AN ACCURATE DESCRIPTION OF WATER USING THE MANY-BODY EXPANSION AND PERTURBATION THEORY

OUYANG FENGCONG JOHN

NATIONAL UNIVERSITY OF SINGAPORE

2016

AN ACCURATE DESCRIPTION OF WATER USING THE MANY-BODY EXPANSION AND PERTURBATION THEORY

OUYANG FENGCONG JOHN
(B.Sc.(Hons), NUS)

**A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

DEPARTMENT OF CHEMISTRY

NATIONAL UNIVERSITY OF SINGAPORE

2016

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety, under the supervision of Assoc. Prof. Ryan P.A. Bettens (in the laboratory MD1-05-03), Chemistry Department, National University of Singapore, between August 2012 and August 2016.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

The content of the thesis has been partly published in:

1) Ouyang, J.F.; Cvitkovic, M.W.; Bettens, R.P.A. *J. Chem. Theory Comput.* **2014,** *10,* 3699–3707.

2) Ouyang, J.F.; Bettens, R.P.A. *Chimia* **2015,** *69,* 104–111.

3) Ouyang, J.F.; Bettens, R.P.A. *J. Chem. Theory Comput.* **2015,** *11,* 5132–5143.

Ouyang Fengcong John		August 7, 2016
Name	Signature	Date

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, Associate Professor Ryan P.A. Bettens. An excellent mentor, he taught me a lot on both theory and programming through our numerous fruitful discussions and these insights will benefit me for many years to come. Technicalities aside, his optimistic outlook towards doing science has always been an inspiration for many budding scientists including myself. His enthusiasm in research and his trust in me have always spurred me on during my PhD journey, which culminated in the completion of this thesis.

I would also like to thank the many professors and lecturers at the National University of Singapore (NUS) who have honed my chemical intuition. This is especially so for my thesis committee—Prof. Richard Wong, A/P Zhang Chun and A/P Kang Hway Chuan—who have offered much encouragement and insightful comments. Their guidance and counsel have definitely made me appreciate my project and Chemistry better. Also, many thanks to Dr. Adrian Lee for the many discussions on quantum chemistry and advice on academic life.

My sincere thanks also goes to the NUS non-academic staff who had helped me tremendously during my PhD candidature. To Miss Suriawati Binte Sa'ad and the Chemistry Department Office for their help in handling the many administrative load. To Yung Shing Gene at the NUS Centre for Computational Science and Engineering and Wang Junhong at the NUS Computer Centre for

their prompt response and extensive help regarding computational resources.

I am also honoured to be a recipient of the President's Graduate Fellowship scholarship, without which it would be almost impossible for me to pursue my graduate studies in NUS. I am also grateful for the Ministry of Education Academic Research Fund for funding me to go to an overseas conference and computational resources, which are crucial to my PhD work.

In my daily work, I have been blessed with a cheerful group of lab mates. In particular, many thanks to my neighbour, Yee Hong, where we have spent countless hours discussing different physics topics that have intrigued us both. To Amelia, Hwee Jia, Michael, Milan and Ryo, thank you for always being there, making the lab a more lively place. I also extend my thanks to my undergraduate mentees: Simin, Chia Ling, Yee Ching, Estee and Danielle. Their fresh ideas always amazed me and I treasure the friendship we have forged. Furthermore, many thanks to the NUS Special Program in Science, which I have been a part of since my undergraduate years. The drive and enthusiasm in science of my juniors in the Program have been an invaluable source of inspiration.

Finally, I am indebted to my family who have shaped me into who I am. No words can express the gratitude to my Dad and Mum for their decades of care and love. Their unwavering support in every decision I make is always reassuring and heart-warming. To my wife Wendy, she has been a pillar of strength since we began dating 10 years back. Your unyielding love has always kept me going in times of distress and made me feel proud in times of success.

In retrospect, this PhD journey has been an exciting and memorable one. I will cherish these memories, whatever the future may bring.

CONTENTS

SUMMARY

This thesis attempts to provide an accurate description of the energetics of large water clusters. In particular, we seek to reduce the computational cost, preferably achieving linear scaling with system size. Firstly, the many-body expansion (MBE) was employed to decompose the total energy of the system up to four-body contributions. Secondly, perturbation theory is then used to select the important many-body contributions. The methods employed in this thesis can be extended to many other large chemical systems and we have included examples of polypeptides in the penultimate chapter.

The first chapter of the thesis provides a succinct review of the status of water modelling. From this mini-review, we identified several ideal properties of water models. Chapter two then highlights some of the theoretical background of quantum chemistry that is relevant to this thesis. The information is presented in an unconventional manner where we proposed and answered five questions pertaining to electronic structure methods and basis sets.

In Chapter three, we found that the MBE exhibits poor convergence with respect to the number of bodies in certain cases. This was attributed to the basis set superposition effect (BSSE) from diffuse basis functions. To restore the rapid MBE convergence, one can either omit the diffuse functions or employ larger basis sets. Alternatively, the electronic structure calculations can be performed using the same set of basis functions centred on the cluster, i.e., the cluster basis.

We further investigated this many-body BBSE using our proposed many-ghost many-body expansion (MGMBE) in Chapter four. The MGMBE separates the many-body BSSE into two components. Firstly, there is an undesirable basis set imbalance error due to a different number of basis functions across different total energy calculations. The second component, named the basis set extension effects, involves the borrowing of basis functions from the cluster basis to improve the description of many-body effects. With the MGMBE, we identified that the poor MBE convergence reported earlier is primarily due to extension effects in the one-body energy.

In Chapter five, we proceeded to utilise perturbation theory to screen for significant many-body interactions. We derived the leading three-body and four-body terms in many-body induction and used these leading terms to estimate the maximum possible many-body effects in a given arrangement. This was then used to identify the significant many-body effects. Consequently, we successfully reproduced the total three-body and four-body interaction energies using a tiny fraction of the individual interactions. More importantly, we identified that extended linear arrangements are favoured to give significant many-body effects. This allows many-body effects to be extended over large distances but only in a directional manner.

The final chapter concludes our findings and suggests some potential future work. This is followed by the Appendices which included tedious mathematical derivations. Other supporting information such as the Cartesian Coordinates of the clusters can be founded in an accompanying CD.

LIST OF TABLES

LIST OF FIGURES

List of Symbols

E	Energy, see equation (2.1) 24
$E_{\text{xc}}^{\text{GGA}}[\rho]$	DFT exchange-correlation energy under GGA, see equation (2.22) 44
$E_{\text{xc}}^{\text{LDA}}[\rho]$	DFT exchange-correlation energy under LDA, see equation (2.22) 44
$E_{\text{xc}}^{\text{nonlocal}}[\rho]$	DFT non-local exchange-correlation energy, see equation (2.23) 45
$E_{\text{BSSE}}^{(n,k)}$	Difference between the k-mer total energy in the cluster basis and that in the nuclei-centered basis, see equation (3.6) 54
E_{el}	Electronic energy, see equation (2.2) 26
$E_{\text{ext.}}^{(k)}$	Basis set extension effects in the k-body interaction, see equation (4.8) . 82
$E_{\text{corr},\infty}$	Correlation energy in the CBS limit, see equation (2.19) 42
$E_{\text{corr},n}$	Correlation energy in the cc-pVnZ basis set, see equation (2.19) 42
E_{tot}	Total energy of the system, see equation (3.1) 52
M	Number of basis functions, see equation (2.11) 32
Ψ	Wavefunction, see equation (2.1) 24
Ψ_0	Reference wavefunction, see equation (2.16) 37
Ψ_{CC}	Coupled Cluster wavefunction, see equation (2.16) 37
Ψ_{HP}	Hartree Product, see equation (2.4) 27
Ψ_{el}	Electronic wavefunction, see equation (2.2) 26
$\varepsilon_{\text{max}}^{(3)}$	Maximum three-body interaction in an arrangement, see equation (5.3) 110
$\varepsilon_{\text{max}}^{(4)}$	Maximum four-body interaction in an arrangement, see equation (5.4) 110
ε_{max}	Maximum many-body interaction in an arrangement, see equation (5.3) 110
ε_i	Energy of orbital i, see equation (2.6) 28
$\hat{F}_i$	Fock operator, see equation (2.6) 28

$\hat{H}$	Hamiltonian operator, see equation (2.1) 24
$\hat{H}_0$	Reference Hamiltonian, see equation (2.13) 35
$\hat{H}_{\mathrm{el}}$	Electronic Hamiltonian, see equation (2.2) 26
$\hat{J}_i$	Coloumb operator, see equation (2.7) 28
$\hat{K}_i$	Exchange operator, see equation (2.7) 28
$\hat{T}_{\mathrm{e}}$	Kinetic energy operator, see equation (2.3) 26
$\hat{V}$	Perturbation operator, see equation (2.13) 35
$\hat{V}_{\mathrm{ee}}$	Electron-electron repulsion operator, see equation (2.3) 26
$\hat{V}_{\mathrm{ne}}$	Nuclear-electron attraction operator, see equation (2.3) 26
$\hat{V}_{\mathrm{nn}}$	Nuclear-nuclear repulsion operator, see equation (2.3) 26
$\hat{h}_i$	One-electron operator, see equation (2.7) 28
$\mathbf{R}$	Nuclear coordinates, see equation (2.2) 26
$\mathbf{T}_i$	Excitation operator exciting i electrons, see equation (2.16) 37
$\mathbf{r}$	Electronic coordinates, see equation (2.2) 26
$\phi^{\mathrm{GTO}}_{\zeta,n,l,m}$	Gaussian-type orbitals, see equation (2.9) 30
$\phi^{\mathrm{STO}}_{\zeta,n,l,m}$	Slater-type orbital, see equation (2.8) 30
$\phi_1(\mathbf{r}_1)$	One-electron wavefunctions, see equation (2.4) 27
ϕ_i	One-electron orbital i, see equation (2.6) 28
$\varepsilon'_{\mathrm{A}\cdots\mathrm{K}}$	k-body interaction in the nuclei-centred basis, see equation (4.2) 77
$\varepsilon^{(n,k)}$	Total k-body energy, see equation (3.1) 52
$\varepsilon_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{N}}}$	k-body interaction in the cluster basis, see equation (4.6) 81
$\varepsilon_{\mathrm{A}\cdots\mathrm{K}}$	k-body interaction in the subcluster basis, see equation (4.4) 80
$\varepsilon_{\mathrm{tot}}$	Cluster binding energy in the nuclei-centred basis, see equation (4.3) . . 77
$\varepsilon^{\mathrm{C}}_{\mathrm{tot}}$	Cluster binding energy in the cluster basis, see equation (4.7) 81
$\varepsilon^{\mathrm{S}}_{\mathrm{tot}}$	Cluster binding energy in the subcluster basis, see equation (4.5) 80
$\xi_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$	Basis set extension effects from m-ghost-body $\mathrm{L}\cdots\mathrm{M}$ in the k-body interaction of $\mathrm{A}\cdots\mathrm{K}$, see equation (4.9) 86
ζ	Exponent of basis function, see equation (2.8) 30

1 | Modelling Water: A Lifetime Enigma

The first attempt to describe water dates back to 1933 with the Bernal–Fowler model and it would take another forty years before the first computer simulation of liquid water in 1969. Since then, over a hundred different water models have been proposed. Despite being widely studied, water remains poorly understood. Examining the evolution of water models, we identified three distinct philosophies in water modelling, namely (i) the employment of effective point charges in pioneering empirical models, (ii) the incorporation of polarization to describe many-body inductive effects and (iii) the extensive use of ab initio calculations to describe short-range effects. In doing so, we can appraise the current understanding of water and identify attributes that a water model should possess to capture the intricate interactions between water molecules.

1.1 Why Study Water?

Considering the rich history of water modelling, it would be prudent to ask why scientists across different disciplines are enthralled by water. An obvious motivation would be its abundance which suggests that water is undeniably important in the grand scheme of nature. The strange properties associated with water also spur academic curiosity to unravel the mysteries behind this small molecule. Most importantly, deciphering the interactions between water

molecules would lead to the basic understanding of intermolecular forces, which govern many dynamic processes in nature.

Given its ubiquity in nature, water has been the subject of extensive research. On Earth, water is the central solvent for naturally occurring chemical processes. In particular, water is the medium for biochemical interactions, widely recognized as the "matrix of life".[1] Its place in biology goes beyond a passive solvent, having many active roles in molecular biology.[2–4] Water-mediated hydrogen bonding provides exchangeable and extensible linkages to manoeuvre the peptide backbone during protein folding, allowing proteins to achieve their active conformation rapidly.[5] Hydration changes can induce modification in DNA conformation and interfacial water possess unique sequence-dependent hydration structure, acting as a "hydration fingerprint" for the recognition of the DNA sequence.[6] On a cosmic scale, detection of water vapour in the atmosphere of an extrasolar gas-giant planet suggest that the presence of water is common in gas-giants.[7] Closer to home, studies on the isotopic composition of water in meteorites help us gain insights about the origins of the early solar system.[8,9] Interestingly, most water in the universe exist as different forms of amorphous ice and their transitions in cold dense interstellar molecular clouds causes radical recombination, resulting in the synthesis of complex organic molecules.[10] The role of water in many chemical and biological processes that are responsible for sustaining life, is the driving force behind understanding its behaviour under different conditions, and in various environments.

Being one of the most studied substances, many physical properties of water are accepted as international standards such as its triple point and density.[11] Even so, many of these physical properties are considered anomalous as they contradict the general theories of the liquid state of matter. The most-widely known property would be the maximum density of water at 4°C, making water the only liquid to expand upon cooling. Other anomalies include the non-monotonic behaviour of its isothermal compressibility and specific heat.[12,13]

Furthermore, water exhibits a very high boiling point and dielectric constant for a simple liquid. Although the aforementioned anomalies were known for some time, new anomalous behaviours are constantly uncovered. It was found that supercooled water becomes more diffusive as pressure is increased to about 200 MPa at room temperature.[14] Also, the discovery of another supercooled liquid water state at 150 K challenges the notion of a single supercooled regime at ambient pressure[15] and this newly discovered supercooled state may lead to the identification of a possible second critical point in supercooled confined water.[16] If the liquid state is strange, the solid state would be bizarre with water having fifteen known forms of ice, many of which were only discovered recently.[17,18] It is ironic that while better technology has allowed us to probe the properties of water further, these observed phenomena can exacerbate confusion as they remain unexplained.

The wealth of knowledge on water, many of which deemed anomalous, imposes severe tests on any newly proposed water model. Despite being a chemically simple molecule, water is notoriously hard to model. Firstly, water can give rise to extensive hydrogen bonding networks.[19] As early as 1920, hydrogen bond is first suggested to occur in water[20] and it is commonly agreed that these fleeting hydrogen bonds makes water unique from most other liquids. Dimer interactions are dominated by a deep minimum at the hydrogen bonded configuration,[21–23] implying that certain configurations are preferred in water clusters and bulk water. The strong directionality of hydrogen bonding is the reason for the inclusion of explicit water molecules in simulating water-mediated processes such as protein folding.[3] However, the hydrogen bond minima is not overly stabilising, making dynamic hydrogen bonding rearrangements possible in bulk water.[19] Secondly, the description of water is complicated by strong non-additive inductive effects that manifest in water due to the large dipole and polarizability of water. Such inductive effects can enhance the dipole moment of water molecules by more than 60% in the condensed phase.[24] This is further

complicated by the fact that the introduction of polarizability can be deceptive,[25] compounded by reasons which will be covered in Section 1.3. All in all, water is especially sensitive to the desctiption of the forces between molecules and thus demand a thorough and basic understanding of intermolecular forces.

1.2 Water Models

The Bernal–Fowler (BF) model can be considered the first realistic water model, describing water as a collection of point charges and a repulsion-dispersion term.[26] A similar representation would be used later in the first Monte Carlo simulation of water by Barker and Watts[27] and the first Molecular Dynamics (MD) simulation of water by Rahman and Stillinger.[28] Since the first computer simulation of water, a myriad of water models, exceeding a hundred to date, have been proposed. While there already exist several excellent reviews on the progress of modelling water,[25,29–32] we still wish to survey the water modelling scene with the aim of highlighting the qualities of a good water model.

In the aforementioned reviews, water models are categorized based on (i) the interaction between water monomers and (ii) the treatment of water monomers. Polarizable models treat many-body inductive effects explicitly using point polarizabilities whereas non-polarizable models describe this polarization in an averaged manner in the pairwise interactions. Rigid water models constraint the intramolecular degrees of freedom, typically to that of the vibrational averaged geometry while flexible counterparts relaxes all degrees of freedom. Due to ab initio calculations approaching experimental accuracy, water models can also be classified based on the nature of the data (ab initio or experimental or both) used to parametrise the model.

Instead of following these traditional and possibly restricting classifications, we analysed the evolution of water models and broadly identified three distinct philosophies in the saga of water modelling, namely (i) the employment of enhanced point charges in pioneering models to effectively describe induc-

tion in a pairwise potential, (ii) the incorporation of polarization in later models to describe explicitly the many-body inductive effects and (iii) the extensive use of ab initio data in state-of-the-art models to accurately describe water-water interaction at all ranges (Figure 1.1). Water models are not necessarily grouped based on chronological order as our demarcations represent distinct principles of water modelling rather than actual time periods. In doing so, we have alluded to the long history of water modelling and its coming of age.

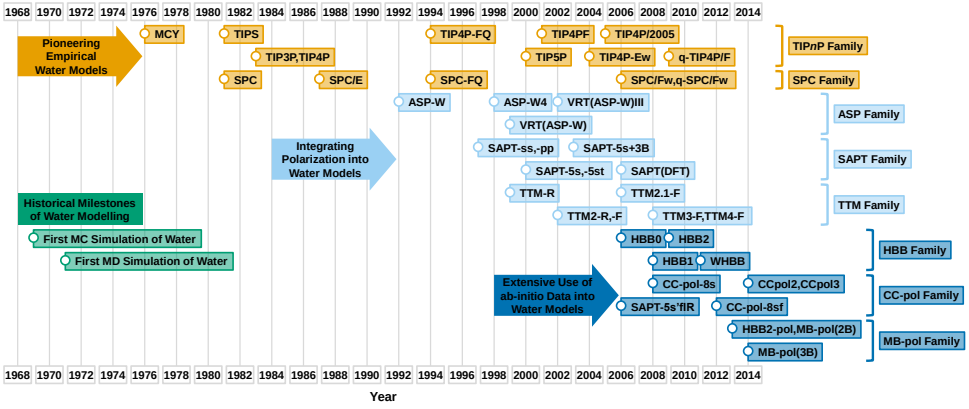


Figure 1.1: Timeline showing the year of implementation of various water models reviewed in this chapter. Water models are grouped (using different colour schemes) according to the three distinct philosophies of modelling water. Within each class of water models, the models are further subdivided into different families of water model that share similar traits.

1.2.1 Pioneering Empirical Water Models

This class of water models has its origins in legacy water models, aimed at describing water with a low computational cost and thus often utilise a rigid water monomer. Similar to the BF model, these models are empirical and non-polarizable, using point charges to represent electrostatics and a Lennard-Jones term for dispersion and repulsion. Induction effects are effectively described by increasing the point charges to simulate an enhanced dipole moment found in the condensed phase. Parameters are fitted to reproduce macroscopic experimental data such as the liquid density and heat of vaporization. The reliance on experimental data can be reconciled by noting that these models flourished

in the 1980s while highly accurate ab initio tools such as the Coupled Cluster Single and Double excitations with perturbative Triples [CCSD(T)] level of theory[33] and Dunning correlation consistent basis sets[34] were only developed in 1989 and became computationally feasible many years later. Consequently, these models only work well at reproducing macroscopic properties of the condensed phase near the conditions under which they are parameterized, typically ambient pressure and temperature, and are targeted towards applications such as biomolecular simulations rather than basic scientific enquiry about the anomalous properties of water. The low computational cost associated with these models would make them remain the preferred choice for the most computationally demanding applications. For example, the TIP3P model is the default water model used in the CHARMM force field for biomolecular simulations. One of the earliest water models in this class is the MCY model,[35] well-known for being constructed entirely from ab initio Hartree–Fock (HF) calculations. We will also look further into two families of these pioneering water models, namely the TIP*n*P and SPC water models.

TIP*n*P Family. First developed by Jorgensen in 1981 as the transferable intermolecular potential functions (TIPS),[36] it is later refined into the TIP3P and TIP4P model[37] which most water scientists are familiar with. Here, the *n*P refers to the number of point sites in the model where point charges and/or Lennard-Jones terms are placed. In the simplest case, the atomic sites were used as seen in TIP3P. An additional M-site along the H–O–H angle bisector is introduced in TIP4P to displace the negative charge towards the hydrogens as placing the negative charge on the oxygen would lead to an excessively high dipole moment.[27] In an attempt to describe inductive effects, TIP4P-FQ (FQ for fluctuating charge) was introduced where the point charges fluctuates in response to the environment to equalize the electronegativities of the sites.[38] Later, Mahoney and Jorgensen would introduce more TIP*n*P variants, namely TIP5P[39] and TIP4PF.[40] TIP5P replaced the M-site with two tetrahedral negative

charges to mimic the lone pairs on water but this resulted in a overly structured water in simulations. TIP4PF is a flexible version of TIP4P where intramolecular stretching and bending are described by quadratic terms and the same study showed that the inclusion of quantum effects improve the predictions made by this flexible water model.

Surprisingly, the models mentioned thus far truncate long-range electrostatics at a certain cut-off distance. The TIP4P-Ew model is designed for use with Ewald techniques to account for long-range electrostatics, commonly employed in biomolecular simulations.[41] Numerous other parameterization attempts were made such as TIP4P/2005[42] and q-TIP4P/F[43] which are optimised to better reproduce the thermodynamic properties of water and to account for quantum effects respectively.

SPC Family. Apart from the TIP*n*P family, another family of water models is the Single Point Charge (SPC) model, which only uses the three atomic sites to place point charges and/or Lennard-Jones terms.[44] Simple values were used for its parameters such as 1.0 Å for the O–H bond length and an ideal tetrahedral angle of 109.5° instead of the experimental gas-phase values used in TIP4P. Shortly, the improved SPC/E model was proposed to account for polarization self-energy.[45] Similar to TIP4P-FQ, SPC-FQ was introduced to incorporate induction effects.[38] Likewise, flexible monomer versions such as SPC/Fw[46] and variants parameterized to account for quantum effects such q-SPC/Fw[47] model have been introduced.

1.2.2 Integrating Polarization into Water Models

The increase in computational power saw a transition towards increasingly complicated water models with an emphasis on the non-additivity of water-water interactions, in particular induction/polarization effects. Polarization is often incorporated explicitly via central or distributed point dipole-dipole polarizabilities, derived from the use of perturbation theory to treat intermolecular forces.[48]

Despite the rigorous theoretical background, such an implementation may lead to deceptive results as we shall see in Section 1.3. Furthermore, higher-rank multipoles, typically up to quadrupoles, are employed to represent electrostatics instead of point charges in recognition of the anisotropic nature of the electron distribution. This led to more elaborate analytic potentials that required more parameters that would come from a mix of ab initio and experimental spectroscopic data. This class of water models flourished in the 1990s and 2000s when accurate ab initio second-order Møller–Plesset perturbation theory (MP2) and later CCSD(T) calculations become amendable. As the majority of the parameters are essentially monomer properties such as the dipole moment and polarizability, highly accurate ab initio calculations can be performed on the small water monomer system. In some cases, the Vibration-Rotation-Tunnelling (VRT) spectroscopic data was used in the parameterization as they represent information at the atomistic level as opposed to bulk water properties. Using these water models, there would be more studies devoted towards water clusters, underscoring the importance of microscopic understanding of water. As the functional form of these water models grew more complex, it would naturally encompass a larger variety of models and some of the notable water models include the ASP, SAPT and TTM family of water models.

ASP Family. The Anisotropic Site Potential with Wormer's dispersion (ASP-W) model is one of the earliest rigid water models to adopt higher-rank multipoles.[49] For electrostatics, distributed multipoles are present on both the Oxygen and Hydrogen atomic sites, up to quadrupole and dipole respectively whereas induction is computed at first order (instead of full iterative) using point polarizabilites on Oxygen up to quadrupole. Site anisotropy was likewise incorporated into the dispersion and repulsion terms. Further refinements by the same group led to the inclusion of a new charge transfer term, creating the ASP-W2 and ASP-W4 models, used to study the stationary structures of the water dimer.[50] The difference between both models lies in the order of mul-

tipoles used with the original multipoles being retained in ASP-W2 and up to hexadecapole present on each atom in ASP-W4.

The ASP-W functional form was also fitted to the $(D_2O)_2$ VRT spectra, giving rise to the VRT(ASP-W) model.[51] VRT(ASP-W) is the first water model to achieve spectroscopic accuracy, able to reproduce most of the tunnelling barriers in the water dimer. While this is not surprising given the use of same experimental data in constructing the model, it is worthwhile to note that the use of the rigid monomer approximation can still lead to accurate predictions at the atomistic level. Later improvements would give rise to the VRT(ASP-W)II and VRT(ASP-W)III model, where induction is full iterative.[52]

SAPT Family. Both the SAPT-ss and SAPT-pp water models,[53] employing rigid water monomers, were developed based on Symmetry-Adapted Perturbation Theory (SAPT).[54] SAPT-ss comprises a site-site form, with a similar placement of sites as TIP4P but instead uses the functional form of the MCY model. Point charges and exponential terms are fitted to 1056 SAPT energies. The SAPT-pp is more complicated, describing the intermolecular interactions using expansions of functions in interatomic vectors and euler angles, again fitted to the same 1056 SAPT energies.

Due to its complexity, SAPT-pp would fall into disuse and the site-site form is evolved to the SAPT-5s model.[55] To reflect the anisotropy of electron distribution, two new symmetry distinct sites, representing lone pairs and out-of-plane charges, were added, giving a total of five symmetry distinct sites (eight different sites in total). An elaborate function form was adopted using a polynomial-exponential terms to represent exchange-repulsion and an inverse power (6-8-10) series to describe induction and dispersion. Consequently, no iteration of the induced dipole is required in calculating the induction energy as it is represented by fitted coefficients. Later, the exchange-repulsion parameters were tuned to better reproduce the water dimer's acceptor tunnelling splitting, giving the revised SAPT-5st.[56]

All the SAPT models mentioned above only contains a pair potential. Thus, the three-body SAPT(HF) energies were incorporated into SAPT-5s, giving the SAPT-5s+3B model.[57] This new three-body potential is the first to include functional forms to model three-body exchange effects using a combination of exponential and Legendre polynomial terms. Long-range effects are described using a damped induced dipole model. Later, the SAPT-5s functional form is refitted using SAPT(DFT) energies and this new SDFT-5s model[58] gives more accurate results, attributed to the faster basis set convergence with DFT.

TTM Family. TTM-R,[59] the first of Thole-Type Model (TTM) water models, is based on Thole's idea of using smeared out multipoles to mirror the diffuse picture of electron distribution.[60] TTM-R utilises TIP4P-style point charges for electrostatic interactions and an inverse power (6-10-12) series to represent dispersion and repulsion. Smeared charges and dipole are present on all atomic sites so that induction and intramolecular polarization can occur, accounting for charge transfer. As the TTM-R model consistently over-binds small water clusters, the TTM2-R model was proposed by refitting the inverse power (6-10-12) series to minimum energy pathways connecting the global minimum and other stationary points of the water dimer.[61]

Monomer flexibility is then incorporated using the Partridge–Schwenke intramolecular Potential Energy Surface (PES) and Dipole Moment Surface (DMS)[62] resulting in the TTM2-F model, the first water model to properly reproduce an increase in the monomer bending angle in water clusters.[63] A revised TTM2.1-F model,[64] intended for simulations, was proposed by modifying the inverse power (6-10-12) series that decreases unphysically below 2.5 Å as such repulsive regions may be sampled during condensed-phase simulations.

Two unrelated updates, the TTM3-F[65] and TTM4-F model[66] were also reported. Aimed at describing the vibrational spectra of water clusters and bulk water, TTM3-F has modified partial charges to reflect the behaviour that water dissociates to H^+/OH^- in liquid as opposed to radical formation in the gas phase.

On the other hand, TTM4-F is reparameterized to better reproduce the polarizability surface. Notably, the popular AMOEBA water model uses a Thole-type induction model.[67]

1.2.3 Extensive Use of ab initio Data in Water Models

As ab initio methods mature into reliable tools rivalling experimental accuracy, we ushered in an era of water models empowered by ab initio data. This class of water models relies on high-quality large datasets (in the order of 10^5 data points) of CCSD(T) energies, the gold standard of quantum chemistry. The water models are deeply rooted in the many-body expansion (MBE) where the total energy of a system can be decomposed into one-body (monomer contribution), two-body (pairwise interactions), three-body contributions and so on. Separate PES are constructed for each of these k-body terms, fitted to large datasets which sample the important configuration space encountered in water clusters and during condensed phase simulations. The extensive amount of high-quality data required can only be fulfilled by large volumes of accurate ab initio calculations which only became amendable recently. The shift towards large datasets and complicated PES construction techniques stems from the realization that short-range effects such as charge transfer and exchange cannot be accurately described by simple analytic forms. Thus, sufficiently flexible functional forms are required to map the accurate ab initio dataset into high-quality PES for on-the-fly evaluation of energies. Water monomer flexibility is a another common feature in these models although a rigid monomer constraint is often imposed in the most demanding calculations such as condensed phase simulations and VRT spectra prediction. As a result, these models are mainly focussed water clusters with only a few examples of condensed phase simulations. As the construction of these water models is laborious, there were only three families of such ab initio water models, namely the HBB, CC-pol and MB-pol family of water models.

HBB Family. The HBB water models describe each of the k-body PES using permutationally invariant polynomials involving inter-atomic distances, incorporating the permutation symmetry of identical atoms. This alleviates the steep computational cost in evaluating high-dimensional PES and reduces the number of data points required for fitting. The first HBB0 model uses polynomials of Morse-type exponential functions, fitted to 19805 CCSD(T)/AVTZ energies.[68] Like all HBB models, all $\binom{N}{2}$ interatomic distances were used to preserve the permutational symmetry, which is more than the actual $3N-6$ degrees-of-freedom present in the system. In the next revision HBB1, the same functional form is refitted to an additional 10227 CCSD(T)/AVTZ energies to better describe the low-energy configuration space below 10000 cm^{-1}.[69] This refitting led to the RMS fitting error to drop by a factor by two, suggesting that the quality of the functional form is previously not maximized in HBB0.

A hybrid pair potential was developed in the new HBB2 model, comprising long-range and short-range components.[70] The short-range component remains to be described by permutationally invariant polynomials while the long-range component is described using the cheaper TTM3-F model. The HBBn models only contain a pair potential and cannot be used to describe water clusters where higher-body effects have to be considered. Thus, the WHBB model is introduced where a three-body potential is constructed using permutationally invariant polynomials, fitted to 40000 MP2/AVTZ energies.[71] Interestingly, it was mentioned that the three-body potential is shorter in range than the two-body counterpart and a cutoff was implemented when the maximum O–O distance is greater than 8 Å. Four-and-higher-body effects are described by induction using the TTM3-F model. For all the water models in the HBB family, the one-body potential is provided by the Partridge–Schwenke intramolecular PES.[62]

CC-pol Family. The CC-pol family of water models is the successor of the SAPT family, utilising ab initio energies computed at CCSD(T) instead of SAPT energies. The first CC-pol model is similar to the SAPT-5s model except

that the induction is now explicitly iterated instead of a fitted inverse power series.[72] CC-pol is able to reproduce the water dimer VRT spectra except for the interchange splitting transition, attributed to the rigid monomer approximation.

The CC-pol-8s model increased the number of interaction sites to eight symmetry distinct sites (25 different sites).[73] The three-dimensional Cartesian space was scanned in regular intervals, followed by finer subgrids to ensure that the most optimal positions were chosen. As only point charges were used (as opposed to higher-rank multipoles), the presence of more interaction sites better represents the anisotropy of the electron distribution and led to a four-fold decrease in the fitting errors. A flexible variant, CC-pol-8sf,[74] was developed where monomer contribution to the interaction energy is obtained from an earlier flexible SAPT-5s'fIR water model.[75]

Feeling that the order of 10^5 data points is inadequate to build an accurate full 21-dimensional flexible-monomer three-body PES, the authors reverted to a rigid monomer model, consisting of the pair potential CCpol2 and three-body potential CCpol3.[76] CCpol2 is essentially the same as CC-pol-8s, except that short-range damping is included to improve the description at very small intermolecular distances as these regions may be sampled during condensed phase simulations. The CCpol3 model, fitted to 71456 CCSD(T) energies, gives improved polarization from the use of three atomic polarization centres, instead of one. Four-and-higher-body interactions are described using a simple polarization model. Surprisingly, the polarization model gives accurate four-body energies to within a few percent, whereas such models are known to have significant errors for three-body interactions.

MB-pol Family. The MB-pol family incorporates many features from the HBB family of ab initio based water models. The prototype HBB2-pol model borrows from the HBB2 model using a hybrid pair potential and the Partridge–Schwenke intramolecular PES.[77] The same HBB2 PES was used for the short-range component of the pair potential while the long-range component was re-

placed with the TTM4-F model. Furthermore, a three-body hybrid potential is included where the short-range component again incorporates the permutational symmetry, fitted to 8019 CCSD(T) trimer energies, while the long-range counterpart, as well as four-and-higher-body effects, are described by induction in the TTM4-F model. The TTM4-F component greatly reduced the order of the permutationally invariant polynomials and the associated computational cost, making HBB2-pol amendable to condensed phase simulations. TTM4-F was chosen after careful comparison with two other polarizable flexible water models, namely TTM3-F and AMOEBA.

The eventual MB-pol model is described in two papers, detailing the hybrid pair potential[78] and higher-body effects separately.[79] The hybrid pair potential MB pol(2B) was improved with the addition of two new sites to represent the lone pairs of water, which greatly improved the flexibility of the functional form in the short-range component. Thus, the permutationally invariant polynomials now involves intersite distances between the atomic sites and/or the lone pair sites, fitted to 42508 CCSD(T) dimer energies. The three-body potential MB pol(3B) is described in a similar fashion as in HBB2-pol but fitted to a larger dataset of 12347 energies. All long range effects are handled by induction using the TTM4-F model. It was noted that short-range corrections are not required at the four-and-higher-body level, in agreement with CCpol3 authors' observation that a simple polarization model is sufficient.

On a final note, both the HBB2-pol and MB-pol are the first few water models constructed from extensive CCSD(T) energies dataset to be employed in classical and quantum simulations of liquid water.[80,81] In both instances, many structural and dynamic properties of liquid water under ambient conditions were reproduced, such as the radial distribution functions, density and diffusion coefficient.

1.3 Qualities of a Good Water Model

After reviewing the plethora of water models shaped by different philosophies, we identified several key features for the proper description of water. They are namely (i) the inclusion of polarizability to account for non-additive effects, (ii) fitting or interpolating energies to account for short-range effects, (iii) incorporation of monomer flexibility, (iv) accounting for quantum effects in simulations and (v) transferability and dissociable water model.

1.3.1 Inclusion of Polarizability

From Section 1.2.2, we witness that the inclusion of polarizability is crucial in describing the significant many-body inductive effects in water. Neglecting polarization effects in empirical point charge water models (Section 1.2.1) prevents an accurate description of virial coefficients, vapour pressures, critical pressure and dielectric constant.[82] The first three quantities involve gas phase properties which are very sensitive to changes in the environment. Clearly, the degree of polarization in the gas phase would differ greatly from that in the condensed phase for which the empirical models are calibrated for. Likewise, polarization is required to reproduce the enhanced dipole moment in condensed phase to properly reproduce the dielectric constant.

There are several excellent reviews[29,83–85] on the implementation of polarization as it found importance not only in water models but also in ion solvation, other small molecules and protein simulations. Three methods for incorporating polarization exist, namely fluctuating charge, Drude oscillator and induced point dipole models. While the first two methods have been implemented in water models, (eg. TIP4P-FQ, SPC-FQ[38] for fluctuating charge and SWM4-DP[86] for Drude oscillator) the induced point dipole model remains the most-implemented for water models. In fact, the ASP, SAPT and TTM families of water models in Section 1.2.2 all uses some kind of induced point dipole

model. In principle, higher-rank multipoles such as the quadrupole can also be induced as seen in the ASP water models but they see little action elsewhere (SAPT and TTM families only involve inducible dipole) perhaps due to the laborious theoretical expressions involved. While the introduction of inducible dipole models is increasingly prevalent, Guillot cautions that poor implementation can lead to deceptive results.[25] The induced dipole is given as the product of the polarizability with the electric field. The electric field is often represented by the point charges/multipoles present in the model and this may be inadequate if higher-rank multipoles are not considered.[87] Furthermore, there is also dipole-quadrupole and quadrupole-quadrupole polarizabilities which are often neglected and these inductive effects can be significant given that water has a strong quadrupole.

Finally, Thole[60] and Applequist *et al.*[88] have pointed out that the point induced dipole may become infinite at small distances, which is commonly known as the "polarization catastrophe". This can be avoided by screening the dipole-dipole interaction at short distances, either using a Tang–Toennies damping function[89] as seen in the ASP and SAPT models or using smeared out charges and dipoles in TTM models. This screening is an indication that point multipoles cannot properly describe the electronic distribution at small distances, underscoring the importance of accounting for short-range effects.

1.3.2 Short Range Effects

At short intermolecular distances R, the R^{-n} power series which define the point multipole diverges, causing the failure of point multipoles at short-range. Furthermore, there is a charge penetration effect as the electrons are "not fully felt" within the electron cloud. Physically, this can be interpreted as the unrealistic representation of the electronic distribution as if it was concentrated at a point. Possible remedies include the use of damping functions smeared out multipoles as seen in Section 1.2.2 as well as partitioning the electronic distribution us-

ing distributed multipoles.[48] Despite these corrections, other short-range interactions such as exchange-repulsion and charge transfer have to be explicitly accounted for. The distinction between short-range and long-range interactions (electrostatic, induction and dispersion) is rooted in their different physical character where short-range effects vary exponentially with intermolecular distance while long-range effects behave as some inverse power of intermolecular distance.[48] Thus, it would be prudent to separate the total interaction energy into short-range and long-range components due to their intrinsically different nature as seen in the HBB2, WHBB, HBB2-pol and MB-pol water models.

Unfortunately, unlike long-range interactions which can be described using perturbation theory, no exact analytic form exists for short-range interactions. Otherwise, high quality ab initio methods which can describe these subtle short-range effects up to any desired numerical precision would have been developed in vain. For the ASP, SAPT and TTM families of models, short-range exchange-repulsion effects were modelled by simple exponential and/or polynomial-exponential terms. As these approaches proved inadequate, large ab initio data sets are fitted to more complicated functional forms to accurately describe these exchange-repulsion effects (Section 1.2.3). Currently, two such functional forms have been implemented. The permutationally invariant polynomials in HBB and MB-pol families of models incorporate the permutational symmetry of identical nuclei into exponential terms involving interatomic distances. On the other hand, CC-pol models uses simple polynomial-exponential terms but applied between a large number of symmetry-distinct sites, greatly increasing the flexibility of the functional form. Inevitably, both methods incorporate some form of symmetry which serves to alleviate the high computational cost. Furthermore, both methods involve fitting of the coefficients of the terms from ab initio data. An alternative to fitting methods would be interpolation methods. Examples include the Shepard interpolation[90] as well as simpler methods such as cubic splines. While interpolation methods ensure that the PES

passes exactly through the dataset, care has to be taken that the asymptotic behaviour of the PES is enforced which are otherwise naturally incorporated into the functional forms used in fitting models. Nonetheless, it would be interesting to see new ab initio based water model based on interpolation methods and compare their accuracy with existing ones.

An essential formalism employed to describe short-range effects would be the MBE. Without the use of MBE, the dimensionality of the system would be too large for any fitting or interpolation method to be feasible. Using the MBE, large water clusters or even bulk water can be decomposed into many-body contributions, truncated at the four-body level. However, basis set superposition effects causes poor convergence of the MBE when diffuse basis functions are involved[91] and these diffuse functions are crucial in accurately describing the hydrogen bonding between water molecules.

1.3.3 Monomer Flexibility

In the MBE formalism, the one-body contribution corresponds to intramolecular distortions of the water monomer. Due to computational limitations, pioneering empirical water models often employ rigid monomers. While later models would incorporate flexible monomers, a rigid monomer approximation is still preferred for computationally demanding calculations. Also, a large dataset is required to fit flexible monomer potentials which can disfavour their use as seen in the CCpol2 and CCpol3 water models. If a rigid monomer approximation is employed, it is recommended that the vibrational averaged geometry be used over the equilibrium geometry.

