{-1} except for one instance

Table 4.4: Error statistics for finite-difference vibrational frequencies in water clusters.

Cluster	max. error (cm^{-1})		MUE[a] (cm^{-1})	
	B3LYP/ 6-311G**	ωB97X-D/ aug-cc-pVTZ	B3LYP/ 6-311G**	ωB97X-D/ aug-cc-pVTZ
$(H_2O)_2$	0.03	0.02	0.01	0.01
$(H_2O)_3$	0.01	0.03	0.01	0.01
$(H_2O)_4$	0.04	0.08	0.01	0.02
$(H_2O)_5$	0.04	0.04	0.01	0.01
$(H_2O)_6$ (book)	−0.98	0.03	0.03	0.01
$(H_2O)_6$ (cage)	0.04	0.07	0.01	0.01
$(H_2O)_6$ (prism)	−0.03	0.03	0.01	0.01
$(H_2O)_6$ (ring)	0.24	0.11	0.03	0.02
Average			0.02	0.01

[a]Averaged over all vibrational modes.

where the FD result deviates by 0.34 cm^{-1}. Errors of such small magnitude imply that the FD approach is capable of distinguishing subtle frequency shifts due to changes in the hydrogen-bonding environment of a particular water molecule.

The 1,2-diphenoxyethane (DPOE) molecule, (C_6H_5)–$O(CH_2)_2O$–(C_6H_5), is a model of a flexible bi-chromophore whose central aliphatic linkage serves as the repetitive unit of polyethylene or poly(ethylene oxide) polymers. The symmetries of the two most abundant conformational isomers of DPOE were previously determined to be C_2 and C_{2h}.[110] Analytic harmonic frequencies for the modes related to the aforementioned linkage are 1500.09, 1500.38, 1530.97, and 1532.74 cm^{-1} (C_2 symmetry) and 1519.54, 1522.64, 1534.76, and 1535.66 cm^{-1} (C_{2h} symmetry). The FD procedure reproduces not only the frequencies but also the frequency *shifts* quite faithfully, with errors in the shifts of only ∼0.01 cm^{-1}; see Table 4.6.

The cysteine residue's side chain is essential for protein structure due to its flexibility and ability to form disulfide bonds with other cysteine residues. The vibrational

spectroscopy of this molecule has been studied, and it is found that the S–H stretching frequency is quite sensitive to hydrogen bonding.[111] As a model system to study this effect, we selected conformational isomers of ethanethiol that are classified by the C^{α}–C^{β}–S–H dihedral angel: two local minima with angles of ∼60° (G isomer) and ∼180° (T isomer), and structures that represent local maxima along the torsional potential, with the angles of ∼0° (C isomer) and ∼120° (S isomer). Analytic frequencies ν_{SH} are 2832.26, 2696.91, 2826.27, and 2831.07 cm^{-1} for isomers G, T, C, and S, respectively. FD errors are again ≤ 0.01 cm^{-1} (Table 4.6), much smaller than the resolution needed to distinguish between these isomers using vibrational spectroscopy.

Isotopic Shifts

Isotopic substitution is an important means of assigning experimental vibrational spectra. The (tryptamine)$\cdots$(H_2O) complex provides an example that has conformational flexibility, with at least two conformers that are spectroscopically accessible in the gas phase, and isotopic frequency shifts (replacing H_2O with D_2O) have been measured.[3] The O–H stretching frequencies ν_1 and ν_2 are listed in Table 4.7, and shift from 3474.91 and 3491.04 cm^{-1} to 2553.44 and 2823.69 cm^{-1} upon deuteration (B3LYP/6-311G** level). Errors in the FD calculation of the isotopic frequency *shift* are a mere 0.01 cm^{-1} (ν_1) and 0.04 cm^{-1} (ν_2), at the level of B3LYP/6-311G**. The corresponding errors at the ωB97X-D/6-31+G* level are -0.03 and -0.02 cm^{-1}.

In contrast to the rather large frequency shifts upon deuteration, isotopic shifts for ^{35}Cl versus ^{37}Cl in tetrachlorodibenzo-*p*-dioxins (TCDDs, Fig. 4.4) amount to a mere 1–2 cm^{-1} in some cases.[112] The frequencies themselves, corresponding to stretching

modes involving Cl, are also much smaller, and shift from 330.14 and 331.63 cm^{-1} to 322.13 and 323.46 cm^{-1} in the case of 1,4,6,9-TCDD (B3LYP/6-311G** level). In 2,3,7,8-TCDD, there is only one mode that is clearly a C–Cl stretch is both isotopologues; this mode shifts from 328.36 to 322.76 cm^{-1} upon isotopic substitution. Despite these rather small shifts, the FD error in the calculated frequency shift is ≤ 0.01 cm^{-1} in magnitude for both molecules (see Table 4.7), such that the shift is clearly resolvable in the FD calculation.

Finally, high-resolution gas-phase spectra of SF_6 reveal isotopic shifts in the ν_3 and ν_4 fundamentals that range from a few cm^{-1} up to 17 cm^{-1}.[113] Calculated isotopic shifts agree well: -2.26 and -17.11 cm^{-1}. FD errors (Table 4.7) are ≤ 0.01 cm^{-1}.

Hydrogenase Active Site Model

Hydrogenase enzymes have attracted much attention because they use an H_2-based energy cycle rather than a CO_2-based cycle. Recently, a model of 5,10-methenyltetrahydromethanopterin hydrogenase (Hmd) has been studied with density functional theory (see Fig. 4.5),[114] with the results suggesting that charge transfer from Fe $3d$ orbitals into unoccupied orbitals can lead to variations in the observed C≡O stretching frequencies. The ligand binding process is coupled with protonation of a thiolate ligand, hence protonated structures were included in our analysis. Harmonic frequencies are computed at the B3LYP/cc-VTZ level of theory, but with g functions removed from Fe.

Errors in the two C≡O stretching frequencies are both 0.03 cm^{-1} for the resting state (Hmd–H_2O), and are 0.05 and 0.04 cm^{-1} for the protonated state. For Hmd–CO,

errors in the three C≡O stretching modes are 0.01, 0.02, and 0.00 cm^{-1}, and for the protonated species (Hmd–CO + H^+) they are 0.01, 0.01, and 0.06 cm^{-1}. Although the C≡O modes are the primary ones of experimental interest, error statistics for all frequencies of the model system in Fig. 4.5 are listed in Table 4.8. None of the errors exceed 0.65 cm^{-1}.

4.4 Conclusion

The finite-difference approach to harmonic frequencies was studied at the level of DFT, in the interest of obtaining a highly-parallelizable, low-memory approach that does not require derivation and implementation of analytic second derivatives. Perhaps contrary to established conventional wisdom, we find that finite-difference results differ from those obtained using an analytic Hessian by < 0.1 cm^{-1} in most cases. Even frequencies in the 500–1000 cm^{-1} range are accurately reproduced, as are frequency shifts arising either from conformational changes or isotopic substitution. Vibrational red-shifts in the O–H stretching modes of water clusters, due to changes in the hydrogen-bonding environment, are also accurately reproduced by the finite-difference approach. The only significant errors that we find are in low-frequency non-bonded modes in dispersion-bound complexes, where the potential surface is very flat. In these cases, our "standard" finite-difference approach, based on displacements of ±0.001 Å, results in errors as large as 32 cm^{-1}, but can be reduced to < 0.1 cm^{-1} by appropriate choice of the displacement in conjunction with a five-point stencil that are requires four energy gradient calculations per degree of freedom.