The first water models to include flexible monomers use quadratic terms to describe the stretching and bending motions, modelling the vibrational modes as harmonic oscillators. This is overly simplistic in dealing with the quantum mechanical effects that arises when the electron clouds of the two hydrogens overlap during the bending motion. Thus, more sophisticated intramolecular PES

were constructed, the most popular being the Partridge–Schwenke intramolecular PES, which is used in the TTM, HBB and MB-pol families of water models. This PES is also accompanied with an intramolecular DMS which supplies the dipole moment required in the calculation of long-range interactions. This could be the reason why higher-rank multipoles are not involved in the long-range components of these models as there is no accurate quadrupole moment surface in the literature.

It is important to realize that these intramolecular vibrations are quantum mechanical in nature and their treatment within classical simulations may not yield satisfactory results.[92–94] The representative example would be the harmonic oscillator where the classical probability would be greatest away from the equilibrium while the quantum counterpart has the maximum probability at the equilibrium position. Thus, flexible water models should be simulated using methods that incorporate quantum effects.

1.3.4 Quantum Effects

Nuclear quantum effects and monomer flexibility are intertwined since nuclei motions obey the laws of quantum mechanics rather than the classical counterpart. This is especially so for water due to the presence of the light hydrogen nuclei and extensive hydrogen bonding, both of which exhibits strong nuclear quantum effects. Thus, processes involving the hydrogen nuclei such as Grotthuss proton shuttling[95] require nuclear quantum effects to be accounted for.[96]

Furthermore, disregarding nuclear quantum effects can lead to a poor description of the heat capacity of the condensed phase[97,98] and low-temperature properties such as the densities of ice polymorphs.[92] In addition, when nuclear quantum effects are neglected, isotopic effects cannot be probed which can have a significant influence on bulk properties. For example, the enthalpy of vaporization is a measure of the strength of the hydrogen bonding within liquid water. Classically, there should be no isotopic effects present. However, it has been

shown experimentally that the isotopic effects on the vaporization enthalpy is important, increasing by 0.4 kcal mol^{-1} from water to tritiated water.[82]

A variety of quantum simulation methods exist and some of the computational methodologies have been reviewed.[99] The most commonly employed method would be Path Integral Molecular Dynamics (PIMD),[100–102] which exploits the isomorphism between the quantum partition function expressed in path integral formalism and the classical partition function of a ring-polymer. This isomorphism provides a way to sample the quantum nuclear configuration through modifications of the classical MD technique. Other quantum simulation methods would include Path Integral Monte Carlo (PIMC), Path Integral Hybrid Monte Carlo (PIHMC), Centroid Molecular Dynamics (CMD) and Ring Polymer Molecular Dynamics (RPMD).

While PIMD simulations have been performed for the HBB2-pol and MB-pol ab initio based models at ambient conditions,[80,81] extreme conditions (low temperatures, critical point) have not been explored to elucidate the anomalous behaviour of water. On a side note, studies on the quantum effects of water performed on empirical water models such as TIP4P should be interpreted with caution. As such water models are parametrized to reproduce experimental values using classical simulations, quantum effects are included in these models in an effective manner. Thus, performing quantum simulations on these water models to investigate quantum effects seems counter-productive unless the model has been re-parametrized for such purposes.

1.3.5 Transferability and Ability to Dissociate

While less discussed in literature, it is ideal to develop a water model to be used outside pure water systems for applications such as explicit solvation of proteins. The empirical and polarizable models (Section 1.2.1 and 1.2.2) are highly transferable due to the use of point multipoles which share the same functional form regardless of the molecular species. This is not the case for ab

initio based water models (Section 1.2.3) that rely on the MBE and new PES have to be constructed for new combinations of k-body interactions.

Finally, very few models in literature are able to dissociate into H^+/OH^- ions. Water dissociation is difficult to handle as the products (charged ions) are very different from the reactant (neutral molecules). This is complicated by the fact that water dissociates homolytically into radicals in the gas phase. It would be optimal to use on-the-fly ab initio simulation techniques such as Car-Parrinello Molecular Dynamics (CPMD)[103] to study water dissociation as these ab initio methods do not make any distinction between H^+/OH^- ions and neutral water molecules.

1.4 Future Outlook for Water Modelling

The scene of water modelling remains a vibrant one where countless water models of distinct modelling philosophies were developed with the sole aim of understanding this mysterious liquid. The strengths and (more often) inadequacies of these water models have provided useful information on the essential ingredients of a universal water model.

It is only very recently, with the extensive use of ab initio data and availability of quantum simulations, that water models possess the right qualities to accurately describe water at both the microscopic and macroscopic level. Yet, there still leaves room for development, in seeking new ways to describe short-range effects using interpolation techniques and employing higher-rank multipoles in long-range interactions.

Nonetheless, it is due time to put these state-of-the-art water models to more rigorous tests to reproduce experimental results at extreme conditions. If these water models succeed at these trials, then perhaps it is ready to explain the many anomalies of water, fulfilling the role of computations in assisting experiments to dispel confusion and eventually pushing the boundaries of science.

2 | FIVE PIECES OF QUANTUM CHEMISTRY

All self-respecting PhD theses in quantum chemistry ought to describe the theory behind the quantum chemical methods employed. Doing away with the typical textbook-styled format, we suggested five questions pertaining to electronic structure methods and basis sets. The questions entail (i) the Hartree–Fock theory being the core of electronic structure methods, (ii) the construction of basis sets, (iii) the link between Møller–Plesset perturbation theory and Coupled Cluster theory, (iv) the poor basis set convergence of electron correlation methods and (v) the deficiency of Density Functional Theory in describing dispersion. In doing so, we hope to provide an unconventional viewpoint and focus on some of the subtleties within the theories.

2.1 The Evolution of Quantum Chemistry

Quantum chemistry is the application of quantum mechanics to study chemical systems. The development of both branches of science are often intertwined especially in the early days of quantum mechanics. One of the important results shaping quantum mechanics would be the Planck's law,[104] proposed in 1900 to explain the black-body radiation. Planck introduced the concept that the total energy is restricted to integer multiples of some definite unit of energy. This concept of quantization would resurface in 1905 when Einstein explained the

photoelectric effect by postulating that all electromagnetic radiation consists of discrete quantized packets, i.e., photons.[105] Quantization would be a central theme in quantum mechanics. Fast forward another twenty years, 1925 marks the beginning of modern quantum mechanics with the development of Heisenberg's matrix mechanics[106] and Schröndinger's wave mechanics.[107] Wave mechanics would give rise to the time-independent Schröndinger equation

$$\hat{H}\Psi = E\Psi \tag{2.1}$$

where $\hat{H}$ is the Hamiltonian operator, which acts on the wavefunction of the system, Ψ, to give the energy of the system, E, multiplied by the wavefunction itself. In fact, there is a third formalism of quantum mechanics attributed to Dirac, involving the use of the bra-ket notation.[108]

It did not take long before quantum mechanics was applied onto chemical systems. In 1927, Heitler and London studied the dihydrogen system using quantum mechanics.[109] They found that bonding only occurs between the hydrogen atoms when the spins of the electron were anti-parallel to each other. This seminal work provided a quantum mechanical understanding of the covalent bond, which lies at the heart of chemistry. Indeed, quantum mechanics would be the *de facto* choice for elucidating enigmatic problems in chemistry. By 1929, Dirac even declared that the theory of quantum mechanics is almost complete and "the whole of chemistry are thus completely known".[110] From there, he stressed the importance of finding practical methods to apply quantum mechanics to chemical systems. Despite so, quantum chemistry would be restricted to systems with a few atoms and electrons until the prevalence of digital computers in the 1970s.

In the 1970s, much would be devoted towards the efficient implementation of electronic structure methods with the emergence of many ab initio computer programs. A prominent figure in this era is John Pople, who developed new and

faster algorithms for the Hartree–Fock method and beyond.[111] Pople also popularized the use of the mathematically efficient gaussian-type orbitals[112,113] over the more physical Slater-type orbitals although the idea was first proposed by Boys in 1950.[114] The 1990s would see another boom in computational chemistry due to the emergence of personal computers. Electronic structure methods have matured to rival or even challenge the accuracy of experiments with the implementation of the Coupled Cluster Singles Doubles excitation with perturbative Triples [CCSD(T)] method[33] and the development of the correlation-consistent basis sets.[34]

Apart from performing calculations, computers have taken up the important role of chemical visualization in recent years. Such visualization allows us to animate the evolution of chemical systems over time and render the precise geometry of enzyme-ligand binding sites. Furthermore, hardware-driven techniques flourished, exploiting the untapped computational power in graphical processing units (GPU), which can perform $100-1000\times$ more computations per second than traditional central processing units (CPU).[115] We also see the interdisciplinary integration of techniques and algorithms. One such example is the use of machine learning in data science to handle large volumes of chemical data.[116] Giving the rapid development of computer technology which drives chemical computation, this is an exciting time to be a quantum chemist.

Most, if not all, self-respecting PhD theses in quantum chemistry include a textbook-styled chapter outlining the various common electronic structure methods and basis sets. Instead of the usual textbook-styled chapter, we came up with five questions pertaining to electronic structure methods and basis sets, which would cover much of the theoretical background that most quantum chemists are familiar with. In answering the questions, we aim to provide a more focussed writing, addressing some of the subtleties within the theories that might have been glossed over otherwise. Inspired by Feynman's book "Six Easy Pieces",[117] this chapter is titled the "Five Pieces of Quantum Chemistry".

2.2 Five Essential Pieces

2.2.1 Why is Hartree–Fock the basis of most electronic structure methods?

The Hartree–Fock (HF) theory lies at the core of quantum chemistry, often serving as a starting point for other quantum chemical methods. Additional approximations can be introduced to achieve speed-ups in computation, resulting in semi-empirical methods while post-HF corrections can be applied to correct for the mean field approximation (which will be elaborated on later), creating the electron correlation methods. Thus, the natural question is: "Why is the HF method the basis of most electronic structure methods?".

To address this question, we need to ask another question: "Why are there so many electronic structure methods?". The goal of all electronic structure methods is to solve the Schröndinger equation in eq (2.1) to obtain the electronic energy of the chemical system, E_{el}. In particular, we wish to solve for the electronic part of the wavefunction, $\Psi_{\text{el}}(\mathbf{R},\mathbf{r})$, which depends on both the nuclear coordinates, $\mathbf{R}$, and electron coordinates, $\mathbf{r}$. This leads to the electronic Schröndinger equation

$$\hat{H}_{\text{el}}\Psi_{\text{el}}(\mathbf{R},\mathbf{r}) = E_{\text{el}}\Psi_{\text{el}}(\mathbf{R},\mathbf{r}) \tag{2.2}$$

where the electronic Hamiltonian, $\hat{H}_{\text{el}}$, can be further broken down into

$$\begin{aligned}\hat{H}_{\text{el}} &= \hat{T}_{\text{e}} + \hat{V}_{\text{ne}} + \hat{V}_{\text{ee}} + \hat{V}_{\text{nn}} \\ &= -\frac{1}{2}\sum_i \nabla_i^2 - \sum_{\text{A},i}\frac{Z_{\text{A}}}{r_{\text{A}i}} + \sum_{i>j}\frac{1}{r_{ij}} + \sum_{\text{A>B}}\frac{Z_{\text{A}}Z_{\text{B}}}{r_{\text{AB}}}\end{aligned} \tag{2.3}$$

the kinetic energy operator, $\hat{T}_{\text{e}}$, the nuclear-electron attraction operator, $\hat{V}_{\text{ne}}$, the electron-electron repulsion operator, $\hat{V}_{\text{ee}}$, and the nuclear-nuclear repulsion

operator, $\hat{V}_{\text{nn}}$. Under the Born-Oppenheimer approximation, the nuclear coordinates becomes parameters and the $\hat{V}_{\text{nn}}$ part simply follows Coloumb's law. The $\hat{T}_{\text{e}}$ and $\hat{V}_{\text{ne}}$ acts on only one electron and these one-electron operators can be easily handled through the separation of variables in the wavefunction. The problem lies with $\hat{V}_{\text{ee}}$, which acts on two electrons simultaneously. Currently, there exist no mathematical treatment that allows us to handle such two-electron operators. This is why the Schröndinger equation has no analytical solutions for systems with more than one electron. To solve the Schröndinger equation approximately, we can either modify the Hamiltonian or rewrite the wavefunction in a special form. An example of the former approach would be to ignore the $\hat{V}_{\text{ee}}$ completely, making the Hamiltonian separable. Obviously, this is not a good idea as electrons do interact strongly with each other. Instead, the HF method does the latter by rewriting the wavefunction in a special form.

Hartree tackled the insolubility of the Schröndinger equation by writing the wavefunction as a product of one-electron wavefunctions

$$\Psi_{\text{HP}}(\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r}_N) = \phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2)\cdots\phi_N(\mathbf{r}_N) \tag{2.4}$$

known as the Hartree Product.[118] However, Slater and Fock independently pointed out that eq (2.4) does not satisfy the antisymmetric nature of the wavefunction when two electrons are exchanged.[119] Interestingly, a determinant satisfies this antisymmetric property. Thus, we can write a Slater determinant[120]

$$\Psi = \frac{1}{\sqrt{N_{\text{el}}!}} \begin{vmatrix} \phi_1(\mathbf{r}_1) & \phi_2(\mathbf{r}_1) & \cdots & \phi_N(\mathbf{r}_1) \\ \phi_1(\mathbf{r}_2) & \phi_2(\mathbf{r}_2) & \cdots & \phi_N(\mathbf{r}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{r}_N) & \phi_2(\mathbf{r}_N) & \cdots & \phi_N(\mathbf{r}_N) \end{vmatrix} \tag{2.5}$$

that assumes that the electrons are independent of each other and satisfy the antisymmetry principle. By minimizing the energy of the Slater determinant in eq (2.5) with respect to the one-electron wavefunctions (an application of the variational principle), we arrive at the canonical Hartree–Fock equations

$$\hat{F}_i \phi_i(\mathbf{r}_1) = \varepsilon_i \phi_i(\mathbf{r}_1) \tag{2.6}$$

where ε_i is the energy of orbital ϕ_i and $\hat{F}_i$ is the Fock operator comprising

$$\hat{F}_i = \hat{h}_i + \sum_{j=1}^{N} \left(\hat{J}_i - \hat{K}_i \right) \tag{2.7a}$$

$$\hat{h}_i = -\frac{1}{2}\nabla_i^2 - \sum_{\mathrm{A}} \frac{Z_{\mathrm{A}}}{|\mathbf{R}_A - \mathbf{r}_i|} \tag{2.7b}$$

$$\hat{J}_i \phi_j(\mathbf{r}_2) = \left[\int \phi_i^*(\mathbf{r}_1) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \phi_i(\mathbf{r}_1) d\mathbf{r}_1 \right] \phi_j(\mathbf{r}_2) \tag{2.7c}$$

$$\hat{K}_i \phi_j(\mathbf{r}_2) = \left[\int \phi_i^*(\mathbf{r}_1) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \phi_j(\mathbf{r}_1) d\mathbf{r}_1 \right] \phi_i(\mathbf{r}_2) \tag{2.7d}$$

the one-electron operator, $\hat{h}_i$, the coulomb operator, $\hat{J}_i$, and the exchange operator, $\hat{K}_i$. Note that the $\hat{V}_{\mathrm{nn}}$ contribution is not shown as it can be trivially added to the total energy. The Fock operator is an effective one-electron operator with the $\hat{h}_i$ part containing the kinetic energy of an electron and the attraction of the electron with all the nuclei while both $\hat{J}_i$ and $\hat{K}_i$ describe the repulsion of the electron with all the other electrons. Clearly, both $\hat{J}_i$ and $\hat{K}_i$ depend on the one electron wavefunctions, $\phi_j(\mathbf{r}_2)$, which are the solutions to the Fock operator itself. Thus, the Hartree–Fock equations have to be solved iteratively. The Hartree–Fock equations can be solved numerically on a set of grid points representing the wavefunction. Alternatively, a basis set can be used to express the wavefunction and the Hartree–Fock equations are transformed into the Roothaan–Hall equations.[121,122] The iterative procedure to solve the Roothaan–Hall equations is known as the self-consistent field (SCF) procedure.

Throughout the derivation of the HF method, there is only one assumption: writing down the total wavefunction as a Hartree Product. Consequently, this leads to a mean-field approximation where each electron behaves independently of all the other electrons and experiences an averaged interaction with the other electrons through the Coulomb and exchange operator. It should be emphasized that no modifications were made to the Hamiltonian although the description of the mean-field approximation seems to suggest otherwise. In reality, individual electrons are affected by each other and such interactions cannot be described in an averaged manner. Thus, the HF method can be improved via corrections to account for the correlated motions of electrons.

The reason behind the HF method being the basis for other electronic structure methods probably lies in the elegance of the theory. Other than the Born–Oppenheimer approximation that is common to all electronic structure methods, there is only a single approximation in the HF method, making it easier to propose corrections. Furthermore, the HF method provides a well-defined energy and wavefunction. The HF energy can be converged to the complete basis set (CBS) limit and the difference between the HF-CBS energy and true energy clearly defines the electron correlation energy. The HF wavefunction is obtained after the SCF procedure and often served as a reference wavefunction in post-HF methods. Thus, such independent particle models are easily understood, which may explain why the Kohn–Sham approach[123] in Density Functional Theory (DFT) bears much resemblance to the HF method.

2.2.2 How are basis sets being constructed?

The basis set is a set of mathematical functions used to describe the wavefunction. Wavefunctions in electronic structure methods are commonly described using molecular orbitals built from the linear combination of atomic orbitals, i.e., the LCAO-MO approach. The atomic orbitals are typically atom-centred one-electron functions. Since the Schröndinger equation cannot be solved an-

alytically for systems with more than two electrons, the functional form of the atomic orbitals are not known except for hydrogenic atoms, i.e., systems with one electron. Then, how do we choose the functional form for these atomic orbitals and optimise the associated parameters?

From the solutions of hydrogenic atoms, atomic orbitals take the form of a slater-type orbital (STO)

$$\phi^{\mathrm{STO}}_{\zeta,n,l,m}(r,\vartheta,\varphi) = NY_{l,m}(\vartheta,\varphi)r^{n-1}\mathrm{e}^{-\zeta r} \tag{2.8}$$

where N is the normalization constant and $Y_{l,m}$ are the spherical harmonics. The integers l and m control the angular momentum while ζ is the exponent, which determines the spread of the orbital. Despite having the correct exponential decay behaviour with increasing r, these STO are rarely used in quantum chemistry as it is difficult to numerically evaluate the associated two-electron integrals. Instead, gaussian-type orbitals (GTO) are frequently employed

$$\phi^{\mathrm{GTO}}_{\zeta,n,l,m}(r,\vartheta,\varphi) = NY_{l,m}(\vartheta,\varphi)r^{2n-2-l}\mathrm{e}^{-\zeta r^2} \tag{2.9}$$

as the product of two Gaussian functions of different exponents at different centres can be expressed as a single Gaussian located at an intermediate location

$$G_{\mathrm{A}}(\mathbf{r}) = \left(\frac{2\alpha}{\pi}\right)^{3/4}\mathrm{e}^{-\alpha(\mathbf{r}+\mathbf{R}_{\mathrm{A}})^2} \tag{2.10a}$$

$$G_{\mathrm{B}}(\mathbf{r}) = \left(\frac{2\beta}{\pi}\right)^{3/4}\mathrm{e}^{-\beta(\mathbf{r}+\mathbf{R}_{\mathrm{B}})^2} \tag{2.10b}$$

$$G_{\mathrm{A}}(\mathbf{r})G_{\mathrm{B}}(\mathbf{r}) = \left(\frac{2}{\pi}\right)^{2}(\alpha\beta)^{3/4}\mathrm{e}^{\frac{\alpha\beta}{\alpha+\beta}(\mathbf{R}_{\mathrm{A}}-\mathbf{R}_{\mathrm{B}})^2}\mathrm{e}^{-(\alpha+\beta)(\mathbf{r}+\frac{\alpha\mathbf{R}_{\mathrm{A}}+\beta\mathbf{R}_{\mathrm{B}}}{\alpha+\beta})^2} \tag{2.10c}$$

This Gaussian product theorem allows the two-electron integrals to be evaluated efficiently. A Gaussian function, with its r^2 exponent, decays more rapidly than an exponential function. To remedy this, we can take linear combinations of

GTO to mimic a STO. Some of the linear combinations are predetermined to reduce the computational cost. For example, one can fix the linear combination of GTO describing core electrons as we expect these core electron orbitals to be hardly changed during chemical bonding. This process is known as basis set contraction and the resulting functions are termed contracted GTO. In essence, more GTO are required to achieve a certain accuracy as compared to STO but the former is still preferred due to computational efficiency.

After choosing GTO as the functions, we need to decide on the number and types of GTO to be included in a basis set. By the type of the function, we are referring to its angular momentum. For example, for an Oxygen atom, we would require two s-type and three p-type functions to describe the 1s, 2s, $2p_x$, $2p_y$ and $2p_z$ core-valence orbitals. Apart from this minimal basis, we can have additional sets of basis functions to better describe the valence orbitals, giving rise to split valence basis sets. Furthermore, basis functions of higher angular momentum (polarization functions) can be included to allow the orbitals to change shape and basis functions with small ζ (diffuse functions) can be added to account for electrons that are relatively far from the nuclei.

Next, we need to understand the notion of basis set balance. Let us consider a minimal basis augmented with many sets of polarization functions. Due to the insufficient number of sp-type functions, the valence orbitals will be inadequately described. The many sets of polarization functions may then be included to compensate for this inadequacy albeit in an inefficient manner. The inclusion of polarization functions to describe the valence space necessarily places small amounts of electron density at undesirable locations, resulting in artefacts. Similarly, a split valence basis set with little or no polarization functions cannot capture changes in orbital shapes during chemical bonding. Thus, the number of valence and polarization functions should complement each other. A rough guide is that the number of sets of basis functions of a particular angular momentum, n_l, should be one less than that of one lower angular momentum,

n_{l-1}, i.e., $n_l = n_{l-1} - 1$. For example, a triply-split valence basis set should be accompanied by two sets of d-type functions and one set of f-type functions, which can be denoted as a 4s3p2d1f basis for an Oxygen atom. Note that despite being triply split valence, we have "4s" due to an additional set of s-type functions describing the core orbitals. The Dunning correlation-consistent basis sets, labelled cc-pVnZ, follow this recipe and the user only needs to specify the cardinal number, n, which determines the number of sets of split valence functions.[34] The cc-pVnZ basis sets were targeted towards reproducing the correlation energy of the valence electrons, which contribute most significantly to chemical bonding. The cc-pVnZ basis sets can also be augmented with diffuse functions, giving the augmented form, labelled aug-cc-pVnZ.[124] Table 2.1 gives the composition of the cc-pVnZ and aug-cc-pVnZ basis sets and the corresponding number of basis functions, M, in a water molecule. Interestingly, M depends on the cardinal number n as such

$$M_{\mathrm{cc-pVnZ}} = \frac{1}{3}(n+1)(n+\frac{3}{2})(n+2) \tag{2.11a}$$

$$M_{\mathrm{aug-cc-pVnZ}} = M_{\mathrm{cc-pVnZ}} + (n+1)^2 \tag{2.11b}$$

As we increase the cardinal number in the cc-pVnZ and aug-cc-pVnZ series, M approximately doubles, limiting the use of the largest basis sets to the smallest systems. The DZ and TZ basis sets are routinely used while the larger QZ and 5Z basis sets are usually employed in benchmark calculations.

With the number and types of basis functions determined, we need to specify the exponents for the basis functions for the basis set to be well defined. To this end, we shall discuss how Dunning arrived at the exponents for the cc-pVnZ basis sets.[34] The s-type and p-type functions were taken from a previous optimization using atomic HF calculations. The exponents of these functions were re-optimized but had little effect on the energy. Thus, Dunning

Table 2.1: Composition of the standard Dunning correlation consistent cc-pVnZ basis sets and the augmented form aug-cc-pVnZ, as well as the corresponding number of basis functions, M.

Basis	Contraction		M		
	Hydrogen	First row elements	H	O	H_2O
cc-pVDZ	2s1p	3s2p1d	5	14	24
cc-pVTZ	3s2p1d	4s3p2d1f	14	30	58
cc-pVQZ	4s3p2d1f	5s4p3d2f1g	30	55	115
cc-pV5Z	5s4p3d2f1g	6s5p4d3f2g1h	55	91	201
cc-pV6Z	6s5p4d3f2g1h	7s6p5d4f3g2h1i	91	140	322
aug-cc-pVDZ	3s2p	4s3p2d	9	23	41
aug-cc-pVTZ	4s3p2d	5s4p3d2f	23	46	92
aug-cc-pVQZ	5s4p3d2f	6s5p4d3f2g	46	80	172
aug-cc-pV5Z	6s5p4d3f2g	7s6p5d4f3g2h	80	127	287
aug-cc-pV6Z	7s6p5d4f3g2h	8s7p6d5f4g3h2i	127	189	443

focussed on the polarization functions. Using the aforementioned (sp) functions as a starting point, sets of d-type functions were added incrementally until there is little change in the calculated configuration interaction (CI) correlation energy. Sets of f-type orbitals are then added to the "d-orbital saturated" basis set until the correlation energy has stabilised and this is repeated for the g-type orbitals. The exponents of the basis functions for each angular momentum follows an even-tempered expansion

$$\zeta_i = \alpha\beta^i \tag{2.12}$$

where α and β are parameters that were optimised to achieve the maximum lowering of the correlation energy. With this, Dunning found that the polarization functions can be grouped according to the extent of energy lowering from the addition of new functions. The grouping of the polarization functions follows the rough guide mentioned in the previous paragraph where the polar-

ization functions should be added in sets of (1d), (2d1f), (3d2f1g) and so on. This laborious optimization process would give rise to the cc-pVnZ basis sets. Since the polarization functions are grouped according to the extent of lowering of correlation energy, an increase in the cardinal number leads to a systematic convergence of the total energy towards the CBS limit.

2.2.3 What is the relationship between MP and CC methods?

Electron correlation methods are post-HF methods used to correct for the mean-field approximation in the HF method. Two common classes of electron correlation methods include the Møller–Plesset perturbation theory to the n-th order (MPn) and the Coupled Cluster (CC) method. The MPn method is based on many-body perturbation theory and treat the difference between twice the true electron-electron repulsion and electron-electron repulsion of the Fock operator as a perturbation. The CC theory introduces corrections in an exponential manner so that corrections of a given type is treated to infinite order. Both the MPn and CC methods are very different approaches to solve the correlation problem and they possess distinct implementations, with the former being perturbative in nature while the latter is iterative. To compare the accuracy between MPn and CC methods, it is important to find the relationship between the two methods.

To establish the connection between MPn and CC methods, we need to look at the Configuration Interaction (CI) method. Let us assume that we performed a restricted HF calculation, with all the electrons paired in the MOs, on a system containing N_{el} electrons using a basis set containing M basis functions. We obtained the solutions of the corresponding Roothaan–Hall equations which contain $N_{\mathrm{el}}/2$ occupied MOs and $(M - N_{\mathrm{el}}/2)$ unoccupied virtual MOs. With that, we can write down a series of excited Slater determinants by taking electrons from occupied MOs and placing these electrons into virtual MOs. This excitation process recovers electron correlation and the CI method variationally optimize the contribution of each excited determinant. If we include all possible

excited determinants, we recover the full electron correlation and this is known as the Full CI method. It seems counter-intuitive that exciting electrons lowers the total energy. However, recall that the HF energy only depends on the occupied MOs. Thus, the only way to improve the wavefunction within the limits of the basis set is to use the virtual MOs. The electron excitation process allows the system to relax itself to a lower energy as the electrons can better avoid each other. Physically, this correspond to the correlated motion of electrons, i.e., electron correlation! Thus, we can use the number of electron excitations as a measure of the amount of electron correlation.

Let us quickly revisit the MPn methods.[125] In the MPn methods, we introduce a perturbation $\hat{V}$ to the sum over Fock operators, previously defined in eq (2.7), which serves as the reference Hamiltonian, $\hat{H}_0$

$$\hat{H}_{\text{el}} = \hat{H}_0 + \hat{V} = \sum_{i=1}^{N_{el}} \hat{F}_i + \hat{V} \tag{2.13}$$

At zero-order, we get the sum of the orbital energies, $\sum_{i=1}^{N_{el}} \varepsilon_i$, by solving the Hartree–Fock equations in eq (2.6). At first-order perturbation, we obtain the Hartree–Fock energy, which accounts for double counting the electron-electron repulsion in the sum of the orbital energies. Thus, electron correlation is first recovered at the second-order perturbation. The energy expressions for the second-order through fifth-order perturbation[126] are

$$E^2 = \sum_{s}^{\text{D}} \hat{V}_{0s} a_s^1 \tag{2.14a}$$

$$E^3 = \sum_{st}^{\text{D}} a_s^1 \bar{V}_{st} a_t^1 \tag{2.14b}$$

$$E^4 = \sum_{s}^{\text{D}} \sum_{t}^{\text{SDTQ}} a_s^1 \bar{V}_{st} a_t^2 - E^2 \sum_{s}^{\text{D}} |a_s^1|^2 \tag{2.14c}$$

$$E^5 = \sum_{st}^{\text{SDTQ}} a_s^2 \bar{V}_{st} a_t^2 - 2E^2 \sum_{s}^{\text{D}} a_s^1 a_s^2 - E^3 \sum_{s}^{\text{D}} |a_s^1|^2 \tag{2.14d}$$

where a_s^i is the amplitude associated with the s excitation (which can be S, D, T or Q denoting Singles, Doubles, Triples or Quadruples respectively) in the i-th order perturbed wavefunction and $\bar{V}$ is a shorthand for the difference between the perturbation and Hartree–Fock energy, $\bar{V} = \hat{V} - E^1$. We can then partition the energies according to the excitations involved as follows

$$E^2 = E^2_{\mathrm{D}} \tag{2.15a}$$

$$E^3 = E^3_{\mathrm{D}} \tag{2.15b}$$

$$E^4 = E^4_{\mathrm{S}} + E^4_{\mathrm{D}} + E^4_{\mathrm{T}} + E^4_{\mathrm{Q}} \tag{2.15c}$$

$$E^5 = \left(E^5_{\mathrm{SS}} + E^5_{\mathrm{SD}} + E^5_{\mathrm{DD}}\right) + \left(E^5_{\mathrm{ST}} + E^5_{\mathrm{DT}} + E^5_{\mathrm{DQ}}\right) + \left(E^5_{\mathrm{TT}} + E^5_{\mathrm{TQ}} + E^5_{\mathrm{QQ}}\right) \tag{2.15d}$$

Only the double excitations (D subscript) are present in the E^2 and E^3. The E^4 involves the S, D, T and Q excitations, evident from the summation in the first term in eq (2.14c). The second term in E^4 in eq (2.14c), a re-normalization term, is included into E^4_{Q} due to partial cancellation of terms. The double summations in the first term in E^5 in eq (2.14d) can be partitioned in a similar manner. Similarly, due to partial cancellation of terms, we have included the second and third term in E^5 in eq (2.14d) into E^5_{DQ} and E^5_{QQ} respectively. Note that we have grouped the terms into terms not containing T or Q, terms that are linear in T or Q and terms that are quadratic in T or Q. These energy partitions will be used later to discuss the similarities of the MPn and CC methods.

The MPn methods add electron correlation from different excitations (S, D, T, Q and so on) up to a particular order (in E^2, E^3, E^4, E^5 and so on). Instead, the CC methods include all the electron correlation from a given excitation to infinite order.[127,128] This is achieved by applying the exponential operator, $\mathrm{e}^{\mathbf{T}}$, to the excitations

$$\Psi_{\mathrm{CC}} = \mathrm{e}^{\mathbf{T}}\Psi_0 \tag{2.16a}$$

$$\mathbf{T} = \mathbf{T}_1 + \mathbf{T}_2 + \cdots + \mathbf{T}_{N_{\mathrm{el}}} \tag{2.16b}$$

$$\mathrm{e}^{\mathbf{T}} = \mathbf{1} + \mathbf{T} + \frac{1}{2}\mathbf{T}^2 + \frac{1}{6}\mathbf{T}^3 + \cdots = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{T}^k \tag{2.16c}$$

where $\mathbf{T}_i$ is an excitation operator that generates all Slater determinants with i excited electrons. Here, Ψ_{CC} and Ψ_0 are the CC correlated wavefunction and HF reference wavefunctions respectively. Due to the exponential operator, we can obtain products of excitation operators, which represents disconnected excitations. For example, a $\mathbf{T}_2^2$ operator corresponds to exciting two non-interacting pairs of interacting electrons. Notably, when we include all the excitations up to N_{el} electrons in eq (2.16b), we obtain the Full CI method. However, this is computationally infeasible in practice. Thus, we truncate eq (2.16b) by the second term to get the Coupled Cluster Single and Double excitation (CCSD) method. Truncation by the third term yield the CCSDT method and so on. Due to the Brillouin's theorem, the single excitations do not contribute directly to the correlation energy. Thus, a CCS method does not recover any electron correlation and is equivalent to the HF method. However, the single excitations do contribute indirectly via coupling with the double excitations in the CCSD method. Sometimes, a perturbative treatment of higher excitations can be augmented. One example is the CCSD with perturbative Triples [CCSD(T)],[33] which is the gold standard of quantum chemistry. Table 2.2 summarizes the list of electron correlation methods covered and their computational cost.[129]

So far, we established that the number of excited Slater determinants included gives a measure of electron correlation and also provided a brief outline of the MPn and CC theories. Furthermore, we partitioned the MPn correlation energy according to the excitations involved and these energetic components will serve as the basis of comparison between the MPn and CC methods. In

Table 2.2: Computational scaling in terms of the number of basis functions, M, for the MPn and CC methods. Hybrid CC methods involve performing an iterative CC procedure, followed by a perturbative treatment of higher excitations. The computational cost of HF and Full CI (FCI) is also included for completeness.

Scaling	MPn methods	CC methods	Hybrid CC	HF and FCI
M^4	MP1=HF	CCS=HF		HF
M^5	MP2			
M^6	MP3, MP4(SDQ)	CCSD		
M^7	MP4		CCSD(T)	
M^8	MP5	CCSDT	CCSDT(Q)	
M^9	MP6			
M^{10}	MP7	CCSDTQ		
$M!$				FCI

Table 2.3, we compared MPn methods up to the fourth order with some of the commonly used CC methods. The excitations included in the MPn methods are listed exhaustively while there are higher-order terms included in the CC methods that are not shown here. The comparison table tells us the correlation energies that we obtain "for free" from a particular calculation. For example, performing a CCSD calculation also gives the MP2, MP3 and MP4(SDQ) correlation energy. The astute reader would have realized that the CCSD method includes the E_Q^4 energy which corresponds to a quadruple excitation. This is because the MP4 quadruples contributions originate from the disconnected $\mathbf{T}_2^2$ excitations. The exponential nature of the CC methods ensure that all disconnected $\mathbf{T}_2^n$ excitations for $n = 1, 2, \cdots, \infty$ are included in the CCSD method.

From the order of appearance of excitations in the MPn methods (Table 2.3), we can deduce the importance of different excitations. As mentioned earlier, the effect of $\mathbf{T}_1$ is small, originating indirectly from the coupling with double excitations. Thus, the most important contribution to the correlation energy comes from $\mathbf{T}_2$. This is echoed by the observation that only double excitations are present in the E^2 and E^3. The connected $\mathbf{T}_3$ and disconnected $\mathbf{T}_2^2$ excitations

Table 2.3: Comparison of MPn and CC methods up to the fourth-order perturbation theory. ✓indicates that the term is included completely.

Method	E_{D}^2	E_{D}^3	E_{S}^4	$E_{\mathrm{D}}^4 + E_{\mathrm{Q}}^4$	E_{T}^4
MP2	✓				
MP3	✓	✓			
MP4(SDQ)	✓	✓	✓	✓	
MP4	✓	✓	✓	✓	✓
CCSD	✓	✓	✓	✓	
CCSD(T)	✓	✓	✓	✓	✓

are probably the next most important contribution to the correlation energy, appearing in the E_T^4 and E_Q^4 energy respectively. This explains the need for at least a perturbative treatment of the triple excitations in the CCSD(T) method. From Table 2.3, we can clearly see that the most important double excitations and associated higher order terms, $\mathbf{T}_2$ and $\mathbf{T}_2^2$, are accounted for in the CCSD part. The perturbative treatment of triple excitations is then required to account for effects of $\mathbf{T}_3$. Notably, the MP4 method also accounts for the aforementioned excitations. To further our analysis, we compared CCSD(T) with other MPn and CC methods of similar accuracy (Table 2.4). Comparing CCSD(T) and MP4, the former contains fifth-order correlation energy terms that are absent in the latter. Thus, the CCSD(T) method is preferred over the MP4 method since both methods have an identical M^7 scaling (Table 2.2). We also arrive at a similar conclusion when comparing the CCSD and MP4(SDQ) methods, both of which have a M^6 scaling. That said, the MP2 method is also frequently used in the literature, being the cheapest electron correlation method. Through the comparison of the MPn and CC methods, we realized that the CC methods recover more electron correlation than the MPn methods with similar costs. Together with the importance of triple excitations, the CCSD(T) method represents the most effective method to recover all the important electron correlation effects. Thus, the CCSD(T) method is the gold standard of quantum chemistry.

Table 2.4: Comparison of MPn and CC methods in the fifth-order perturbation theory. The methods are accurate up to fourth-order perturbation theory, containing all the contributions up to E^4, with the exception of the CCSD method which does not include the E_{T}^4 energy. ✓indicates that the term is included completely while $\approx$ denotes that the term is included partially. Note that none of the fifth-order perturbation theory terms are included at MP4 and that row is supposed to be empty.

Method	$E_{\mathrm{SS}}^5 + E_{\mathrm{SD}}^5 + E_{\mathrm{DD}}^5$	E_{ST}^5	E_{DT}^5	E_{DQ}^5	E_{TT}^5	E_{TQ}^5	E_{QQ}^5
MP4							
MP5	✓	✓	✓	✓	✓	✓	✓
CCSD	✓	$\approx$		✓		$\approx$	$\approx$
CCSD(T)	✓	✓	✓	✓		$\approx$	$\approx$
CCSDT	✓	✓	✓	✓	✓	$\approx$	$\approx$

2.2.4 How do we tackle the slow basis set convergence in post-HF methods?

The CCSD(T) method is able to yield highly accurate results for many molecular properties and reproduce energies to chemical accuracy. However, the method—and electron correlation methods in general—suffers from two major limitations. Firstly, the CCSD(T) method has a very steep M^7 scaling, a consequence of the summations over the various included excitations. Secondly, the correlation energy converges very slowly with respect to the basis set size. We will discuss the various strategies employ to accelerate this basis set convergence.

Before discussing the strategies, we need to understand the cause of the slow basis set convergence in electron correlation methods. As mentioned in Section 2.2.3, we can recover the effects of correlation by exciting electrons from the occupied MOs to the virtual MOs. Since the occupied MOs represent the best linear combination of basis functions that minimizes the HF energy, the virtual MOs are more sensitive to the quality of the basis sets. Physically, the electrons move in correlated manner to minimize electron-electron repulsion and this would demand greater flexibility of the wavefunction.

Mathematically, this is due to the electron-electron repulsion operator

$$\hat{V}_{\mathrm{ee}} = \sum_{i>j} \frac{1}{r_{ij}} \tag{2.17}$$

since the electron-electron repulsion drives the correlated motion of electrons. Note that there are two forms of electron correlation, namely the Fermi and Coulomb correlation, arising from electron pairs of parallel and anti-parallel spins respectively. The former is accounted for in the antisymmetrization of the wavefunction using Slater determinants while electron correlation methods seek to recover the latter. The electron-electron repulsion operator, $\hat{V}_{\mathrm{ee}}$, has a singularity when any of the inter-electronic distance, say r_{12}, becomes zero. Kato showed that this results in the exact wavefunction containing a cusp, i.e., a discontinuous derivative[130] and the discontinuity can be expressed as

$$\left(\frac{\partial \Psi(r_{12})}{\partial r_{12}}\right)_{r_{12}=0} = \frac{1}{2}\Psi(r_{12}) \tag{2.18}$$

The slow basis set convergence is attributed to the fact that the product of one electron functions is incompatible with this cusp requirement. The product of orbitals remains smooth when two electrons are close, i.e., small r_{12}, which is undesirable.[131,132] Thus, a lot of basis functions are required to mimic the cusps in the wavefunction. It should be emphasized that this cusp behaviour is essentially a two-electron effect and thus is absent in the HF method which assumes that the particles are independent of each other.

One possible strategy to avoid the slow convergence problem is to perform an extrapolation to the CBS limit. Since there are no modifications made to the theory or wavefunction, this can be easily performed by applying some formula to energies calculated using different basis sets. Obviously, the basis sets should exhibit a systematic convergence and the Dunning cc-pVnZ basis sets (and their augmented counterparts) make excellent choices. This is because the

polarization functions are grouped according to their contribution to the correlation energy, allowing for a balanced treatment of electron correlation (Section 2.2.2). Furthermore, we need to know how the correlation energy changes with the quality of the basis set to derive an extrapolation formula. Schwartz found that the MP2 correlation energy for the Helium atom converges asymptotically as $(l+\frac{1}{2})^{-4}$ where l is the highest angular momentum quantum number of the atomic orbitals employed.[133] In Schwartz's work, the radial part of the basis functions were saturated while limiting the highest angular momentum of the functions to l. Thus, the trend may not apply directly to the cc-pVnZ basis sets. Nonetheless, this suggests that the correlation energy has an inverse power relationship with the highest angular momentum function in the basis set. Indeed, Helgaker *et al.* found that the the correlation energy can be expressed as

$$E_{\mathrm{corr},n} = E_{\mathrm{corr},\infty} + An^{-3} \tag{2.19}$$

where $E_{\mathrm{corr},n}$ is the correlation energy in the cc-pVnZ basis set, $E_{\mathrm{corr},\infty}$ is the correlation at the CBS limit and A is a parameter to be determined.[134] Since there are two unknowns, namely $E_{\mathrm{corr},\infty}$ and A, two correlation energies calculated at different n are required. We can solve eq (2.19) in terms of $E_{\mathrm{corr},\infty}$ and obtain a CBS extrapolation formula

$$E_{\mathrm{corr},\infty} = \frac{x^3 E_{\mathrm{corr},x} - y^3 E_{\mathrm{corr},y}}{x^3 - y^3} \tag{2.20}$$

Another approach would be to modify the wavefunction such that it better satisfies the cusp condition. Notably, the cusp condition given in eq (2.18) implies that the exact wavefunction depends linearly with r_{12} when r_{12} is small

$$\Psi(r_{12}) = k + \frac{1}{2} r_{12} + \cdots \tag{2.21}$$

This suggest that the electron correlation methods can benefit from the inclusion

of interelectronic distance dependence in the wavefunction. Based on this idea, Kutzelnigg and Klopper developed the R12 methods[135] where HF determinants multiplied with the interelectronic distance, $r_{12}\Psi_{\mathrm{HF}}$, are included in the wavefunction. However, this leads to computational difficulties from the introduction of integrals varying with three and four electron coordinates. Fortunately, these integrals can be avoided through the use of the Resolution of the Identity technique with an auxiliary basis set. With that, the R12 methods can be calculated with reasonable cost and the correlation energy converges rapidly as $(l+1)^{-7}$. Other than the R12 methods, Ten-no proposed that the r_{12} factor is replaced with a Slater-type function $F_{12} = \mathrm{e}^{-\gamma r_{12}}$ as the linear r_{12} factor is unphysical at large distances.[136] These F12 methods are more numerically stable and provided better convergence of the correlation energy with respect to the basis set.

2.2.5 Why can't DFT describe dispersion?

Electron correlation methods provide a systematic way to approach the exact Schröndinger equation. However, these methods are limited by their poor computational scaling and slow basis set convergence. Fundamentally, this is due to the high dimensionality of these wavefunction-based methods where there are four degrees of freedom (three spatial and one spin) associated with each electron. Density functional theory (DFT) offers an attractive alternative, based on the Hohenberg–Kohn theorem that the ground state electronic energy depends solely on the electron density, ρ, a three-dimensional quantity.[137] However, early applications of DFT to study intermolecular interactions were unsatisfactory.[138–140] It was then identified that common DFT functionals, such as B3LYP, cannot describe the dispersion interaction. Here, we shall attempt to unravel the reason and discuss possible remedies to this DFT dispersion problem.

Often described as a "instantaneous dipole-induced dipole" effect in high-school textbooks, dispersion is a quantum mechanical effect where the correlated motions of electrons in two molecules lowers the interaction energy. An

electron correlation effect, dispersion is clearly missing in the HF method. Furthermore, dispersion can be derived from the second order perturbation theory where the two interacting molecules are excited.[48] Thus, we can establish that dispersion is an electron correlation effect and occurs at long-range due to the fluctuation of two well separated charge densities.

The inability of DFT in reproducing the dispersion interaction stems from the choice of functionals.[141] Common DFT functionals employ either the local density approximation (LDA) or some form of the generalized gradient approximations (GGA). The exchange-correlation functional only depends on the electron density under the LDA and contributions from the gradient of the density are incorporated in the GGA functionals. Thus, the exchange-correlation energies for both LDA and GGA can be written as

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[\rho] = \int \rho(\mathbf{r})\varepsilon_{\mathrm{xc}}(\rho(\mathbf{r}))d\mathbf{r} \tag{2.22a}$$

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[\rho] = \int \rho(\mathbf{r})\varepsilon_{\mathrm{xc}}(\rho(\mathbf{r}),\nabla\rho(\mathbf{r}))d\mathbf{r} \tag{2.22b}$$

where $\varepsilon_{\mathrm{xc}}$ is the exchange-correlation energy. Higher derivatives, $\nabla^2\rho(\mathbf{r})$, can be included to give the meta-GGA methods. The GGA (and meta-GGA) functionals are often described as "non-local" due to the inclusion of the derivatives of the charge density. However, the derivatives can only describe the immediate vicinity of a particular point, making the description of exchange-correlation effects localized, i.e., semi-local. This is clearly incompatible with the dispersion interactions which depend on the fluctuation of charge densities at two well-separated points. Thus, only a truly non-local functional that depends on the density at two different points, say $\mathbf{r}$ and $\mathbf{r}'$, can describe dispersion properly

$$E_{\mathrm{xc}}^{\mathrm{nonlocal}}[\rho] = \int\int \rho(\mathbf{r})\rho(\mathbf{r}')\varepsilon_{\mathrm{xc}}(\rho(\mathbf{r}),\nabla\rho(\mathbf{r}),\rho(\mathbf{r}'),\nabla\rho(\mathbf{r}'))d\mathbf{r}d\mathbf{r}' \tag{2.23}$$

It should be emphasized that DFT itself is capable of approaching the exact Schröndinger equation provided that the universal exchange-correlation functional is known. Unfortunately, this is not the case at present, which explains the plethora of functionals in the literature. Nonetheless, it is clear that the universal functional should be truly non-local but this would likely increase the computational cost due to the extra degrees of freedom involved.

There are several approaches to correct for the lack of dispersion interactions in DFT, which have been reviewed extensively elsewhere.[142,143] We shall discuss four different approaches according to the ease of implementation. The simplest way to incorporate dispersion into DFT is to add a dispersion energy correction, E_{disp} to the DFT energy, E_{DFT}, as follows

$$E_{\text{tot}} = E_{\text{DFT}} + E_{\text{disp}} \tag{2.24}$$

Interestingly, this type of corrections has been applied even earlier to treat the similar lack of dispersion in the HF method. The correction is motivated by the initial observations that DFT functionals are unable to reproduce the R^{-6} distance dependence that is characteristic of dispersion. Thus, the dispersion correction term often bears the form

$$E_{\text{disp}} = -\frac{C_6}{R^6} - \frac{C_8}{R^8} - \frac{C_{10}}{R^{10}} \tag{2.25}$$

where R denotes an interatomic distance and C_n are the dispersion coefficients. The C_n coefficients are usually related to the static polarizability derived from the DFT density. However, the effects of dispersion are often not included in the optimization of the density, making the treatment not self-consistent. Note that the inverse-distance expansion in eq (2.25) diverges rapidly and gives unphysical values at short-range. To resolve this, one can multiply damping functions[89] onto the dispersion terms to gradually remove the dispersion contributions with

decreasing R. The DFT-D3 method[144,145] and Tkatchenko–Scheffler model[146] are some examples of methods that employ a dispersion correction term.

Secondly, current functionals can be re-parametrized to better describe dispersion interactions. Most of the earlier and commonly used functionals, such as B3LYP, were focussed on chemical bonding and thus not designed for noncovalent interactions. While these LDA- or GGA-based functionals are intrinsically local or semi-local, the functionals can be tuned to produce additional attraction at an intermediate range where molecules are in close contact. This can be achieved by fitting elaborate reference data from experiments or highly-accurate quantum chemical calculations. One example is the M06 suite of functionals,[147] which comprises four different functionals (M06-L, M06, M06-2X and M06-HF) calibrated for different types of calculations and molecules. A third approach would be to employ double hybrid functionals. Hybrid functionals include a part of exact exchange from the HF theory. On top of that, double hybrid functionals introduce an additional portion of correlation energy, typically from an MP2-type calculation. However, this would increase the computational cost to that of MP2 calculations, which goes against our initial motivation of the low computational cost of DFT. Some examples of double hybrid functionals are the B2PLYP[148] and mPW2PLYP.[149] Finally, truly non-local functionals can be developed. The first truly non-local functional, vdW-DF,[150] contains a non-local correlation term, expressed as a density-density interaction, following the form of eq (2.23). The ε_{xc} is based on a analytically derived frequency-dependent response function and depends on both the density and its gradient. This non-local term is then combined with an LDA correlation functional to recover the total functional. One important feature of the vdW-DF is the seamless integration of the LDA and non-local terms as the latter term does not have a local contribution.

The lack of dispersion in many DFT functionals can be traced to a bigger problem plaguing the DFT theory—the universal exchange-correlation func-

tional is unknown. This leads to a variety of functionals parametrized for different applications. Despite this, DFT methods still have a large appeal due to their low computational cost. This allows for the study of larger chemical systems, expanding the chemical space that can be studied theoretically.

2.3 Remarks on Computations and Experiments

By answering the five questions, we covered a substantial part of the theoretical background of quantum chemistry. The creativity and dedication of generations of quantum chemists have led to the constant development of new theories and methodologies. Coupled with the advancement of computer hardware, quantum chemistry has matured to a stage where the accuracy of computational studies rival that of experiment. Thus, experiment and computation have become inextricably intertwined, akin to the relationship between sword and shield. Experiments are likened to swords, providing the most definitive understanding of the nature itself. Yet, these swords can get dull with repeated use and require tremendous resources for maintenance. This is where computations serve as the shield, guiding experiments by predicting the most likely outcomes, reducing unnecessary experiments. Similar to how shields cannot win the battle alone, computations cannot stand on their own. This is because the observational quality of science demands real-life, physical experiments to uncover the different mysteries of nature. Therefore, the marriage of experiments and computations would be the most optimal way to approach science for many years to come.

3 | TROUBLE WITH THE MANY-BODY EXPANSION

Longstanding conventional wisdom dictates that the widely-used many-body expansion (MBE) converges rapidly by the four-body term when applied to large chemical systems. We have found, however, that this is not true for calculations using many common, moderate-sized basis sets such as 6-311++G** and aug-cc-pVDZ. Energy calculations performed on water clusters using these basis sets showed a deceptively small error when the MBE was truncated at the three-body level, while inclusion of four- and five-body contributions drastically increased the error. Moreover, the error per monomer increases with system size, showing that the MBE is unsuitable to apply to large chemical systems when using these basis sets. Through a systematic study, we identified the cause of the poor MBE convergence to be a many-body basis set superposition effect exacerbated by diffuse functions. This was verified by analysis of MO coefficients and the behavior of the MBE with increasing monomer-monomer separation. We also found poor convergence of the MBE when applied to valence-bonded systems, which has implications for molecular fragmentation methods. The findings in this chapter suggests that calculations involving the MBE must be performed using the full-cluster basis set, using basis sets without diffuse functions, or using a basis set of at least aug-cc-pVTZ quality.

3.1 Introduction

The many-body expansion (MBE) is a useful and ubiquitous formalism in the theoretical study of large chemical systems.[151–162] The MBE expresses the total energy, E_{tot}, of an n-body system as the sum of one-body, two-body, etc., up to n-body energy contributions (Section 3.2). Calculating E_{tot} directly for large systems is often computationally unaffordable. The benefit of the MBE is that for many systems E_{tot} can be well approximated by truncating the expansion to just the first few terms. Truncated MBEs have found especially widespread use in the study of water clusters, in which most intermolecular interactions are assumed to be pairwise additive (i.e. completely captured in an MBE truncated after the two-body term). The remaining (mostly inductive) interaction energy is accounted for by the rest of the terms in the MBE.

A longstanding and crucial question in modeling aqueous systems is how many terms of the MBE are necessary to adequately approximate the total energy. The earliest studies addressing this question were performed on water dimers, trimers, and tetramers. They found that three-body effects accounted for about 10% of the interaction energy, and that four- and higher-body effects were negligible.[163,164] Subsequent work on slightly larger clusters agreed that four- and higher-body energy contributions were minute.[165–170] The most thorough examination of many-body effects was performed on water hexamers by Xantheas in 1994, in which he found "the contribution from four-body and higher terms to be negligible for these systems."[171]

The results from these studies eventually coalesced into an oft-cited piece of conventional wisdom: that the many-body expansion for water converges rapidly by the four-body term and in a well behaved manner.[31,172,173] Indeed, most current ab-initio-based simulation models use MBEs truncated at three or occasionally four bodies.[71,77,174,175]

Despite this, we decided to verify the rapid convergence of the MBE for

a few water clusters. We calculated all the terms in the MBEs of four $(H_2O)_{16}$ clusters with HF/6-31++G**, expecting, per conventional wisdom, to observe convergence to the true cluster energy by at most the five-body term (convergence herein defined as consistently having an error less than 1 m-E_h). Instead, not only did these MBEs not converge by anywhere near the five-body term, the convergence was notably erratic (Figure 3.1). Particularly worrying was that while truncating the many-body expansion at the four-body term led to a decent result (3.2–4.2 m-E_h error for 4444-a,c1b,cie), inclusion of the five-body term *increased* the error (4.6–4.7 m-E_h error for 4444-a,c1b,cie), rather than further converging the MBE towards the true cluster energy.

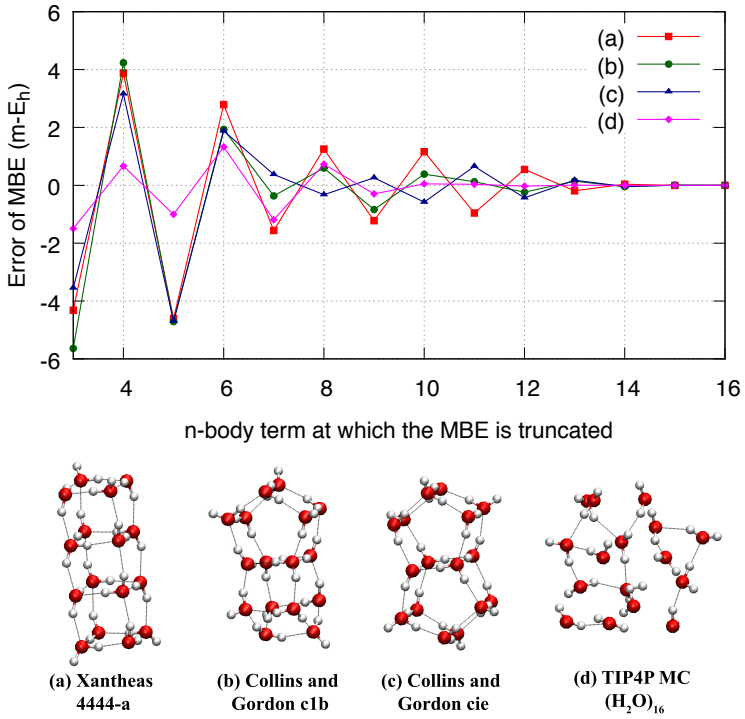


Figure 3.1: Slow, erratic convergence of the MBE towards the full-cluster energy for four $(H_2O)_{16}$ clusters calculated at HF/6-31++G**. The linear fused cube 4444-a (a) was obtained from Xantheas[176] while the fused pentameric structures c1b (b) and cie (c) were obtained from Collins and Gordon.[177] The TIP4P MC structure (d) was obtained by taking a random fragment of 16 water molecules from a TIP4P Monte Carlo simulation of 400 water molecules.

We are by no means the first to observe problems with the MBE.[169,178,179] To our knowledge, however, there has been no thorough examination of under

what circumstances the MBE fails to converge rapidly. In the following sections we will demonstrate the wide extent of the MBE convergence problem and show that a many-body Basis Set Superposition Effect (BSSE) is its cause. We stress that we do not question the accuracy or validity of the many methods that use the MBE, or question previous studies of many-body effects in water. Our purpose in this chapter is simply to refine the conventional wisdom about the MBE.

3.2 Computational Details

For a chemical system comprising n monomers, the MBE of the total energy of the system, E_{tot}, is the finite sum

$$E_{\mathrm{tot}} = \sum_{k=1}^{n} \varepsilon^{(n,k)} \tag{3.1}$$

where $\varepsilon^{(n,k)}$ is the total k-body energy of the system. The total k-body energy of the system is the component of the total energy due to all k-body effects, so $\varepsilon^{(n,1)}$ is the sum of the energies of all isolated monomers; $\varepsilon^{(n,2)}$ is the sum of the energies of all dimers minus the energies of the monomers they comprise, i.e., it is the sum of all the pairwise interaction energies. Thus the binding energy of the system is given by

$$\varepsilon_{\mathrm{tot}} = E_{\mathrm{tot}} - \varepsilon^{(n,1)} \tag{3.2}$$

$\varepsilon^{(n,k)}$ can be expressed recursively in terms of the total energy and lower-body energies[180]

$$\varepsilon^{(n,k)} = \sum_{\alpha} E_{\alpha} - \sum_{i=1}^{k-1} \left[\frac{(n-i)!}{(n-k)!(k-i)!} \right] \varepsilon^{(n,i)} \tag{3.3}$$

where E_{α} is the total energy of the k-mer sub-system α of which there are $\binom{n}{k}$.

We also wish to clarify our use of the term Basis Set Superposition Effect. When small basis sets are used in ab initio calculations of molecular clusters,

basis functions from one molecule can be utilized by other molecules to compensate for the incompleteness of their basis set. This improves the description of the wavefunction of all molecules in the cluster, which leads to a lowering of the total energy known as the BSSE. One way of quantifying the BSSE is the counterpoise (CP) method.[181] In the CP method, the familiar expression for the BSSE in the interaction energy of two molecules, A and B, is given as

$$E_{\mathrm{BSSE}} = (E_A(\mathbf{ab}) - E_A(\mathbf{a})) + (E_B(\mathbf{ab}) - E_B(\mathbf{b})) \tag{3.4}$$

where **a**, **b** and **ab** are the basis sets of molecule A, molecule B, and the cluster AB, respectively. Applying the CP method to the interaction energy of A and B, the BSSE-free interaction energy is

$$\begin{aligned} \varepsilon_{\mathrm{tot}}^{\mathrm{CP}} &= \varepsilon_{\mathrm{tot}} - E_{\mathrm{BSSE}} \\ &= E_{AB}(\mathbf{ab}) - (E_A(\mathbf{a}) + E_B(\mathbf{b})) - E_{\mathrm{BSSE}} \\ &= E_{AB}(\mathbf{ab}) - (E_A(\mathbf{ab}) + E_B(\mathbf{ab})) \end{aligned} \tag{3.5}$$

Now, all the quantities are calculated consistently in the same basis, namely the basis set of the cluster AB. The brilliance of the CP method lies in that it does not try to remove the lowering of energy in the total energy of the cluster AB due to sharing of basis functions which is a natural consequence of the variational principle. Instead, it does the opposite where the constituents A and B are calculated in the basis set of the cluster AB so as to achieve a similar lowering of energy.

In the spirit of the CP method, we define the BSSE for the sum of the total energies of all $\binom{n}{k}$ k-mers in the cluster containing n monomers as

$$E_{\mathrm{BSSE}}^{(n,k)} = E^{(n,k)}(\mathbf{n\text{-}mer}) - E^{(n,k)}(\mathbf{k\text{-}mer}) \tag{3.6}$$

where **n-mer** and **k-mer** are the basis set of the full molecular cluster (cluster basis) and the basis set of the k monomers considered (nuclei-centered basis) respectively. Using our definition, the removal of the BSSE in the total energies when performing a many-body energy decomposition will result in all the quantities being calculated consistently in the cluster basis, ensuring that the MBE remains formally exact. Indeed, the use of a consistent cluster basis has been employed previously[171,182,183] to obtain BSSE-free many-body energies. From eq (3.6), BSSE can be seen as a lowering of the total energy of the k-mer in the cluster due to the sharing of basis functions from the remaining $n-k$ monomers. Thus, when $k=n$, there is no BSSE, i.e. $E_{\mathrm{BSSE}}^{(n,n)}=0$. Notably, this definition of BSSE reduces to familiar expression in eq (3.4) in the context of the interaction energy of a cluster where $E_{\mathrm{BSSE}}=E_{\mathrm{BSSE}}^{(n,1)}$.

All quantum chemical calculations were performed using the Gaussian 09 package[184] or the MOLPRO suite of programs[185] at the Hartree-Fock (HF) or second-order Møller-Plesset perturbation (MP2) level of theory. A variety of Pople split-valence basis sets were used, along with the series of Dunning correlation-consistent cc-pVXZ basis sets, $X=2-4$, labeled VDZ, VTZ, and VQZ. An "A" or "dA" prepended to these basis sets indicate they are augmented or doubly-augmented, respectively, with diffuse functions.

3.3 Results and Discussion

3.3.1 Extent of the Poor Convergence of the MBE

Prompted by our initial results (Figure 3.1), we attempted to ascertain the extent of the MBE convergence problem. We calculated the MBE up to the five-body term for a variety of $(H_2O)_n$, $n=6-57$, clusters (Table 3.1). Some geometries are optimized structures from the literature, others were taken from TIP4P Monte Carlo simulations. These latter structures were included as they are representative of geometries encountered in simulations using MBE-based water

models. All calculations were performed using the AVDZ basis set, which is a better yet computationally manageable basis set compared to the 6-31++G** basis used in Figure 3.1. The MBE in Table 3.1 were only calculated to at most the five-body term due to the steep computational cost of calculating high-body terms for large clusters: the number of additional calculations required to obtain the k-body energy of an n-body system is $\frac{n!}{k!(n-k)!}$. Table 3.1 shows that the MBE of small clusters ($n = 6-8$) do converge by the three-body term, as shown in previous studies.[171] But for larger clusters, while MBEs truncated at the three-body term appear converged, inclusion of four- and five-body energies *increases* the error in the MBE. A notable oscillatory behavior also occurs wherein the error changes sign when three-and-higher-body energies are included.

More alarmingly, Figure 3.2 shows that for the clusters studied in Table 3.1, the four-body and five-body MBE truncation errors per monomer increase with system size. Note that this was also noticed previously in a smaller sample of water clusters by Gadre, who called for further examination.[169] This is concerning as the error-per-monomer should be an intensive, not extensive, property. Otherwise, the scalability of MBE-based computational methods, such as fragmentation methods and bulk material simulations, becomes questionable.

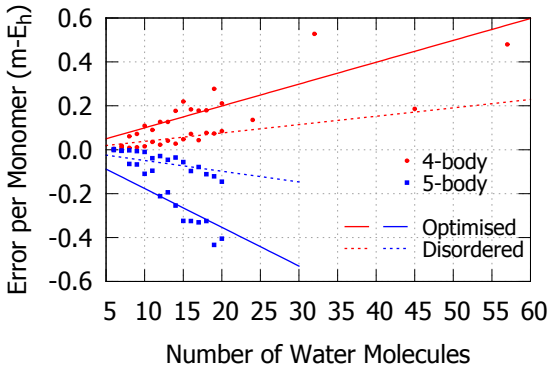


Figure 3.2: Error-per-monomer of the MBE truncated at the four-body level (red solid circles) and at the five-body level (blue solid squares) for optimized water clusters (solid lines) and disordered water clusters (dashed lines) from Table 1.

Table 3.1: Error (m-E_h) in the total energy of water clusters $(H_2O)_n$, $n = 6-57$, as approximated by an MBE truncated after the 2- through 5-body term.

$(H_2O)_n$	2-body	3-body	4-body	5-body
6[a]	-9.769	-0.488	-0.006	0.017
8[a]	-16.905	-0.426	0.491	-0.518
10[a]	-23.227	-1.444	1.092	-1.099
12[a]	-26.745	-0.399	1.519	-2.541
14[a]	-27.613	-1.302	2.478	-3.563
16[a]	-37.045	-0.631	2.940	-5.200
18[a]	-37.060	-0.563	3.227	-5.848
20[a]	-46.844	-0.741	4.219	-8.101
24[b]	-106.289	-10.762	3.252	
32[c]	-130.794	-6.612	16.895	
57[d]	-97.126	-5.401	27.334	
6[e]	-1.674	-0.024	0.022	-0.009
8[e]	-2.591	-0.072	0.072	-0.023
10[e]	-1.704	0.041	0.157	-0.098
12[e]	0.189	0.266	0.274	-0.344
14[e]	-6.870	-0.003	0.391	-0.495
16[e]	-11.364	-0.378	1.153	-1.555
18[e]	-6.524	-0.191	1.382	-2.009
20[e]	-14.251	-0.721	1.692	-2.910
45[e]	-36.931	-2.117	8.373	

All calculations done at HF/AVDZ level of theory. Optimized water clusters obtained from [a]Gadre,[186] [b]Szalewicz,[173] [c]Collins and Gordon,[177] and [d]Herbert.[187] [e]Disordered random fragments of $(H_2O)_n$ obtained from a TIP4P Monte Carlo simulations. For these disordered fragments, the MBE truncation errors were averaged over four different random fragments for $n = 6-20$.

3.3.2 *Cause of the Poor Convergence of the MBE*

Having found the MBE convergence problem to be widespread, we sought to determine its cause. Initially, we thought the poor convergence was due to the MBE not properly capturing the many-body induction energy. Since induction energy is closely related to the polarizability of the bodies, we performed MBE calculations for the optimized clusters studied in Table 3.1 using a series of basis sets of increasing polarizability (Table 3.2). If the poor MBE convergence were due to induction, the error for truncating the MBE should worsen with increasing polarizability. It is clear from Table 3.2 that this is not the case.

Table 3.2: Error in approximating the total energy of optimized water clusters $(H_2O)_n$, $n = 8-20$, with an MBE truncated at the four-body and five-body term (given in order, separated by a comma) using basis sets of increasing isotropic dipole-dipole polarizability $\bar{\alpha}$.

	Basis Set			
	P2[a]	P3[a]	D[b]	T[b]
$\bar{\alpha}$ (a.u.)	4.87	6.50	7.97	8.23
$(H_2O)_n$	Error (m-E_h)			
8	0.07 , 0.01	1.24 , -0.67	0.49 , -0.52	0.05 , -0.04
10	-0.01 , 0.01	0.91 , -0.65	1.09 , -1.10	0.04 , -0.09
12	0.21 , -0.02	1.13 , -0.55	1.52 , -2.54	0.01 , -0.36
14	0.25 , 0.02	1.57 , -1.12	2.48 , -3.56	0.19 , -0.47
16	0.32 , 0.02	2.82 , -1.75	2.94 , -5.20	0.00 , -0.66
18	0.38 , -0.02	1.32 , -0.66	3.23 , -5.85	
20	0.48 , -0.04	2.00 , -1.15	4.22 , -8.10	

All calculations performed at HF level. Water geometries are from Gadre[186] (the same geometries as used in Table 3.1). Pople basis set P2: 6-31G** and P3: 6-311+G(2d,p); [b]Dunning basis set AV*X*Z where X= D or T.

Instead, poor MBE convergence only occurred when using small, incomplete basis sets augmented with diffuse functions, namely 6-311++G** (P3) and AVDZ (D). This led us to suspect that the convergence problem was due to

BSSE. As a preliminary test of this, the MBEs in Table 3.2 were recalculated using the cluster basis, as opposed to the usual nuclei-centered basis, in all k-mer calculations. This eliminated the BSSE in the MBE calculations as explained in Section 3.2. As shown in Table 3.3, the poor MBE convergence observed in Table 3.2 disappeared when BSSE was removed. Note that not all terms were recalculated due to the computational cost of using the full-cluster basis. In fact, the BSSE present in the many-body energies can be easily computed as the difference between the errors in both tables.

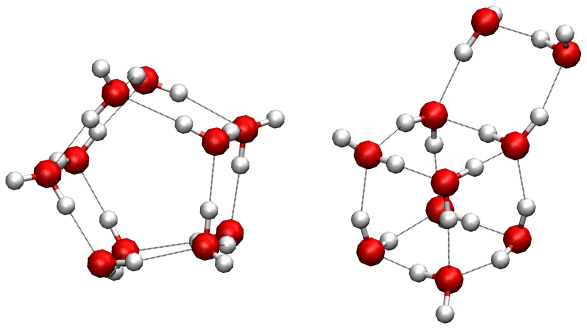
Table 3.3: Error for truncating the MBE at the four-body and five-body term (given in order, separated by a comma) using the cluster basis, as opposed to the nuclei-centered basis used in Table 3.2, in all k-body calculations. Results shown for a series of optimized water clusters $(H_2O)_n$, $n = 8-16$, at HF level using various basis set of increasing isotropic dipole-dipole polarizability $\bar{\alpha}$.

	Basis Set			
	P2[a]	P3[a]	D[b]	T[b]
$\bar{\alpha}$ (a.u.)	4.87	6.50	7.97	8.23
$(H_2O)_n$	Error (m-E_h)			
8	-0.02 , 0.01	0.01 , 0.00	0.04 , 0.00	0.04 , 0.00
10	-0.03 , 0.00	-0.02 , -0.01	-0.02 , -0.01	-0.01 , -0.01
12	0.00 , -0.01	-0.01 , -0.02	0.00 , -0.03	
14	0.02 , 0.00	0.03 , -0.01	0.05 , -0.01	
16	-0.05 , 0.02	-0.02 , 0.00	0.02 , -0.01	

[a]Pople basis P2: 6-31G** and P3: 6-311+G(2d,p); [b]Dunning basis AVXZ where X= D or T.

To verify that BSSE was the cause of the poor MBE convergence, we calculated the full MBEs for two $(H_2O)_{10}$ clusters, **10PP** and **10OB** (Figure 3.3), using basis sets of increasing quality and diffusiveness and using both the nuclei-centred and cluster bases. MP2 calculations were also performed for the 6-31G** and the 6-311G** series to investigate the effects of electron correlation on the MBE convergence. As the results were similar for both **10PP** and **10OB** clusters, only the errors of the MBE for **10PP** are presented in Figure 3.4.

Results for **10OB** are in the Appendix.



Pentagonal Prism
10PP

Open Box
10OB

Figure 3.3: $(H_2O)_{10}$ clusters chosen for a more detailed study on the cause of poor MBE convergence. Both the pentagonal prism (**10PP**) and open box (**10OB**) were obtained from Gadre.[186]

It is clear from Figure 3.4 that when the nuclei-centered basis is used (solid lines), the more diffuse functions are present the worse the MBE convergence (red and orange solid lines), with the exception of AVTZ and AVQZ. This is all precisely what one would expect if BSSE were the cause of the poor MBE convergence: more diffuse functions lead to more overlap of basis functions between water molecules, increasing BSSE, except for basis sets like AVTZ and AVQZ which are so complete that water monomers need not rely on diffuse functions from their neighbors to describe their wavefunctions. Indeed, Truhlar and co-workers have made a similar observation by examining the effects of increasing augmentation in the Dunning basis sets.[188–190] It should be noted that there are still tiny oscillations in the MBE truncation errors for AVTZ and AVQZ in the range of 10–50 μ-E_h, which are hard to see in the Figure. Moreover, when the cluster basis is used (dashed lines) and BSSE is eliminated, the MBE converges by the four-body term regardless of the presence of diffuse functions. Figure 3.4c-f further show that when electron correlation is included, the MBE errors are amplified. This can be attributed to additional BSSE associated with electron correlation—it is known that correlation energy converges more slowly towards the complete basis set limit than the SCF energy.[191]

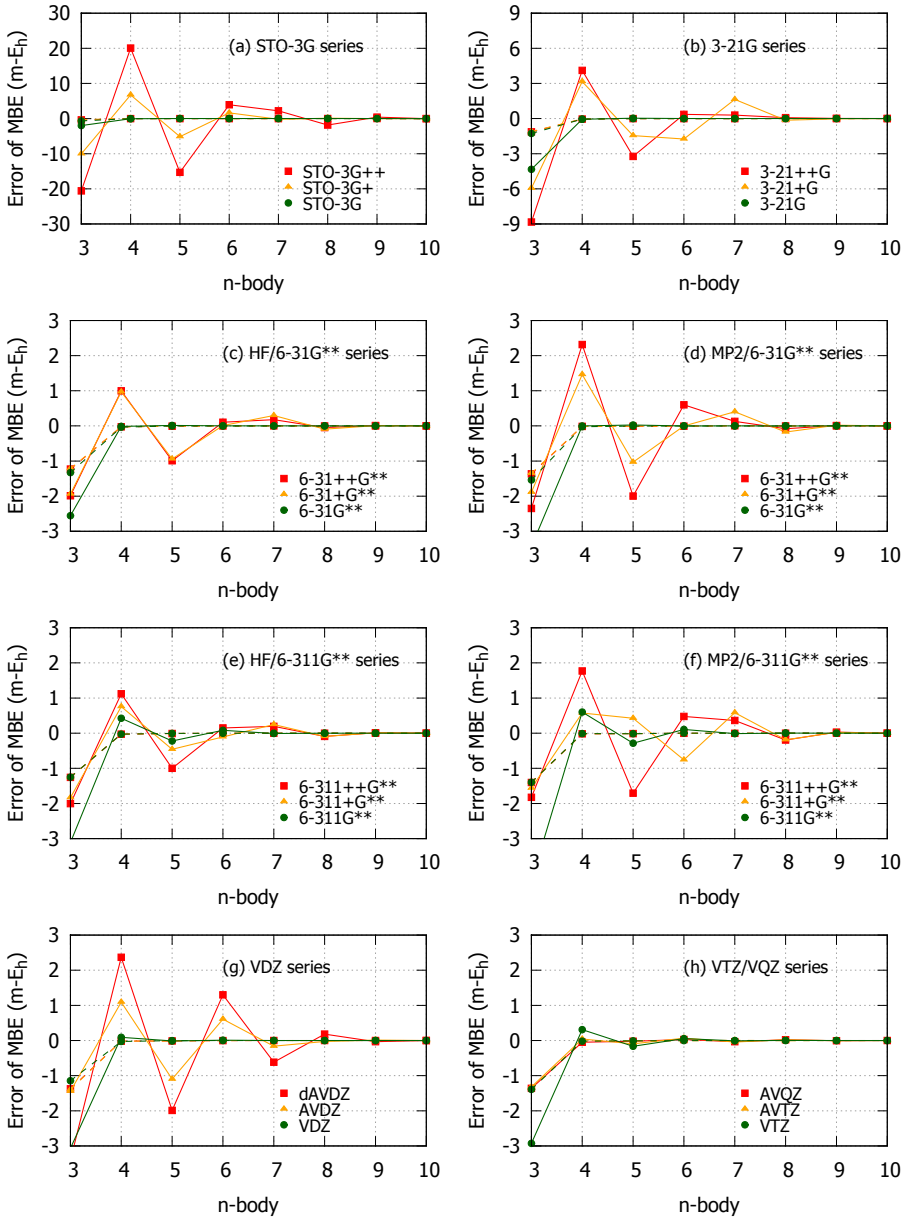


Figure 3.4: Convergence of the MBE for **10PP** using various basis sets: (a) STO-3G series, (b) 3-21G series, (c) HF/6-31G** series, (d) MP2/6-31G** series, (e) HF/6-311G** series, (f) MP2/6-311G** series, (g) VDZ series and (h) VTZ/VQZ series. Solid lines represent MBE calculated using the nuclei-centered basis while dashed lines represent MBEs calculated using the cluster basis. It should be noted that the diffuse functions of the 6-31++G** basis set were used as the diffuse functions for the STO-3G and 3-21G basis sets, as these basis sets have no defined diffuse functions.

We also examined the MO coefficients in these calculations to specifically check whether the BSSE originated from the diffuse functions. HF calculations were performed using VDZ, AVDZ, and AVTZ basis sets on an arbitrarily chosen monomer from **10PP** with the ghost basis functions of all other waters in the cluster present. The choice of monomer does not significantly affect the results due to the symmetry of the cluster. The distribution of MO coefficients for the occupied MOs are shown in Figure 3.5. By performing calculations on a single monomer in the cluster basis, all observations are solely due to BSSE and not physical interaction between molecules. If the diffuse functions were causing the BSSE—that is, if water molecules were using diffuse basis functions centered on other molecules to improve the description of their own wavefunction—then there should be significant MO coefficients for basis functions centered on the ghost molecules. Similarly, a many-body BSSE effect can be inferred if there are significant non-zero MO coefficients arising from *many* of these ghost molecules simultaneously.

For VDZ (Figure 3.5a), significant non-zero MO coefficients, represented by red or blue colored regions, are only found for a few water molecules' basis functions. In contrast, the AVDZ basis set (Figure 3.5b) has significant non-zero MO coefficients on all the water molecules' basis functions. As the ghost water molecules in Figure 3.5 are ordered by their proximity to the monomer under study, the colored regions become fainter across the horizontal axis due to decreasing overlap of the basis functions from more distant ghosts. Nonetheless, the non-zero coefficients imply that the BSSE is many-body in nature, with contributions from all monomers in the system. The contributions come primarily from the diffuse functions (denoted D in the figure) of both oxygen and hydrogen, again implicating diffuse functions in causing the BSSE. The MO coefficient distribution for AVTZ (Figure 3.5c) also shows contributions from diffuse functions, but less so than those for AVDZ. Again, this is due to AVTZ being a more complete basis set: the wavefunction of the monomer can be de-

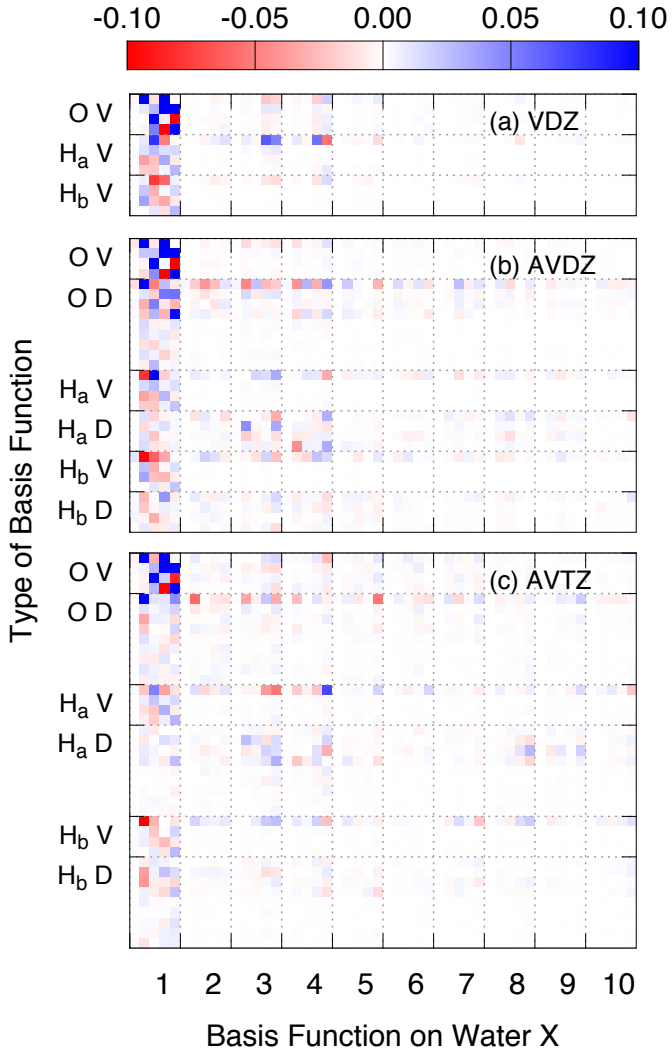


Figure 3.5: Distribution of MO coefficients of an arbitrarily chosen monomer of **10PP** calculated with the cluster basis using (a) VDZ, (b) AVDZ, and (c) AVTZ basis sets. The vertical axis shows the basis functions arranged according to the nuclei (O or H). Only valence functions (denoted V), functions with the smallest exponent, and diffuse functions (denoted D), are shown. The horizontal axis shows which water molecule the basis functions are centered on: Water 1 is the monomer under study and the rest are ghost molecules.

scribed using its own core, valence, and diffuse basis functions without the need for the basis functions of its neighbors. In fact, AVTZ's BSSE contribution to the total energy of the monomer is low at 2.7 ppm of the total energy of the monomer (-0.2 m-E_h) in contrast to the higher contribution from both VDZ (68 ppm, -5.2 m-E_h) and AVDZ (10 ppm, -0.8 m-E_h).

As a final test of our hypothesis, we investigated how MBE convergence is affected by the average nearest-neighbor distances of water molecules in a cluster. HF/AVDZ MBE calculations using the nuclei-centered basis were performed on a series of progressively expanded structures derived from **10PP** (Figure 3.6). The expanded structures were constructed by scaling the distance between the center-of-mass of each water and the center-of-mass of the entire **10PP** cluster. This ensures the nearest-neighbor distance of all water molecules are increased by the same factor. As the mean nearest-neighbor distance increases, the oscillations in the MBE error gradually disappear. This is because when the waters are farther apart the overlap between diffuse functions on different waters decreases exponentially and so does the BSSE. This can be seen explicitly in Figure 3.7 where an exponential fit captures the decay of BSSE with increasing inter-water distance.