In view of their accuracy, easy parallelizability and low memory footprint, we see no reason not to recommend the finite-difference approach to DFT harmonic frequency calculations. This should extend harmonic analysis to cases where analytic Hessian calculations are cumbersome, intractable, or where the Hessian simply has not been implemented.

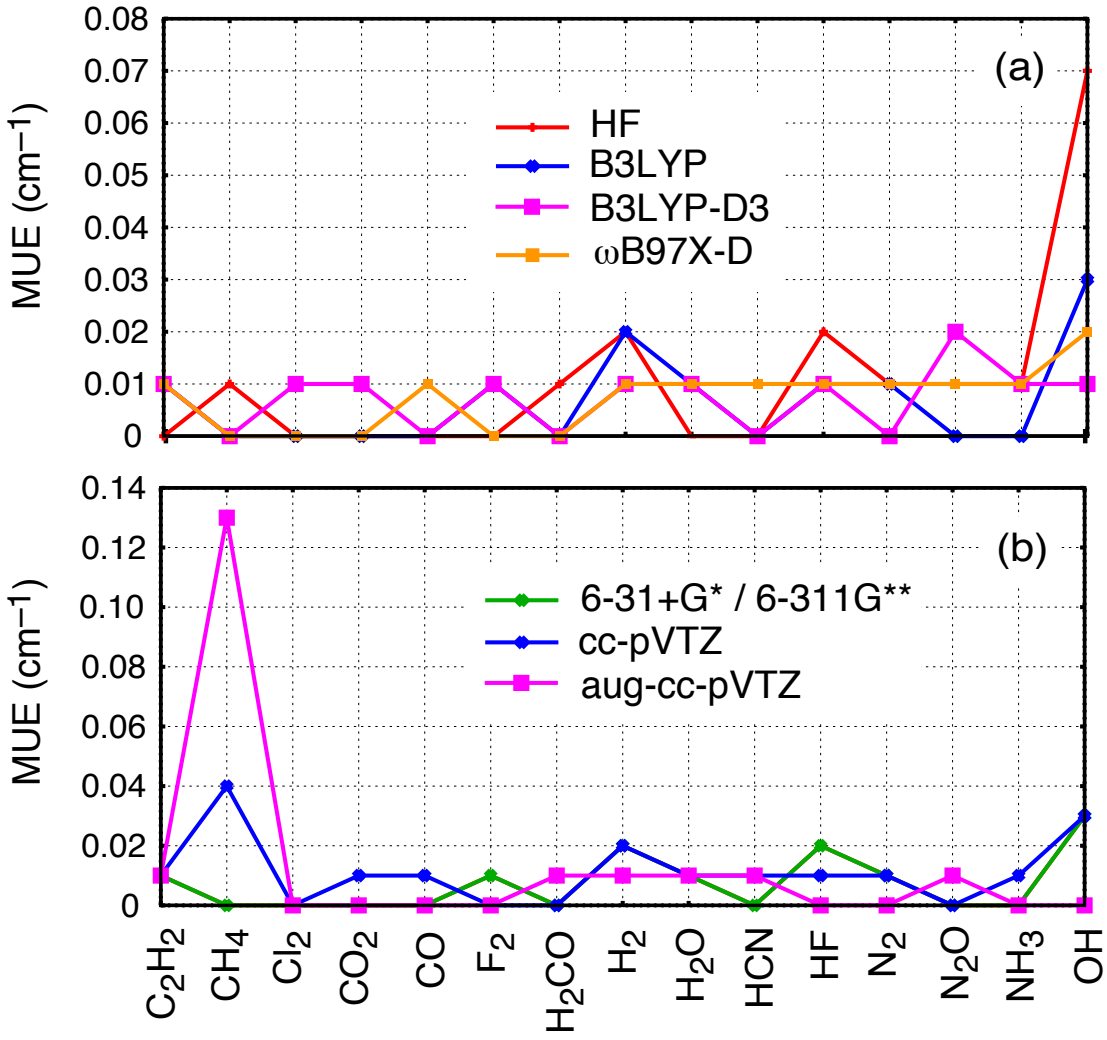

Figure 4.1: (a) MUEs for finite-difference errors for the F38 data set, averaged across five different theoretical models and all vibrational modes, with all calculations using the 6-311G** basis set. (b) MUEs for B3LYP finite-difference frequencies for F38 in various basis sets, averaged across all vibrational modes in each molecule. The 6-31+G* and 6-311G** results in (b) are indistinguishable on this scale.

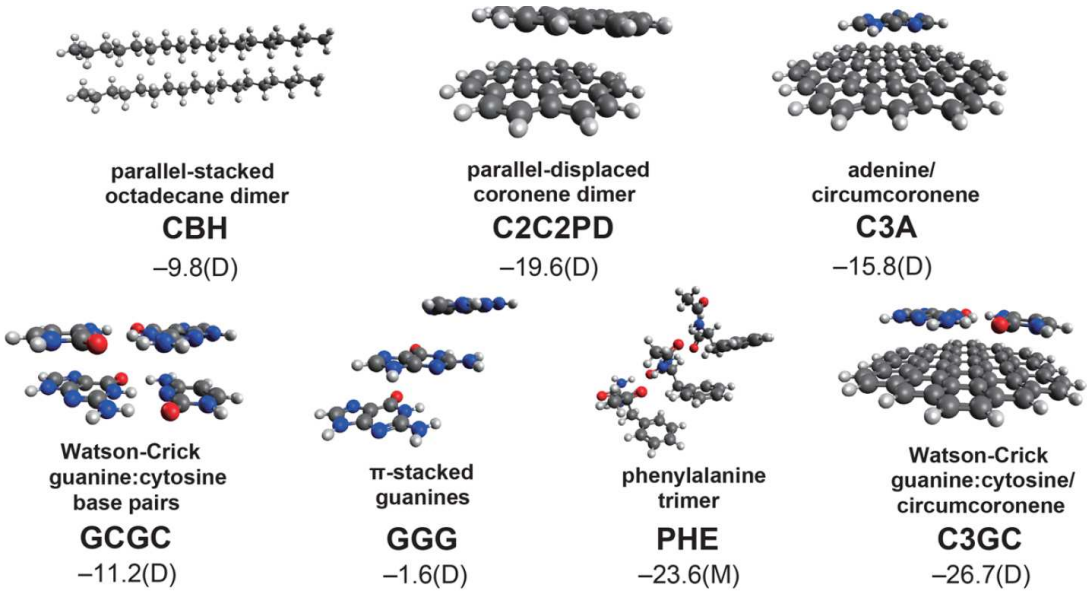

Figure 4.2: Complexes from the L7 data set of Ref. 2.