The curious reader may wonder why the error oscillates from positive to negative in nearly all the poorly convergent MBEs we have shown. Our best explanation is that this behavior is related to the inclusion/exclusion principle inherent in the MBE. To obtain a system's k-body energy, the total energy of each k-mer in the system has the total energy of all its constituent $(k-1)$-mers subtracted from it. But this results in over-subtraction of $(k-2)$-body energies, so the $(k-2)$-mer total energies have to be added back, and so on. When subsequently obtaining the $(k+1)$-body energy, the signs of the terms in the expression switch: k-body energies are subtracted where they were previously added, etc. (in addition to there being many more terms in the calculation). So if a particular k-body energy is underestimated—perhaps due to an inadequate

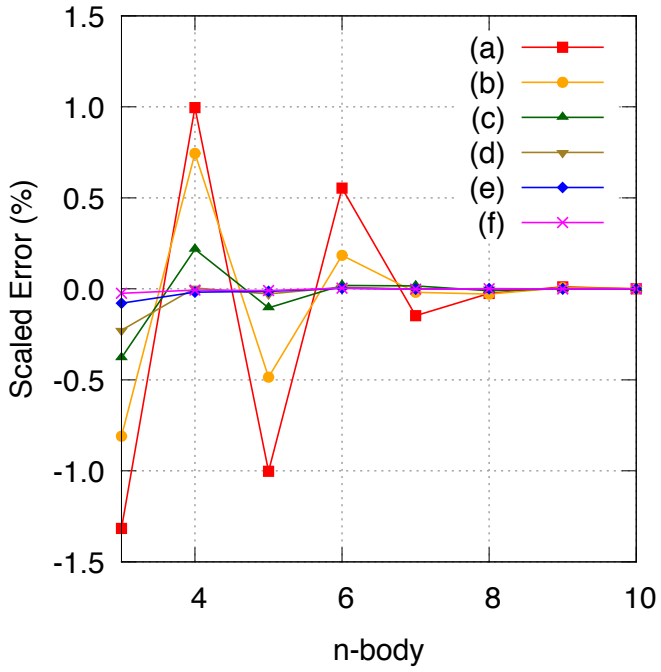


Figure 3.6: Convergence of the MBE for expanded structures derived from **10PP** with the following mean nearest-neighbour distance: (a) 5.25 a_0, (b) 5.67 a_0, (c) 6.61 a_0, (d) 7.56 a_0, (e) 8.50 a_0 and (f) 9.45 a_0. The MBE truncation error has been normalised to the one-body error (i.e. the interaction energy of the cluster).

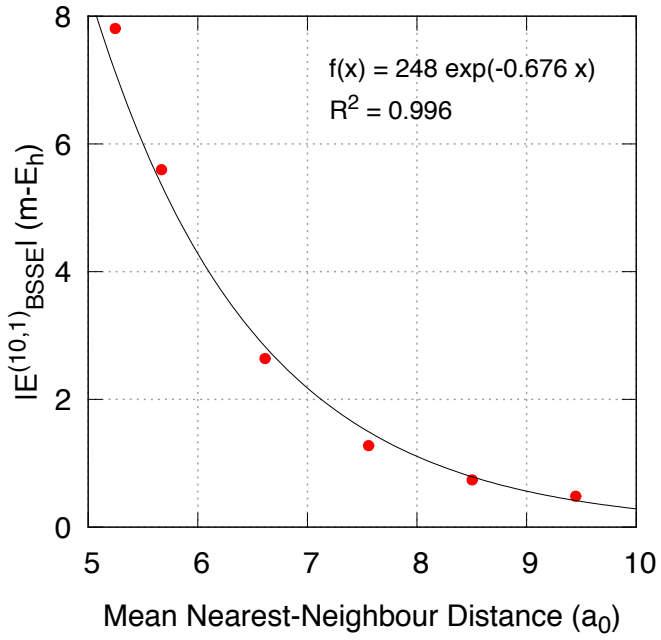


Figure 3.7: Magnitude of the BSSE in the interaction energy $|E^{(10,1)}_{\mathrm{BSSE}}|$ of the expanded structures derived from **10PP** shown in Figure 3.6 shows an exponential decay with increasing mean nearest-neighbour distance.

basis set—it will contribute to the overestimation of the $(k+1)$-body energy, and then the underestimation of the $(k+2)$-body energy, and so on, leading to oscillation in the error. Of course, the MBE by definition converges by the final term. And each subsequent term in the MBE contains less BSSE-derived error, $E_{\mathrm{BSSE}}^{(n,k+1)} < E_{\mathrm{BSSE}}^{(n,k)}$, since the more waters monomers are in a calculation the closer the basis set is to the correct, full-cluster basis. But in the early MBE terms there is much utilization of neighboring waters' diffuse basis functions, and more error in each calculated energy, and thus oscillations which diminish as more terms are added.

3.3.3 Methods to Improve MBE Convergence

We have shown that rapid convergence of the MBE can be guaranteed by performing calculations in the cluster basis or with a high-quality basis set. These are, of course, fairly dispiriting solutions as they greatly increase computational cost, so we have examined several alternatives.

Since basis function overlap is distance-dependent, we tested a distance-cutoff basis (d-c basis) which includes ghost functions only from waters within a specified cutoff distance of the water molecules in a calculation. Testing this d-c basis on the $(H_2O)_{16}$ 4444-a cluster using various cutoff distances gave poor results (Table 3.4), however. This is likely due to the small MBE truncation errors involved in the m-E_{h} range. With slight changes in cutoff distances, one more or one fewer water's ghost functions might be included in a calculation, which could lead to significant changes in calculated energies, drastically affecting the MBE truncation errors. That said, we think the distance-cutoff basis might work with a larger cutoff distance, but in those cases it would be more economical to use a high-quality basis set instead.

Another workaround that has been proposed, albeit for a different problem, is the *k*-mer-centerd basis set (*k*CBS) approach of Szalewicz.[173] The *k*CBS approach attempts to remove BSSE in an MBE calculation by calculating each

Table 3.4: Error for truncating the MBE up to the five-body term, performed using various methods to improve MBE convergence. Calculations were performed at HF/6-31++G** on the $(H_2O)_{16}$ 4444-a cluster shown in Figure 3.1. Nuclei-centred basis and cluster basis are shown for reference.

	Error (m-E_h)			
Method	2-body	3-body	4-body	5-body
nuclei-centred basis	-42.418	-4.763	4.588	-5.563
cluster basis	-56.170	-1.217	-0.043	-0.056
d-c basis, 3 Å	-35.325	-28.202	10.527	9.549
d-c basis, 4 Å	-41.468	-46.804	85.769	-108.259
d-c basis, 5 Å	-48.275	-34.555	87.929	-161.926
*k*CBS	-50.464	4.779	6.268	6.279
Charge Field	-12.466	-0.449	2.073	-2.054

k-mer *and all its sub-calculations* using the *k*-mer's basis set. That is, a dimer's two-body contribution would be computed as total energy of the dimer minus the total energy of its constituent monomers, all calculated with the dimer basis set. This results in substantially more calculations to compute the MBE since calculations from previous terms cannot be reused, but it does mean each calculation has no BSSE. We applied the *k*CBS approach to the $(H_2O)_{16}$ 4444-a cluster. From Table 3.4, we see that the MBE does converge rapidly, but to an incorrect value. This is likely because the *k*CBS approximation is not formally exact: the terms in the MBE do not cancel due to the different numbers of basis functions used in each term's calculations. The *k*CBS approach certainly does converge correctly when a high-quality basis set is used, as has been demonstrated in the literature, but this seems to be due to the high quality of the basis set, not the *k*CBS method.

Strategies unrelated to BSSE for improving MBE convergence are widely used. Many MBE-based computational methods incorporate a charge field to approximate higher-body effects by interacting the one-body or two-body fragments with a charge field mimicing the rest of the system.[152–154,156,160,192,193]

While we have not done a thorough analysis, preliminary results using embedded charges from Stone's distributed multipole analysis[194] indicate that embedded charges dampen, but do not remove, the oscillatory MBE behaviour (Table 3.4). This is not surprising as the embedded charges only serve to approximate the *physical* higher-body effects arising from induction and thus do not remove the many-body BSSE.

Other methods incorporate higher-body effects by performing a low-level ab initio calculation on the full system.[153,154,195,196] Such methods capture many-body effects far better than methods using only a truncated MBE.[197] The full-system calculations in these methods are not susceptible to BSSE-based MBE convergence issues since they use full-system basis, but lower-body calculations performed using only the nuclei-centred basis are still susceptible.

Thus, unhappily, we have found no alternative for avoiding poor MBE convergence that is more efficient than using the full-cluster basis or a high-quality basis set. As the use of the cluster basis is computationally prohibitive, our recommendation is to use a high-quality basis set for MBE calculations; our results indicate that at least AVTZ-quality is prudent.

3.3.4 Extension to Valence-bonded Systems

So far we have only presented data for noncovalent water clusters, but MBE convergence problems also arise in valence-bonded systems. This has great implications for fragmentation methods, which in most cases use a truncated MBE, or something analogous to it, to approximate the total energies of large chemical systems.[187,198–200]

In fragmentation methods, small groups of adjacent atoms are treated as bodies. Using our CFM algorithm[161] to define groups/bodies, we calculated the MBEs for a $C_{22}H_{24}$ conjugated alkene and α-cyclodextrin (Figure 3.8). Slow MBE convergence is observed in both systems when incomplete basis sets with diffuse functions are used, as seen in the case of AVDZ for

$C_{22}H_{24}$ and 6-31++G** for α-cyclodextrin. This is no surprise as the same borrowing of basis functions from adjacent groups that causes poor convergence in water clusters occurs in these valence-bonded systems. The errors in the $C_{22}H_{24}$ MBE are small (even negligible) because the molecule's linear shape minimizes basis function overlap. Compare this to the MBE of the more compact α-cyclodextrin, where the errors are beyond chemical accuracy until the inclusion of the 9-body term. Since fragmentation methods rarely include 5-or-higher-body effects, it seems likely that fragmentation calculations using BSSE-prone basis sets are liable to, and have in the past been afflicted by, preventable, BSSE-based errors. On an interesting related point, due to the above mentioned affects, we expect that any calculation that is performed in order to predict bond-breaking energies would be overestimated.

It should be noted, though, that poor MBE convergence in a valence-bonded system depends on the definition of "body". Another type of MBE that our group has examined is to treat distortions in the internal degrees-of-freedom of a molecule as bodies. Using the equilibrium geometry as a reference, an MBE can be used to calculate the distortion energy of a molecule. We demonstrate a proof-of-concept using the methanol molecule (Figure 3.9). We distorted the molecule randomly in all twelve degrees of freedom, yielding a total distortion energy of about 140 m-E_h. The degree-of-freedom MBE converges by the four-body term, even when a BSSE-prone basis set is used. This is expected as a consistent basis set is used in all calculations, essentially equivalent to the use of a full-cluster basis. Apart from intramolecular degrees-of-freedom, intermolecular degrees-of-freedom or a combination of both could be treated in the same manner. The utility of such an approach is obvious. A high-dimensional system is broken down into to numerous, completely independent (and thus highly parallelizable) much lower-dimensional function evaluations. Future work will explore how degree-of-freedom MBEs can be used to construct accurate, high-dimensional potential energy surfaces from many lower-dimensional surfaces.

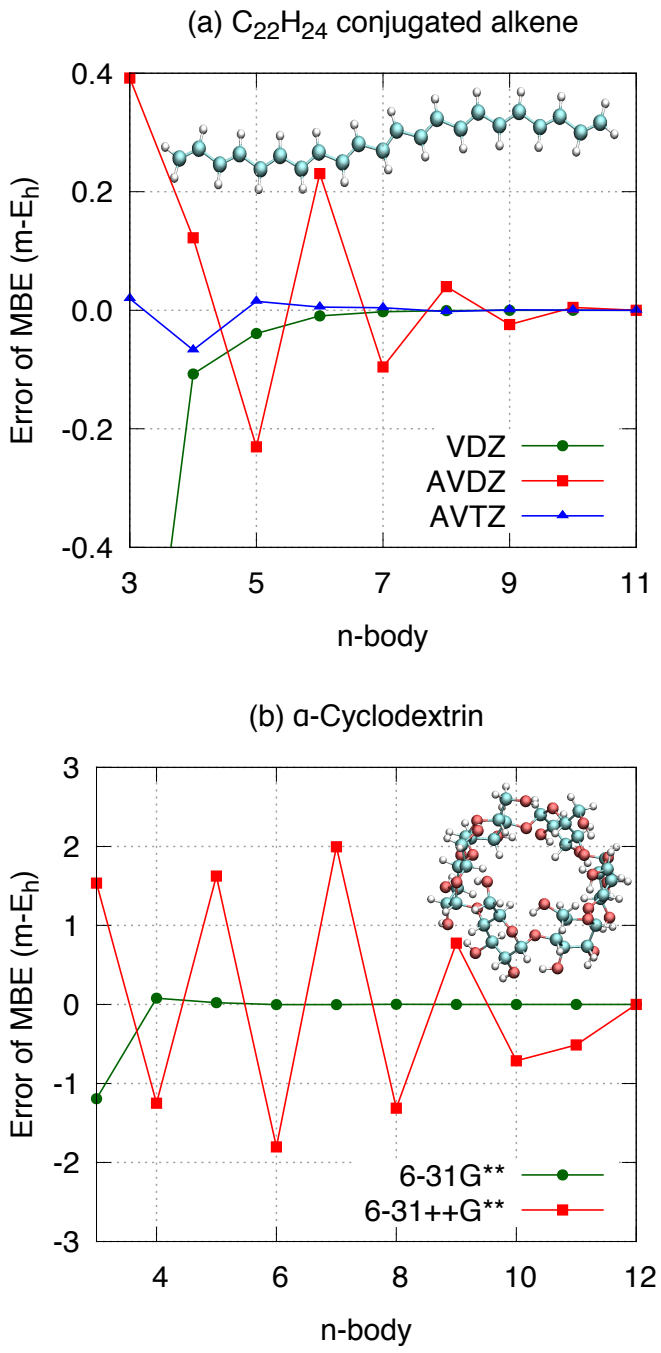


Figure 3.8: Convergence of the MBE for the total energy of (a) $C_{22}H_{24}$ conjugated alkene and (b) α-cyclodextrin at HF level of theory for various basis sets. Here, the CFM algorithm was used to define the bodies in the MBE. The inset shows the structures of the molecules.

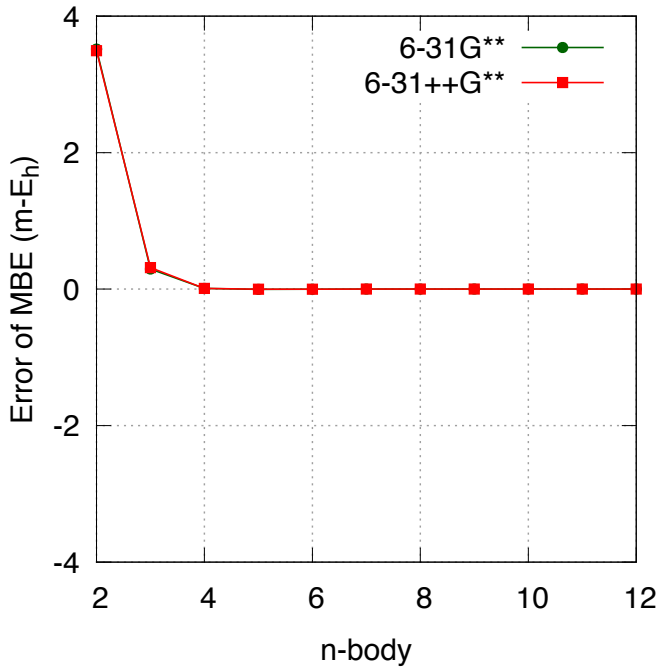


Figure 3.9: Convergence of the MBE for the distortion energy of methanol molecule at HF level of theory for various basis sets. Here, distortions in the 12 internal degrees-of-freedom are treated as bodies in the MBE. Convergence is clearly independent of the choice of basis set—the results from both basis sets overlap almost perfectly.

3.4 Summary

There is no question that the many-body expansion is a theoretically sound and extremely useful formalism in the study of large molecular systems. But it is likewise clear from our observations that care must be taken in its implementation. Rapid convergence at the four-body term of the MBE cannot be assumed, even when convergence appears to have occurred. Incautious use of MBEs with systems and levels of theory susceptible to BSSE is liable to yield errors well beyond chemical accuracy. Moreover, the error-per-monomer worsens extensively with system size. Such concerns are relevant in valence-bonded and noncovalent systems alike. We conclude that the use of a consistent basis set, either in the form of the full-cluster basis or a high-quality basis set (at least AVTZ quality), is necessary to avoid poor MBE convergence due to BSSE.

4 | MANY-BODY BASIS SET SUPERPOSITION EFFECT

The basis set superposition effect (BSSE) arises in electronic structure calculations of molecular clusters when questions relating to interactions between monomers within the larger cluster are asked. The binding energy, or total energy, of the cluster may be broken down into many smaller subcluster calculations and the energies of these subsystems linearly combined to, hopefully, produce the desired quantity of interest. Unfortunately BSSE can plague these smaller fragment calculations. In this chapter we carefully examine the major sources of error associated with reproducing the binding energy and total energy of a molecular cluster. In order to do so we decompose these energies in terms of a many-body expansion (MBE), where a "body" here refers to the monomers that make up the cluster. In our analysis we found it necessary to introduce something we designate here as a many-ghost many-body expansion (MGMBE). The chapter presented here produces some surprising results, but perhaps the most significant of all is that BSSE effects up to the order of truncation in a MBE of the total energy cancel exactly. In the case of the binding energy the only BSSE correction terms remaining arise from the removal of the one-body monomer total energies. Nevertheless, our earlier chapter indicated that BSSE effects continued to remain in the total energy of the cluster up to very high truncation order in the MBE. We show in this chapter that the

vast majority of these high-order many-body effects arise from BSSE associated with the one-body monomer total energies. Also we found that remarkably the complete basis set limit values for the three-body and four-body interactions differed very little from that at the MP2/aug-cc-pVDZ level for the respective subclusters embedded within a larger cluster.

4.1 Introduction

A major barrier to the theoretical study of large chemical systems is the fact that the computational effort of electronic structure methods increases drastically with system size (Section 2.2.3). To circumvent this, one can look to fragmentation methods[187,201–203] where a large chemical system is broken up into numerous small subsystems. From there, only certain important interactions between these subsystems are considered for electronic structure calculations so as to recover the total energy of the system. Fundamentally, fragmentation methods are rooted in the many-body expansion (MBE), which have been discussed in some detail in Chapter 3.

The basis set superposition effect[204] (BSSE) comes into play as energy differences are involved in computing the many-body interactions. BSSE arises when monomers within a molecular cluster borrow basis functions from other monomers to compensate for their basis set incompleteness. The same applies to any subcluster within the molecular cluster. Thus, when the total energies of interacting monomers and their isolated counterparts are compared, there is an imbalance in the computed many-body interactions. To eliminate this basis set imbalance error (BSIE) for a dimer system, Boys and Bernardi proposed the counterpoise (CP) method to compute the binding energy where the monomer energies are calculated in the dimer basis.[181] To clarify our use of terminology, we use the term "*location* basis" to describe the placement of basis functions at the specified *location* in the cluster. The CP method was extended for many-monomer molecular clusters to give the Site-Site Function Counterpoise

(SSFC) method by calculating the monomer energies in the cluster basis.[205] The many-body counterpoise (MBCP) method[196,206] was proposed later to approximate the expensive calculation of the monomer energies in the cluster basis by performing a MBE-like decomposition of the effects of the ghost functions present. Nonetheless, the consistent use of the cluster basis in the SSFC method allows for a meaningful decomposition of the binding energy into its many-body contributions. The SSFC method is not a unique extension of the CP method.[205,207] Valiron and Mayer proposed that the many-body interaction of a subcluster can be instead computed using the set of basis functions centred on the subcluster of interest, i.e., the subcluster basis.[182] These many-body interactions can then be summed to give the Valiron–Mayer Function Counterpoise (VMFC) corrected binding energy. While both the SSFC and VMFC methods eliminate BSIE through the use of a consistent basis, the use of the cluster basis in the former incorporates an additional basis set extension effect (BSEE) where the monomers surrounding the subcluster of interest can extend their basis functions—functions present in the cluster basis but not the subcluster basis—to improve the quality of the computed many-body interactions.

The crucial point from the aforementioned counterpoise methods is that, many-body BSSE can be divided into two components, namely the BSIE and BSEE (See Section 4.2.2 for a detailed description). The BSIE (where the E stands for error) is undesirable, causing pairwise interactions and consequently the binding energy of large clusters to be over-stabilizing. On the other hand, the BSEE (where the E stands for effect) is necessary to reproduce the binding energy and total energy of molecular clusters because all the monomers are better electronically described with the additional external basis functions. In Chapter 3, we showed that the use of the cluster basis leads to rapid convergence of the MBE,[91] indicating that the BSEE is indeed present in the total energy. When the subcluster basis is used, we observed that the MBE converged rapidly, but to an "incorrect value" (Section 3.3.3). There is a significant differ-

ence between this "incorrect value" and the total energy, which is essentially the BSEE. More importantly, the rapid convergence associated with the subcluster basis suggested that the BSEE diminishes as rapidly as the many-body interactions. This is of relevance as we noticed that the many-body interactions computed using the subcluster basis are commonly employed in the construction of ab initio based potential energy surface (PES) in the literature.[71,76,77,79]

In this chapter we examine the amount of BSSE, in particular BSEE, present in many-body interactions to identify the major sources of error associated with reproducing the binding energy and total energy of a molecular cluster via an MBE. Firstly, we investigate whether the BSEE is significant in the many-body interactions up to the four-body level. Secondly, we introduce the many-ghost many-body expansion (MGMBE) to precisely and quantitatively account for both the BSIE and BSEE. Remarkably we found that the oscillatory behaviour of the MBE when diffuse functions are involved can be traced to the BSEE in the one-body interactions, i.e., the monomer total energies. Thirdly, with the removal of the monomer total energies and associated BSEE, the MGMBE is able to accurately reproduce the binding energies of molecular clusters using the energies of numerous subclusters that are no larger than four monomers. Notably the utilization of embedded charges, or a coulomb field, is entirely unnecessary to accomplish this.

4.2 Theory

Before discussing the theory behind the MBE, many-body BSSE and MGMBE, we need to define the following terms and quantities which will be constantly used throughout this chapter. From here on, we denote the molecular cluster of interest simply as the "cluster" while a "subcluster" refers to a collection of monomers taken from the cluster. In the counterpoise methods, additional basis functions are placed on the locations of nuclei in the cluster, but without the nuclei being present in the electronic structure calculation and these functions

are called "ghost functions". We also use the term "*location* basis" to describe the *location* at which basis functions are placed in the calculation. For example, the cluster basis refers to the placement of basis functions at the locations of all nuclei present in the cluster. Each of the "bodies" in the many-body interactions refers to a monomer from the cluster, which is taken to be an individual water molecule in this chapter. When discussing the BSEE, we denote a "ghost-body" as the set of ghost functions centred on a monomer surrounding the subcluster of interest. Table 4.1 summarises the relevant quantities described in this chapter.

Table 4.1: List of important quantities presented in this chapter, followed by a brief definition and the equation in which it first appeared.

Quantity	Definition	Eq
E_{tot}	Total energy of a cluster.	(4.1)
$E_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{M}}}$	Total energy of k-mer subcluster $\text{A}\cdots\text{K}$ calculated in the presence of ghost functions centred on $\text{L}\cdots\text{M}$	(4.4)[a]
$\varepsilon_{\text{tot}}^{\text{C}}$	Binding energy computed using the cluster basis	(4.7)
$E_{\text{ext.}}^{(k)}$	Basis set extension effect (BSEE) in the total k-body interaction	(4.8)
$\xi_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{M}}}$	BSEE from m-ghost-body $\text{L}\cdots\text{M}$ in the k-body interaction of $\text{A}\cdots\text{K}$	(4.9)
$\varepsilon_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{M}}}$	k-body interaction of $\text{A}\cdots\text{K}$ computed using total energies calculated with basis functions centred on $\text{A}\cdots\text{KL}\cdots\text{M}$	(4.10)[b]

[a] $E_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{M}}}$ is mentioned much earlier in text at the beginning of Section 4.2.2. [b] $\varepsilon_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{M}}}$ is defined and explained much earlier in text at the second paragraph of Section 4.2.3.

4.2.1 Many-body Expansion

For a cluster containing n monomers, the MBE allows us to decompose the total energy of the cluster, E_{tot}, into its many-body contributions

$$E_{\text{tot}} = \sum_{\text{A}}^{\binom{n}{1}} \varepsilon'_{\text{A}} + \sum_{\text{A}<\text{B}}^{\binom{n}{2}} \varepsilon'_{\text{AB}} + \sum_{\text{A}<\text{B}<\text{C}}^{\binom{n}{3}} \varepsilon'_{\text{ABC}} + \sum_{\text{A}<\text{B}<\text{C}<\text{D}}^{\binom{n}{4}} \varepsilon'_{\text{ABCD}} + \cdots + \varepsilon'_{\text{A}\cdots\text{N}} \quad (4.1)$$

where $\varepsilon'_{\mathrm{A}\cdots\mathrm{K}}$ is the k-body interaction of the k-mer subcluster $\mathrm{A}\cdots\mathrm{K}$, of which there are $\binom{n}{k}$ of such terms (Figure 4.1). Eq (4.1) is the expanded form of eq (3.1) expect that there is a prime symbol in $\varepsilon'_{\mathrm{A}\cdots\mathrm{K}}$, which indicates that the basis functions are placed exclusively at the location of the nuclei, i.e., no ghost functions are involved. In this chapter, we truncate the MBE at the four-body level and thus only provide the relevant equations up to the four-body level.

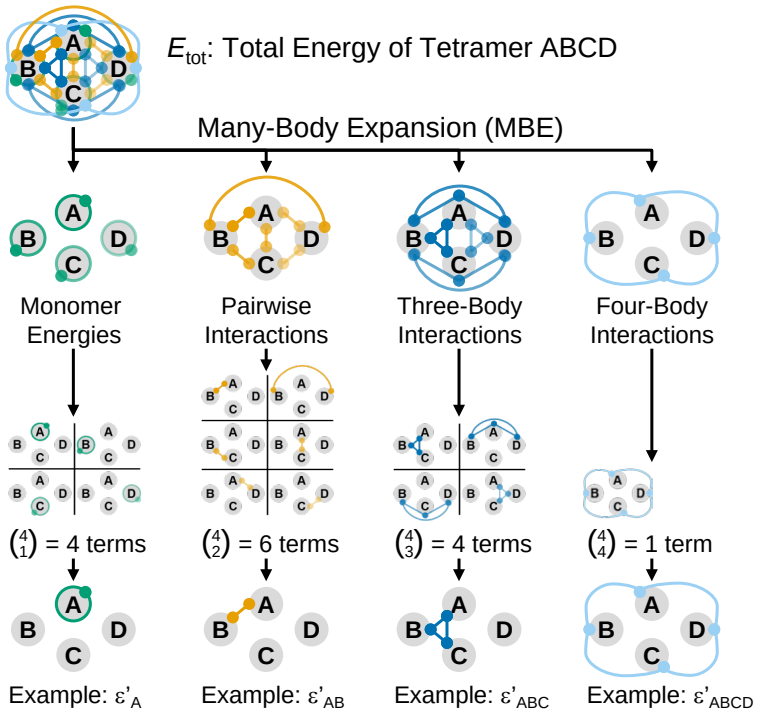


Figure 4.1: The MBE allows us to easily identify the numerous interactions between monomers that is encompassed within the total energy of the cluster. The MBE is illustrated here for a tetramer ABCD ($n = 4$) where the total energy, E_{tot}, is decomposed into the total k-body interactions, $k = 1-4$, which comprises $\binom{n}{k}$ individual terms. The explicit formulae for calculating each individual k-body interaction is given in eq (4.2a–d) in the text.

$\varepsilon'_{\mathrm{A}\cdots\mathrm{K}}$ cannot be obtained directly from electronic structure calculations, which only gives the total energy, $E_{\mathrm{A}\cdots\mathrm{K}}$, of the k-mer subcluster of interest. Thus, we need to write the many-body interactions in terms of the total energies. $\varepsilon'_{\mathrm{A}\cdots\mathrm{K}}$ is defined recursively using lower-body interactions,[171,178,180,187] and can be expressed in terms of total energies

$$\varepsilon'_{\mathrm{A}} = E_{\mathrm{A}} \tag{4.2a}$$

$$\varepsilon'_{\mathrm{AB}} = E_{\mathrm{AB}} - (E_{\mathrm{A}} + E_{\mathrm{B}}) \tag{4.2b}$$

$$\begin{aligned}\varepsilon'_{\mathrm{ABC}} = {} & E_{\mathrm{ABC}} - (E_{\mathrm{AB}} + E_{\mathrm{AC}} + E_{\mathrm{BC}}) \\ & + (E_{\mathrm{A}} + E_{\mathrm{B}} + E_{\mathrm{C}})\end{aligned} \tag{4.2c}$$

$$\begin{aligned}\varepsilon'_{\mathrm{ABCD}} = {} & E_{\mathrm{ABCD}} - (E_{\mathrm{ABC}} + E_{\mathrm{ABD}} + E_{\mathrm{ACD}} + E_{\mathrm{BCD}}) \\ & + (E_{\mathrm{AB}} + E_{\mathrm{AC}} + E_{\mathrm{AD}} + E_{\mathrm{BC}} + E_{\mathrm{BD}} + E_{\mathrm{CD}}) \\ & - (E_{\mathrm{A}} + E_{\mathrm{B}} + E_{\mathrm{C}} + E_{\mathrm{D}})\end{aligned} \tag{4.2d}$$

The one-body interaction, $\varepsilon'_{\mathrm{A}}$, is the total energy of isolated monomer A while the two-body interaction, $\varepsilon'_{\mathrm{AB}}$, gives the pairwise interaction between monomers A and B. The three-body interaction, $\varepsilon'_{\mathrm{ABC}}$, can be understood as the effect of a third monomer C on the interaction between the other two monomers A and B, and the higher-body interactions can be interpreted similarly. In order to employ the MBE, the k-body interaction of individual subclusters computed using eq (4.2a–d) have to be collected to give the total k-body interaction. Many total energy terms are repeated during this collection process and a compact expression of the total k-body interaction can be found in eq (3.3).

Other than expressing the total energy of the cluster using the MBE, another quantity of interest is the binding energy of the cluster

$$\begin{aligned}\varepsilon_{\mathrm{tot}} &= E_{\mathrm{tot}} - \sum_{\mathrm{A}}^{\binom{n}{1}} E_{\mathrm{A}} \\ &= \sum_{\mathrm{A<B}}^{\binom{n}{2}} \varepsilon'_{\mathrm{AB}} + \sum_{\mathrm{A<B<C}}^{\binom{n}{3}} \varepsilon'_{\mathrm{ABC}} + \sum_{\mathrm{A<B<C<D}}^{\binom{n}{4}} \varepsilon'_{\mathrm{ABCD}} + \cdots + \varepsilon'_{\mathrm{A\cdots N}}\end{aligned} \tag{4.3}$$

Note that this is an expanded form of eq (3.2) and the expansion is neccesary for later discussion.

4.2.2 Many-body Basis Set Superposition Effect

In many-body systems, BSSE can be divided into two components, namely the basis set imbalance error (BSIE) and basis set extension effect (BSEE). The distinction between these two components becomes clear when we compare the various placement of basis functions, i.e., the *location* basis, in computing the many-body interactions (Table 4.2).

Table 4.2: Comparison of the choice of basis in computing the many-body interactions and binding energy, together with the name of the method reported in the literature. Furthermore, the absence of the BSIE and presence of BSEE in the many-body interactions are compared with their effects on the convergence of the MBE of the total energy.

Location basis	Nuclei-centred	Subcluster	Cluster
Many-body interactions	$\varepsilon'_{\mathrm{A}\cdots\mathrm{K}}$	$\varepsilon_{\mathrm{A}\cdots\mathrm{K}}$	$\varepsilon_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{N}}}$
	Eq (4.2)	Eq (4.4)	Eq (4.6)
Binding energy	$\varepsilon_{\mathrm{tot}}$	$\varepsilon^{\mathrm{S}}_{\mathrm{tot}}$	$\varepsilon^{\mathrm{C}}_{\mathrm{tot}}$
	Eq (4.3)	Eq (4.5)	Eq (4.7)
Name in literature	Uncorrected[a]	VMFC/*k*CBS	SSFC/CP
Absence of BSIE?[b]	No	Yes	Yes
Presence of BSEE?[b]	No	No	Yes
MBE converge to E_{tot}?	Yes	No	Yes
MBE convergence	Slow, oscillatory	Rapid	Rapid
Further remarks	Default basis	Agree with PT	Expensive

[a] There is no formal name for the nuclei-centred basis as it is the default basis in electronic structure calculations. [b] As summarised towards the end of Section 4.2.2, the *absence* of BSIE and the *presence* of BSEE is desirable, as in the case of the cluster basis.

So far, the total energy, $E_{\mathrm{A}\cdots\mathrm{K}}$, is written such that it is determined by the identity of the k-mer subcluster or more specifically, the location of the nuclei constituting the subcluster. In the context of molecular orbital based electronic structure calculations, the total energy also depends on the placement of basis functions. For example, the total energy of monomer A calculated in the nuclei-centred basis centred on A alone would be different from that using the set of

basis functions centred on the cluster A$\cdots$N. From here on, we use the more general notation, $E_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$. The overline in the subscript indicates the presence of ghost functions centred on L$\cdots$M in the electronic structure calculations.

The most straightforward method to compute the many-body interactions is to place basis functions exclusively at the locations of nuclei in the electronic structure calculations. This is the usual way of calculating an electronic energy of a molecule and the k-body interaction follows eq (4.2a–d) presented earlier. We emphasize again that the prime symbol in the k-body interaction, $\varepsilon'_{\mathrm{A}\cdots\mathrm{K}}$, indicates the nuclei-centred basis where the number of basis functions are different across the different total energy terms. This is in contrast to the "consistent" subcluster and cluster basis which will be introduced shortly. Similarly, the binding energy computed using the nuclei-centred basis, $\varepsilon_{\mathrm{tot}}$, follows eq (4.3). The $\varepsilon_{\mathrm{tot}}$ is often called the uncorrected binding energy for reasons that will be obvious in the following discussion. In computing $\varepsilon'_{\mathrm{AB}}$ using eq (4.2b), it is clear that E_{AB} is calculated using more basis functions as compared to E_{A} and E_{B}. In calculating E_{AB}, monomer A can utilize basis functions centred on monomer B to improve the description of its wave function and vice versa. This is obviously absent in the calculation of E_{A} and E_{B}. This imbalance in the number of basis functions used in the three different calculations of the total energy is the origin of the BSIE. The same BSIE manifests in higher-body interactions in eq (4.2c–d) and the binding energy in eq (4.3). For the two-body interactions the BSIE leads to the interactions being over-stabilizing. In Chapter 3, we also found that MBEs using $\varepsilon'_{\mathrm{A}\cdots\mathrm{K}}$ exhibit slow and oscillatory convergence especially when diffuse basis functions are present.[91]

To remove the BSIE in many-body interactions, we need to ensure that there is a common set of basis functions employed in each of the total energy calculation. The smallest common set is one that is centred on the subcluster for which the many-body interaction is computed. We denote this as the subcluster basis and the k-body interaction can be written as

$$\varepsilon_{\mathrm{A}} = E_{\mathrm{A}} \tag{4.4a}$$

$$\varepsilon_{\mathrm{AB}} = E_{\mathrm{AB}} - \left(E_{\mathrm{A}\overline{\mathrm{B}}} + E_{\mathrm{B}\overline{\mathrm{A}}}\right) \tag{4.4b}$$

$$\begin{aligned}\varepsilon_{\mathrm{ABC}} = {} & E_{\mathrm{ABC}} - \left(E_{\mathrm{AB}\overline{\mathrm{C}}} + E_{\mathrm{AC}\overline{\mathrm{B}}} + E_{\mathrm{BC}\overline{\mathrm{A}}}\right) \\ & + \left(E_{\mathrm{A}\overline{\mathrm{BC}}} + E_{\mathrm{B}\overline{\mathrm{AC}}} + E_{\mathrm{C}\overline{\mathrm{AB}}}\right)\end{aligned} \tag{4.4c}$$

$$\begin{aligned}\varepsilon_{\mathrm{ABCD}} = {} & E_{\mathrm{ABCD}} - \left(E_{\mathrm{ABC}\overline{\mathrm{D}}} + E_{\mathrm{ABD}\overline{\mathrm{C}}} + E_{\mathrm{ACD}\overline{\mathrm{B}}} + E_{\mathrm{BCD}\overline{\mathrm{A}}}\right) \\ & + \left(E_{\mathrm{AB}\overline{\mathrm{CD}}} + E_{\mathrm{AC}\overline{\mathrm{BD}}} + E_{\mathrm{AD}\overline{\mathrm{BC}}} + E_{\mathrm{BC}\overline{\mathrm{AD}}} + E_{\mathrm{BD}\overline{\mathrm{AC}}} + E_{\mathrm{CD}\overline{\mathrm{AB}}}\right) \\ & - \left(E_{\mathrm{A}\overline{\mathrm{BCD}}} + E_{\mathrm{B}\overline{\mathrm{ACD}}} + E_{\mathrm{C}\overline{\mathrm{ABD}}} + E_{\mathrm{D}\overline{\mathrm{ABC}}}\right)\end{aligned} \tag{4.4d}$$

For each of the many-body interactions, the same set of basis functions centred on the subcluster of interest is employed for all the total energy calculations and this introduces ghost functions, denoted by the overline in the subscript of total energy terms. Note that, unlike the nuclei-centred basis, the prime symbol is absent here. The binding energy computed using the subcluster basis is

$$\varepsilon_{\mathrm{tot}}^{\mathrm{S}} = \sum_{\mathrm{A}<\mathrm{B}}^{\binom{n}{2}} \varepsilon_{\mathrm{AB}} + \sum_{\mathrm{A}<\mathrm{B}<\mathrm{C}}^{\binom{n}{3}} \varepsilon_{\mathrm{ABC}} + \sum_{\mathrm{A}<\mathrm{B}<\mathrm{C}<\mathrm{D}}^{\binom{n}{4}} \varepsilon_{\mathrm{ABCD}} + \cdots + \varepsilon_{\mathrm{A}\cdots\mathrm{N}} \tag{4.5}$$

The $\varepsilon_{\mathrm{tot}}^{\mathrm{S}}$ is known as the Valiron–Mayer function counterpoise (VMFC) corrected binding energy.[182] In Section 3.3.3, we referred to the subcluster basis as the *k*CBS method. The subcluster basis is the standard way of predicting many-body interactions in ab intito based PES as it is free of BSIE.[31,71,76,77,79] Furthermore, these many-body interactions are reproduced with high accuracy using multipoles and perturbation theory[48]—which are BSSE-free by definition—at intermediate to long intermolecular separations. Unlike the $\varepsilon_{\mathrm{tot}}$ in eq (4.3), we cannot express $\varepsilon_{\mathrm{tot}}^{\mathrm{S}}$ as the difference between the total energy and the monomer total energies. This is because the sum of the $\varepsilon_{\mathrm{A}\cdots\mathrm{K}}$ does not add up to the E_{tot}, i.e. eq (4.1) does not hold true. This is due to the incompatibility of the to-

tal energies between different $\varepsilon_{\mathrm{A}\cdots\mathrm{K}}$. For example, the different E_{A} and $E_{\mathrm{A}\overline{\mathrm{B}}}$ are involved in computing ε_{A} and $\varepsilon_{\mathrm{AB}}$ respectively whereas the nuclei-centred counterpart would only require the same E_{A} in both cases. Thus, when the $\varepsilon_{\mathrm{A}\cdots\mathrm{K}}$ are summed in a MBE according to eq (4.1), the total energy terms do not cancel to give the exact total energy eventually. This implies that there are some effects present in the total energy that are not accounted for in the subcluster basis. In fact, this is due to the second component of BSSE—the BSEE.