Table 4.5: Finite-difference errors (in cm^{-1}) for vibrational O–H red shifts in water clusters.

Cluster	B3LYP/ 6-311G**	ωB97X-D/ aug-cc-pVTZ
$(H_2O)_2$	0.01	0.02
$(H_2O)_3$	0.00	0.02
	0.01	0.01
	0.00	0.01
$(H_2O)_4$	0.02	0.01
	0.01	0.03
	0.01	−0.02
	0.01	0.00
$(H_2O)_5$	0.01	−0.01
	0.01	0.01
	0.01	0.01
	0.01	0.00
	0.01	−0.01
$(H_2O)_6$ (book)	0.01	0.01
	0.00	0.01
	0.01	0.02
	0.01	0.00
	0.01	0.00
	0.01	0.01
	0.01	0.00
$(H_2O)_6$ (cage)	0.00	0.01
	0.01	0.00
	0.00	−0.01
	0.00	−0.01
	0.01	−0.01
$(H_2O)_6$ (prism)	0.01	0.00
	0.01	0.01
	0.01	0.00
	0.01	0.00
	0.00	0.01
	0.00	0.01
	0.01	0.00
	0.00	0.00
	0.00	−0.01
$(H_2O)_6$ (ring)	0.01	0.01
	−0.06	0.03
	0.05	0.34
	0.08	0.01
	0.07	0.01
	−0.07	0.11

Table 4.6: Error statistics for finite-difference calculations of structure-dependent frequency shifts.

Molecule	error (cm^{-1})	
	max	MUE[a]
DPOE (C_{2h})	0.04	0.01
DPOE (C_2)	0.03	0.01
ethanethiol (C)	−0.07	0.01
ethanethiol (G)	−0.02	0.01
ethanethiol (S)	−0.01	0.01
ethanethiol (T)	−0.01	0.01

[a]Averaged over vibrational modes.

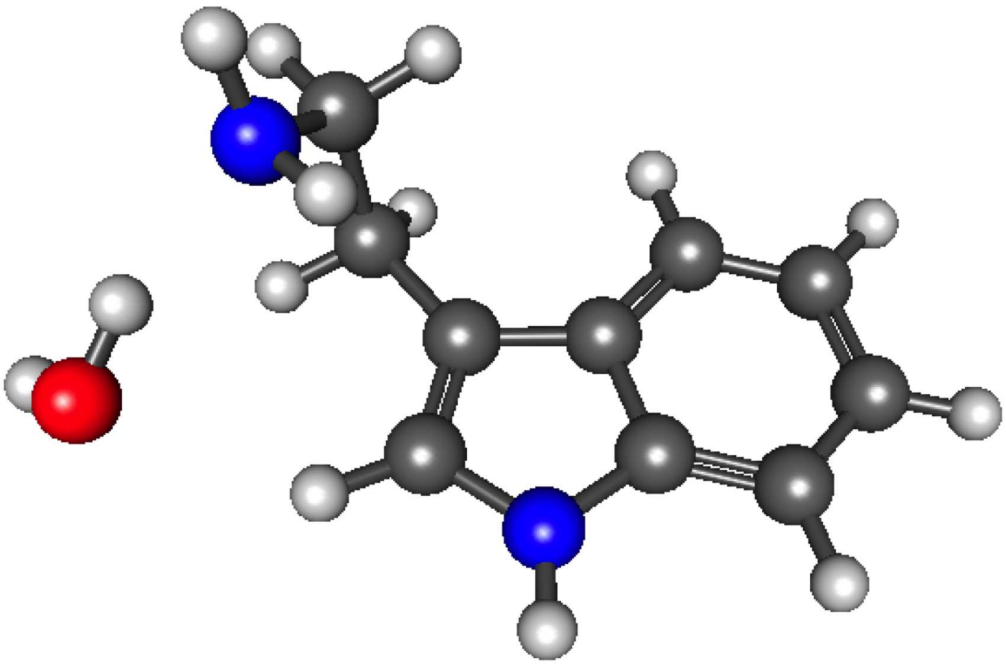

Figure 4.3: The (tryptamine)$\cdots$(H_2O) complex of Ref. 3.

Table 4.7: Errors (in cm^{-1}) in selected isotopic shifts.

System	B3LYP/ 6-311G**	ωB97X-D/ 6-31+G*
tryptamine + H_2O[a]		
ν_1	0.01	−0.03
ν_2	0.04	−0.02
1,4,6,9-TCDD[b]		
ν_1	−0.01	−0.01
ν_2	−0.01	−0.01
2,3,7,8-TCDD[b]		
ν_1	0.00	0.01
SF_6[c]		
ν_3	0.00	0.00
ν_4	0.01	0.00

[a] H_2O to D_2O. [b] ^{35}Cl to ^{37}Cl. [c] ^{32}S to ^{34}S.

Table 4.8: Error statistics for finite-difference harmonic frequencies in a model of the Hmd active site.[a]

System	error (cm^{-1})	
	max	MUE[b]
Hmd–H_2O	0.20	0.05
Hmd–H_2O + H^+	0.65	0.06
Hmd–CO	0.40	0.04
Hmd–CO + H^+	0.42	0.05

[a] B3LYP/cc-pVTZ level, with g functions removed from Fe. [b] Averaged over vibrational modes.

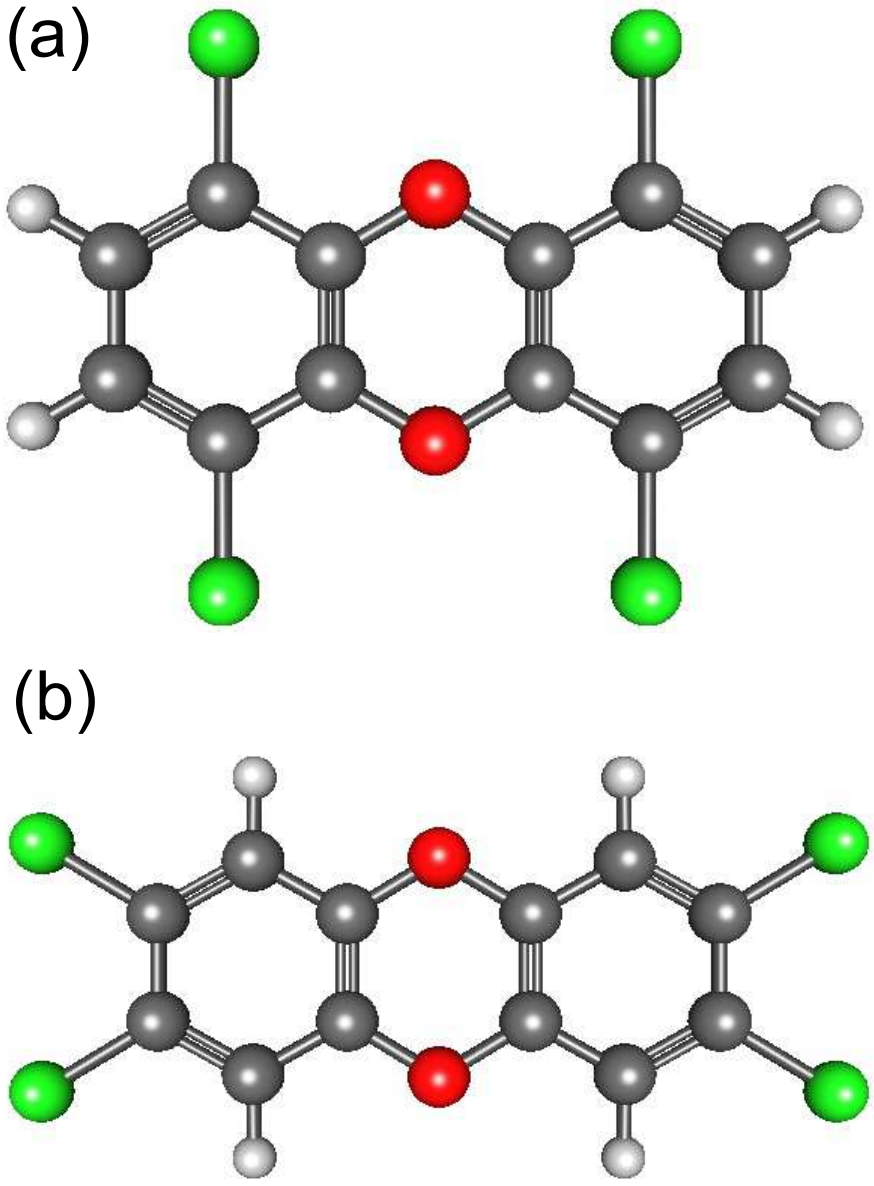

Figure 4.4: (a) 1,4,6,9-TCDD and (b) 2,3,7,8-TCDD.

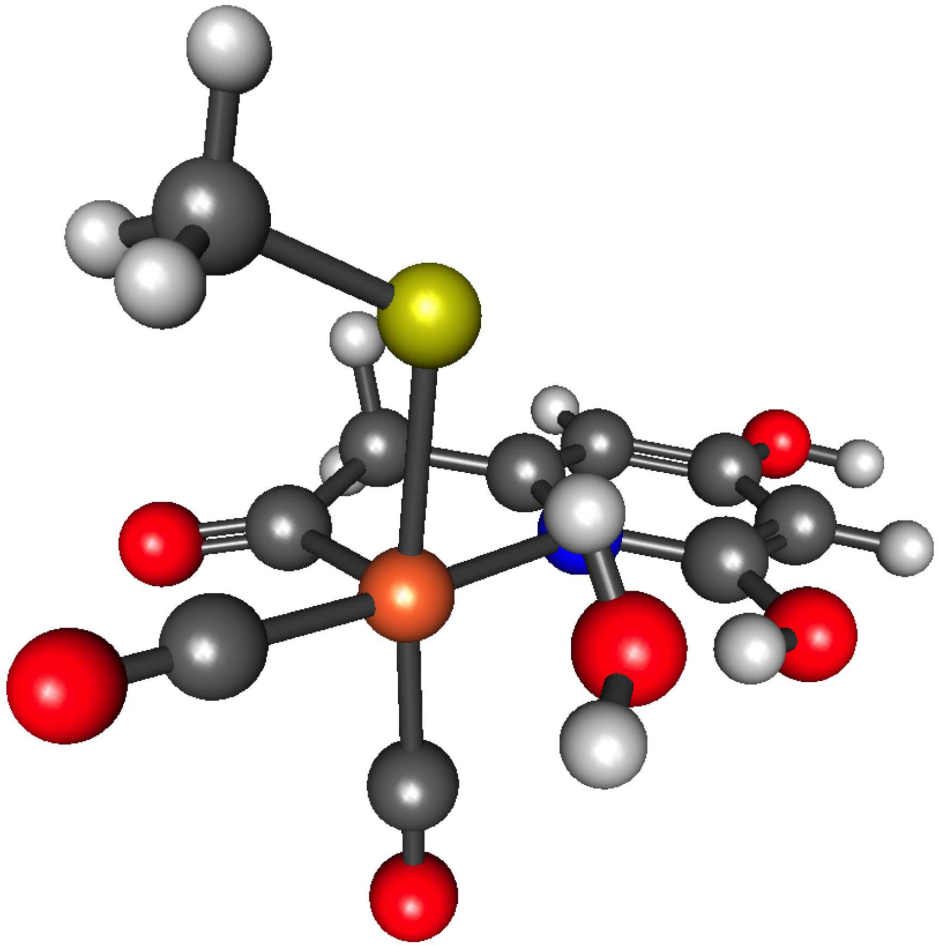

Figure 4.5: Model of the Hmd active site in its resting state, Hmd–H_2O, from Ref. 114.

CHAPTER 5

Conclusions and Outlook

In this work we have explored various aspects of the GMBE. The most prominent results would be the non-embedding method via screening through energy space and the embedding method in conjunction with the SAPT. With the energy thresholding, we can achieve linear scaling for non-additive three-body interactions and high accuracy for the total interaction energies. In other words, the cost for the total interaction energies would be close to $\mathcal{O}(N^2)$. The new charge embedding scheme, CM5, for the embedding method enable the XSAPT methodologies to calculate non-bonded interactions accurately and efficiently.

Despite the success that the GMBE can apply to larger systems and afford high accuracy, the GMBE still suffer from $\mathcal{O}(N^2)$ for the non-embedding method. Instead, one would apply distance-based thresholding to pairwise interactions combined with the energy-based threshold on non-additive interactions to reduce cost. Further studies are required to validate this approach. The potential of the GMBE would extend to larger systems with the capability of linear scaling.

The ultimate goal for many of *ab initio* methods is characterization and simulations of condensed-phase systems or bio systems. The GMBE provides essential

elements to reach the goal by reducing the unphysically exponential cost. Alongside the variational embedding, the XSAPT, as an interpretative method (energy decomposition anaylysis), more insights can be drawn from large systems.

Bibliography

[1] K. U. Lao and J. M. Herbert, J. Chem. Theory Comput. **14**, 2955 (2018).

[2] R. Sedlak et al., J. Chem. Theory Comput. **9**, 3364 (2013).

[3] J. R. Clarkson, J. M. Herbert, and T. S. Zwier, J. Chem. Phys. **126**, 134306:1 (2007).

[4] K. U. Lao, R. Schäffer, G. Jansen, and J. M. Herbert, J. Chem. Theory Comput. **11**, 2473 (2015).

[5] B. A, L. Shulenburger, N. A. Romero, J. Kim, and O. A. von Lilienfeld, J. Chem. Theory Comput. **10**, 3417 (2014).