Apart from the subcluster basis, another common set of basis functions that remove BSIE is one that is centred on the cluster. We denote this as the cluster basis and the k-body interaction can be written as

$$\varepsilon_{\mathrm{AB}\overline{\cdots\mathrm{N}}} = E_{\mathrm{AB}\overline{\cdots\mathrm{N}}} \tag{4.6a}$$

$$\varepsilon_{\mathrm{ABC}\overline{\cdots\mathrm{N}}} = E_{\mathrm{ABC}\overline{\cdots\mathrm{N}}} - \left(E_{\mathrm{A}\overline{\mathrm{BC}\cdots\mathrm{N}}} + E_{\mathrm{B}\overline{\mathrm{AC}\cdots\mathrm{N}}}\right) \tag{4.6b}$$

$$\begin{aligned}\varepsilon_{\mathrm{ABCD}\overline{\cdots\mathrm{N}}} = {} & E_{\mathrm{ABCD}\overline{\cdots\mathrm{N}}} - \left(E_{\mathrm{AB}\overline{\mathrm{CD}\cdots\mathrm{N}}} + E_{\mathrm{AC}\overline{\mathrm{BD}\cdots\mathrm{N}}} + E_{\mathrm{BC}\overline{\mathrm{AD}\cdots\mathrm{N}}}\right) \\ & + \left(E_{\mathrm{A}\overline{\mathrm{BCD}\cdots\mathrm{N}}} + E_{\mathrm{B}\overline{\mathrm{ACD}\cdots\mathrm{N}}} + E_{\mathrm{C}\overline{\mathrm{ABD}\cdots\mathrm{N}}}\right)\end{aligned} \tag{4.6c}$$

Here we omit the four-body term, $\varepsilon_{\mathrm{ABCD}\overline{\mathrm{E}\cdots\mathrm{N}}}$, to reduce clutter as it can be easily obtained from eq (4.4) by appending $\overline{\mathrm{E}\cdots\mathrm{N}}$ to the subscripts of each term. The basis functions centred on other monomers surrounding the subcluster of interest are involved, indicated by the overline in the many-body interaction. For example, computing $\varepsilon_{\mathrm{ABC}\overline{\cdots\mathrm{N}}}$ in eq (4.6b) requires total energies involving basis functions centred on $\mathrm{C}\cdots\mathrm{N}$ surrounding the subcluster AB. The binding energy computed using the cluster basis is

$$\varepsilon^{\mathrm{C}}_{\mathrm{tot}} = E_{\mathrm{tot}} - \sum_{\mathrm{A}}^{\binom{n}{1}} E_{\mathrm{A}\overline{\cdots\mathrm{N}}} = \sum_{\mathrm{A}<\mathrm{B}}^{\binom{n}{2}} \varepsilon_{\mathrm{AB}\overline{\cdots\mathrm{N}}} + \sum_{\mathrm{A}<\mathrm{B}<\mathrm{C}}^{\binom{n}{3}} \varepsilon_{\mathrm{ABC}\overline{\cdots\mathrm{N}}} + \cdots + \varepsilon'_{\mathrm{A}\cdots\mathrm{N}} \tag{4.7}$$

The $\varepsilon^{\mathrm{C}}_{\mathrm{tot}}$ is named the site-site function counterpoise (SSFC) corrected binding

energy.[205] It is commonly referred to simply as the counterpoise (CP) method, being a direct generalization of the CP method for a dimer system.[181] The cluster basis ensures that a common set of basis functions is employed in each of the total energy calculations, removing the undesirable BSIE. Furthermore, all the total energy terms employ the same basis, allowing for the sum of these many-body interactions to add up to the E_{tot} according to eq (4.1). Comparing the subcluster basis and cluster basis, there is an additional effect in the latter where the ghost functions surrounding the subcluster, e.g. functions centred on C$\cdots$N in the case of $\varepsilon_{\text{AB}\overline{\text{C}\cdots\text{N}}}$, improve the many-body interaction associated with the subcluster. This is the BSEE. Mathematically, we define the BSEE in the total k-body interaction, $E^{(k)}_{\text{ext.}}$, as the difference between the total k-body interaction computed using the cluster basis and subcluster basis

$$E^{(1)}_{\text{ext.}} = \sum_{\text{A}}^{\binom{n}{1}} \left(\varepsilon_{\text{A}\overline{\text{B}\cdots\text{N}}} - \varepsilon_{\text{A}}\right) \tag{4.8a}$$

$$E^{(2)}_{\text{ext.}} = \sum_{\text{A<B}}^{\binom{n}{2}} \left(\varepsilon_{\text{AB}\overline{\text{C}\cdots\text{N}}} - \varepsilon_{\text{AB}}\right) \tag{4.8b}$$

$$E^{(3)}_{\text{ext.}} = \sum_{\text{A<B<C}}^{\binom{n}{3}} \left(\varepsilon_{\text{ABC}\overline{\text{D}\cdots\text{N}}} - \varepsilon_{\text{ABC}}\right) \tag{4.8c}$$

$$E^{(4)}_{\text{ext.}} = \sum_{\text{A<B<C<D}}^{\binom{n}{4}} \left(\varepsilon_{\text{ABCD}\overline{\text{E}\cdots\text{N}}} - \varepsilon_{\text{ABCD}}\right) \tag{4.8d}$$

Unlike the BSIE, the BSEE is important in reproducing the total energy of a cluster. In Chapter 3, we shown that the MBEs using the cluster basis exhibit rapid convergence to the total energy by the four-body term.[91] This indicates that the total energy contains the BSEE as part of the variational optimization and/or the perturbative treatment of electron correlation in the total energy. The borrowing of basis functions from other monomers surrounding the subcluster does improve the flexibility of the wavefunction of the subcluster and consequently the quality of the many-body interaction. On a side note, this is like-

wise true in valence-bonded systems where the bonding between atoms can be improved by the basis functions from other surrounding atoms. This importance of BSEE also applies to the binding energy where the BSEE should be incorporated into the many-body interactions used to compute the binding energy. Therefore, we consider the $\varepsilon_{\text{tot}}^{\text{C}}$, and *not* $\varepsilon_{\text{tot}}^{\text{S}}$, to be the best estimate of the binding energy at a given level of theory and basis set.

To summarize many-body BSSE, there are two components, namely the BSIE and BSEE. The first component is undesirable, arising from an imbalance in the number of basis functions when computing energy differences in the many-body interactions and the binding energy. This BSIE can be removed by using a common set of basis functions in each of the total energy calculations, using either the subcluster basis or the cluster basis. The second component originates from the extension of the subcluster basis due to the presence of monomers surrounding the subcluster in the cluster. This BSEE is necessary to reproduce the binding energy and total energy of the cluster and can only be accounted for using the cluster basis. However, computing the many-body interactions in the cluster basis is very expensive and defeats the usefulness of the MBE in decomposing a large many-body system into manageable few-body subsystems. Thus, we wish to analyse the amount of many-body BSSE present so as to accurately yet cheaply reproduce the binding energy and total energy.

4.2.3 Many-ghost Many-body Expansion

To account for both the BSIE and BSEE (Section 4.2.2), we introduce the many-ghost many-body expansion (MGMBE). The MGMBE defines these two components of many-body BSSE up to the order of truncation of the many-body interactions, allowing us to establish the amount of many-body BSSE present in the many-body interactions. The MGMBE performs a two-dimensional many-body decomposition with each decomposition accounting for one component of many-body BSSE (Figure 4.2).

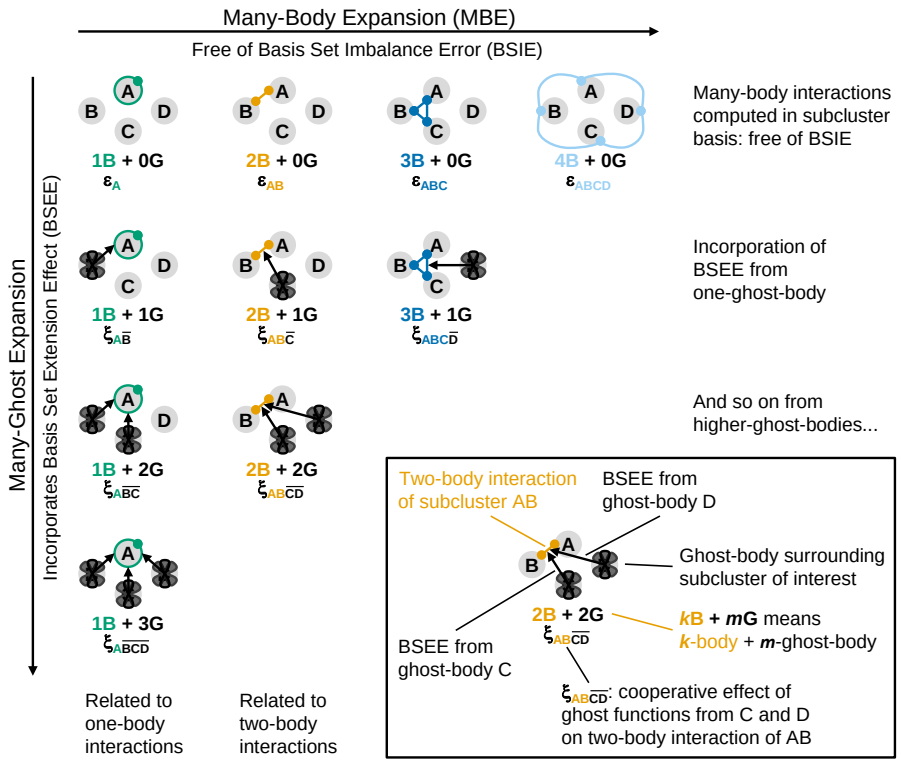


Figure 4.2: The MGMBE performs a two-dimensional many-body decomposition with the first being an MBE (left to right) up to the k-body interaction computed using the subcluster basis, $\varepsilon_{\mathrm{A}\cdots\mathrm{K}}$, using eq (4.4a–d). It is important to note that the BSIE is removed by performing this calculation. A second many-ghost expansion (top to bottom) then decomposes the BSEE from the cluster basis into contributions from m-ghost-bodies, represented by the black lobes. These BSEE terms, $\xi_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$, are shown here for $k+m \leq 4$, which can be computed using eq (4.10a–f). The inset explains the various symbols in the figure while the comments at the right and bottom edge summarizes the components of the MGMBE along each row and column respectively. We also note that along the diagonal where $k+m$ is constant, the BSIE and BSEE cancels as these terms share the same basis functions.

The first decomposition involves the MBE (Section 4.2.1) using the many-body interactions computed in the subcluster basis, $\varepsilon_{\mathrm{A}\cdots\mathrm{K}}$, which are free of the BSIE. The second decomposition, denoted the many-ghost expansion, breaks down the BSEE present in the cluster basis into contributions from one-ghost-body, two-ghost-body and so on, up to $(n-k)$-ghost-body. To reiterate, a ghost-body refers to the set of ghost functions centred on a monomer surrounding the subcluster of interest. Both decompositions can be truncated at a low order to hopefully reproduce the binding energy and total energy of the cluster. We note that the MGMBE is a logical extension of the earlier many-body counterpoise

(MBCP) method.[196,206] The MBCP method seeks to cheaply approximate the $\varepsilon_{\text{tot}}^{\text{C}}$ in eq (4.7) by performing two separate many-body decompositions on both the E_{tot} and $E_{\text{A}\overline{\text{B}\cdots\text{N}}}$. The former is an MBE computed using the nuclei-centred basis while the latter is essentially the many-ghost expansion performed on the monomer total energies. In the MGMBE, we extend the many-ghost expansion for any arbitrary k-body interaction to identify the BSEE present.

To recap, we denoted $E_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{M}}}$ as the total energy of subcluster $\text{A}\cdots\text{K}$ calculated in the presence of additional ghost functions centred on $\text{L}\cdots\text{M}$. The same situation applies to the many-body interactions, evident from the discussion on the nuclei-centred, subcluster and cluster basis in Section 4.2.2. Here, we denote $\varepsilon_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{M}}}$ as the k-body interaction of the k-mer subcluster $\text{A}\cdots\text{K}$ computed using total energies calculated with the set of basis functions centred on $\text{A}\cdots\text{KL}\cdots\text{M}$. The overline in the subscript denotes the ghost-bodies, namely the set of ghost functions centred on monomers $\text{L}\cdots\text{M}$ surrounding the subcluster. For example, $\varepsilon_{\text{AB}\overline{\text{CD}}} = E_{\text{AB}\overline{\text{CD}}} - (E_{\text{A}\overline{\text{BCD}}} + E_{\text{B}\overline{\text{ACD}}})$. In particular, the many-body interaction computed using the subcluster basis, $\varepsilon_{\text{A}\cdots\text{K}}$, and the cluster basis, $\varepsilon_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{N}}}$, are specific cases of this general notation. In the former, there are no ghost-bodies involved while the entire cluster (excluding the subcluster of interest) constitute all the ghost-bodies in the latter case. Now, we can write the MGMBE of the total energy as

$$
\begin{aligned}
E_{\text{tot}} = &\sum_{\text{A}}^{\binom{n}{1}} \varepsilon_{\text{A}} + \sum_{\text{A}<\text{B}}^{\binom{n}{2}} \varepsilon_{\text{AB}} + \sum_{\text{A}<\text{B}<\text{C}}^{\binom{n}{3}} \varepsilon_{\text{ABC}} + \sum_{\text{A}<\text{B}<\text{C}<\text{D}}^{\binom{n}{4}} \varepsilon_{\text{ABCD}} + \cdots \\
&\sum_{\text{A,B}}^{\binom{n}{1}\cdot\binom{n-1}{1}} \xi_{\text{A}\overline{\text{B}}} + \sum_{\text{A}<\text{B,C}}^{\binom{n}{2}\cdot\binom{n-2}{1}} \xi_{\text{AB}\overline{\text{C}}} + \sum_{\text{A}<\text{B}<\text{C,D}}^{\binom{n}{3}\cdot\binom{n-3}{1}} \xi_{\text{ABC}\overline{\text{D}}} + \cdots \\
&\sum_{\text{A,B}<\text{C}}^{\binom{n}{1}\cdot\binom{n-1}{2}} \xi_{\text{A}\overline{\text{BC}}} + \sum_{\text{A}<\text{B,C}<\text{D}}^{\binom{n}{2}\cdot\binom{n-2}{2}} \xi_{\text{AB}\overline{\text{CD}}} + \cdots \\
&\sum_{\text{A,B}<\text{C}<\text{D}}^{\binom{n}{1}\cdot\binom{n-1}{3}} \xi_{\text{A}\overline{\text{BCD}}} + \cdots + \xi_{\text{higher}}
\end{aligned}
\tag{4.9}
$$

where $\xi_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$ is the BSEE from m-ghost-body $\mathrm{L}\cdots\mathrm{M}$ in the k-body interaction of k-mer subcluster $\mathrm{A}\cdots\mathrm{K}$, of which there are $\binom{n}{k}\cdot\binom{n-k}{m}$ of such terms. The first line in eq (4.9) gives the MBE using many-body interactions computed in the subcluster basis using eq (4.4a–d). These interactions are free of BSIE but lack the important BSEE. The missing BSEE terms are added in the following lines with each line introducing contributions from a different number of ghost-bodies. For cases where $k+m \leq 4$, $\xi_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$ can be expressed as

$$\xi_{\mathrm{A}\overline{\mathrm{B}}} = \varepsilon_{\mathrm{A}\overline{\mathrm{B}}} - \varepsilon_{\mathrm{A}} \tag{4.10a}$$

$$\xi_{\mathrm{AB}\overline{\mathrm{C}}} = \varepsilon_{\mathrm{AB}\overline{\mathrm{C}}} - \varepsilon_{\mathrm{AB}} \tag{4.10b}$$

$$\xi_{\mathrm{ABC}\overline{\mathrm{D}}} = \varepsilon_{\mathrm{ABC}\overline{\mathrm{D}}} - \varepsilon_{\mathrm{ABC}} \tag{4.10c}$$

$$\xi_{\mathrm{A}\overline{\mathrm{BC}}} = \varepsilon_{\mathrm{A}\overline{\mathrm{BC}}} - \left(\varepsilon_{\mathrm{A}\overline{\mathrm{B}}} + \varepsilon_{\mathrm{A}\overline{\mathrm{C}}}\right) + \varepsilon_{\mathrm{A}} \tag{4.10d}$$

$$\xi_{\mathrm{AB}\overline{\mathrm{CD}}} = \varepsilon_{\mathrm{AB}\overline{\mathrm{CD}}} - \left(\varepsilon_{\mathrm{AB}\overline{\mathrm{C}}} + \varepsilon_{\mathrm{AB}\overline{\mathrm{D}}}\right) + \varepsilon_{\mathrm{AB}} \tag{4.10e}$$

$$\xi_{\mathrm{A}\overline{\mathrm{BCD}}} = \varepsilon_{\mathrm{A}\overline{\mathrm{BCD}}} - \left(\varepsilon_{\mathrm{A}\overline{\mathrm{BC}}} + \varepsilon_{\mathrm{A}\overline{\mathrm{BD}}} + \varepsilon_{\mathrm{A}\overline{\mathrm{CD}}}\right) + \left(\varepsilon_{\mathrm{A}\overline{\mathrm{B}}} + \varepsilon_{\mathrm{A}\overline{\mathrm{C}}} + \varepsilon_{\mathrm{A}\overline{\mathrm{D}}}\right) - \varepsilon_{\mathrm{A}} \tag{4.10f}$$

The meaning of these terms can be better understood by looking at specific examples. For example, $\xi_{\mathrm{ABC}\overline{\mathrm{D}}}$ in eq (4.10c) quantifies the amount by which the ghost functions centred on D affect the three-body interaction of ABC, i.e., the BSEE from D on $\varepsilon_{\mathrm{ABC}}$. Likewise, $\xi_{\mathrm{AB}\overline{\mathrm{CD}}}$ in eq (4.10e) gives the cooperative effect of the ghost functions centred on both C and D on two-body interaction of AB and higher ghost-body BSEE can be interpreted similarly. These terms represent the many-body decomposition of the BSEE present in the cluster basis. As such, the eq (4.10a–c), eq (4.10d–e) and eq (4.10f) resembles eq (4.2a), eq (4.2b) and eq (4.2c) respectively. However, there is an additional $\varepsilon_{\mathrm{A}\cdots\mathrm{C}}$ term (last term in each equation) in eq (4.10). This is the 0-ghost-body term where there is no BSEE and the equivalent in a MBE correspond to a 0-body interaction which is zero and thus omitted in the many-body interaction expressions.

In order to compute the $\xi_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$ terms, all the $\varepsilon_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$ terms have to be expressed in terms of total energies that can be readily obtained from electronic structure calculations. Here, we give an example where we express $\xi_{\mathrm{AB}\overline{\mathrm{CD}}}$ in terms of total energies

$$\begin{aligned}\xi_{\mathrm{AB}\overline{\mathrm{CD}}} &= \varepsilon_{\mathrm{AB}\overline{\mathrm{CD}}} - \varepsilon_{\mathrm{AB}\overline{\mathrm{C}}} - \varepsilon_{\mathrm{AB}\overline{\mathrm{D}}} + \varepsilon_{\mathrm{AB}} \\ &= \left(E_{\mathrm{AB}\overline{\mathrm{CD}}} - E_{\mathrm{A}\overline{\mathrm{BCD}}} - E_{\mathrm{B}\overline{\mathrm{ACD}}}\right) - \left(E_{\mathrm{AB}\overline{\mathrm{C}}} - E_{\mathrm{A}\overline{\mathrm{BC}}} - E_{\mathrm{B}\overline{\mathrm{AC}}}\right) \\ &\quad - \left(E_{\mathrm{AB}\overline{\mathrm{D}}} - E_{\mathrm{A}\overline{\mathrm{BD}}} - E_{\mathrm{B}\overline{\mathrm{AD}}}\right) + \left(E_{\mathrm{AB}} - E_{\mathrm{A}\overline{\mathrm{B}}} - E_{\mathrm{B}\overline{\mathrm{A}}}\right) \end{aligned} \tag{4.11}$$

From eq (4.11), the maximum number of basis functions is limited to four monomers in computing $\xi_{\mathrm{AB}\overline{\mathrm{CD}}}$. In fact, the maximum number of basis functions is limited to $(k+m)$ monomers in computing $\xi_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$. It is also obvious that rewriting the $\xi_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$ terms in terms of total energies can be cumbersome. Fortunately, many total energy terms are repeated and can be collected to give a more compact expression when all the $\xi_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$ terms are summed. The derivation of these working equations is presented in the Appendix.

Given that the two decompositions are independent, the BSEE present in each of k-body interactions can be truncated at a different m-ghost-body. A prudent choice would be to truncate at order (k,m) such that $k+m=\alpha$, keeping the maximum number of basis functions in each electronic structure calculation to that of α monomers. For example, truncating the MGMBE at $\alpha=2$ would include the ε_{A}, $\varepsilon_{\mathrm{AB}}$ and $\xi_{\mathrm{A}\overline{\mathrm{B}}}$ terms while truncation at $\alpha=3$ includes the previously mentioned terms as well as $\varepsilon_{\mathrm{ABC}}$, $\xi_{\mathrm{AB}\overline{\mathrm{C}}}$ and $\xi_{\mathrm{A}\overline{\mathrm{BC}}}$ terms.

A surprising result surfaced when the truncation order of the MGMBE is such that $k+m=\alpha$. Careful analysis of the working equations revealed that all the total energies involving any ghost functions vanishes when we sum the $\varepsilon_{\mathrm{A}\cdots\mathrm{K}}$ and $\xi_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$ terms with $k+m=\alpha$, where α is a constant. Consequently,

we obtain the many-body interactions computed using the nuclei-centred basis from this summation. This implies that an MBE using the nuclei-centred basis truncated at α bodies incorporates some BSEE, in particular contributions from up to $m = (\alpha - k)$-ghost-bodies in each of the k-body interactions. We stress that this surprising result only occurs when the MGMBE terms are summed *across* different k number of interacting bodies to obtain either the binding energy or total energy. To illustrate this cancellation, let us consider the sum of $\xi_{\mathrm{A}\overline{\mathrm{B}}}$, $\xi_{\mathrm{B}\overline{\mathrm{A}}}$ and $\varepsilon_{\mathrm{AB}}$. The first two terms would be $\xi_{\mathrm{A}\overline{\mathrm{B}}} = E_{\mathrm{A}\overline{\mathrm{B}}} - E_{\mathrm{A}}$ and $\xi_{\mathrm{B}\overline{\mathrm{A}}} = E_{\mathrm{B}\overline{\mathrm{A}}} - E_{\mathrm{B}}$ respectively and the total energies involving ghost functions would be eliminated when we include the $\varepsilon_{\mathrm{AB}} = E_{\mathrm{AB}} - E_{\mathrm{A}\overline{\mathrm{B}}} - E_{\mathrm{B}\overline{\mathrm{A}}}$. This leaves us with $\varepsilon'_{\mathrm{AB}} = E_{\mathrm{AB}} - E_{\mathrm{A}} - E_{\mathrm{B}}$. In essence, the BSEE in $\xi_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$ replaces the total energy terms involving ghost functions in the $\varepsilon_{\mathrm{A}\cdots\mathrm{K}}$ with corresponding ghost-free terms, "transforming" it into the nuclei-centred counterpart, $\varepsilon'_{\mathrm{A}\cdots\mathrm{K}}$. Expressed alternatively—the BSIE for a higher-body interaction (something that must be subtracted) is an BSEE for a lower-body interaction (something that must be added)—and the two effects cancel each other exactly!

4.3 Results and Discussion

All quantum chemical calculations were performed using the MOLPRO suite of programs.[185] Calculations were carried out at the second-order Møller-Plesset perturbation (MP2) level of theory using the augmented correlation-consistent basis sets aug-cc-pVnZ, labelled AVnZ, $n = \mathrm{D}, \mathrm{T}, \mathrm{Q}, 5$.[34,124] The explicitly correlated MP2 (MP2-F12) theory[208] was employed with the AVDZ basis set.

4.3.1 Basis Set Extension Effect in Many-body Interactions

In Chapter 3, we observed rapid convergence in the MBE using either the subcluster or cluster basis. This indirectly suggest that the difference between these two MBE—the BSEE—should converge rapidly with the number of bodies. We computed the $E^{(k)}_{\mathrm{ext.}}$ and total k-body interaction for the $(H_2O)_6$ cage and prism

isomers up to the four-body term with increasing basis set quality. Both isomers are taken from Richard *et al.*[206] and showed similar trends. Thus, the data for the cage isomer are shown in Figure 4.3 while the prism isomer counterpart are in the Appendix. Similar studies exist in literature but are performed on small trimer and tetramer clusters[209,210] or focused on the binding energy.[206,211] Instead, we choose to separately examine the BSEE in each of the total k-body interaction, especially between the two-body and the three-and-higher-body interactions, because they are dominated by different intermolecular interactions.[48]

At the two-body level (Figure 4.3a), the $E^{(2)}_{\text{ext.}}$ is always negative, indicating that the additional ghost functions in the cluster basis help to lower the two-body interactions. As expected, increasing the quality of the basis set decreases this borrowing of basis functions to improve the two-body interactions. These BSEE are generally small, below 1 m-E_{h}, because the additional basis functions in the cluster basis are not centred on the nuclei or on regions between nuclei where the interaction occurs. This is in contrast to the use of midbond functions where the placement of basis functions at regions between interacting molecules improves the description of the interaction.[212] We point out that the $E^{(k)}_{\text{ext.}}$ also serves as an error indicator of how well the many-body interactions computed using the subcluster basis can be used in place of the cluster basis counterpart to reproduce the binding energy or total energy. Thus, the $E^{(2)}_{\text{ext.}}$ can still be substantial if very high accuracy is demanded. For the higher-body interactions (Figure 4.3b,c), the $E^{(3)}_{\text{ext.}}$ and $E^{(4)}_{\text{ext.}}$ are minuscule—smaller than 0.045 m-E_{h}—and we can treat the many-body interactions computed in both the subcluster and cluster basis to be practically the same. This is of comfort as the use of the subcluster basis render the construction of MBE-based ab initio water potentials[31,213] possible. The reduction in dimensionality from applying the MBE is preserved unlike the cluster basis which depends on the geometry of the cluster. Indeed, the many-body interactions computed using the subcluster basis were used to construct ab initio based PES to study large water clusters and bulk water.[71,76,77,79]

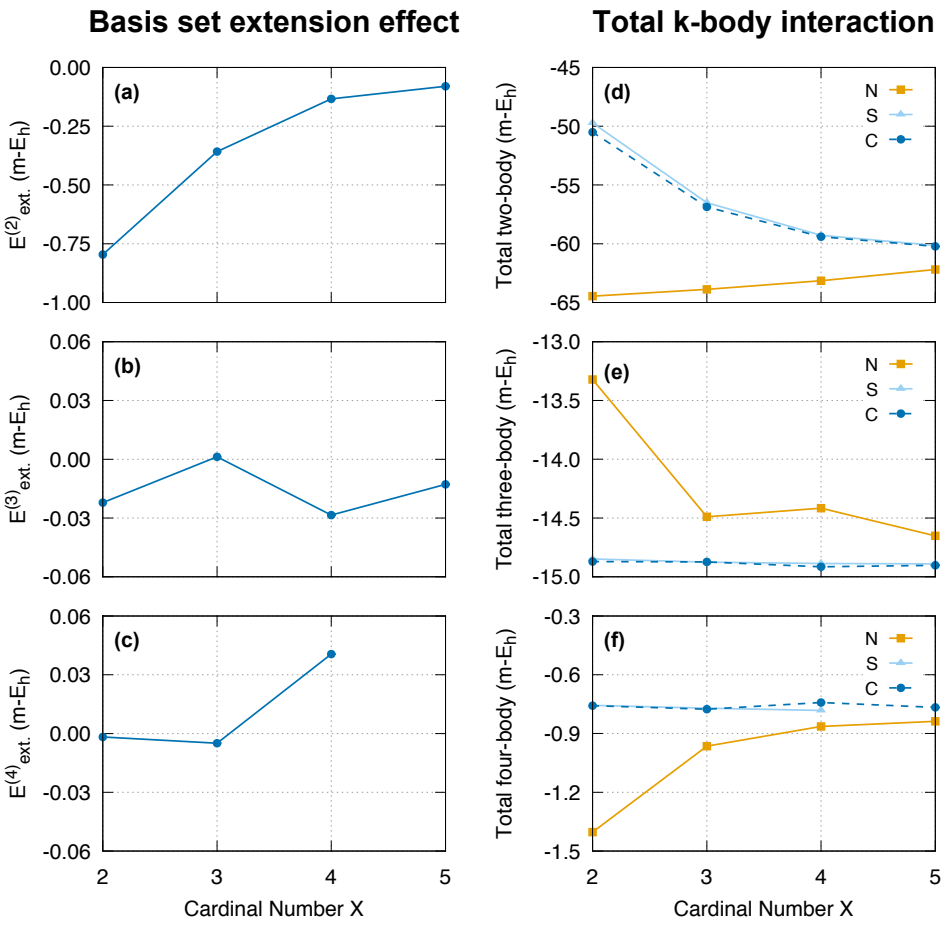


Figure 4.3: The (a–c) BSEE in the total k-body interaction, $E^{(k)}_{\text{ext.}}$, as defined in eq (4.8) and the (d–f) total k-body interaction for the cage isomer of $(H_2O)_6$ with increasing basis set quality at MP2/AVXZ. The total k-body interaction are computed using various *location* basis, namely the nuclei-centred (N), subcluster (S) and cluster (C) basis described in Section 4.2.2 to determine the effects of many-body BSSE on the many-body interactions. In particular, the lines for the cluster basis are dashed to show clearly the similarities between that and the subcluster basis results. The $E^{(4)}_{\text{ext.}}$ and total four-body interaction computed using the subcluster basis at MP2/AV5Z are omitted due to steep computational cost.

The tiny $E^{(3)}_{\text{ext.}}$ and $E^{(4)}_{\text{ext.}}$ brings us to an unrelated but important result. At the complete basis set (CBS) limit, there is no BSSE, i.e., $E^{(k)}_{\text{ext.}} = 0$. While the converse is not necessarily true, it is worthwhile to investigate if the CBS limit can be approximated using moderate-sized basis sets. Clearly, this is true for the three-body (Figure 4.3e) and four-body interactions (Figure 4.3f). Both the total three-body and four-body interaction computed using the subcluster or cluster basis (light and dark blue lines) appear to have converged, presumably to the CBS limit, varying by 0.005–0.015 m-E_{h}. This was mentioned in passing recently in the construction of an ab initio water PES where the three-body interactions computed using the subcluster basis at CCSD(T)/AVTZ are very similar to the CBS limit values.[79] With the removal of BSIE, we only require an AVDZ basis set to obtain CBS limit three-body and four-body interactions. This result implies that primarily the convergence of the total energies with increasing basis set quality comes from changes in the one-body and two-body interactions. Thus, we can obtain the total energies of water clusters with increasing basis set quality by recalculating the one-body and two-body interactions at the respective basis sets. The fact that three-and-higher-body interactions do not require a large basis set to achieve the CBS limit eliminates the need for extrapolation or ad hoc measures as commonly employed for two-body interactions. The ad hoc methods involve taking a fraction of the two-body interaction computed using the subcluster and nuclei-centred basis,[214–217] motivated by the well-documented trend[191,214,216,217] that these two quantities converge to the CBS limit from above and below respectively (Figure 4.3d).

As with our previous chapter, there is no guarantee that the observations made on small $(H_2O)_6$ clusters still hold true in larger clusters.[91] Thus, we computed the BSEE up to the four-body interaction for a homologous series of optimized $(H_2O)_{8\text{-}16}$ clusters taken from Maheshwary *et al.*[186] which is presented together with the hexamer results (Figure 4.4). Since various cluster sizes are involved, all the energies reported henceforth will be on a per monomer basis.

Calculations were performed at the MP2/AVDZ and MP2/AVTZ. The explicitly correlated MP2-F12 theory model[208] was also employed with the AVDZ basis set as this combination typically yielded results of MP2/AVQZ quality,[218] complementing the MP2/AVDZ and MP2/AVTZ results.

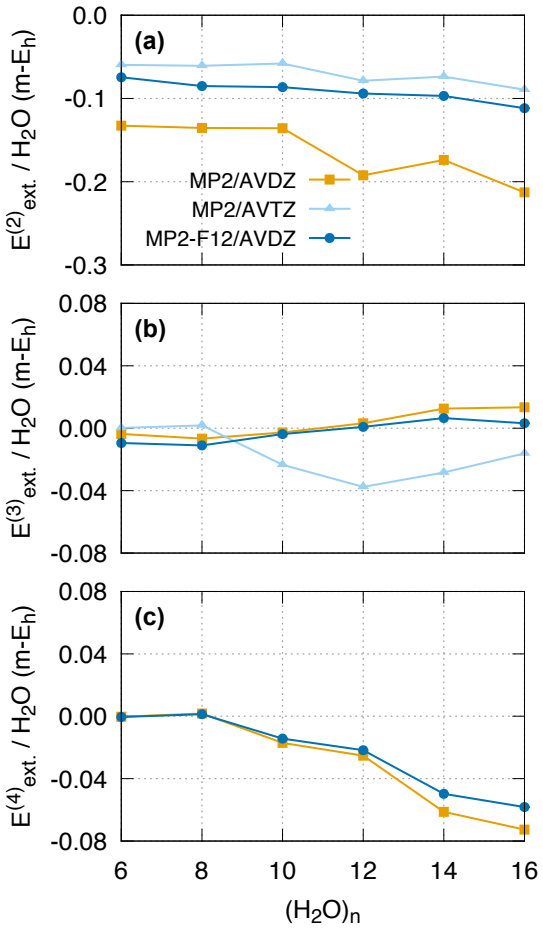


Figure 4.4: The BSEE in the total k-body interaction, $E^{(k)}_{\text{ext.}}$, per H_2O monomer for water clusters of increasing size, $(H_2O)_{6\text{-}16}$, computed at MP2 levels of theory with various basis sets. The results for $E^{(4)}_{\text{ext.}}$ are not available at the MP2/AVTZ level due to the steep computational cost involved in computing the four-body interactions.

As mentioned earlier, the $E^{(k)}_{\text{ext.}}$ serves as an error indicator of how well the cheaper subcluster basis can reproduce the more expensive cluster basis. Here, we propose an acceptable value for $E^{(k)}_{\text{ext.}}$. Studies on atomization energies and reaction enthalpies often require calculations to agree with experiments within chemical accuracy, which is 4.2 kJ mol^{-1} or 1.6 m-E_h.[219] However, the MBE is often used to study the dynamical evolution of large molecular clusters and

do not involve bond breaking. As such, we introduce the "dynamical accuracy" where the error for large clusters is computed on a per monomer basis as the properties derived from dynamical simulations are intensive in nature. A suitable dynamical accuracy might be 10 % of the thermal uncertainty at room temperature, kT, which is about 0.10 m-E_h or 0.25 kJ mol^{-1}.

From Figure 4.4a, we again observe that the $E^{(2)}_{\mathrm{ext.}}$ decreases with increasing basis set quality where the use of the higher quality MP2/AVTZ (light blue line) or the explicitly correlated MP2-F12/AVDZ (dark blue line) halved the small $E^{(2)}_{\mathrm{ext.}}$ present in MP2/AVDZ (orange line). While the $E^{(2)}_{\mathrm{ext.}}$ per monomer at MP2/AVTZ falls within dynamical accuracy, the BSEE exhibits a slow increase with increasing cluster size. Fortunately, due to the small system size, the $E^{(2)}_{\mathrm{ext.}}$ can be practically eliminated through the use of larger basis sets or CBS extrapolation. At the higher-body level, we confirmed that the $E^{(3)}_{\mathrm{ext.}}$ and $E^{(4)}_{\mathrm{ext.}}$ are, if not negligible, then acceptable. The $E^{(3)}_{\mathrm{ext.}}$ is insignificant, always below 0.040 m-E_h per monomer (Figure 4.4b). The $E^{(4)}_{\mathrm{ext.}}$ shows an increasing trend with cluster size (Figure 4.4c). Nonetheless, the value is quite small ($<$ 0.080 m-E_h per monomer) and would be even smaller if a larger basis set such as AVTZ is used. Furthermore, there would be some partial cancellation of the BSEE when the three-body and four-body interactions are summed. Therefore, we conclude that the subcluster basis can be employed in computing three-body and four-body interactions in place of the more expensive cluster basis.

Next, we overlaid the total k-body interaction computed using the subcluster basis at different basis set quality (Figure 4.5). The cluster basis counterpart shows identical trends and is presented in the Appendix. It is clear that the total three-body and four-body interactions remain the same regardless of the basis set used (Figure 4.5b,c). The total four-body interaction at MP2/AVTZ (light blue line) appears to be different due to the scale of the energy axis which exaggerates the small difference ($<$ 0.055 m-E_h) between the MP2/AVTZ and MP2/AVDZ values. The total two-body interaction (Figure 4.5a) becomes more

stabilising with increasing basis set quality, echoing the hexamer results.

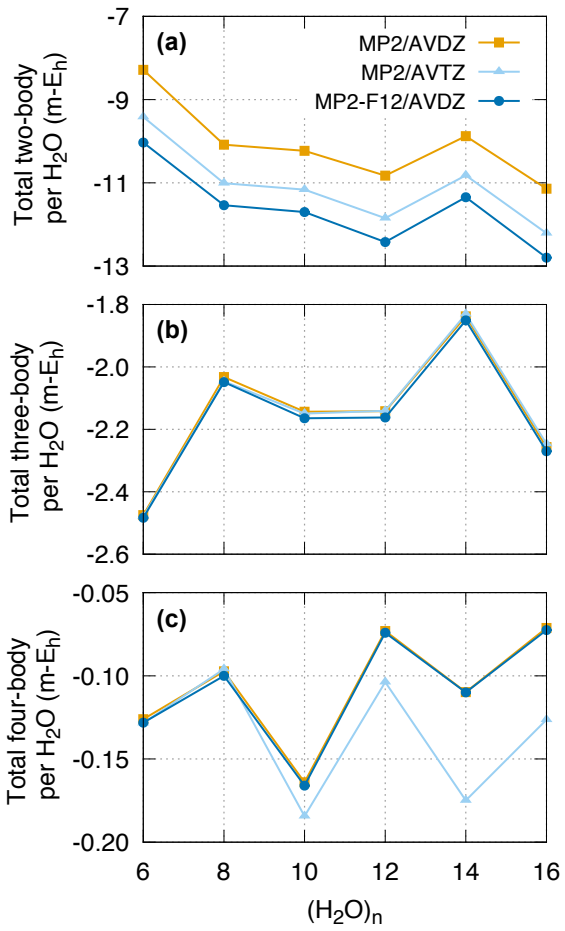


Figure 4.5: Comparison of the total k-body interaction computed using the subcluster basis per H_2O monomer for water clusters of increasing size, $(H_2O)_{6\text{-}16}$, computed at various levels of theory and basis sets.

In summary, we made three key observations: (i) the $E^{(2)}_{\text{ext.}}$ is small but significant and diminishes with increasing basis set quality, (ii) the $E^{(3)}_{\text{ext.}}$ and $E^{(4)}_{\text{ext.}}$ are much smaller, supporting the use of the cheaper subcluster basis to compute the three-body and four-body interactions and (iii) the three-body and four-body interactions computed using the subcluster basis have converged to the CBS limit using an AVDZ basis set.

4.3.2 *Many-ghost Many-body Expansion of Total Energy*

CBS extrapolation at the two-body level would eliminate the $E_{\text{ext.}}^{(2)}$, which can be coupled with AVDZ-quality three-body and four-body interactions to yield binding energies of CBS quality. However, one may also be interested in reproducing the total energy at a particular basis set. This is useful in assessing the accuracy of fragmentation methods where small groups of adjacent atoms are treated as bodies and selected many-body interactions are computed to approximate the total energies of large chemical systems.[187,201–203] We employed the MGMBE truncated at different order in an attempt to reproduce the total energy (Table 4.3). We can then determine whether the omission of certain BSEE affects the accuracy of the predicted total energy.

Table 4.3: Root Mean Square of the Error per H_2O monomer (RMSE, m-E_h) and Maximum Absolute Error per H_2O monomer (MxAE, m-E_h), in reproducing the total energy for a series of optimised water clusters from Figure 4.4 calculated at MP2/AVDZ, MP2/AVTZ and MP2-F12/AVDZ. The MGMBE includes up to the four-body term (k=1–4) of which the BSEE are truncated at different m-ghost-body.[a]

m-ghost-body in k-body[b]				MP2/AVDZ		MP2/AVTZ		MP2-F12/AVDZ	
k=1	k=2	k=3	k=4	RMSE	MxAE	RMSE	MxAE	RMSE	MxAE
0	0	0	0	2.505	2.779	1.286	1.402	0.811	0.868
1	0	0	0	0.122	0.162	0.065	0.110	0.431	0.518
2	1	0	0	0.101	0.150	0.032	0.060	0.267	0.343
1	2	1	0	0.280	0.324	0.091	0.105	0.510	0.595
2	2	1	0	0.145	0.197	0.066	0.095	0.338	0.461
3	2	1	0	0.320	0.463	0.088	0.141	0.298	0.462
All	All	All	All	0.044	0.075	—[c]	—[c]	0.037	0.062

[a] Error here is defined as the total energy of the cluster minus the MGMBE-predicted total energy. The error per H_2O monomer is first obtained before the RMS or maximum is taken. [b] The digits give the highest number of ghost-bodies, m, that is incorporated into the k-body interaction using the MGMBE and "All" refers to the cluster basis which includes all the BSEE. For example, the second entry, $\{1,0,0,0\}$, indicates that the BSEE from up to one-ghost-body is incorporated in the one-body interactions and there are no BSEE included for the two-to-four-body interactions. [c] As the four-body interaction computed using the cluster basis is computationally expensive at MP2/AVTZ, an estimate of the total energy is unavailable.

From Table 4.3, including the BSEE from one-ghost-body into the one-body interaction decreased the error by one order of magnitude as seen in the entry $\{1,0,0,0\}$. This suggests that the one-body interaction is very sensitive to the BSEE. This is not surprising as the one-body interaction constitutes the majority ($\approx 99.98\%$) of the total energy. The error decreased again when more BSEE is incorporated (entry $\{2,1,0,0\}$). However, from entry $\{2,1,0,0\}$ to $\{3,2,1,0\}$, further inclusion of BSEE resulted in a larger error. Hypothesizing that this could be due to the BSEE in the one-body interaction, we varied the truncation order of the BSEE in the one-body interaction (entry $\{1,2,1,0\}$, $\{2,2,1,0\}$ and $\{3,2,1,0\}$) and observed a fluctuation in the error. While the data is not shown here, the error actually oscillates wildly, changing in sign from positive (entry $\{1,2,1,0\}$) to negative (entry $\{2,2,1,0\}$) and back to positive again (entry $\{3,2,1,0\}$). Recall the surprising result in Section 4.2.3 that the MBE using the nuclei-centred basis truncated at the α-body term contains the BSEE from up to $m = (\alpha - k)$-ghost-bodies in each of the k-body interactions. This suggests that the similar oscillatory behaviour reported previously[91] in the MBEs using the nuclei-centred basis could be due to the BSEE present in the one-body interaction. To determine if the two oscillatory behaviours are related, we compared the convergence of the MBE using the nuclei-centred basis to the total energy of the cluster with that of the MGMBE of the one-body interaction to the total one-body interaction in the cluster basis (Figure 4.6).

It is clear from Figure 4.6 that the two many-body decompositions are practically identical except for the first two data points. It appears to be the case that the poor convergence of the MBE using the nuclei-centred basis is almost completely caused by the BSEE in the one-body interaction. The differences in the first two data points is because the MBE (Figure 4.6a) includes the actual many-body interactions together with the BSEE. In the first two data points, there are additional errors in the MBE associated with neglecting these many-body interactions. From the four-body term onwards, the majority of the

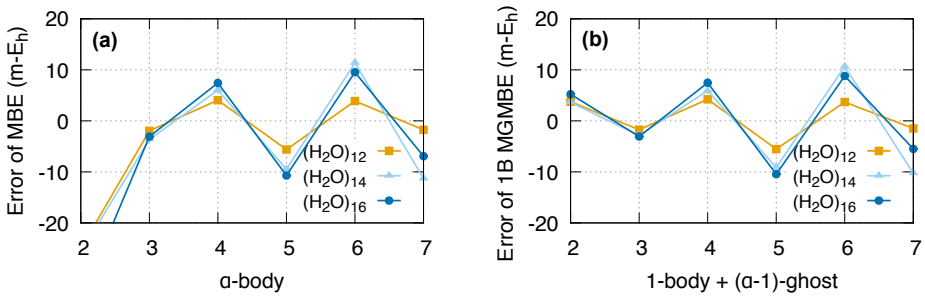


Figure 4.6: (a) The error of the MBE of the total energy using the nuclei-centred basis truncated at the α-body term follows an almost identical trend as (b) the error of the MGMBE of the total one-body interaction in the cluster basis truncated at the $(\alpha-1)$-ghost-body term. The calculations were performed at MP2/AVDZ for the $(H_2O)_{12}$, $(H_2O)_{14}$ and $(H_2O)_{16}$ from Figure 4.4. The error of the MBE is defined as the difference between the total energy, E_{tot}, and the sum of the total one-body interaction up to the total α-body interaction. Similarly, the error of the MGMBE is defined as the difference between the one-body interaction computed using the cluster basis, $E_{\text{A}\overline{\text{B}\cdots\text{N}}}$, and the sum of the BSEE from up to $(\alpha-1)$-ghost-bodies summed across all the monomers.

many-body interactions are accounted for and virtually all the remaining error is apparently due to BSEE in the one-body interaction.

The errors in the MBE associated with the BSEE not only applies to the brute force computation of all the $\binom{n}{k}$ individual k-body interactions but also to "internally consistent" selected many-body interactions virtually always employed in fragmentation methods. These interactions are "internally consistent" in a sense that the many-body interactions of the selected fragments (interacting groups of atoms) and their constituent lower-body interactions are included and only included once. This allows for the BSIE and BSEE to cancel. The poor convergence of the MBE/MGMBE allows us to explain certain observations in fragmentation methods. "Grafting" is employed in some fragmentation methods[220–222] where the total energy of the system is calculated at a lower level of theory or basis set to serve as a correction to the predicted total energy. Such grafting approaches not only correct for missing important many-body interactions but also account for the BSEE to a large extent, explaining the low errors associated with these methods. Since the BSEE converges poorly with respect to

the number of ghost-bodies, the expensive one-body interactions computed using the cluster basis is required to accurately reproduce the total energy. Future investigations to develop cheaper alternatives to the cluster basis will be undertaken. One possibility includes the omission of certain basis functions from the basis set, in particular the tight valence-type functions (i.e., not diffuse functions), on ghost-bodies that are far away from the monomer of interest. Thus, only the contributing diffuse functions remain.

4.3.3 Many-ghost Many-body Expansion of Binding Energy

We have shown that the poor convergence of the MBE using the nuclei-centred basis is due to the BSEE in the one-body interactions. With the removal of the one-body interactions and its associated BSEE, we expect the remaining energy to converge rapidly with the number of bodies. This remaining energy is the binding energy, $\varepsilon_{\rm tot}^{\rm C}$, and the accuracy of the MGMBE is evaluated in Table 4.4.

Table 4.4: Root Mean Square of the Error per H_2O monomer (RMSE, m-E_h) and Maximum Absolute Error per H_2O monomer (MxAE, m-E_h), in reproducing the $\varepsilon_{\rm tot}^{\rm C}$ for the same water clusters in Table 4.3. The MGMBE includes up to the four-body term (k=2–4) of which the BSEE are truncated at different m-ghost-bodies.[a]

m-ghost-body in k-body[b]			MP2/AVDZ		MP2/AVTZ		MP2-F12/AVDZ	
k=2	k=3	k=4	RMSE	MxAE	RMSE	MxAE	RMSE	MxAE
0	0	0	0.160	0.197	0.046	0.058	0.090	0.104
1	0	0	0.046	0.070	0.073	0.114	0.060	0.090
2	1	0	0.009	0.015	0.019	0.043	0.015	0.028
All	All	All	0.044	0.075	—[c]	—[c]	0.037	0.062

[a] Error here is defined as the $\varepsilon_{\rm tot}^{\rm C}$ of the cluster minus the MGMBE-predicted $\varepsilon_{\rm tot}^{\rm C}$. The error per H_2O monomer is first obtained before the RMS or maximum is taken. [b] The digits give the highest number of ghost-bodies, m, that is incorporated into the k-body interaction using the MGMBE and "All" refers to the cluster basis which includes all the BSEE. For example, the second entry, $\{1,0,0\}$, indicates that the BSEE from up to one-ghost-body is incorporated in the two-body interactions and there are no BSEE included for the three-body and four-body interactions. [c] As the four-body interaction computed using the cluster basis is computationally expensive at MP2/AVTZ, an estimate of the $\varepsilon_{\rm tot}^{\rm C}$ is unavailable.

From Table 4.4, the incorporation of the BSEE greatly reduces the error in

reproducing the $\varepsilon_{\text{tot}}^{\text{C}}$, eventually giving a tiny error per monomer of below 0.015–0.043 m-E_{h} (entry $\{2,1,0\}$), which is well within dynamical accuracy. In fact, entry $\{2,1,0\}$ gives a lower error than entry $\{\text{All},\text{All},\text{All}\}$ which incorporates all the BSEE in the two-to-four-body interactions. This can be attributed to a reversal in the sign of the error. In entry $\{0,0,0\}$, the absence of BSEE which stabilizes the binding energy results in negative errors. On the other hand, the errors from the incorporation of all the BSEE up to the four-body interaction in entry $\{\text{All},\text{All},\text{All}\}$ are positive due to the neglect of higher-than-four-body interactions. Thus, there is some form of error cancellation between the two factors when the majority of the BSEE is accounted for in entry $\{2,1,0\}$. Furthermore, the maximum number of basis functions ever employed in any total energy calculation is limited to that of four monomers in entry $\{2,1,0\}$, originating from either the interacting bodies or ghost-bodies. This allows for expensive theoretical models such as the Coupled Cluster Singles and Doubles with perturbative Triples [CCSD(T)] to be applied to obtain highly accurate $\varepsilon_{\text{tot}}^{\text{C}}$ for large clusters or even bulk-water simulations. It should be emphasized that no charge embedding scheme[152,153,223] was used although they are commonly applied to water clusters. The use of such schemes is prevalent in the literature due to the belief that the water-water interactions are highly many-body in nature. However, our results indicate that we only require up to the four-body interactions. It is likely that any apparent higher-than-four-body effects are caused by the BSEE in the one-body interactions which we have shown to be highly many-body in nature (Figure 4.6).

Notably, the calculations involved in entry $\{2,1,0\}$ is equivalent to that in a MBCP(4) calculation.[196,206] A MBCP(4) calculation would involve a MBE using the nuclei-centred basis truncated at the four-body term minus the one-body interactions with the BSEE truncated at the $(4-1)=3$-ghost-body level. This is equivalent to a "$\{3,2,1,0\}$" MGMBE calculation of the total energy minus the one-body interactions and its associated BSEE, i.e. the entry $\{2,1,0\}$

in Table 4.4. Thus, an "$\{\alpha-2,\alpha-1,\cdots,0\}$" MGMBE calculation of the $\varepsilon^{\mathrm{C}}_{\mathrm{tot}}$ is identical to an MBCP(α) calculation. Note that for the MBCP method, the MG1BE have to be truncated at one-order less than that of the MBE. This is important to ensure that all the BSEE in the one-body interactions are properly removed and this requirement only becomes obvious with the analysis of the BSEE using the MGMBE presented in this chapter.

4.4 Summary

Through a systematic study of water clusters with improving basis set and increasing cluster size, we concluded that one has to account for many-body BSSE in order to reproduce the many-body interactions computed using the cluster basis. There are two distinct components to the many-body BSSE. The first arises due to an imbalance in the number of basis functions used to compute a particular k-body interaction. In this case the k-body total energy calculation utilizes many more basis functions than does the lower-body counterparts which are necessary to extract the k-body interaction. The second arises due to the fact that a k-body within a much larger cluster is further stabilized by the basis functions of the surrounding bodies denoted as the BSEE. If one wants to reproduce the binding energy and/or the total energy through a many-body approach, the first BSIE is undesirable as it leads to erroneous many-body interactions. However, the BSEE 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. Thus, the best estimate of the binding energy at a given basis set would be the total energy minus the one-body intramolecular interactions computed using the cluster basis.

We found that both components of the many-body BSSE are accounted for in the three-body and four-body interactions computed using the subcluster basis and that these interactions appear to have converged to the CBS limit using the AVDZ basis set. For the two-body interactions, and particularly for

the one-body intramolecular interactions, important BSEE are significant and have to be accounted for, thus making the use of the subcluster basis insufficient. To account for both the BSIE and the BSEE, we introduce the MGMBE in this chapter. The MGMBE performs a two-dimensional many-body decomposition with each decomposition accounting for one component of many-body BSSE. Through the MGMBE of the total energy, we found that the oscillatory behaviour encountered in MBEs using diffuse functions is caused by the BSEE in the one-body interactions. With the adequate removal of the one-body interactions and the associated BSEE, the MGMBE successfully reproduces the binding energies of clusters using numerous small calculations that involves no more than four monomers.

Despite the utility of decomposing a large cluster into small subsystems, the MBE and the MGMBE comes with a limitation. The number of four-body calculations increases quartically with the cluster size, substantially hindering the scalability of these methods. To circumvent this, the next chapter will establish a rigorous criterion to select out all potentially significant many-body interactions.

5 | When are Many-body Effects Significant?

Many-body effects are required for an accurate description of both structure and dynamics of large chemical systems. However, there are numerous such interactions to consider and it is not obvious which ones are significant. We provide a general and fast method for establishing which small set of three-body and four-body interactions are important. This is achieved by estimating the maximum many-body effects, ε_{max}, that can arise in a given arrangement of bodies. Through careful analysis of ε_{max} we find two overall causes for significant many-body interactions. Firstly, many-body induction propagates in non-branching paths, i.e, in a chain-like manner. Secondly, linear arrangements of bodies promote the alignment of the dipoles to reinforce the many-body interaction. Compact arrangements are favoured, not because of dipole alignment, but rather because there are many short non-branching paths connecting the bodies. Extended linear arrangements are favoured because dipoles can align well while maintaining at least one short non-branching path. The latter result is not intuitive as these linear arrangements can lead to significant many-body effects extending over large distances. This chapter provides a rigorous explanation as to how cooperative effects provide enhanced stability in helices making them one of the most common structures in biomolecules. Not only do these helices promote linear dipole alignment but their chain-like structure is consistent with

the way many-body induction propagates. Finally, using $\varepsilon_{\max}$ to screen for significant many-body interactions, we are able to reproduce the total three-body and four-body interaction energies using a small number of individual many-body interactions.

5.1 Introduction

The interplay of numerous noncovalent interactions often underpin the dynamics and structure of large chemical systems. Initial theoretical studies of noncovalent interactions often assumed that interactions only occur between two bodies, described by classical electrostatics and Lennard-Jones potentials.[32,224] Here, the body refers to a subunit within the large chemical system, for example, a monomer in molecular clusters or an amino acid residue in proteins. However, the interaction picture is more intricate than initially assumed where a third body can alter the interaction between two bodies. This "third-party effect" has acquired several names in the literature—cooperativity,[225–227] nonadditive effects,[48,228] and many-body effects.[163,180] In this chapter, we refer to these effects as many-body effects or more precisely k-body effects, where k is the number of interacting bodies. When the bodies are highly polar or exhibit hydrogen bonding, these many-body effects manifest strongly. One example of this is the drastic enhancement of the dipole moment of water molecules in the condensed phase as compared to the gas phase.[24,229] In biology, structural changes such as the shortening of O···H hydrogen bonds in α-helices of increasing length also illustrate the extent of these many-body effects.[230–232]

The importance of many-body effects means that we need to consider the interactions between all triple and quadruple of bodies and so on, a seemingly insurmountable task given the sheer number of such interactions. Fortunately, due to the nearsightedness of electronic matter,[233,234] it is possible to neglect some, if not the majority, of the many-body effects. Based on this nearsightedness notion, fragmentation methods were developed to reproduce the total

energies of large chemical systems by breaking these systems into small fragments.[187,201–203,235] The small fragments are then selectively interacted based on the inter-fragment distance or connectivity. Likewise, several parameters based on the distances between the bodies in a large chemical system have been proposed to identify important many-body effects.[77,79,179,236–238] However, the choice of interactions/parameters are often based on chemical intuition, making the aforementioned studies seem highly empirical. Thus, an analysis of the selected many-body effects from these studies does not lend itself to a clear picture of how many-body effects manifest.

To rigorously identify the significant many-body effects, we look to the theory of intermolecular interactions. We focus on long-range interactions as they persist at large separations, causing many-body effects to remain significant even when the bodies are moderately apart. At long-range, intermolecular interactions can be separated into contributions from electrostatic, induction and dispersion.[48] Electrostatic interaction arises from the interaction between static charge distributions while dispersion is a stabilization due to the correlated motions of the electrons in different molecules. Electrostatic and dispersion interactions exhibit no and negligible many-body effects respectively. Induction originates from the polarization of the electron cloud of a molecule by the electric field of the neighbouring molecules. As the electric field exerted by a molecule can be enhanced or negated by that of another molecule, induction is highly many-body in nature and constitutes the majority of many-body effects.

This chapter revolves around estimating the maximum many-body interaction energy, $\varepsilon_{\max}$, that can arise from a particular arrangement of three or four bodies. To keep $\varepsilon_{\max}$ simple, we are only interested in the most important contribution to the many-body effects. Thus, we derived the leading terms in the many-body induction interaction. We then used $\varepsilon_{\max}$ to identify significant many-body effects in water clusters and secondary structures of polyglycine, both of which contain highly polar bodies. By including the many-body contri-

butions above a certain cut-off, we successfully reproduce the total many-body interactions using a small number of the possible contributions. Furthermore, we refined some of the included many-body interactions and found that the many-body effects in 3_{10}- and α-helices of polyglycine extend up to 12 and 18 residues respectively. Other than its utility in many-body-based applications, we separated ε_{max} into distance and orientational components to understand the propagation of many-body effects. Careful analysis of the distance component revealed that many-body induction propagates in non-branching paths. On the other hand, the orientational counterpart is related to the alignment of the dipoles to maximise induction. With that, we identified that compact and extended linear arrangements tend to possess significant many-body effects due to their strong distance and orientational components respectively. The latter implies that many-body effects can extend over large distances. These insights help us rethink how many-body effects propagate.

5.2 Computational Details

5.2.1 Many-body Interactions

Following Chapter 4,[239] the three-body and four-body interactions are computed in the subcluster basis using the following formulae

$$\varepsilon_{\text{ABC}} = E_{\text{ABC}} - \left(E_{\text{AB}\overline{\text{C}}} + E_{\text{AC}\overline{\text{B}}} + E_{\text{BC}\overline{\text{A}}}\right) + \left(E_{\text{A}\overline{\text{BC}}} + E_{\text{B}\overline{\text{AC}}} + E_{\text{C}\overline{\text{AB}}}\right) \tag{5.1}$$

$$\begin{aligned}\varepsilon_{\text{ABCD}} = {} & E_{\text{ABCD}} - \left(E_{\text{ABC}\overline{\text{D}}} + E_{\text{ABD}\overline{\text{C}}} + E_{\text{ACD}\overline{\text{B}}} + E_{\text{BCD}\overline{\text{A}}}\right) \\ & + \left(E_{\text{AB}\overline{\text{CD}}} + E_{\text{AC}\overline{\text{BD}}} + E_{\text{AD}\overline{\text{BC}}} + E_{\text{BC}\overline{\text{AD}}} + E_{\text{BD}\overline{\text{AC}}} + E_{\text{CD}\overline{\text{AB}}}\right) \\ & - \left(E_{\text{A}\overline{\text{BCD}}} + E_{\text{B}\overline{\text{ACD}}} + E_{\text{C}\overline{\text{ABD}}} + E_{\text{D}\overline{\text{ABC}}}\right)\end{aligned} \tag{5.2}$$

where $E_{\text{A}\cdots\text{D}}$ are total energies obtained from quantum chemical calculations and the subscript indicate the bodies being calculated. The overline in the sub-

script, for example $E_{\mathrm{ABC\overline{D}}}$, indicates the presence of ghost functions centred on body D where basis functions are placed on the locations of nuclei, but without the nuclei being present in the quantum chemical calculation. In eq (5.1) and (5.2), the use of a consistent set of basis functions centred on the subcluster (ABC and ABCD respectively) removes the undesirable basis set imbalance error in many-body BSSE (Section 4.2.2).[239]

5.2.2 Quantum Chemical Calculations

Quantum chemical calculations of all the total energies were performed using the MOLPRO suite of programs.[185] Calculations were carried out at either the second-order Møller-Plesset perturbation (MP2) or Hartree-Fock (HF) level of theory. Either the correlation-consistent basis sets, cc-pV*n*Z, labelled V*n*Z, *n*= D or T, or the augmented form, aug-cc-pVDZ, labelled AVDZ was used.[34,124] In particular, the three-body and four-body interactions in the water clusters are computed at MP2/AVDZ and HF/AVDZ respectively, which is of CBS quality according to Section 4.3.1.[239] For the polyglycine, the three-body and four-body interactions were initially computed at MP2/VDZ and HF/VDZ respectively (Figure 5.5). Selected three-body and four-body interactions are later refined to a higher quality at MP2/VTZ (Figure 5.6). The magnitude of the dipole, μ, and isotropic polarizability, α, were calculated using the Gaussian 09 package[184] at the respective level of theory and basis set. For the calculation of μ and α, we used the equilibrium geometry of the water monomer at the respective level of theory and basis set for the water clusters and the optimised geometry of the repeating unit for the polyglycine. The magnitude of the dipole vector is given as $\mu = \sqrt{\mu_x^2 + \mu_y^2 + \mu_z^2}$ while the isotropic dipole-dipole polarizability is the average of the trace of the dipole-dipole polarizability matrix, $\alpha = (\alpha_{xx} + \alpha_{yy} + \alpha_{zz})/3$.

The $(H_2O)_{16}$, $(H_2O)_{20}$, $(H_2O)_{32}$ and $(H_2O)_{40}$ clusters were obtained from Yoo *et al.*,[176] Wang *et al.*,[240] Pruitt *et al.*[177] and Saha *et al.*[222] respectively.

There are five clusters for each cluster size and the Cartesian coordinates are found in the accompanying CD. For the $H(C(O)NHCH_2)_{24}H$ polyglycine, the repeating unit approach[230] is used where each repeating $C(O)NHCH_2$ body has the same geometry. To obtain the geometry of the repeating unit, geometry optimization using the Gaussian 09 package at HF/VDZ was performed on a shorter $H(C(O)NHCH_2)_{18}H$ polyglycine model with a constraint that the geometry of each $C(O)NHCH_2$ repeating unit is kept identical. The optimised geometry of the repeating unit is found in the accompanying CD. In order to define the bodies and compute the many-body interactions in polyglycine, covalent bonds have to be severed. All severed covalent bonds were capped with hydrogen atoms, a common practice in the fragmentation of valence-bonded systems.[187,201–203] The placement of capping hydrogen atoms follows that of a previous study.[161] Furthermore, to ensure that a consistent set of basis functions are used in each total energy calculation, ghost functions are placed at all the locations of the capping hydrogens when these capping hydrogens are absent in the calculation.

5.2.3 Genetic Algorithm

To obtain the dipole orientations giving ε_{max}, we applied a genetic algorithm (GA). A GA dataset is created and fitted to simple models so that the orientational components can be evaluated quickly. The GA method is adapted from Guimarães *et al.*[241] where the energy function is replaced with the leading many-body terms given in eq (5.3) and eq (5.4). In the GA method, the initial population, N_{pop}, the fraction of clusters mutated per generation, f_{mut}, the number of generations with no improvement in energy, N_{conv} and the number of iteration cycles for the history operator, N_{cycles} are required. For all the GA runs, we specified that f_{mut}=0.1, N_{conv}=10 and N_{cycles}=15. N_{pop} were set to 26 and 50 for the three-body and four-body cases respectively. Triplicates of the GA runs were performed and identical results were obtained.

To generate the geometries for the three-body GA dataset, two of the inter-

body distances were fixed to 4 Å and the arrangement is solely dependent on the angle, θ (Figure 5.1a). For the four-body counterpart, the arrangement can be described using three lengths, two bending angles and one dihedral angles, similar to the construction of a Z-matrix of a four-atom molecule (Figure 5.1b). The angles in the above arrangements are chosen such that the fixed inter-body distances form the shortest non-branching path. Thus, when presented with a three-body or four-body arrangement, the shortest non-branching path will be determined first. The corresponding angles will then be computed and plugged into the models/functional fits to obtain the orientational component. The models/functional fits will be discussed next.

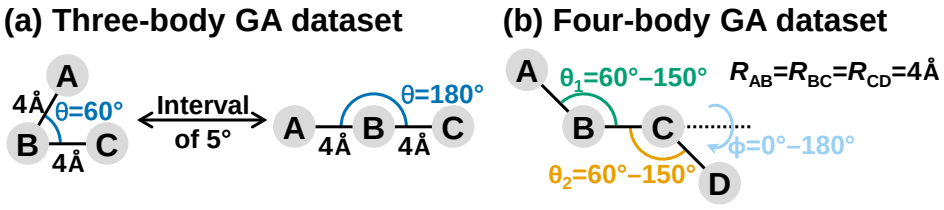


Figure 5.1: (a) To generate the three-body arrangements for the GA dataset, two of the inter-body distances are fixed to 4 Å and the angle, θ is varied from $\theta = 60°$ to $\theta = 180°$. (b) Similarly, the four-body arrangements are generated by fixing three of the inter-body distances to 4 Å and varying the two bending angles, θ_1 and θ_2, and one dihedral angle, ϕ.

From this three-body GA dataset, we observed that the dipole orientations exhibit two different behaviour. When the bodies are close to each other (small θ), the dipoles point to the circumcentre of the triangle formed by the three bodies ABC. Conversely, when the bodies are arranged linearly, the dipoles become aligned along AC with some offset related to the angles of the triangle ABC. Thus, when presented with a three-body arrangement, both the "circumcentre" and "align dipole" models will be applied and the larger three-body interaction will be taken as $\varepsilon^{(3)}_{\text{max}}$. This procedure was tested on the geometries of the three-body GA dataset and gave a mean absolute error (MAE) of 6%. A figure detailing the two models and a comparison of the $\varepsilon^{(3)}_{\text{max}}$ obtained from the two models and the actual GA results are found in the Appendix. For the four-body

interactions, the orientational components obtained from GA are fitted to a sum of product of cosine functions using Mathematica[242] and the fitted coefficients are found in the Appendix.

5.3 Results and Discussion

5.3.1 Nature of Many-body Induction

To describe the induction interactions between the many bodies, we apply the Rayleigh–Schrödinger perturbation theory.[48] For each body, the charge distribution is described as the magnitude of its dipole, μ, and the degree of distortion of this dipole by the electric field of neighbouring molecules is given by an isotropic dipole-dipole polarizability, α. Notice that the higher-rank multipoles (the quadrupoles, octopoles and so on) are ignored as we are only interested in the most significant contribution to the many-body induction. With that, the leading three-body and four-body terms in the induction interactions were obtained in eq (5.3) and (5.4) respectively (see Appendix for a detailed derivation)

$$\varepsilon_{\text{max}}^{(3)} = \sum_{\text{3paths}} \frac{\mu^{\text{A}} \mathscr{T}^{\text{AB}} \alpha^{\text{B}} \mathscr{T}^{\text{BC}} \mu^{\text{C}}}{R_{\text{AB}}^3 R_{\text{BC}}^3} \tag{5.3}$$

$$\varepsilon_{\text{max}}^{(4)} = \sum_{\text{12paths}} \frac{\mu^{\text{A}} \mathscr{T}^{\text{AB}} \alpha^{\text{B}} \mathscr{T}^{\text{BC}} \alpha^{\text{C}} \mathscr{T}^{\text{CD}} \mu^{\text{D}}}{R_{\text{AB}}^3 R_{\text{BC}}^3 R_{\text{CD}}^3} \tag{5.4}$$

where $\mathscr{T}^{\text{AB}}$ gives a measure of alignment of the dipoles of body A and B and R_{AB} is the distance between body A and B. These leading terms are characterized by an orientational component (the $\mathscr{T}$) in the numerator and a distance component (the R) in the denominator. These leading many-body terms will serve as the formulae to compute ε_{max}.

From the distance component, we observe that the bodies are coupled in non-branching paths. This can be understood by following the polarization of the bodies (Figure 5.2a). Firstly, the electric field of body A polarizes body B,

creating a first-order induced dipole, $\mu^{\mathrm{A}}T^{\mathrm{AB}}\alpha^{\mathrm{B}}$. This alters the electric field of B, which in turn changes the polarization of body C. The second-order induced dipole, $\mu^{\mathrm{A}}T^{\mathrm{AB}}\alpha^{\mathrm{B}}T^{\mathrm{BC}}\alpha^{\mathrm{C}}$, can then interact with the electric field of body D to give the second-order induction interaction, $\mu^{\mathrm{A}}T^{\mathrm{AB}}\alpha^{\mathrm{B}}T^{\mathrm{BC}}\alpha^{\mathrm{C}}T^{\mathrm{CD}}\mu^{\mathrm{D}}$, which involves ABCD. This polarization process can be extended to any arbitrary *k*th-order. Notice that the path followed by the polarization process can only be extended by coupling the terminal body of the path with another body. This ensures that the path is always non-branching, a key finding in this chapter. In general, k polarizations lead to $(k+1)$ couplings, allowing for at most $(k+2)$ bodies to be involved. This general trend also allows us to list down the possible couplings between the bodies at first-order and second-order induction (Figure 5.2b). We observe that the leading two-body and three-body terms are identical in order in distance, i.e., both having an overall dependence of L^{-6} where L is one of the inter-body distances. Furthermore, the leading three-body induction term decays slower than the Axilrod–Teller–Muto three-body dispersion term which varies as L^{-9}.[243,244] This is because the three-body dispersion depends simultaneously on all three inter-body distances. On the other hand, the three-body induction depends on the non-branching paths, which only involves two distances. The leading four-body term is only found in the second-order induction, having a L^{-9} overall distance dependence. In the leading many-body terms, there are 3 and 12 possible non-branching paths joining three bodies and four bodies respectively. The other cases shown in Figure 5.2b show back-polarization to previously-involved bodies and these paths give rise to non-leading terms. More importantly, the results show that the many-body effects can extend over large distances as these effects do not depend simultaneously on all the inter-body distances but rather the length of the non-branching path.

The orientational component of $\varepsilon_{\mathrm{max}}$ gives a measure of how well the dipoles can align themselves to reinforce the induction interaction. We constructed different three-body (Figure 5.3a) and four-body (Figure 5.3b) arrange-

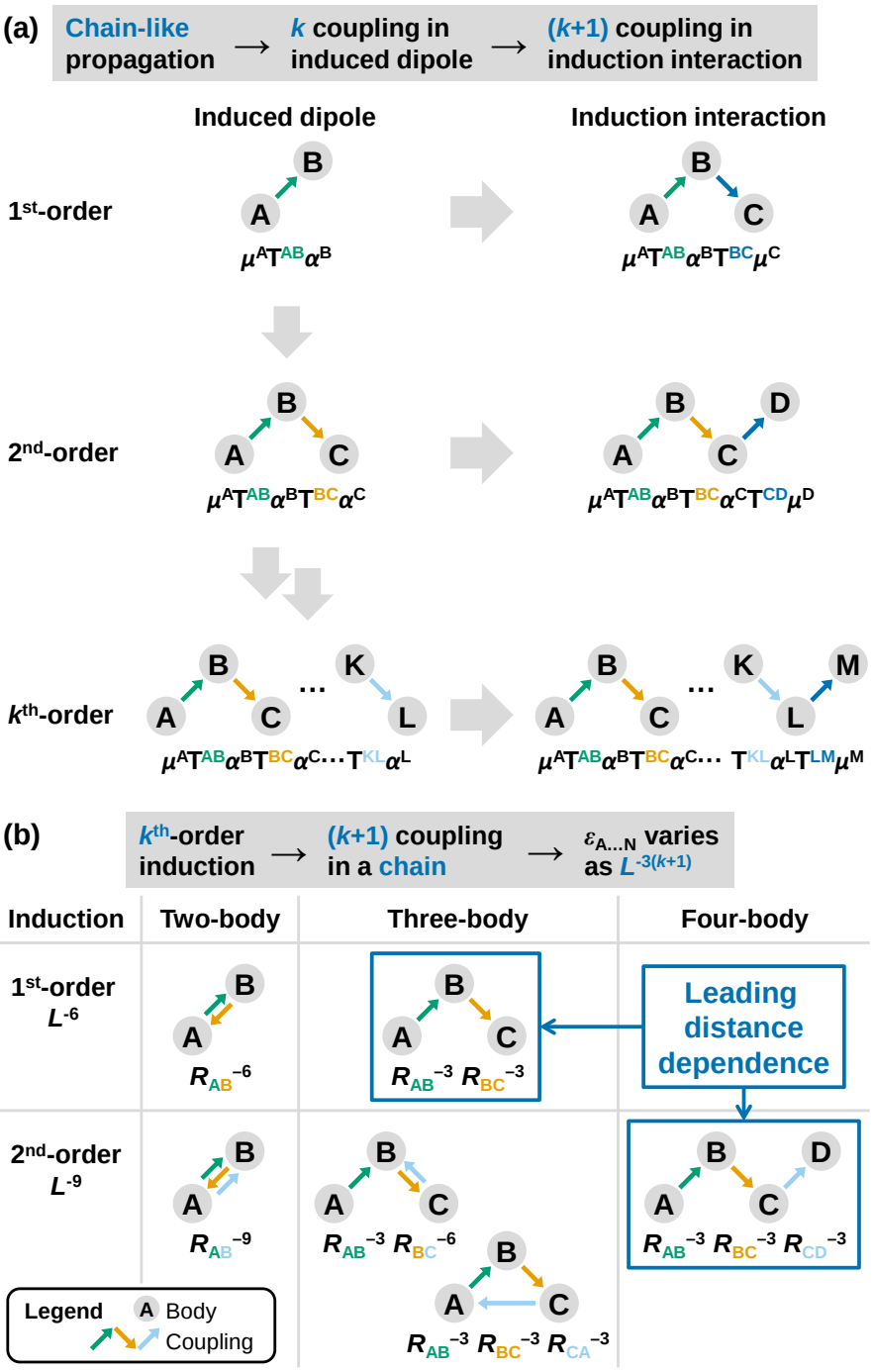


Figure 5.2: The distance component of $\varepsilon_{\max}$ reveals that many-body effects connect the bodies in a chain-like manner. (a) Pictorial representation of the polarization of the bodies to illustrate the chain-like coupling between bodies. Here, we use a shorthand for the coupling between AB, $T^{\mathrm{AB}} = \mathscr{T}^{\mathrm{AB}}/R_{\mathrm{AB}}^3$, and the terms on the right-hand side are defined in eq (5.3) and (5.4). In general, at the k-order induction, there are $(k+1)$ pairwise couplings. (b) List of possible couplings, and the corresponding distance dependence, at the first-order and second-order induction.

ments to investigate which ones favour dipole alignment, i.e, a large orientational component. The size of these arrangements is determined by some of the inter-body distances, L, while the shape is controlled by the angles θ, θ_{Ch} and θ_{Td}. The dipole orientations are obtained from simple models fitted to a GA dataset (Section 5.2.3).

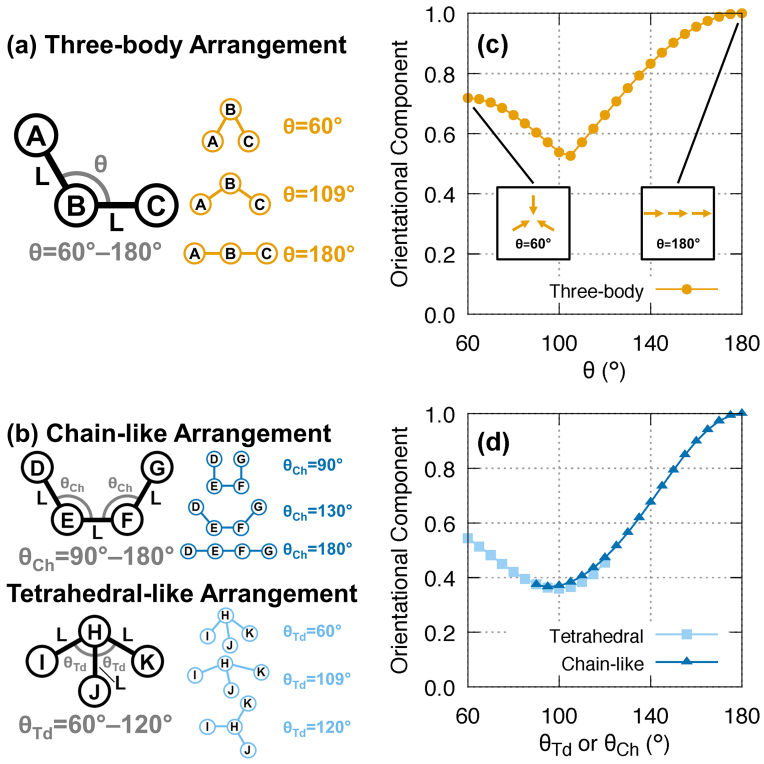


Figure 5.3: Compact and extended linear arrangement tend to possess significant many-body effects. (a) The geometry of three bodies changes from a compact arrangement to a linear arrangement as the angle between the bodies, θ, increases. (b) For four bodies, the chain-like arrangement becomes more extended as θ_{Ch} increases while the tetrahedral-like arrangements become more compact as θ_{Td} decreases. The orientational components of $\varepsilon_{\mathrm{max}}$ at the (c) three-body and (d) four-body level are given with varying angles. Note that for both the orientational and distance components are normalized against the largest possible value.

We observe that the extended linear arrangements give rise to large orientational components, corresponding to large θ values for the three bodies (Figure 5.3c) and large θ_{Ch} values at the four-body level (Figure 5.3d). Such linear arrangements allow all the dipoles to point in the same direction (see "$\theta = 180°$" inset in Figure 5.3c), reinforcing the electric field and maximiz-

ing the many-body effects. On the other hand, compact arrangements (small θ and θ_{Td}) provide a modest orientational component with the dipoles pointing towards a common point (see "$\theta = 60°$" inset in Figure 5.3c).

Altogether, compact arrangements give significant many-body effects because there are many short non-branching paths connecting the bodies, giving a large distance component while having a modest orientational component. Extended linear arrangements also give significant many-body effects because dipoles can align well, resulting in a large orientational component while maintaining at least one short non-branching path. The latter result is counter-intuitive as it implies that many-body effects can propagate over large distances but only in a directional manner. This explains why linear-type structures such as helices are common stable structures in biomolecules. Furthermore, knowledge of the geometries that can give rise to significant many-body effects is important in a wide variety of many-body-based applications. It was shown previously that the inclusion of linear arrangements in a three-body water potential energy surface was critical to achieve high accuracy.[77,79] Furthermore, a recent analysis of the three-body effects in water clusters proposed looking at the "shell sum" of a trimer, determined by the distances between a central water molecule and the two other waters.[237] This is similar to the non-branching paths which we derived have rigorously from perturbation theory whereas not much explanation was given by the authors for their use of "shell sum". Consequently, a straightforward extension to four-body interactions is not clear based on their work. Thus, to our knowledge, it appears that the work presented here is the first to propose a method to establish which four-body interactions are significant.

It should be pointed out that the calculation of $\varepsilon_{\mathrm{max}}$ can be extended to charged and non-polar species. We can describe the charge distribution of the bodies by the first non-zero multipole, i.e. the charge or the quadrupole. Consequently, the R_{AB}^{-3} coupling term is modified to a more general $R_{\mathrm{AB}}^{-(l_{\mathrm{A}}+l_{\mathrm{B}}+1)}$ dependence where the exponent is determined by the rank of the multipole, l.

Most importantly, the bodies are still coupled in a non-branching path regardless of the multipoles involved. This non-branching nature of many-body induction will be a recurring theme in this chapter.

5.3.2 *Identifying the Significant Many-body Effects*

To determine if ε_{max} can identify significant many-body interactions, we computed the ε_{max} for all possible three and four bodies in a series of $(H_2O)_{16\text{-}40}$ clusters (Figure 5.4). Water clusters are of interest as they contain strong many-body interactions,[25,29,32,213] originating from the large dipole of water. The wide range of conformations of the water molecules in water clusters also provide a thorough test. Note that there were five clusters being investigated for each cluster size. To prevent clutter, we only present the results for one of the five clusters in Figure 5.4a,b but the full dataset can be found in the Appendix.

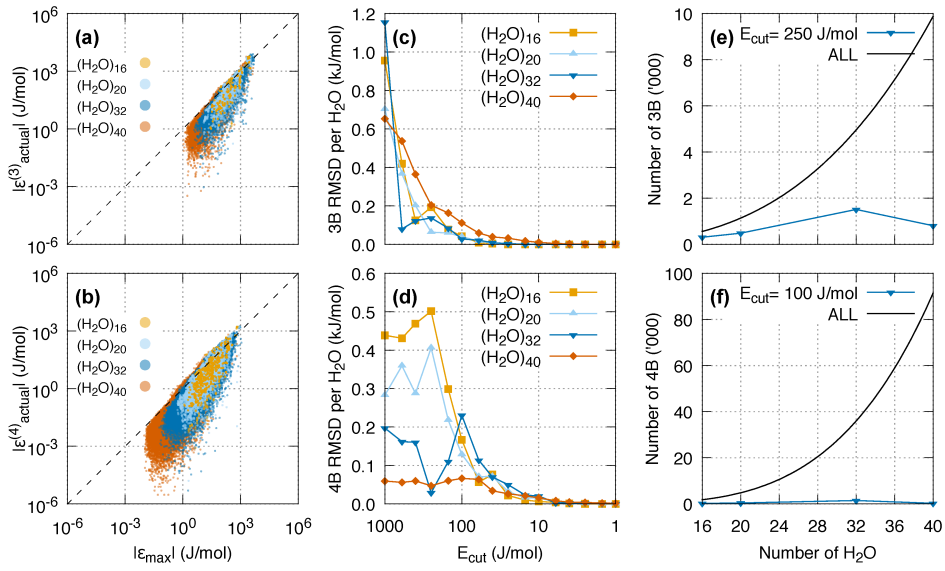


Figure 5.4: The ability of ε_{max} to identify significant many-body interactions is evaluated in water clusters. (a,b) The magnitude of the actual many-body interactions, $|\varepsilon_{actual}|$, were compared with that of the predicted maximum many-body interactions, $|\varepsilon_{max}|$. (c,d) Root mean square deviation (RMSD) per H_2O in reproducing the total many-body interactions using truncated values by including many-body interactions with $|\varepsilon_{max}| > E_{cut}$. The desired accuracy of 0.25 kJ mol^{-1} is achieved using E_{cut} of 250 J mol^{-1} and 100 J mol^{-1} at the three-body and four-body level respectively. (e,f) The total number of many-body interactions (ALL) and the number of interactions included at the above-mentioned E_{cut}.