[6] K. Carter-Fenk, K. U. Lao, K.-Y. Liu, and J. M. Herbert, J. Phys. Chem. Lett. **10**, 2706 (2019).

[7] K. U. Lao and J. M. Herbert, J. Phys. Chem. A **119**, 235 (2015).

[8] K. U. Lao, K.-Y. Liu, R. M. Richard, and J. M. Herbert, J. Chem. Phys. **144**, 164105:1 (2016).

[9] M. Stiborova et al., Cancer Res. **64**, 8374 (2004).

[10] M. N. Ucisik, D. S. Dashti, J. C. Faver, and K. M. Merz, Jr., J. Chem. Phys. **135**, 085101:1 (2011).

[11] K. Fukuzawa, K. Kitaura, K. Nakata, T. Kaminuma, and T. Nakano, Pure Appl. Chem. **75**, 2405 (2003).

[12] E. K. Kurbanov, H. R. Leverentz, D. G. Truhlar, and E. A. Amin, J. Chem. Theory Comput. **9**, 2617 (2013).

[13] T. Otsuka, N. Okimoto, and M. Taiji, J. Comput. Chem. **36**, 2209 (2015).

[14] R. M. Parrish, D. F. Sitkoff, D. L. Cheney, and C. D. Sherrill, Chem. Eur. J. **23**, 7887 (2017).

[15] D. G. Fedorov and K. Kitaura, Modeling and visualization for the fragment molecular orbital method with the graphical user interface FU, and analyses of protein–ligand binding, in *Fragmentation: Toward Accurate Calculations on Complex Molecular Systems*, edited by M. S. Gordon, chapter 3, pages 119–140, Wiley, Hoboken, 2017.

[16] Y. Wang, J. Liu, J. Li, and X. He, J. Comput. Chem. **39**, 1617 (2018).

[17] J. Choi et al., Sci. Rep. **8**, 13063:1 (2018).

[18] B. Thapa, D. Beckett, J. Erickson, and K. Raghavachari, J. Chem. Theory Comput. **14**, 5143 (2018).

[19] Y. Okiyama et al., J. Phys. Chem. B **123**, 957 (2019).

[20] L. D. Jacobson and J. M. Herbert, J. Chem. Phys. **134**, 094118:1 (2011).

[21] J. M. Herbert, L. D. Jacobson, K. U. Lao, and M. A. Rohrdanz, Phys. Chem. Chem. Phys. **14**, 7679 (2012).

[22] K. U. Lao and J. M. Herbert, J. Phys. Chem. Lett. **3**, 3241 (2012).

[23] K. U. Lao and J. M. Herbert, J. Chem. Phys. **139**, 034107:1 (2013), Erratum: *ibid.* **140**, 119901 (2014).

[24] W. Xie, L. Song, D. G. Truhlar, and J. Gao, J. Chem. Phys. **128**, 234108:1 (2008).

[25] R. Z. Khaliullin, E. A. Cobar, R. C. Lochan, A. T. Bell, and M. Head-Gordon, J. Phys. Chem. A **111**, 8753 (2007).

[26] E. G. Hohenstein and C. D. Sherrill, WIREs Comput. Mol. Sci. **2**, 304 (2012).

[27] K. U. Lao and J. M. Herbert, J. Chem. Phys. **140**, 044108:1 (2014).

[28] M. M. Francl and L. E. Chirlian, The pluses and minuses of mapping atomic charges to electrostatic potentials, in *Reviews in Computational Chemistry*, edited by K. B. Lipkowitz and D. B. Boyd, volume 14, chapter 1, pages 1–32, Wiley-VCH, New York, 2000.

[29] Z. C. Holden, R. M. Richard, and J. M. Herbert, J. Chem. Phys. **139**, 244108:1 (2013), Erratum: *ibid.* **142**, 059901:1–2 (2015).

[30] Z. C. Holden, B. Rana, and J. M. Herbert, J. Chem. Phys. **150**, 144115:1 (2019).

[31] A. V. Marenich, S. V. Jerome, C. J. Cramer, and D. G. Truhlar, J. Chem. Theory Comput. **8**, 527 (2012).

[32] F. L. Hirshfeld, Theor. Chem. Acc. **44**, 129 (1977).

[33] E. R. Davidson and S. Chakravorty, Theor. Chem. Acc. **83**, 319 (1992).

[34] S. Dasgupta and J. M. Herbert, J. Comput. Chem. **38**, 869 (2017).

[35] P. Jurečka, J. Šponer, J. Černý, and P. Hobza, Phys. Chem. Chem. Phys. **8**, 1985 (2006).

[36] J. Řezác, K. E. Riley, and P. Hobza, J. Chem. Theory Comput. **7**, 2427 (2011), Erratum: *ibid.* **10**, 1359–1360 (2014).

[37] J. Řezác and P. Hobza, J. Chem. Theory Comput. **8**, 141 (2012).

[38] R. M. Fogarty et al., J. Chem. Phys. **148**, 193817:1 (2018).

[39] R. Sure and S. Grimme, J. Chem. Theory Comput. **11**, 3785 (2015), Erratum: *ibid.* **11**, 5990 (2015).

[40] M. Katouda, A. Naruse, Y. Hirano, and T. Nakajima, J. Comput. Chem. **37**, 2623 (2016).

[41] D. J. Cole, C.-K. Skylaris, E. Rajendra, A. R. Venkitaraman, and M. C. Payne, Europhys. Lett. **91**, 37004:1 (2010).

[42] S. J. Fox, J. Dziedzic, T. Fox, C. S. Tautermann, and C.-K. Skylaris, Proteins **82**, 3335 (2014).

[43] W. Li, Z. Ni, and S. Li, Mol. Phys. **114**, 1447 (2016).

[44] H. J. Kulik, N. Luehr, I. S. Ufimtsev, and T. J. Martinez, J. Phys. Chem. B **116**, 12501 (2012).

[45] G. D. Fletcher, D. G. Fedorov, S. R. Pruitt, T. L. Windus, and M. S. Gordon, J. Chem. Theory Comput. **8**, 75 (2012).

[46] D. G. Fedorov, T. Ishida, M. Uebayasi, and K. Kitaura, J. Phys. Chem. A **111**, 2722 (2007).

[47] D. G. Fedorov, Y. Alexeev, and K. Kitaura, J. Phys. Chem. Lett. **2**, 282 (2011).

[48] N. Sahu, S. D. Yeole, and S. R. Gadre, J. Chem. Phys. **138**, 104101:1 (2013).

[49] N. Sahu and S. R. Gadre, Acc. Chem. Res. **47**, 2739 (2014).

[50] S. Li, W. Li, and J. Ma, Acc. Chem. Res. **47**, 2712 (2014).

[51] X. He, T. Zhu, X. W. Wang, J. F. Liu, and J. Z. H. Zhang, Acc. Chem. Res. **47**, 2748 (2014).

[52] A. Saha and K. Raghavachari, J. Chem. Theory Comput. **11**, 2012 (2015).

[53] J. Liu and J. M. Herbert, J. Chem. Theory Comput. **12**, 572 (2016).