In Figure 5.4a,b, the majority of the many-body interactions are bound by the $|\varepsilon_{\text{max}}|$ as the data points are found below the $y = x$ dashed line. However, a small number of interactions still exceed $|\varepsilon_{\text{max}}|$ slightly due to the neglect of short-range effects in our treatment. This is confirmed by the observation that the cases where $|\varepsilon_{\text{actual}}| > |\varepsilon_{\text{max}}|$ are frequently associated with large many-body interactions where the bodies are likely to be close to each other. Fortunately, this is not a problem as these cases are frequently associated with large many-body interactions, which are always deemed significant given a reasonable cutoff. Furthermore, we obtain a larger number of small many-body interactions with increasing cluster size, indicating that a lot of these interactions are insignificant and can be neglected. By summing the many-body interactions above a cutoff, E_{cut}, we can cheaply reproduce the total many-body interaction. Clearly, the value of E_{cut} depends on the desired accuracy. In dynamical simulations where many-body-based methods are employed, an acceptable error for the root mean square deviation (RMSD) per H_2O might be a small fraction, say 10%, of the thermal uncertainty, kT. This desired accuracy of 0.25 kJ mol^{-1} at room temperature is achieved at E_{cut} of 250 J mol^{-1} and 100 J mol^{-1} at the three-body and four-body level respectively (Figure 5.4c,d). Except for very large E_{cut}, the RMSD decreases rapidly with smaller E_{cut} regardless of the cluster size, suggesting that the above E_{cut} is applicable to larger water clusters. Only a tiny number of interactions are included at the aforementioned E_{cut}, unlike the drastic increase in the total possible number of many-body interactions (Figure 5.4e,f). This effect is more pronounced at the four-body level due to the faster decay of the the four-body interactions, which has an L^{-9} overall distance dependence as compared to the L^{-6} dependence in the three-body (and two-body) counterpart. More importantly, the number of many-body interactions included increases linearly with cluster size, making this approach a linear-scaling method and therefore highly amenable for large chemical systems.

We also applied the same analysis to different secondary structures of the

$H(C(O)NHCH_2)_{24}H$ polyglycine with each $C(O)NHCH_2$ repeating unit being a body (Figure 5.5). It is important to preserve the amide linkage within each body as it is responsible for the large dipole moment and polarizability in polyglycine. Again, we observe that the majority of $|\varepsilon_{\text{actual}}|$ are bound by $|\varepsilon_{\text{max}}|$ except for the tiny four-body interactions in the β-strands of $\approx 10^{-3}$ J mol^{-1} (Orange dots in Figure 5.5b). This is because these interactions are essentially zero as they are at the limits of our numerical precision. These results support our aim to include only the significant many-body effects. Otherwise, in even larger chemical systems, there will be an overwhelming number of tiny and numerically unstable interactions, which can affect the accuracy of many-body-based applications.[245] Chemical accuracy of 4.2 kJ mol^{-1} is achieved using E_{cut} of 25 J mol^{-1} and 10 J mol^{-1} at the three-body (Figure 5.5c) and four-body (Figure 5.5d) level respectively. Furthermore, the number of many-body interactions required still exhibits linear growth with the number of residues (Figure 5.5e,f). As a side observation, the polarizability of the $C(O)NHCH_2$ repeating unit decreases dramatically with the removal of a methylene group and vice versa. This implies that the many studies[246–248] using chains of formamide (which has one less methylene group than $C(O)NHCH_2$) may have underestimated the many-body effects in polypeptides although qualitative trends should still hold true. Conversely, it was shown that the presence of alkyl groups on the side chains of the amino acids favours helix formation.[249]

Due to the extended planar geometry of the peptide bond, there is significant anisotropy in the one-body *N*-methyl formamide. However, ε_{max} is still able to identify the significant many-body effects. This suggests that the use of an isotropic polarizability does not undermine the accuracy of ε_{max}. The success in the identification of significant many-body effects in both water clusters and polyglycine demonstrates the universality of the method. The water clusters provide a variety of conformations while our isotropic polarizability assumption is tested with the significant aniostropy in the peptide bonds.

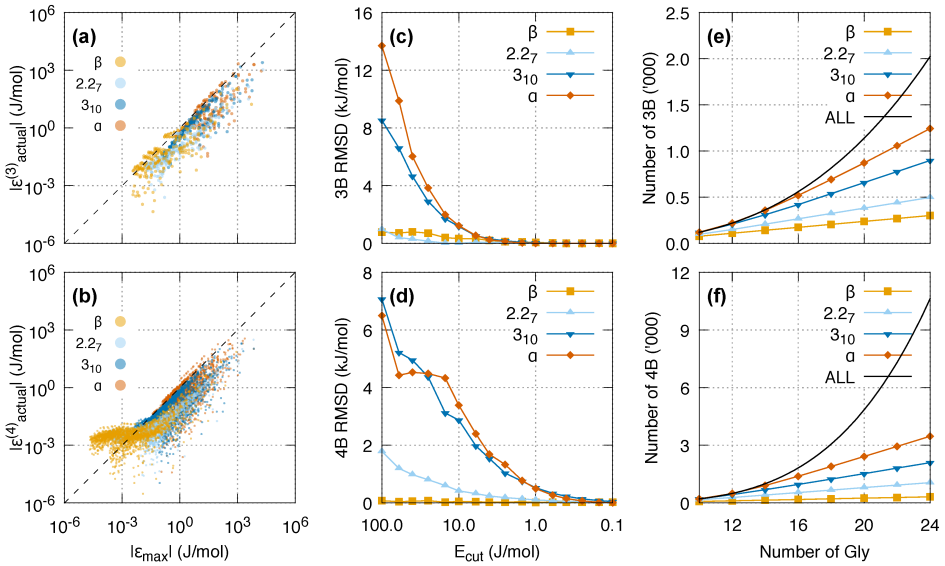


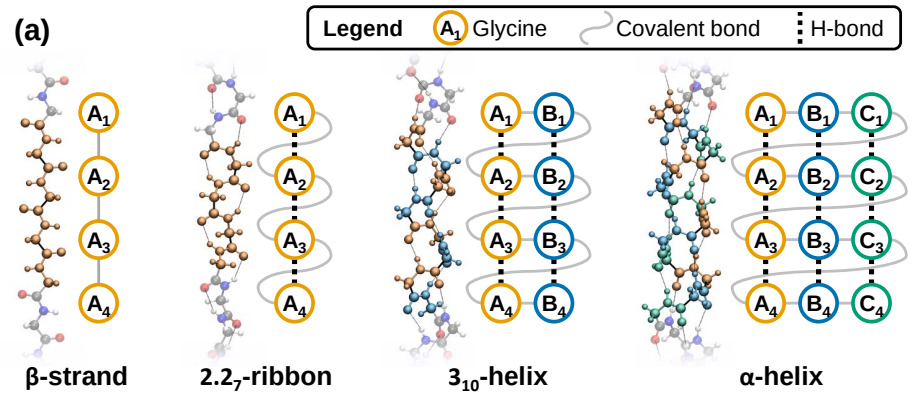
Figure 5.5: The ability of $\varepsilon_{\max}$ to identify significant many-body interactions is evaluated in secondary structure of polyglygine. (a,b) The $|\varepsilon_{\text{actual}}|$ were compared with the $|\varepsilon_{\max}|$ for the fully extended β-strand (β), 2.2_7-ribbon (2.2_7), 3_{10}-helix (3_{10}) and α-helix (α) structures. (c,d) RMSD in reproducing the total many-body interactions using truncated values by including many-body interactions with $|\varepsilon_{\max}| > E_{\text{cut}}$. Chemical accuracy of 4.2 kJ mol^{-1} is achieved using E_{cut}s of 25 J mol^{-1} and 10 J mol^{-1} at the three-body and four-body level respectively. (e,f) The total number of many-body interactions (ALL) and the number of interactions included at the above E_{cut}.

5.3.3 Many-body Effects in Helical Structures

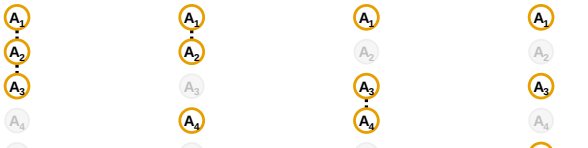
Amongst the different secondary structures, the helical structures display the most many-body effects. This is because the dipoles are aligned along the helical axis, reinforcing the many-body interactions. Consequently, the $|\varepsilon_{\text{actual}}|$ of the 3_{10}- and α-helix structures tend to be near the $|\varepsilon_{\max}|$ (Figure 5.5a,b). To determine the extent of the many-body effects, we focus on the interactions within the same and across different hydrogen-bonding chains (Figure 5.6a). The hydrogen-bonding chains were chosen as a basis since they are consistent with the way many-body induction manifests, i.e., in chains.

The $A_1A_2A_3$-interactions are the largest amongst the selected many-body interactions and it accounts for the preference of the α-helix over the 3_{10}-helix at moderate to long peptide lengths (Figure 5.6b). Note that the $A_1A_2A_3$-interaction is small in the case of the β-strand, echoing the earlier observation

(a)

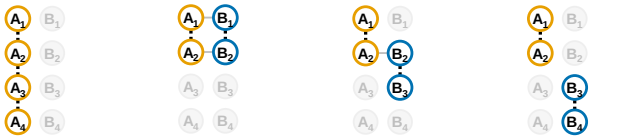


(b)



ε_{ABC} (kJ/mol)	$A_1A_2A_3$	$A_1A_2A_4$	$A_1A_3A_4$	$A_1A_3A_5$
β-strand	0.10	0.08	0.13	-0.07
2.2$_7$-ribbon	-0.93	-0.14	0.02	-0.13
3$_{10}$-helix	-1.63	-0.17	-0.07	-0.01
α-helix	-2.41	-0.26	-0.23	-0.02

(c)



ε_{ABCD} (kJ/mol)	$A_1A_2A_3A_4$	$A_1A_2B_1B_2$	$A_1A_2B_2B_3$	$A_1A_2B_3B_4$
3$_{10}$-helix	-0.11	-0.16	0.10	-0.07
α-helix	-0.21	0.32	0.03	-0.10

ε_{ABCD} (kJ/mol)	$A_1A_2A_3B_1$	$A_1A_2A_3B_2$	$A_1A_2A_3B_3$
3$_{10}$-helix	0.29	0.24	0.06
α-helix	0.25	0.22	0.10

Figure 5.6: The extent of many-body effects in helical stuctures were investigated. (a) Diagrammatic representation of the hydrogen-bonding chains in the different secondary structures of polyglycine. Each residue is labelled X_Y where X identifies the hydrogen-bonding chain that the residue belong to and Y gives the position within the chain. Comparison of selected (b) three-body and (c) four-body interactions in the different secondary structures of polyglycine.

that there is negligible many-body effects. When two of the bodies are separated apart by one residue, the three-body interaction decreases by an order of magnitude (Entry $A_1A_2A_4$ and $A_1A_3A_4$). A rough calculation on ε_{max} suggests similar results where one of the inter-body distances doubles and the R^{-3} in eq (5.3) leads to a decrease by a factor of $2^3 = 8$. Similarly, the three-body interaction decreases by another order of magnitude for the $A_1A_3A_5$-interaction. Comparing the three-body and four-body interactions, the inclusion of an additional body decreases the many-body interactions by an order of magnitude (Figure 5.6c). Thus, the five-body interactions can be safely ignored. We also observe that many-body effects across different hydrogen bonding chains are generally repulsive especially for the selected AAAB-type interactions. We postulate that this could be due to the body in the B-chain disrupting the enhancement of electric field and dipole moment of the other three bodies in the A-chain. Thus, the stability of the helix is driven by the three-body interactions within the same hydrogen bonding chain.

Furthermore, given the drastic decrease in many-body interactions with increasing separation of the residues, we can consider interactions spanning the first to fifth position on the hydrogen bonding chain to be significant. We can be more conservative and include borderline cases spanning the first to sixth position. Since the 3_{10}-helix has two hydrogen bonding chains while the α-helix has three, we would expect the many-body effects to asymptote by the 12th and 18th residue in the 3_{10}- and α-helices respectively. This is in good agreement with previous studies on polyglycine[230] (3_{10}: 14 residues, α: ≈ 20 residues) and studies on polyalanine[231] (3_{10}: 18 residues, α: > 20 residues). The slower asymptotic behaviour in the polyalanines is likely due to higher polarizability of alanine as compared to glycine which enhances the many-body induction effects. Interestingly, Hua *et al.* concluded that the cooperative effects have not reached their asymptotic limits by the 40th residue in both 3_{10}- and α-helices.[232] We attribute this to the many-body BSSE from the diffuse

functions in the 6-311++G** basis set used by the authors, which leads to poor convergence of the MBE.[91]

5.4 Summary

We developed a general and fast method to identify significant many-body effects. This is achieved by estimating the maximum many-body effects, $\varepsilon_{\max}$, that can arise in a given arrangement of bodies. Through careful analysis of the distance and orientational components of $\varepsilon_{\max}$, we find two overall causes for significant many-body interactions. Firstly, many-body induction propagates in non-branching paths, i.e, in a chain-like manner. Secondly, linear arrangements of bodies promote the alignment of the dipoles to reinforce the many-body interaction. Consequently, we identified that compact and extended linear arrangements are preferred to give significant many-body effects. Compact arrangements are favoured due to the presence of many short non-branching paths connecting the bodies. Extended linear arrangements are also preferred as they favour the alignment of dipoles. The latter result is not intuitive as these linear arrangements can extend over large distances but in a directional manner. For the first time, this study provides a rigorous explanation as to how cooperative effects provide enhanced stability in helicical structures.

We also tested the effectiveness of $\varepsilon_{\max}$ in identifying significant many-body effects. By including a small set of many-body interactions above a certain cutoff, we are able to cheaply reproduce the total three-body and four-body interaction energies. The number of many-body interactions included scales linearly with the number of bodies, making our method highly suitable in the study of large chemical systems. Furthermore, the method works for both water clusters and polyglcine secondary structures, demonstrating its universal nature. In conclusion, we have provided an explanation for the propagation of many-body effects, which we believe will have far-reaching impact in the study of large chemical systems, ranging from the condensed phase to large biomolecules.

6 | CONCLUDING REMARKS

Throughout this thesis, we proposed various methodologies with the aim of lowering the computational cost in the study of large water clusters and large chemical systems in general. Our tools of choice are the many-body expansion (MBE) and perturbation theory (PT). The study of basis set superposition effect (BSSE) in the MBE spanned the Chapters 3 and 4 while the use of PT to identify significant many-body effects was covered in Chapter 5.

At the beginning, our efforts were focussed on the refinement of existing fragmentation methods to cheaply reproduce the total energy of large water clusters. During this process, we stumbled upon a distressing observation that the MBE does not converge rapidly by the four-body term as we initially hypothesized. More alarmingly, the errors of truncating the MBE grew with system size. We later identified that this is due to the BSSE associated with diffuse functions, which is highly-many-body in nature. This was verified through the analysis of the molecular orbital coefficients where the contributions of diffuse ghost functions can be found simultaneously at many bodies. Furthermore, the poor convergence of the MBE disappears when the bodies are pulled apart, in agreement that BSSE decays rapidly with inter-boy distances. Thus, two obvious ways to restore the rapid convergence of the MBE is to omit the use of diffuse basis functions or employ a larger basis set that is more complete. Notably, the placement of a charge field, a common practice in fragmentation methods,

do not resolve this problem. This is not surprising as the charges are used to recover physical many-body effects and cannot compensate for BSSE. Furthermore, we found that the consistent use of the cluster basis in all the electronic structure calculations also removes the poor convergence of the MBE.

While we have identified that the diffuse basis functions are the cause for poor MBE convergence, there are still many unanswered questions regarding many-body BSSE. In the previous chapter, we found that the many-body contributions computed in the BSSE-prone nuclei-centred basis exhibit rapid MBE convergence in the absence of diffuse functions. Furthermore, the use of the subcluster basis led to rapid convergence albeit to an incorrect value, i.e., not the total energy. The latter observation suggests that the difference between the many-body contributions computed in the cluster basis and the subcluster basis has to be responsible for the poor MBE convergence. Within the framework of our proposed many-ghost many-body expansion (MGMBE), this difference would be called the basis set extension effects. The basis set extension effects correspond to the borrowing of basis functions outside of the subcluster (from the cluster basis) to improve the quality of the many-body interactions. They are necessary to recover the total energy using a many-body approach. Unfortunately, in the presence of diffuse functions, these extension effects in the one-body energy are highly many-body, extending up to numerous ghost-bodies simultaneously. This is the true cause of the poor convergence of the MBE. Notably, the extension effects only extend up to several bodies in the absence of diffuse functions. In contrast, the BSSE that most quantum chemists are familiar with, where there is an imbalance in the number of basis functions, is called the basis set imbalance error. Surprisingly, in the nuclei-centred basis, the basis set imbalance error cancels exactly the basis set extension effects, resulting in a rapid MBE convergence.

A major drawback of the MBE is the sheer number of many-body interactions to be considered. Intuitively, we expect a large majority of these

many-body contributions to be insignificant especially with increasing cluster size. While it is straightforward to determine whether a two-body interaction is significant based on the inter-body distance, a selection criterion is not obvious for higher number of bodies given the high dimensionality in the arrangements. Thus, we rigorously derived the leading three-body and four-body terms in many-body induction. These terms will serve as a parameter to estimate the maximum possible many-body effects in a given many-body arrangement. We used this new parameter to identify the significant many-body effects and successfully reproduced the total three-body and four-body interactions cheaply. More importantly, we found that many-body induction manifest in a chain-like manner. Consequently, the compact arrangements and extended linear arrangements are favoured to give significant many-body effects. The latter result is not intuitive as it implies that many-body effects can extend over large distances in a directional manner. This can be rationalized by linear arrangements favouring the strong alignment of dipole to reinforce induction and the arrangement of the bodies are consistent with the way many-body induction manifest.

By elucidating the behaviour of many-body BSSE, we provide a sound theoretical foundation for the proper implementation of the MBE in the study of large chemical systems. This is useful in assessing the accuracy of fragmentation methods as we can identify the errors associated with BSSE and errors associated with the method itself. The ability to rigorously identify the many-body interactions also provide a non-empirical approach to design new fragmentation methods. Furthermore, the computational cost of many other MBE-based applications can be drastically reduced. Although the studies presented in this thesis are focussed on water clusters, they are undeniably applicable to other large chemical systems, ranging from large biomolecules in biochemistry to other molecular clusters in condensed phase physics.

Some possible future work includes the use of high-rank multipolar electrostatics to accurately reproduce some of the many-body interactions where

the bodies are separated at an intermediate range. This will further reduce the number of many-body interactions that have to be treated quantum mechanically. Furthermore, we identified that the one-body energy should be computed in the cluster basis to recover all the basis set extension effects. However, the above calculation becomes computationally intractable with increasing cluster size due to the increase in the number of basis functions. Perhaps approximations can be developed to approximate the one-body energy in the cluster basis. Some criterion can be developed to identify the nearer ghost functions that would contribute more significantly to the extension effects.

With the findings in this thesis and the proposed future work, highly accurate quantum chemical calculations of large water clusters become possible, allowing us to better understand the mysteries behind this simple liquid.

BIBLIOGRAPHY

[1] Franks, F. *Water: A Matrix of Life*, 2nd ed.; Royal Society Of Chemistry: Cambridge, England, 2000.

[2] Tait, M. J.; Franks, F. Water in Biological Systems. *Nature* **1971**, *230*, 91–94.

[3] Chaplin, M. Do we Underestimate the Importance of Water in Cell Biology? *Nature Rev.* **2006**, *7*, 861–866.

[4] Ball, P. Water as an Active Constituent in Cell Biology. *Chem. Rev.* **2008**, *108*, 74–108.

[5] Cheung, M. S.; Garcia, A. E.; Onuchic, J. N. Protein Folding Mediated by Solvation: Water Expulsion and Formation of the Hydrophobic Core Occur after the Structural Collapse. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 685–690.

[6] Fuxreiter, M.; Mezei, M.; Simon, I.; Osman, R. Interfacial Water as a "Hydration Fingerprint" in the Noncognate Complex of *Bam*HI. *Biophys. J.* **2005**, *89*, 903–911.

[7] Tinetti, G.; Vidal-Madjar, A.; Liang, M.-C.; Beaulieu, J.-P.; Yung, Y.; Carey, S.; Barber, R. J.; Tennyson, J.; Ribas, I.; Allard, N.; Ballester, G. E.; Sing, D. K.; Selsis, F. Water Vapour in the Atmosphere of a Transiting Extrasolar Planet. *Nature* **2007**, *448*, 169–171.

[8] Hubbard, W. B. Neptune's Deep Chemistry. *Science* **1997**, *275*, 1279–1280.

[9] Cavazzoni, C.; Chiarotti, G. L.; Scandolo, S.; Tosatti, E.; Bernasconi, M.; Parrinello, M. Superionic and Metallic States of Water and Ammonia at Giant Planet Conditions. *Science* **1999**, *283*, 44–46.

[10] Jenniskens, P.; Blake, D. F. Structural Transitions in Amorphous Water Ice and Astrophysical Implications. *Science* **1994**, *265*, 753–756.

[11] Franks, F. In *Water: A Comprehensive Treatise Vol 1 Physics and Physical Chemistry of Water*; Franks, F., Ed.; Plenum Press: New York, USA, 1972; pp 1–20.

[12] Poole, P. H.; Sciortino, F.; Essmann, U.; Stanley, H. E. Phase Behaviour of Metastable Water. *Nature* **1992**, *360*, 324–328.

[13] Stanley, H. E.; Buldyrev, S. V.; Canpolat, M.; Havlin, S.; Mishima, O.; Sadr-Lahijany, M. R.; Scala, A.; Starr, F. The Puzzle of Liquid Water: A Very Complex Fluid. *Physica D* **1999**, *133*, 453–462.

[14] Prielmeier, F. X.; Lang, E. W.; Speedy, R. J.; Lüdemann, H.-D. Diffusion in Supercooled Water to 300 MPa. *Phys. Rev. Lett.* **1987**, *59*, 1128–1131.

[15] Smith, R. S.; Kay, B. D. The Existence of Supercooled Liquid Water at 150 K. *Nature* **1999**, *398*, 788–791.

[16] Liu, L.; Chen, S.-H.; Faraone, A.; Yen, C.-W.; Mou, C.-Y. Pressure Dependence of Fragile-to-Strong Transition and a Possible Second Critical Point in Supercooled Confined Water. *Phys. Rev. Lett.* **2005**, *95*, 117802.

[17] Salzmann, C. G.; Radaelli, P. G.; Hallbrucker, A.; Mayer, E.; Finney, J. L. The Preparation and Structures of Hydrogen Ordered Phases of Ice. *Science* **2006**, *311*, 1758–1761.

[18] Salzmann, C. G.; Radaelli, P. G.; Mayer, E.; Finney, J. L. Ice XV: A New Thermodynamically Stable Phase of Ice. *Phys. Rev. Lett.* **2009**, *103*, 105701.

[19] Ohmine, I.; Tanaka, H. Fluctuation, Relaxations, and Hydration in Liquid Water. Hydrogen-Bond Rearrangement Dynamics. *Chem. Rev.* **1993**, *93*, 2545–2566.

[20] Latimer, W. M.; Rodebush, W. H. Polarity and Ionization from the Standpoint of the Lewis Theory of Valence. *J. Am. Chem. Soc.* **1920**, *42*, 1419–1433.

[21] Smith, B. J.; Swanton, D. J.; Pople, J. A.; Schaefer III, H. F.; Radom, L. Transition Structures for the Interchange of Hydrogen Atoms within the Water Dimer. *J. Chem. Phys.* **1990**, *92*, 1240–1247.

[22] Tschumper, G. S.; Leininger, M. L.; Hoffman, B. C.; Valeev, E. F.; Schaefer III, H. F.; Quack, M. Anchoring the Water Dimer Potential Energy Surface with Explicitly Correlated Computations and Focal Point Analyses. *J. Chem. Phys.* **2002**, *116*, 690–701.

[23] Lane, J. R. CCSDTQ Optimized Geometry of Water Dimer. *J. Chem. Theory Comput.* **2013**, *9*, 316–323.

[24] Batista, E. R.; Xantheas, S. S.; Jónsson, H. Molecular multipole moments of water molecules in ice Ih. *J. Chem. Phys.* **1998**, *109*, 4546–4551.

[25] Guillot, B. A Reappraisal of What we have Learnt during Three Decades of Computer Simulations on Water. *J. Mol. Liq.* **2002**, *101*, 219–260.

[26] Bernal, J. D.; Fowler, R. H. A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions. *J. Chem. Phys.* **1933**, *1*, 515–548.

[27] Barker, J. A.; Watts, R. O. Structure of Water; A Monte Carlo Calculation. *Chem. Phys. Lett.* **1969**, *3*, 144–145.

[28] Rahman, A.; Stillinger, F. H. Molecular Dynamics Study of Liquid Water. *J. Chem. Phys.* **1971**, *55*, 3336–3359.

[29] Halgren, T. A.; Damm, W. Polarizable force fields. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236–242.

[30] Finney, J. L. The Water Molecule and its Interactions: the Interaction between Theory, Modelling, and Experiment. *J. Mol. Liq.* **2001**, *90*, 303–312.

[31] Szalewicz, K.; Leforestier, C.; van der Avoird, A. Towards the complete understanding of water by a first-principles computational approach. *Chem. Phys. Lett.* **2009**, *482*, 1–14.

[32] Demerdash, O.; Yap, E.-H.; Head-Gordon, T. Advanced Potential Energy Surfaces for Condensed Phase Simulation. *Annu. Rev. Phys. Chem.* **2014**, *65*, 149–174.

[33] Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

[34] Dunning, Jr., T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

[35] Popkie, H.; Kistenmacher, H.; Clementi., E. Study of the structure of molecular complexes. IV. The Hartree-Fock potential for the water dimer and its application to the liquid state. *J. Chem. Phys.* **1973**, *59*, 1325–1336.

[36] Jorgensen, W. L. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water. *J. Am. Chem. Soc.* **1981**, *103*, 335–340.

[37] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

[38] Rick, S. W.; Stuart, S. J.; Berne, B. J. Dynamical fluctuating charge force fields: Application to liquid water. *J. Chem. Phys.* **1994**, *101*, 6141–6156.

[39] Mahoney, M. W.; Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **2000**, *112*, 8910–8922.

[40] Mahoney, M. W.; Jorgensen, W. L. Quantum, intramolecular flexibility, and polarizability effects on the reproduction of the density anomaly of liquid water by simple potential functions. *J. Chem. Phys.* **2001**, *115*, 10758–10768.

[41] Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665–9678.

[42] Abascal, J. L. F.; Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123*, 234505.

[43] Habershon, S.; Markland, T. E.; Manolopoulos, D. E. Competing quantum effects in the dynamics of a flexible water model. *J. Chem. Phys.* **2009**, *131*, 024501.

[44] Berendsen, H. J. C.; Postma, J. P. M.; von Gunstaren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, Holland, 1981; pp 331–342.

[45] Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

[46] Wu, Y.; Tepper, H. L.; Voth, G. A. Flexible simple point-charge water model with improved liquid-state properties. *J. Chem. Phys.* **2006**, *124*, 024503.

[47] Paesani, F.; Zhang, W.; Case, D. A.; Cheatham III, T. E.; Voth, G. A. An accurate and simple quantum model for liquid water. *J. Chem. Phys.* **2006**, *125*, 184507.

[48] Stone, A. J. *The Theory of Intermolecular Forces*, 2nd ed.; Oxford University Press: Oxford, England, 2013.

[49] Millot, C.; Stone, A. J. Towards an accurate intermolecular potential for water. *Mol. Phys.* **1992**, *77*, 439–462.

[50] Millot, C.; Soetens, J.-C.; Martins Costa, M. T. C.; Hodges, M. P.; Stone, A. J. Revised Anisotropic Site Potentials for the Water Dimer and Calculated Properties. *J. Phys. Chem. A* **1998**, *102*, 754–770.

[51] Fellers, R. S.; Leforestier, C.; Braly, L. B.; Brown, M. G.; Saykally, R. J. Spectroscopic Determination of the Water Pair Potential. *Science* **1999**, *284*, 945–948.

[52] Goldman, N.; Fellers, R. S.; Brown, M. G.; Braly, L. B.; Keoshian, C. J.; Leforestier, C.; Saykally, R. J. Spectroscopic determination of the water dimer intermolecular potential-energy surface. *J. Chem. Phys.* **2002**, *116*, 10148–10163.

[53] Mas, E. M.; Szalewicz, K.; Bukowski, R.; Jeziorski, B. Pair potential for water from symmetry-adapted perturbation theory. *J. Chem. Phys.* **1997**, *107*, 4207–4218.

[54] Jeziorski, B.; Moszynski, R.; Szalewicz, K. Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes. *Chem. Rev.* **1994**, *94*, 1887–1930.

[55] Mas, E. M.; Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Wormer, P. E. S.; van der Avoird, A. Water Pair Potential of Near Spectroscopic Accuracy. I. Analysis of Potential Surface and Virial Coefficients. *J. Chem. Phys.* **2000**, *113*, 6687–6701.

[56] Groenenboom, G. C.; Wormer, P. E. S.; van der Avoird, A.; Mas, E. M.; Bukowski, R.; Szalewicz, K. Water pair potential of near spectroscopic accuracy. II. Vibration-rotation-tunneling levels of the water dimer. *J. Chem. Phys.* **2000**, *113*, 6702–6715.

[57] Mas, E. M.; Bukowski, R.; Szalewicz, K. Ab initio three-body interactions for water. I. Potential and structure of water trimer. *J. Chem. Phys.* **2003**, *118*, 4386–4403.

[58] Bukowski, R.; Szalewicz, K.; Groenenboom, G.; van der Avoird, A. Interaction potential for water dimer from symmetry-adapted perturbation theory based on density functional description of monomers. *J. Chem. Phys.* **2006**, *125*, 044301.

[59] Burnham, C. J.; Li, J.; Xantheas, S. S.; Leslie, M. The parametrization of a Thole-type all-atom polarizable water model from first principles and its application to the study of water clusters ($n = 2-21$) and the phonon spectrum of ice Ih. *J. Chem. Phys.* **1999**, *110*, 4566–4581.

[60] Thole, B. T. Molecular Polarizabilites Calculated With a Modified Dipole Interaction. *Chem. Phys.* **1981**, *59*, 341–350.

[61] Xantheas, S. S.; Burnham, C. J.; Harrison, R. J. Development of transferable interaction models for water. II. Accurate energetics of the first few water clusters from first principles. *J. Chem. Phys.* **2002**, *116*, 1493–1499.

[62] Partridge, H.; Schwenke, D. W. The determination of an accurate isotope dependent potential energy surface for water from extensive ab initio calculations and experimental data. *J. Chem. Phys.* **1997**, *106*, 4618–4639.

[63] Burnham, C. J.; Xantheas, S. S. Development of transferable interaction models for water. III. Reparametrization of an all-atom polarizable rigid model TTM2-R from first principles. *J. Chem. Phys.* **2002**, *116*, 1500–1510.

[64] Fanourgakis, G. S.; Xantheas, S. S. The Flexible, Polarizable, Thole-Type Interaction Potential for Water (TTM2-F) Revisited. *J. Phys. Chem. A* **2006**, *110*, 4100–4106.

[65] Fanourgakis, G. S.; Xantheas, S. S. Development of transferable interaction potentials for water. V. Extension of the flexible, polarizable, Thole-type model potential (TTM3-F, v. 3.0) to describe the vibrational spectra of water clusters and liquid water. *J. Chem. Phys.* **2008**, *128*, 074506.

[66] Burnham, C. J.; Anick, D. J.; Mankoo, P. K.; Reiter, G. F. The vibrational proton potential in bulk liquid water and ice. *J. Chem. Phys.* **2008**, *128*, 154519.

[67] Ren, P.; Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

[68] Huang, X.; Braams, B. J.; Bowman, J. M. Ab Initio Potential Energy and Dipole Moment Surfaces of $(H_2O)_2$. *J. Phys. Chem. A* **2006**, *110*, 445–451.

[69] Huang, X.; Braams, B. J.; Bowman, J. M.; Kelly, R. E. A.; Tennyson, J.; Groenenboom, G. C.; van der Avoird, A. New ab initio potential energy surface and the vibration-rotation-tunneling levels of $(H_2O)_2$ and $(D_2O)_2$. *J. Chem. Phys.* **2008**, *128*, 034312.

[70] Shank, A.; Wang, Y.; Kaledin, A.; Braams, B. J.; Bowman, J. M. Accurate ab initio and "hybrid" potential energy surfaces, intramolecular vibrational energies, and classical ir spectrum of the water dimer. *J. Chem. Phys.* **2009**, *130*, 144314.

[71] Wang, Y.; Huang, X.; Shepler, B. C.; Braams, B. J.; Bowman, J. M. Flexible, ab initio potential, and dipole moment surfaces for water. I. Tests and applications for clusters up to the 22-mer. *J. Chem. Phys.* **2011**, *134*, 094509.

[72] Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. Polarizable interaction potential for water from coupled cluster calculations. I. Analysis of dimer potential energy surface. *J. Chem. Phys.* **2008**, *128*, 094313.

[73] Cencek, W.; Szalewicz, K.; Leforestier, C.; van Harrevelt, R.; Szalewicz, A. v. K.; Murdachaew, G.; Bukowski, R.; Akin-Ojo, O.; Leforestier, C. An accurate analytic representation of the water pair potential. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4716–4731.

[74] Leforestier, C.; Szalewicz, K.; van der Avoird, A. Spectra of water dimer from a new ab initio potential with flexible monomers. *J. Chem. Phys.* **2012**, *137*, 014305.

[75] Szalewicz, K.; Murdachaew, G.; Bukowski, R.; Akin-Ojo, O.; Leforestier, C. In *Lecture Series on Computer and Computational Science: ICCMSE 2006*; Maroulis, G., Simos, T., Eds.; Brill Academic: Leiden, Holland, 2006; pp 482–491.

[76] Góra, U.; Cencek, W.; Podeszwa, R.; van der Avoird, A.; Szalewicz, K. Predictions for water clusters from a first-principles two- and three-body force field. *J. Chem. Phys.* **2014**, *140*, 194101.

[77] Medders, G. R.; Babin, V.; Paesani, F. A Critical Assessment of Two-Body and Three-Body Interactions in Water. *J. Chem. Theory Comput.* **2013**, *9*, 1103–1114.

[78] Babin, V.; Leforestier, C.; Paesani, F. Development of a "First Principles" Water Potential with Flexible Monomers: Dimer Potential Energy Surface, VRT Spectrum, and Second Virial Coefficient. *J. Chem. Theory Comput.* **2013**, *9*, 5395–5403.

[79] Babin, V.; Medders, G. R.; Paesani, F. Development of a "First Principles" Water Potential with Flexible Monomers. II: Trimer Potential Energy Surface, Third Virial Coefficient, and Small Clusters. *J. Chem. Theory Comput.* **2014**, *10*, 1599–1607.

[80] Babin, V.; Medders, G. R.; Paesani, F. Toward a Universal Water Model: First Principles Simulations from the Dimer to the Liquid Phase. *J. Phys. Chem. Lett.* **2012**, *3*, 3765–3769.

[81] Medders, G. R.; Babin, V.; Paesani, F. Development of a "First-Principles" Water Potential with Flexible Monomers. III. Liquid Phase Properties. *J. Chem. Theory Comput.* **2014**, *10*, 2906–2910.

[82] Vega, C.; Abascal, J. L. F. Simulating water with rigid non-polarizable models: a general perspective. *Phys. Chem. Chem. Phys.* **2011**, *13*, 19663–19688.

[83] Lipkowitz, K. B.; Boyd, D. B.; Rick, S. W.; Stuart, S. J. Potentials and Algorithms for Incorporating Polarizability in Computer Simulations. *Rev. Comp. Chem.* **2002**, *18*, 89–146.

[84] Ponder, J. W.; Case, D. A. Force Fields For Protein Simulations. *Adv. Prot. Chem.* **2003**, *66*, 27–85.

[85] Lopes, P. E. M.; Roux, B.; MacKerell, Jr., A. D. Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability: theory and applications. *Theor. Chem. Acc.* **2009**, *124*, 11–28.

[86] Lamoureux, G.; MacKerell, Jr., A. D.; Roux, B. A simple polarizable model of water based on classical Drude oscillators. *J. Chem. Phys.* **2003**, *119*, 5185–5197.

[87] Cardamone, S.; Hughes, T. J.; Popelier, P. L. A. Multipolar electrostatics. *Phys. Chem. Chem. Phys.* **2014**, *16*, 10367–10387.

[88] Applequist, J.; Carl, J. R.; Fung, K.-K. Atom dipole interaction model for molecular polarizability. Application to polyatomic molecules and determination of atom polarizabilities. *J. Am. Chem. Soc.* **1972**, *94*, 2952–2960.

[89] Tang, K. T.; Toennies, J. P. An improved simple model for the van der Waals potential based on universal damping functions for the dispersion coefficients. *J. Chem. Phys.* **1984**, *80*, 3726–3741.

[90] Ischtwan, J.; Collins, M. A. Molecular potential energy surfaces by interpolation. *J. Chem. Phys.* **1994**, *100*, 8080–8088.

[91] Ouyang, J. F.; Cvitkovic, M. W.; Bettens, R. P. A. Trouble with the Many-Body Expansion. *J. Chem. Theory Comput.* **2014**, *10*, 3699–3707.

[92] McBride, C.; Vega, C.; Noya, E. G.; Ramírez, R.; Sesé, L. M. Quantum contributions in the ice phases: The path to a new empirical model for water–TIP4PQ/2005. *J. Chem. Phys.* **2009**, *131*, 024506.

[93] Noya, E. G.; Vega, C.; Sesé, L. M.; Ramírez, R. Quantum effects on the maximum in density of water as described by the TIP4PQ/2005 model. *J. Chem. Phys.* **2009**, *131*, 124518.

[94] Conde, M. M.; Vega, C.; McBride, C.; Noya, E. G.; Ramírez, R.; Sesé, L. M. Can gas hydrate structures be described using classical simulations? *J. Chem. Phys.* **2010**, *132*, 114503.

[95] Agmon, N. The Grotthuss mechanism. *Chem. Phys. Lett.* **1995**, *244*, 456–462.

[96] Knight, C.; Voth, G. A. The Curious Case of the Hydrated Proton. *Acc. Chem. Res.* **2012**, *45*, 101–109.

[97] Shinoda, W.; Shiga, M. Quantum simulation of the heat capacity of water. *Phys. Rev. E* **2005**, *71*, 041204.

[98] Shiga, M.; Shinoda, W. Calculation of heat capacities of light and heavy water by path-integral molecular dynamics. *J. Chem. Phys.* **2005**, *123*, 134502.

[99] Paesani, F.; Voth, G. A. The Properties of Water: Insights from Quantum Simulations. *J. Phys. Chem. B* **2009**, *113*, 5702–5719.

[100] Tuckerman, M. E.; Berne, B. J.; Martyna, G. J.; Klein, M. L. Efficient molecular dynamics and hybrid Monte Carlo algorithms for path integrals. *J. Chem. Phys.* **1993**, *99*, 2796–2808.

[101] Marx, D.; Parrinello, M. Ab initio path integral molecular dynamics: Basic ideas. *J. Chem. Phys.* **1995**, *104*, 4077–4082.

[102] Shiga, M.; Tachikawa, M.; Miura, S. A unified scheme for ab initio molecular orbital theory and path integral molecular dynamics. *J. Chem. Phys.* **2001**, *115*, 9149–9159.

[103] Laasonen, K.; Sprik, M.; Parrinello, M.; Car, R. "Ab initio" liquid water. *J. Chem. Phys.* **1993**, *99*, 9080–9089.

[104] Planck, M. On the Theory of Energy Distribution Law of the Normal Spectrum Radiation. *Verhandl. Dtsch. Phys. Ges.* **1900**, *2*, 237–245.

[105] Einstein, A. Concerning an Heuristic Point of View Toward the Emission and Transformation of Light. *Ann. d. Physik* **1905**, *17*, 132–148.