[54] H. Nakata and D. G. Fedorov, J. Phys. Chem. A **120**, 9794 (2016).

[55] K. V. J. Jose and K. Raghavachari, Mol. Phys. **113**, 3057 (2015).

[56] K. V. J. Jose and K. Raghavachari, J. Chem. Theory Comput. **13**, 1147 (2017).

[57] J. D. Hartman and G. J. O. Beran, J. Chem. Theory Comput. **10**, 4862 (2014).

[58] J. D. Hartman, G. M. Day, and G. J. O. Beran, Cryst. Growth Des. **16**, 6479 (2016).

[59] G. J. O. Beran, J. D. Hartman, and Y. N. Heit, Acc. Chem. Res. **49**, 2501 (2016).

[60] J. F. Ouyang and R. P. A. Bettens, J. Chem. Theory Comput. **12**, 5860 (2016).

[61] J. Liu, L. Qi, J. Z. H. Zhang, and X. He, J. Chem. Theory Comput. **13**, 2021 (2017).

[62] R. M. Richard and J. M. Herbert, J. Chem. Phys. **137**, 064113:1 (2012).

[63] J. Gauss, Molecular properties, in *Modern Methods and Algorithms of Quantum Chemistry*, edited by J. Grotendorst, volume 3 of *NIC Series*, pages 541–592, John von Neumann Institute for Computing, Jülich, 2nd edition, 2000.

[64] G. A. Cisneros et al., Chem. Rev. **116**, 7501 (2016).

[65] J. Cui, H. Liu, and K. D. Jordan, J. Phys. Chem. B **110**, 18872 (2006).

[66] R. M. Richard, K. U. Lao, and J. M. Herbert, J. Chem. Phys. **141**, 014108:1 (2014).

[67] E. E. Dahlke and D. G. Truhlar, J. Chem. Theory Comput. **3**, 46 (2007).

[68] E. E. Dahlke and D. G. Truhlar, J. Chem. Theory Comput. **3**, 1342 (2007).

[69] E. E. Dahlke, H. R. Leverentz, and D. G. Truhlar, J. Chem. Theory Comput. **4**, 33 (2008).

[70] R. M. Richard, K. U. Lao, and J. M. Herbert, Acc. Chem. Res. **47**, 2828 (2014).

[71] G. D. Chen, J. Weng, G. Song, and Z. H. Li, J. Chem. Theory Comput. **13**, 2010 (2017).

[72] R. M. Richard, K. U. Lao, and J. M. Herbert, J. Chem. Phys. **139**, 224102:1 (2013).

[73] R. M. Richard and J. M. Herbert, J. Chem. Theory Comput. **9**, 1408 (2013).

[74] J. Friedrich et al., J. Phys. Chem. Lett. **5**, 666 (2014).

[75] D. Yuan et al., J. Chem. Theory Comput. **13**, 2696 (2017).

[76] R. M. Richard, K. U. Lao, and J. M. Herbert, J. Phys. Chem. Lett. **4**, 2674 (2013).

[77] J. F. Ouyang, M. W. Cvitkovic, and R. P. A. Bettens, J. Chem. Theory Comput. **10**, 3699 (2014).

[78] M. Kamiya, S. Hirata, and M. Valiev, J. Chem. Phys. **128**, 074103:1 (2008).

[79] J. F. Ouyang and R. P. A. Bettens, J. Chem. Theory Comput. **11**, 5132 (2015).

[80] S. F. Boys and F. Bernardi, Mol. Phys. **19**, 553 (1970).

[81] B. H. Wells and S. Wilson, Chem. Phys. Lett. **101**, 429 (1983).

[82] G. S. Tschumper, Reliable electronic structure computations for weak noncovalent interactions in clusters, in *Reviews in Computational Chemistry*, edited by K. B. Lipkowitz and T. R. Cundari, volume 26, chapter 2, pages 39–90, Wiley-VCH, 2009.

[83] P. Valiron and I. Mayer, Chem. Phys. Lett. **275**, 46 (1997).

[84] F. B. van Duijneveldt, J. G. C. M. van Duijneveldt-van de Rijdt, and J. H. van Lenthe, Chem. Rev. **94**, 1873 (1994).

[85] L. W. Chung et al., Chem. Rev. **115**, 5678 (2015).

[86] N. J. Mayhall and K. Raghavachari, J. Chem. Theory Comput. **8**, 2669 (2012).

[87] K. V. J. Jose, D. Beckett, and K. Raghavachari, J. Chem. Theory Comput. **11**, 4238 (2015).

[88] S. Kazachenko and A. J. Thakkar, J. Chem. Phys. **138**, 194302:1 (2013).

[89] D. J. Wales and M. P. Hodges, Chem. Phys. Lett. **286**, 65 (1998).

[90] N. Mardirossian and M. Head-Gordon, Phys. Chem. Chem. Phys. **16**, 9904 (2014).

[91] P. M. W. Gill, B. G. Johnson, and J. A. Pople, Chem. Phys. Lett. **209**, 506 (1993).

[92] Y. Shao et al., Mol. Phys. **113**, 184 (2015).

[93] S. Kazachenko and A. J. Thakkar, Chem. Phys. Lett. **476**, 120 (2009).

[94] M. S. Gordon et al., J. Phys. Chem. A **105**, 293 (2001).

[95] D. G. Fedorov, L. V. Slipchenko, and K. Kitaura, J. Phys. Chem. A **114**, 8742 (2010).

[96] D. G. Fedorov and K. Kitaura, Chem. Phys. Lett. **597**, 99 (2014).

[97] J. A. Pople, R. Krishnan, H. B. Schlegel, and J. S. Binkley, Int. J. Quantum Chem. Symp. **13**, 225 (1979).

[98] P. P. Korambath, J. Kong, T. R. Furlani, and M. Head-Gordon, Mol. Phys. **100**, 1771 (2002).

[99] O. A. Vydrov and T. Van Voorhis, J. Chem. Phys. **133**, 244103:1 (2010).

[100] N. Mardirossian and M. Head-Gordon, J. Chem. Phys. **142**, 074111:1 (2015).

[101] N. Mardirossian and M. Head-Gordon, J. Chem. Phys. **144**, 214110:1 (2016).

[102] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys. **132**, 154104:1 (2010).

[103] J.-D. Chai and M. Head-Gordon, J. Chem. Phys. **128**, 084106:1 (2008).

[104] J.-D. Chai and M. Head-Gordon, Phys. Chem. Chem. Phys. **10**, 6615 (2008).

[105] M. W. Schmidt et al., J. Comput. Chem. **14**, 1347 (1993).

[106] J. M. Turney et al., WIREs Comput. Mol. Sci. **2**, 556 (2012).

[107] M. Valiev et al., Comput. Phys. Commun. **181**, 1477 (2010).

[108] Y. Zhao and D. G. Truhlar, Acc. Chem. Res. **41**, 157 (2008).

[109] J. C. Howard and G. S. Tschumper, J. Chem. Theory Comput. **11**, 2126 (2015).

[110] E. G. Buchanan, E. L. Sibert, and T. S. Zwier, J. Phys. Chem. A **117**, 2800 (2013).

[111] W. Qian and S. Krimm, J. Comput. Chem. **32**, 1025 (1992).

[112] G. Rauhut and P. Pulay, J. Am. Chem. Soc. **117**, 4167 (1995).

[113] T. D. Kolomiitsova, V. A. Kondaurov, E. V. Sedelkova, and D. N. Shchepkin, Opt. Spectrosc. **92**, 512 (2002).

[114] A. Dey, J. Am. Chem. Soc. **132**, 13892 (2010).

APPENDIX A

Supplementary Material for"Intermolecular energy decomposition analysis in large supramolecular complexes using symmetry-adapted perturbation theory"

Table A.1: Errors in interaction energies for the S22 data set, at the XS-APT(KS)/hpTZVPP level.

	error (kcal/mol)	
	CM5	ChElPG
1 2-pyridoxine 2-aminopyridine complex	0.22	0.20
2 Adenine thymine complex stack	−0.19	−0.19
3 Adenine thymine Watson-Crick complex	0.94	0.92
4 Ammonia dimer	0.15	0.15
5 Benzene - Methane complex	0.08	0.08
6 Benzene ammonia complex	−0.11	−0.11
7 Benzene dimer parallel displaced	0.10	0.10
8 Benzene dimer T-shaped	0.02	0.02
9 Benzene HCN complex	−0.82	−0.82
10 Benzene water complex	−0.31	−0.31
11 Ethene dimer	0.10	0.07
12 Ethene ethyne complex	0.42	0.42
13 Formamide dimer	−1.12	−1.15
14 Formic acid dimer	−1.11	−1.09
15 Indole benzene complex stack	0.60	0.58
16 Indole benzene T-shape complex	−0.19	−0.19
17 Methane dimer	0.14	0.14
18 Phenol dimer	0.36	0.35
19 Pyrazine dimer	−0.08	−0.08
20 Uracil dimer h-bonded	0.97	0.97
21 Uracil dimer stack	−0.28	−0.29
22 Water dimer	−0.01	−0.01
MUE	0.38	0.38

Table A.2: Errors in interaction energies for the S66 data set, at the XSAPT(KS)/hpTZVPP level.

	error (kcal/mol)	
	CM5	ChElPG
1 Water ... Water	−0.14	−0.18
2 Water ... MeOH	−0.10	−0.11
3 Water ... MeNH2	0.13	0.09
4 Water ... Peptide	−0.01	−0.02
5 MeOH ... MeOH	−0.07	−0.07
6 MeOH ... MeNH2	0.01	0.02
7 MeOH ... Peptide	0.00	0.06
8 MeOH ... Water	−0.06	−0.14
9 MeNH2 ... MeOH	0.08	0.08
10 MeNH2 ... MeNH2	−0.04	−0.04
11 MeNH2 ... Peptide	−0.03	0.02
12 MeNH2 ... Water	−0.03	−0.01
13 Peptide ... MeOH	−0.01	0.01
14 Peptide ... MeNH2	−0.09	−0.06
15 Peptide ... Peptide	0.19	0.25
16 Peptide ... Water	0.10	0.07
17 Uracil ... Uracil (BP)	1.04	1.15
18 Water ... Pyridine	−0.10	−0.12
19 MeOH ... Pyridine	−0.15	−0.08
20 AcOH ... AcOH	−0.28	−0.22
21 AcNH2 ... AcNH2	−0.78	−0.63
22 AcOH ... Uracil	0.38	0.39
23 AcNH2 ... Uracil	−0.19	−0.19
24 Benzene ... Benzene (pi-pi)	−0.07	−0.07
25 Pyridine ... Pyridine (pi-pi)	−0.44	−0.43
26 Uracil ... Uracil (pi-pi)	−0.46	−0.40
27 Benzene ... Pyridine (pi-pi)	−0.20	−0.20
28 Benzene ... Uracil (pi-pi)	0.23	0.21
29 Pyridine ... Uracil (pi-pi)	0.09	0.08
30 Benzene ... Ethene	0.36	0.39
31 Uracil ... Ethene	0.07	0.04
32 Uracil ... Ethyne	0.29	0.23
33 Pyridine ... Ethene	0.23	0.25

Table A.3: Continued errors in interaction energies for the S66 data set, at the XS-APT(KS)/hpTZVPP level.

	error (kcal/mol)	
	CM5	ChElPG
34 Pentane ... Pentane	0.81	0.85
35 Neopentane ... Pentane	0.82	0.86
36 Neopentane ... Neopentane	0.79	0.83
37 Cyclopentane ... Neopentane	0.55	0.59
38 Cyclopentane ... Cyclopentane	0.22	0.25
39 Benzene ... Cyclopentane	−0.30	−0.28
40 Benzene ... Neopentane	0.09	0.13
41 Uracil ... Pentane	−1.03	−0.95
42 Uracil ... Cyclopentane	−1.11	−1.07
43 Uracil ... Neopentane	−0.54	−0.49
44 Ethene ... Pentane	0.57	0.55
45 Ethyne ... Pentane	0.15	0.14
46 Peptide ... Pentane	0.11	0.16
47 Benzene ... Benzene (TS)	0.06	0.07
48 Pyridine ... Pyridine (TS)	−0.44	−0.42
49 Benzene ... Pyridine (TS)	−0.10	−0.11
50 Benzene ... Ethyne (CH-pi)	−0.23	−0.26
51 Ethyne ... Ethyne (TS)	0.12	0.05
52 Benzene ... AcOH (OH-pi)	0.13	0.15
53 Benzene ... AcNH2 (NH-pi)	0.15	0.22
54 Benzene ... Water (OH-pi)	−0.31	−0.30
55 Benzene ... MeOH (OH-pi)	−0.16	−0.18
56 Benzene ... MeNH2 (NH-pi)	−0.10	−0.11
57 Benzene ... Peptide (NH-pi)	−0.31	−0.30
58 Pyridine ... Pyridine (CH-N)	0.26	0.28
59 Ethyne ... Water (CH-O)	0.02	−0.05
60 Ethyne ... AcOH (OH-pi)	0.65	0.64
61 Pentane ... AcOH	0.18	0.23
62 Pentane ... AcNH2	0.00	0.04
63 Benzene ... AcOH	−0.02	−0.01
64 Peptide ... Ethene	0.18	0.18
65 Pyridine ... Ethyne	−0.04	−0.04
66 MeNH2 ... Pyridine	−0.50	−0.47
MUE	0.26	0.27

Table A.4: Errors in interaction energies for the IHB data set, at the XSAPT(KS)/def2-TZVPPD level.

	error (kcal/mol)	
	CM5	ChElPG
1 acetate ... methanol	0.82	−1.54
2 acetate ... water	1.52	−3.34
3 acetate ... methylamine	1.24	0.50
4 methylammonium ... formaldehyde	−0.95	−1.48
5 methylammonium ... methylamine	−3.29	−4.22
6 methylammonium ... methanol	−0.90	−1.49
7 methylammonium ... water	−0.22	−0.56
8 guanidinium ... formaldehyde	0.72	0.39
9 guanidinium ... methylamine	−0.30	−0.31
10 guanidinium ... methanol	1.48	1.29
11 guanidinium ... water	1.20	1.28
12 imidazolium ... formaldehyde	−0.45	−1.48
13 imidazolium ... methylamine	−2.39	−5.27
14 imidazolium ... methanol	−0.61	−1.53
15 imidazolium ... water	−0.03	−0.69
MUE	1.08	1.69

Table A.5: Errors in interaction energies for the AHB21 data set, at the XSAPT(KS)/def2-TZVPPD level.

	error (kcal/mol)	
	CM5	ChElPG
1	1.87	1.13
2	0.75	0.09
3	−2.41	−2.72
4	0.71	0.55
5	0.24	0.02
6	1.09	−0.25
7	1.71	−0.28
8	−0.50	−3.37
9	2.71	2.46
10	1.33	−1.17
11	0.14	−0.38
12	−0.26	−0.91
13	−0.39	−2.50
14	−1.06	−2.01
15	1.10	0.60
16	0.98	0.08
17	1.28	0.91
18	1.29	0.33
19	1.38	0.47
20	1.53	0.37
21	1.53	−1.85
MUE	1.15	1.07

Table A.6: Errors in interaction energies for the CHB6 data set, at the XSAPT(KS)/def2-TZVPPD level.

	error (kcal/mol)	
	CM5	ChElPG
22	0.70	1.91
23	0.89	1.29
24	0.75	0.80
25	−2.93	−1.83
26	−0.27	0.20
27	−0.28	−0.68
MUE	0.97	1.12

Table A.7: Errors in interaction energies for theIL16 data set, at the XSAPT(KS)/def2-TZVPPD level.

	error (kcal/mol)	
	CM5	ChElPG
008	−0.11	−6.80
144	−1.30	−4.70
147	−1.15	−6.68
148	−3.72	−11.57
150	−0.93	−6.17
152	−0.98	−6.73
187	−2.90	−13.83
202	−3.79	−15.06
212	−3.71	−13.62
213	−1.52	−6.26
214	−1.38	−5.09
227	−2.94	−9.95
228	−0.36	−3.71
229	−0.86	−4.84
230	−2.00	−2.74
231	−2.81	−5.58
MUE	1.90	7.71

Table A.8: Errors in interaction energies for the S30L data set, at the XSAPT(KS)/def2-TZVPPD level.

	error (kcal/mol)	
	CM5	ChElPG
1 TCNQ@tweezer	−3.47	−3.21
2 DCB@tweezer	−1.14	−0.98
3 TCNB@pincer	8.11	8.00
4 NBD@pincer	4.57	4.10
5 TNF@tweezer2	−2.23	−2.23
6 TCNQ@tweezer2	0.76	0.87
7 5CPPA@8CPPA	−4.13	−6.55
8 6CPPA@9CPPA	−7.96	−9.76
9 C60@catcher	−5.69	2.48
10 C70@catcher	−4.29	5.23
11 C60@CA10	−3.76	−3.61
12 C70@CA10	−3.44	−1.16
13 morpholine@RA4	3.01	1.03
14 tioxane@RA4	1.44	−1.09
15 TMPDA@XB-donor	−2.60	−0.90
16 HHTAP@XB-donor	9.00	7.94
17 BQ@mcycle	0.87	0.84
18 GLH@mcycle	1.65	1.82
19 C5H9OH@β-CD	1.77	2.17
20 C8H15OH@β-CD	−0.30	0.22
21 AdOH@CB7	−8.19	−5.70
22 DAAD@ADDA	15.30	15.63
23 AAAA@DDDD$^+$	6.33	8.26
24 $Ad_2(NMe_3)_2$@CB7	−6.39	−11.03
25 tetraphene@Ex^{2}Box	−6.43	−6.42
26 chrysene@Ex^{2}Box	−6.57	−7.75
27 BuNH4$^+$@CB6	1.41	2.42
28 PrNH4$^+$@CB6	2.80	3.66
29 acetate@CP4	1.01	3.21
30 benzoate@CP4	1.76	3.65
MUE	4.21	4.68

APPENDIX B

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

Table B.1: Comparison of δE^{CP} and MBCP(2) for $(H_2O)_N$ clusters, $N = 6$–37.

N	CP correction (Hartree)		difference
	δE^{CP}	MBCP(2)	(kcal/mol/monomer)
6	0.005	0.005	0.060
7	0.006	0.006	0.027
8	0.007	0.007	0.024
9	0.008	0.008	0.020
10	0.009	0.009	0.020
11	0.011	0.011	0.041
12	0.012	0.013	0.047
13	0.013	0.014	0.033
14	0.014	0.015	0.041
15	0.016	0.017	0.035
16	0.017	0.019	0.062
17	0.018	0.020	0.044
18	0.019	0.021	0.051
19	0.021	0.022	0.041
20	0.022	0.023	0.048
21	0.024	0.026	0.050
22	0.025	0.026	0.042
23	0.026	0.028	0.053
24	0.027	0.029	0.051
25	0.028	0.029	0.034
26	0.030	0.032	0.042
27	0.032	0.034	0.049
28	0.033	0.036	0.056
29	0.034	0.035	0.041
30	0.037	0.040	0.051
31	0.038	0.041	0.060
32	0.040	0.042	0.050
33	0.042	0.044	0.048
34	0.043	0.046	0.048
35	0.044	0.047	0.046
36	0.046	0.049	0.059
37	0.046	0.049	0.043

Table B.2: Interaction energies (in kcal/mol) arising from sub-clusters separated by 8–9Å, for the four structural motifs in $(H_2O)_{20}$ clusters.

Isomer	fused cubes	dodecahedra	face-sharing pentagonal prisms	edge-sharing pentagonal prisms
1	−0.544	0.000	−0.917	−0.455
2	−1.251	0.000	−2.064	−0.455
3	−0.737	0.000	−1.992	−0.454
4	0.521	0.241	−0.963	0.033
5	−0.565	0.000	−1.335	0.513
6	−0.579	0.000	−0.942	−0.455
7	−2.001	0.000	−0.916	−0.454
8	−0.738	0.000	−1.986	−0.028
9	−1.300	0.000	−1.986	−0.454
10	0.024	0.000	−0.917	−4.422
11	−0.615	0.000	−0.917	−0.028
12	−1.813	0.000	−1.064	0.031
13	−1.328	0.000	−1.799	0.032
14	−2.920	0.000	−2.064	−0.248
15	−0.076	0.216	−2.057	−0.029
16	0.508	0.000	−2.063	0.032
17	−0.753	0.000	−0.942	−0.028
18	−2.004	0.000	−2.057	−4.111
19	−0.951	0.000	−1.852	−0.028
20	−0.747	0.000	−1.828	−0.378