[106] Heisenberg, W. Quantum theoretical re-interpretation of kinematic and mechanical relations. *Z. Physik* **1925**, *33*, 879–893.

[107] Schröndinger, E. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Phy. Rev.* **1926**, *3*, 104–1070.

[108] Dirac, P. A. M. A new notation for quantum mechanics. *Math. Proc. Cambridge Philos. Soc.* **1939**, *35*, 416–418.

[109] Heitler, W.; London, F. Wechselwirkung neutraler Atome und homöopolare Bindung nach der Quantenmechanik. *Z. Physik* **1927**, *44*, 455–472.

[110] Dirac, P. A. M. Quantum Mechanics of Many-Electron Systems. *Proc. R. Soc. A* **1929**, *123*, 714–733.

[111] Pople, J. A.; Krishnan, R.; Schlegel, H. B.; Binkley, J. S. Derivative studies in hartree–fock and møller–plesset theories. *Int. J. Quantum Chem.* **1979**, *16*, 225–241.

[112] Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56*, 2257–2261.

[113] Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **1980**, *72*, 650–654.

[114] Boys, S. F. Electronic Wave Functions. I. A General Method of Calculation for the Stationary States of Any Molecular System. *Proc. R. Soc. A* **1950**, *200*, 542–554.

[115] Stone, J. E.; Hardya, D. J.; Ufimtsev, I. S.; Schulten, K. GPU-accelerated molecular modeling coming of age. *J. Mol. Graphics* **2010**, *29*, 116–125.

[116] Gertrudes, J.; Maltarollo, V.; Silva, R.; Oliveira, P.; Honório, K.; da Silva, A. B. F. Machine Learning Techniques and Drug Design. *Curr. Med. Chem.* **2012**, *19*, 4289–4297.

[117] Feynman, R. P. *Six Easy Pieces: Essentials of Physics Explained by Its Most Brilliant Teacher*, 1st ed.; Perseus Books: New York, USA, 1995.

[118] Hartree, D. R. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Math. Proc. Cambridge Philos. Soc.* **1928**, *24*, 89–110.

[119] Fock, V. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Z. Physik* **1930**, *61*, 126–148.

[120] Slater, J. C. The Theory of Complex Spectra. *Phys. Rev.* **1929**, *34*, 1293–1322.

[121] Roothaan, C. C. J. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951**, *23*, 69–89.

[122] Hall, G. G. The Molecular Orbital Theory of Chemical Valency. VIII. A Method of Calculating Ionization Potentials. *Proc. Roy. Soc. A* **1951**, *205*, 541–552.

[123] Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.

[124] Kendall, R. A.; Dunning, Jr., T. H.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

[125] Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618–622.

[126] Raghavachari, K.; Pople, J. A.; Replogle, E. S.; Head-Gordon, M. Fifth-Order Møller–Plesset Perturbation Theory: Comparison of Existing Correiatlon Methods and Implementation of New Methods Correct to Fifth Order. *J. Phys. Chem.* **1990**, *94*, 5579–5586.

[127] Cizek, J. On the Use of the Cluster Expansion and the Technique of Diagrams in Calculations of Correlation Effects in Atoms and Molecules. *Adv. Chem. Phys.* **1966**, *14*, 35–89.

[128] Cizek, J. On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods. *J. Chem. Phys.* **1966**, *45*, 4256–4266.

[129] Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; John Wiley and Sons, Ltd: West Sussex, England, 2006.

[130] Kato, T. On the eigenfunctions of many-particle systems in quantum mechanics. *Comm. Pure Appl. Math.* **1957**, *10*, 151–177.

[131] Noga, J.; Kutzelnigg, W.; Klopper, W. CC-R12, a correlation cusp corrected coupled-cluster method with a pilot application to the Be_2 potential curve. *Chem. Phys. Lett.* **1992**, *199*, 497–504.

[132] Tew, D. P.; Klopper, W.; Helgaker, T. Electron correlation: The many-body problem at the heart of chemistry. *J. Comput. Chem.* **2007**, *28*, 1307–1320.

[133] Schwartz, C. Importance of Angular Correlations between Atomic Electrons. *Phys. Rev.* **1962**, *126*, 1015–1019.

[134] Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-set convergence of correlated calculations on water. *J. Chem. Phys.* **1997**, *106*, 9639–9646.

[135] Kutzelnigg, W.; Klopper, W. Wave functions with terms linear in the interelectronic coordinates to take care of the correlation cusp. I. General theory. *J. Chem. Phys.* **1991**, *94*, 1985–2001.

[136] Ten-no, S. Initiation of explicitly correlated Slater-type geminal theory. *Chem. Phys. Lett.* **2004**, *398*, 56–61.

[137] Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.

[138] Del Bene, J. E.; Person, W. B.; Szczepaniak, K. Properties of Hydrogen-Bonded Complexes Obtained from the B3LYP Functional with 6-31G(d,p) and 6-31+G(d,p) Basis Sets: Comparison with MP2/6-31+G(d,p) Results and Experimental Data. *J. Phys. Chem.* **1995**, *99*, 10705–10707.

[139] Tsuzuki, S.; Lüthi, H. P. Interaction energies of van der Waals and hydrogen bonded systems calculated using density functional theory: Assessing the PW91 model. *J. Chem. Phys.* **2001**, *114*, 3949–3957.

[140] Anderson, J. A.; Tschumper, G. S. Characterizing the Potential Energy Surface of the Water Dimer with DFT: Failures of Some Popular Functionals for Hydrogen Bonding. *J. Phys. Chem. A* **2006**, *110*, 7268–7271.

[141] Strømsheim, M. D.; Kumar, N.; Coriani, S.; Sagvolden, E.; Teale, A. M.; Helgaker, T. Dispersion interactions in density-functional theory: An adiabatic-connection analysis. *J. Chem. Phys.* **2011**, *135*, 194109.

[142] Riley, K. E.; Pitonak, M.; Jurecka, P.; Hobza, P. Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories. *Chem. Rev.* **2010**, *110*, 5023–5063.

[143] Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. Dispersion-Corrected Mean-Field Electronic Structure Methods. *Chem. Rev.* **2016**, *116*, 5105–5154.

[144] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

[145] Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

[146] Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phy. Rev. Lett.* **2009**, *102*, 073005.

[147] Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theo. Chem. Acc.* **2008**, *120*, 215–241.

[148] Grimme, S. Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.* **2006**, *124*, 034108.

[149] Schwabe, T.; Grimme, S. Towards chemical accuracy for the thermodynamics of large molecules: new hybrid density functionals including non-local correlation effects. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398–4401.

[150] Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. Van der Waals Density Functional for General Geometries. *Phy. Rev. Lett.* **2004**, *92*, 264401.

[151] Dahlke, E. E.; Truhlar, D. G. Assessment of the Pairwise Additive Approximation and Evaluation of Many-Body Terms for Water Clusters. *J. Phys. Chem. B* **2006**, *110*, 10595–10601.

[152] Dahlke, E. E.; Truhlar, D. G. Electrostatically Embedded Many-Body Expansion for Large Systems, with Applications to Water Clusters. *J. Chem. Theory Comput.* **2007**, *3*, 46–53.

[153] Dahlke, E. E.; Truhlar, D. G. Electrostatically Embedded Many-Body Correlation Energy, with Applications to the Calculation of Accurate Second-Order Møller-Plesset Perturbation Theory Energies for Large Water Clusters. *J. Chem. Theory Comput.* **2007**, *3*, 1342–1348.

[154] Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. Evaluation of the Electrostatically Embedded Many-Body Expansion and the Electrostatically Embedded Many-Body Expansion of the Correlation Energy by Application to Low-Lying Water Hexamers. *J. Chem. Theory Comput.* **2008**, *4*, 33–41.

[155] Fedorov, D. G.; Kitaura, K. Extending the Power of Quantum Chemistry to Large Systems with the Fragment Molecular Orbital Method. *J. Phys. Chem. A* **2007**, *111*, 6904–6914.

[156] Addicoat, M. A.; Collins, M. A. Accurate treatment of nonbonded interactions within systematic molecular fragmentation. *J. Chem. Phys.* **2009**, *131*, 104103.

[157] Řezáč, J.; Salahub, D. R. Multilevel Fragment-Based Approach (MFBA): A Novel Hybrid Computational Method for the Study of Large Molecules. *J. Chem. Theory Comput.* **2009**, *6*, 91–99.

[158] Weiss, S. N.; Huang, L.; Massa, L. A Generalized Higher Order Kernel Energy. *J. Comput. Chem.* **2010**, *31*, 2889.

[159] Beran, G. J. O.; Nanda, K. Predicting Organic Crystal Lattice Energies with Chemical Accuracy. *J. Phys. Chem. Lett.* **2010**, *1*, 3480–3487.

[160] Wang, X.; Liu, J.; Zhang, J. Z. H.; He, X. Electrostatically Embedded Generalized Molecular Fractionation with Conjugate Caps Method for Full Quantum Mechanical Calculation of Protein Energy. *J. Phys. Chem. A* **2013**, *117*, 7149–7161.

[161] Le, H.-A.; Tan, H.-J.; Ouyang, J. F.; Bettens, R. P. A. Combined Fragmentation Method: A Simple Method for Fragmentation of Large Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 469–478.

[162] Tan, H.-J.; Bettens, R. P. A. Ab initio NMR chemical-shift calculations based on the combined fragmentation method. *Phys.Chem. Chem. Phys.* **2013**, *15*, 7541–7547.

[163] Elrod, M. J.; Saykally, R. J. Many-Body Effects in Intermolecular Forces. *Chem. Rev.* **1994**, *94*, 1975–1997.

[164] Hermansson, K. Many-body effects in tetrahedral water clusters. *J. Chem. Phys.* **1988**, *89*, 2149–2159.

[165] Xantheas, S. S. Significance of higher-order many-body interaction energy terms in water clusters and bulk water. *Philos. Mag. B* **1996**, *73*, 107–115.

[166] Chen, W.; Gordon, M. S. Energy decomposition analyses for many-body interaction and applications to water complexes. *J. Phys. Chem.* **1996**, *100*, 14316–14328.

[167] Hodges, M. P.; Stone, A. J.; Xantheas, S. S. Contribution of Many-Body Terms to the Energy for Small Water Clusters: A Comparison of ab Initio Calculations and Accurate Model Potentials. *J. Phys. Chem. A* **1997**, *101*, 9163–9168.

[168] Xantheas, S. S. Cooperativity and hydrogen bonding network in water clusters. *Chem. Phys.* **2000**, *258*, 225–231.

[169] Kulkarni, A. D.; Ganesh, V.; Gadre, S. R. Many-body interaction analysis: Algorithm development and application to large molecular clusters. *J. Chem. Phys.* **2004**, *121*, 5043–5050.

[170] Xantheas, S. S. Interaction Potentials for Water from Accurate Cluster Calculations. *Struct. Bond* **2005**, *116*, 119–148.

[171] Xantheas, S. S. Ab initio studies of cyclic water clusters $(H_2O)_n$, $n = 1 - 6$. II. Analysis of many-body interactions. *J. Chem. Phys.* **1994**, *100*, 7523–7534.

[172] Christie, R. A.; Jordan, K. D. n-Body Decomposition Approach to the Calculation of Interaction Energies of Water Clusters. *Struct. Bond* **2005**, *116*, 27–41.

[173] Góra, U.; Podeszwa, R.; Cencek, W.; Szalewicz, K. Interaction energies of large clusters from many-body expansion. *J. Chem. Phys.* **2011**, *135*, 224102.

[174] Kumar, R.; Wang, F.-F.; Jenness, G. R.; Jordan, K. D. A second generation distributed point polarizable water model. *J. Chem. Phys.* **2010**, *132*, 014309.

[175] Kumar, R.; Keyes, T. The polarizing forces of water. *Theor. Chem. Acc.* **2012**, *131*, 1197.

[176] Yoo, S.; Aprá, E.; Zeng, X. C.; Xantheas, S. S. High-Level Ab Initio Electronic Structure Calculations of Water Clusters $(H_2O)_{16}$ and $(H_2O)_{17}$: A New Global Minimum for $(H_2O)_{16}$. *J. Phys. Chem. Lett.* **2010**, *1*, 3122–3127.

[177] Pruitt, S. R.; Addicoat, M. A.; Collins, M. A.; Gordon, M. S. The fragment molecular orbital and systematic molecular fragmentation methods applied to water clusters. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7752–7764.

[178] Hermann, A.; Krawczyk, R. P.; Lein, M.; Schwerdtfeger, P. Convergence of the many-body expansion of interaction potentials: From van der Waals to covalent and metallic systems. *Phys. Rev. A* **2007**, *76*, 013202.

[179] Cui, J.; Liu, H.; Jordan, K. D. Theoretical Characterization of the $(H_2O)_{21}$ Cluster: Application of an *n*-body Decomposition Procedure. *J. Phys. Chem. B* **2006**, *110*, 18872–18878.

[180] Kaplan, I. G.; Santamaria, R.; Novaro, O. Non-additive forces in atomic clusters: The case of Ag. *Mol. Phys.* **1995**, *84*, 105–114.

[181] Boys, S. F.; Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **1970**, *19*, 553–566.

[182] Valiron, P.; Mayer, I. Hierarchy of counterpoise corrections for N-body clusters: generalization of the Boys-Bernardi scheme. *Chem. Phys. Lett.* **1997**, *275*, 46–55.

[183] Milet, A.; Moszynski, R.; Wormer, P. E. S.; van der Avoird, A. Hydrogen Bonding in Water Clusters: Pair and Many-Body Interactions from Symmetry-Adapted Perturbation Theory. *J. Phys. Chem. A* **1999**, *103*, 6811–6819.

[184] Frisch, M. J. et al. *Gaussian 09, Revision C.01*; Gaussian, Inc.: Wallingford, CT, 2009.

[185] Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; M., *MOLPRO, version 2012.1*; http://www.molpro.net: Cardiff, U.K., 2012.

[186] Maheshwary, S.; Patel, N.; Sathyamurthy, N.; Kulkarni, A. D.; Gadre, S. R. Structure and Stability of Water Clusters $(H_2O)_n$, $n = 8-20$: An Ab Initio Investigation. *J. Phys. Chem. A* **2001**, *105*, 10525–10537.

[187] Richard, R. M.; Herbert, J. M. A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. *J. Chem. Phys.* **2012**, *137*, 064113.

[188] Papajak, E.; Leverentz, H. R.; Zheng, J.; Truhlar, D. G. Efficient Diffuse Basis Sets: cc-pVxZ+ and maug-cc-pVxZ. *J. Chem. Theory Comput.* **2009**, *5*, 1197–1202.

[189] Papajak, E.; Truhlar, D. G. Convergent Partially Augmented Basis Sets for Post-Hartree-Fock Calculations of Molecular Properties and Reaction Barrier Heights. *J. Chem. Theory Comput.* **2011**, *7*, 10–18.

[190] Papajak, E.; Zheng, J.; Xu, X.; Leverentz, H. R.; Truhlar, D. G. Perspectives on Basis Sets Beautiful: Seasonal Plantings of Diffuse Basis Functions. *J. Chem. Theory Comput.* **2011**, *7*, 3027–3034.

[191] Halkier, A.; Koch, H.; Jørgensen, P.; Christiansen, O.; Nielsen, I. M. B.; Helgaker, T. A systematic ab initio study of the water dimer in hierarchies of basis sets and correlation models. *Theor. Chem. Acc.* **1997**, *97*, 150–157.

[192] Beran, G. J. O. Approximating quantum many-body intermolecular interactions in molecular clusters using classical polarizable force fields. *J. Chem. Phys.* **2009**, *130*, 164115.

[193] Le, H.-A.; Lee, A. M.; Bettens, R. P. A. Accurately Reproducing Ab Initio Electrostatic Potentials with Multipoles and Fragmentation. *J. Phys. Chem. A* **2009**, *113*, 10527–10533.

[194] Stone, A. J. Distributed Multipole Analysis: Stability for Large Basis Sets. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.

[195] Bates, D. M.; Smith, J. R.; Janowski, T.; Tschumper, G. S. Development of a 3-body: many-body integrated fragmentation method for weakly bound clusters and application to water clusters $(H_2O)_{n=3-10,16,17}$. *J. Chem. Phys.* **2011**, *135*, 044123.

[196] Richard, R. M.; Lao, K. U.; Herbert, J. M. Achieving the CCSD(T) Basis-Set Limit in Sizable Molecular Clusters: Counterpoise Corrections for the Many-Body Expansion. *J. Phys. Chem. Lett.* **2013**, *4*, 2674–2680.

[197] Mata, R.; Stoll, H. Incremental expansions for SCF interaction energies: A comparison for hydrogen-bonded clusters. *Chem. Phys. Lett.* **2008**, *465*, 136–141.

[198] Hua, D.; Leverentz, H. R.; Amin, E. A.; Truhlar, D. G. Assessment and Validation of the Electrostatically Embedded Many-Body Expansion for Metal-Ligand Bonding. *J. Chem. Theory Comput.* **2011**, *7*, 251–255.

[199] Kurbanov, E. K.; Leverentz, H. R.; Truhlar, D. G.; Amin, E. A. Electrostatically Embedded Many-Body Expansion for Neutral and Charged Metalloenzyme Model Systems. *J. Chem. Theory Comput.* **2012**, *8*, 1–5.

[200] Kurbanov, E. K.; Leverentz, H. R.; Truhlar, D. G.; Amin, E. A. Analysis of the Errors in the Electrostatically Embedded Many-Body Expansion of the Energy and the Correlation Energy for Zn and Cd Coordination Complexes with Five and Six Ligands and Use of the Analysis to Develop a Generally Successful Fragmentation Strategy. *J. Chem. Theory Comput.* **2013**, *9*, 2617–2628.

[201] Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.* **2012**, *112*, 632–672.

[202] Collins, M. A.; Bettens, R. P. A. Energy-Based Molecular Fragmentation Methods. *Chem. Rev.* **2015**, *115*, 5607–5642.

[203] Raghavachari, K.; Saha, A. Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules. *Chem. Rev.* **2015**, *115*, 5643–5677.

[204] Liu, B.; McLean, A. D. Accurate calculation of the attractive interaction of two ground state helium atoms. *J. Chem. Phys.* **1973**, *59*, 4557–4558.

[205] Wells, B. H.; Wilson, S. van der waals Interaction Potentials: Many-body Basis Set Superposition Effects. *Chem. Phys. Lett.* **1983**, *101*, 429–434.

[206] Richard, R. M.; Lao, K. U.; Herbert, J. M. Approaching the complete-basis limit with a truncated many-body expansion. *J. Chem. Phys.* **2013**, *139*, 224102.

[207] Turi, L.; Dannenberg, J. J. Correcting for Basis Set Superposition Error in Aggregates Containing More Than Two Molecules: Ambiguities in the Calculation of the Counterpoise Correction. *J. Phys. Chem.* **1993**, *97*, 2488–2490.

[208] Werner, H.-J.; Adler, T. B.; Manby, F. R. General orbital invariant MP2-F12 theory. *J. Chem. Phys.* **2007**, *126*, 164102.

[209] Lendvay, G.; Mayer, I. Some difficulties in computing BSSE-corrected potential surfaces of chemical reactions. *Chem. Phys. Lett.* **1998**, *297*, 365–373.

[210] Salvador, P.; Szczęśniak, M. M. Counterpoise-corrected geometries and harmonic frequencies of N-body clusters: Application to $(HF)_n$ ($n = 3, 4$). *J. Chem. Phys.* **2003**, *118*, 537–549.

[211] Kamiya, M.; Hirata, S.; Valiev, M. Fast electron correlation methods for molecular clusters without basis set superposition errors. *J. Chem. Phys.* **2008**, *128*, 074103.

[212] Williams, H. L.; Mas, E. M.; Szalewicz, K.; Jeziorski, B. On the effectiveness of monomer-, dimer-, and bond-centered basis functions in calculations of intermolecular interaction energies. *J. Chem. Phys.* **1995**, *103*, 7374–7391.

[213] Ouyang, J. F.; Bettens, R. P. A. Modelling Water: A Lifetime Enigma. *Chimia* **2015**, *69*, 104–111.

[214] Halkier, A.; Klopper, W.; Helgaker, T.; Jørgensen, P.; Taylor, P. R. Basis set convergence of the interaction energy of hydrogen-bonded complexes. *J. Chem. Phys.* **1999**, *111*, 9157–9167.

[215] Shields, R. M.; Temelso, B.; Archer, K. A.; Morrell, T. E.; Shields, G. C. Accurate Predictions of Water Cluster Formation, $(H_2O)_{n=2-10}$. *J. Phys. Chem. A* **2010**, *114*, 11725–11737.

[216] Burns, L. A.; Marshall, M. S.; Sherrill, C. D. Comparing Counterpoise-Corrected, Uncorrected, and Averaged Binding Energies for Benchmarking Noncovalent Interactions. *J. Chem. Theory Comput.* **2014**, *10*, 49–57.

[217] Brauer, B.; Kesharwani, M. K.; Martin, J. M. L. Some Observations on Counterpoise Corrections for Explicitly Correlated Calculations on Noncovalent Interactions. *J. Chem. Theory Comput.* **2014**, *10*, 3791–3799.

[218] Werner, H.-J.; Adler, T. B.; Knizia, G.; Manby, F. R. In *Recent Progress in Coupled Cluster Methods*; Cársky, P., Paldus, J., Pittner, J., Eds.; Springer: Dordrecht, Holland, 2010; pp 573–619.

[219] Helgaker, T.; Ruden, T. A.; Jórgensen, P.; Olsen, J.; Klopper, W. A priori calculation of molecular properties to chemical accuracy. *J. Phys. Org. Chem.* **2004**, *17*, 913–933.

[220] Furtado, J. P.; Rahalkar, A. P.; Shanker, S.; Bandyopadhyay, P.; Gadre, S. R. Facilitating Minima Search for Large Water Clusters at the MP2 Level via Molecular Tailoring. *J. Phys. Chem. Lett.* **2012**, *3*, 2253–2258.

[221] Sahu, N.; Yeole, S. D.; Gadre, S. R. Appraisal of molecular tailoring approach for large clusters. *J. Chem. Phys.* **2013**, *138*, 104101.

[222] Saha, A.; Raghavachari, K. Dimers of Dimers (DOD): A New Fragment-Based Method Applied to Large Water Clusters. *J. Chem. Theory Comput.* **2014**, *10*, 58–67.

[223] Jiang, N.; Ma, J.; Jiang, Y. Electrostatic field-adapted molecular fractionation with conjugated caps for energy calculations of charged biomolecules. *J. Chem. Phys.* **2006**, *124*, 114112.

[224] Cisneros, G. A.; Karttunen, M.; Ren, P.; Sagui, C. Classical Electrostatics for Biomolecular Simulations. *Chem. Rev.* **2014**, *114*, 779–814.

[225] Williams, D. H.; Westwell, M. S. Aspects of weak interactions. *Chem. Soc. Rev.* **1998**, *27*, 57–63.

[226] Hunter, C. A.; Anderson, H. L. What is Cooperativity? *Angew. Chem. Int. Ed.* **2009**, *48*, 7488–7499.

[227] Mahadevi, A. S.; Sastry, G. N. Cooperativity in Noncovalent Interactions. *Chem. Rev.* **2016**, *116*, 2775–2825.

[228] Farina, C.; Santos, F. C.; Tort, A. C. A simple way of understanding the nonadditivity of van der Waals dispersion forces. *Am. J. Phy.* **1999**, *67*, 344–349.

[229] Gregory, J. K.; Clary, D. C.; Liu, K.; Brown, M. G.; Saykally, R. J. The Water Dipole Moment in Water Clusters. *Science* **1997**, *275*, 814–817.

[230] Wu, Y.-D.; Zhao, Y.-L. A Theoretical Study on the Origin of Cooperativity in the Formation of 3_{10}- and α-Helices. *J. Am. Chem. Soc.* **2001**, *123*, 5313–5319.

[231] Wieczorek, R.; Dannenberg, J. J. Comparison of Fully Optimized α- and 3_{10}-Helices with Extended β-Strands. An ONIOM Density Functional Theory Study. *J. Am. Chem. Soc.* **2004**, *126*, 14198–14205.

[232] Hua, S.; Xu, L.; Li, W.; Li, S. Cooperativity in Long α- and 3_{10}-Helical Polyalanines: Both Electrostatic and van der Waals Interactions Are Essential. *J. Phys. Chem. B* **2011**, *115*, 11462–11469.

[233] Kohn, W. Density Functional and Density Matrix Method Scaling Linearly with the Number of Atoms. *Phys. Rev. Lett.* **1996**, *76*, 3168–3171.

[234] Prodan, E.; Kohn, W. Nearsightedness of electronic matter. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 11635–11638.

[235] Collins, M. A.; Cvitkovic, M. W.; Bettens, R. P. A. The Combined Fragmentation and Systematic Molecular Fragmentation Methods. *Acc. Chem. Res.* **2014**, *47*, 2776–2785.

[236] Qi, H. W.; Leverentz, H. R.; Truhlar, D. G. Water 16-mers and Hexamers: Assessment of the Three-Body and Electrostatically Embedded Many-Body Approximations of the Correlation Energy of the Nonlocal Energy As Ways to Include Cooperative Effects. *J. Phys. Chem. A* **2013**, *117*, 4486–4499.

[237] Cobar, E. A.; Horn, P. R.; Bergman, R. G.; Head-Gordon, M. Examination of the hydrogen-bonding networks in small water clusters ($n = 2-5, 13, 17$) using absolutely localized molecular orbital energy decomposition analysis. *Phys. Chem. Chem. Phys.* **2012**, *14*, 15328–15339.

[238] Červinka, C.; Fulem, M.; Růžička, K. CCSD(T)/CBS fragment-based calculations of lattice energy of molecular crystals. *J. Chem. Phys.* **2016**, *144*, 064505.

[239] Ouyang, J. F.; Bettens, R. P. A. Many-Body Basis Set Superposition Effect. *J. Chem. Theory Comput.* **2015**, *11*, 5132–5143.

[240] Wang, K.; Li, W.; Li, S. Generalized Energy-Based Fragmentation CCSD(T)-F12a Method and Application to the Relative Energies of Water Clusters $(H_2O)_{20}$. *J. Chem. Theory Comput.* **2014**, *10*, 1546–1553.

[241] Guimarães, F. F.; Belchior, J. C.; Johnston, R. L.; Roberts, C. Global optimization analysis of water clusters $(H_2O)_n$ ($11 \leq n \leq 13$) through a genetic evolutionary approach. *J. Chem. Phys.* **2002**, *116*, 8327–8333.

[242] Wolfram Research, I. *Mathematica*, version 10.0 ed.; Wolfram Research, Inc.: Champaign, Illinois, 2014.

[243] Axilrod, B. M.; Teller, E. Interaction of the van der Waals Type Between Three Atoms. *J. Chem. Phys.* **1943**, *11*, 299–300.

[244] Muto, Y. Force Between Nonpolar Molecules. *Proc. Phys. Math. Soc. Japan* **1943**, *17*, 629–631.

[245] Richard, R. M.; Lao, K. U.; Herbert, J. M. Understanding the many-body expansion for large systems. I. Precision considerations. *J. Chem. Phys.* **2014**, *141*, 014108.

[246] Kobko, N.; Paraskevas, L.; del Rio, E.; Dannenberg, J. J. Cooperativity in Amide Hydrogen Bonding Chains: Implications for Protein-Folding Models. *J. Am. Chem. Soc.* **2001**, *123*, 4348–4349.

[247] Kobko, N.; Dannenberg, J. J. Cooperativity in Amide Hydrogen Bonding Chains. Relation between Energy, Position, and H-Bond Chain Length in Peptide and Protein Folding Models. *J. Phys. Chem. A* **2003**, *107*, 10389–10395.

[248] Suhai, S. Density Functional Theory of Molecular Solids: Local versus Periodic Effects in the Two-Dimensional Infinite Hydrogen-Bonded Sheet of Formamide. *J. Phys. Chem.* **1996**, *100*, 3950–3958.

[249] Wieczorek, R.; Dannenberg, J. J. H-Bonding Cooperativity and Energetics of α-Helix Formation of Five 17-Amino Acid Peptides. *J. Am. Chem. Soc.* **2003**, *125*, 8124–8129.

A | Supplementary Figures and Tables

Here, we include supplementary figures which contains similar trends as some of the figures presented in the main text. For the convenience of the reader, the list of supplementary figures and their corresponding counterparts in the main text is listed below.

Figure		
Main text	Appendix	Brief description
3.4	A.1	Convergence of the MBE for the **10OB** (in Main text) and **10OB** (in Appendix) decamer
4.3	A.2	BSEE and the total k-body interaction for the cage (in Main text) and prism (in Appendix) isomer of $(H_2O)_6$
4.5	A.3	Comparison of the total k-body interaction computed using the subcluster basis (in Main text) and cluster basis (in Appendix)
5.1	A.4	The two models used to describe the three-body GA dataset is further described here
5.4a	A.5	Comparison of $\lvert\varepsilon^{(3)}_{\text{actual}}\rvert$ with $\lvert\varepsilon^{(3)}_{\text{max}}\rvert$. Data for one cluster is shown in Main text while data for all five clusters are in the Appendix
5.4b	A.6	Comparison of $\lvert\varepsilon^{(4)}_{\text{actual}}\rvert$ with $\lvert\varepsilon^{(4)}_{\text{max}}\rvert$. Data for one cluster is shown in Main text while data for all five clusters are in the Appendix

Furthermore, tables with long lists of parameters are found here.

A.1 Additional Figures

A.1.1 Convergence of MBE for 10OB

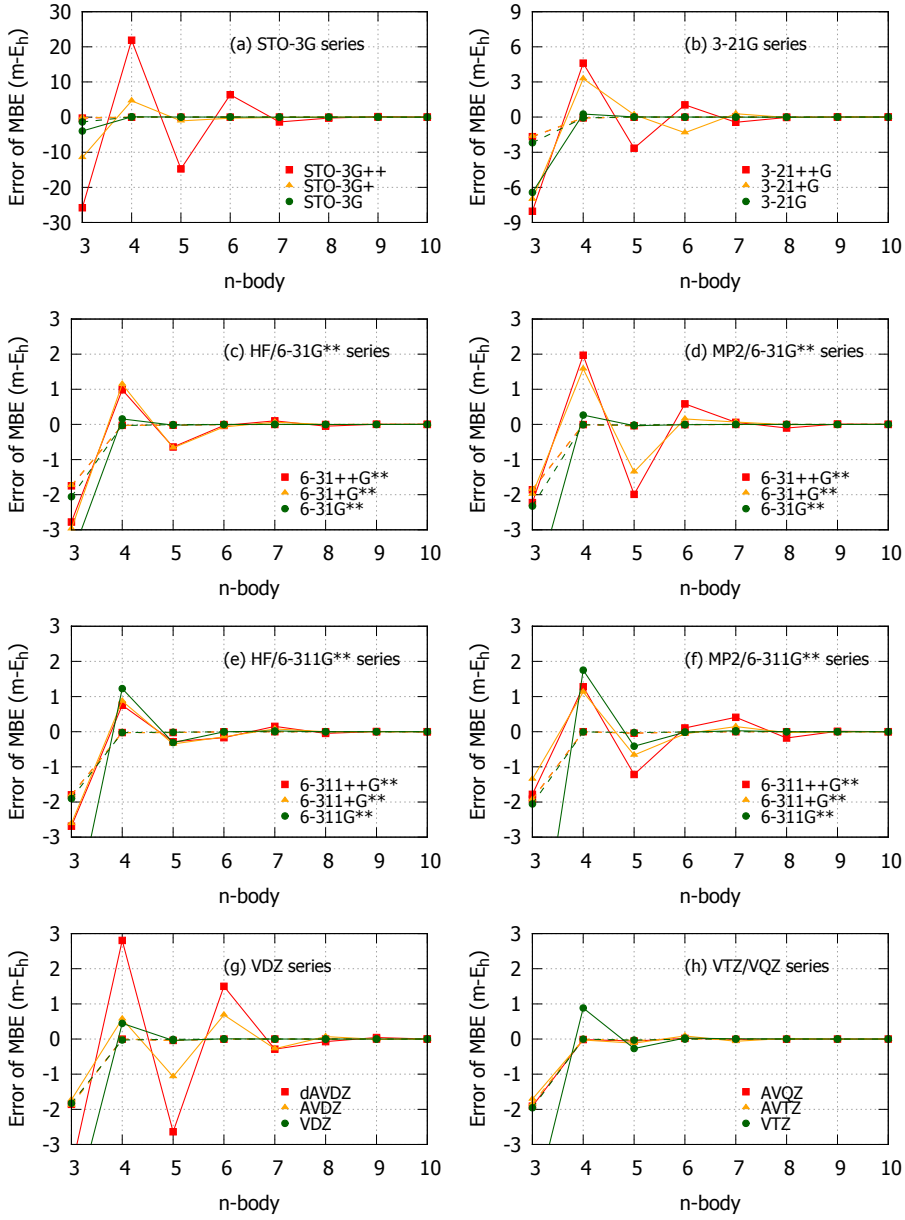


Figure A.1: Convergence of the MBE for **10OB** using various basis sets. Solid lines represent the use of the k-mer basis while dashed lines represent the use of the cluster basis.

A.1.2 *BSEE and Total Many-body Interaction in Prism Water Hexamer*

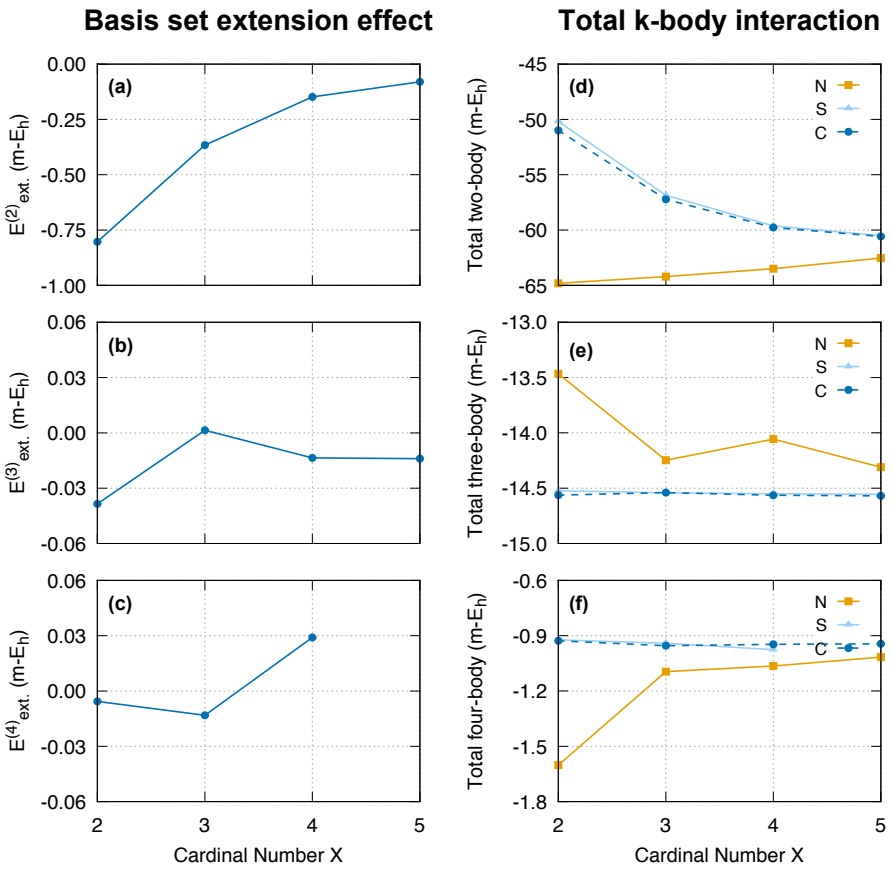


Figure A.2: The (a–c) BSEE in the total k-body interaction, $E^{(k)}_{\text{ext.}}$, and the (d–f) total k-body interaction for the prism isomer of $(H_2O)_6$ with increasing basis set quality at MP2/AVXZ. The total k-body interaction are computed using various *location* basis, namely the nuclei-centred (N), subcluster (S) and cluster (C) basis described in Section 4.2.2 to determine the effects of many-body BSSE on the many-body interactions. In particular, the lines for the cluster basis are dashed to show clearly the similarities between that and the subcluster basis results. The $E^{(4)}_{\text{ext.}}$ and total four-body interaction computed using the subcluster basis at MP2/AV5Z are omitted due to steep computational cost.

A.1.3 Comparison of Total Many-body Energy Computed using Cluster Basis

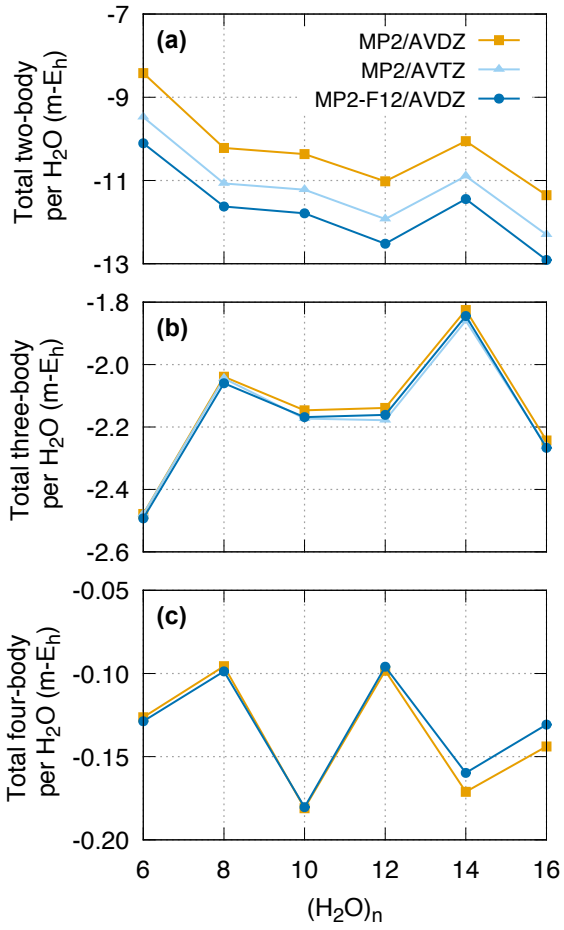


Figure A.3: Comparison of the total k-body interaction computed using the cluster basis per H_2O monomer for water clusters of increasing size, $(H_2O)_{6\text{-}16}$, computed at various levels of theory and basis sets.

A.1.4 Modelling the Three-body GA Dataset

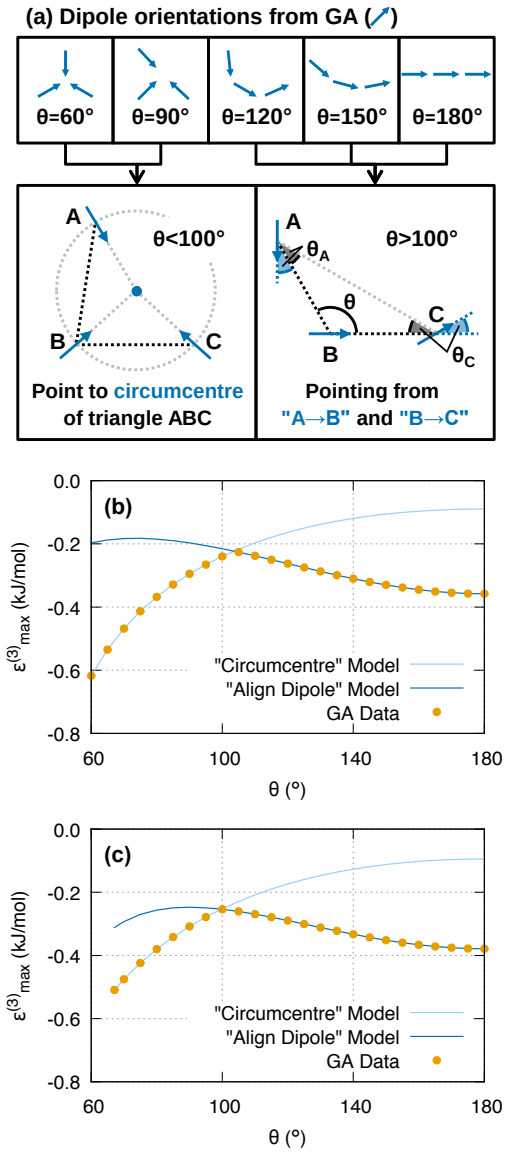


Figure A.4: The dipole orientations in the three-body GA dataset exhibits two different behaviour depending on θ. (a) The dipoles either point to the circumcentre of the triangle formed by the three bodies ABC when θ is small or become aligned along AC with some offset related to the angles of the triangle ABC when θ is large. The two models are then tested on the θ used in the three-body GA dataset with (b) R_{AB}=4.0Å, R_{BC}=4.0Å and (c) R_{AB}=3.5Å, R_{BC}=4.5Å.

A.1.5 Comparison of the Actual and Maximum Possible Many-body Interaction

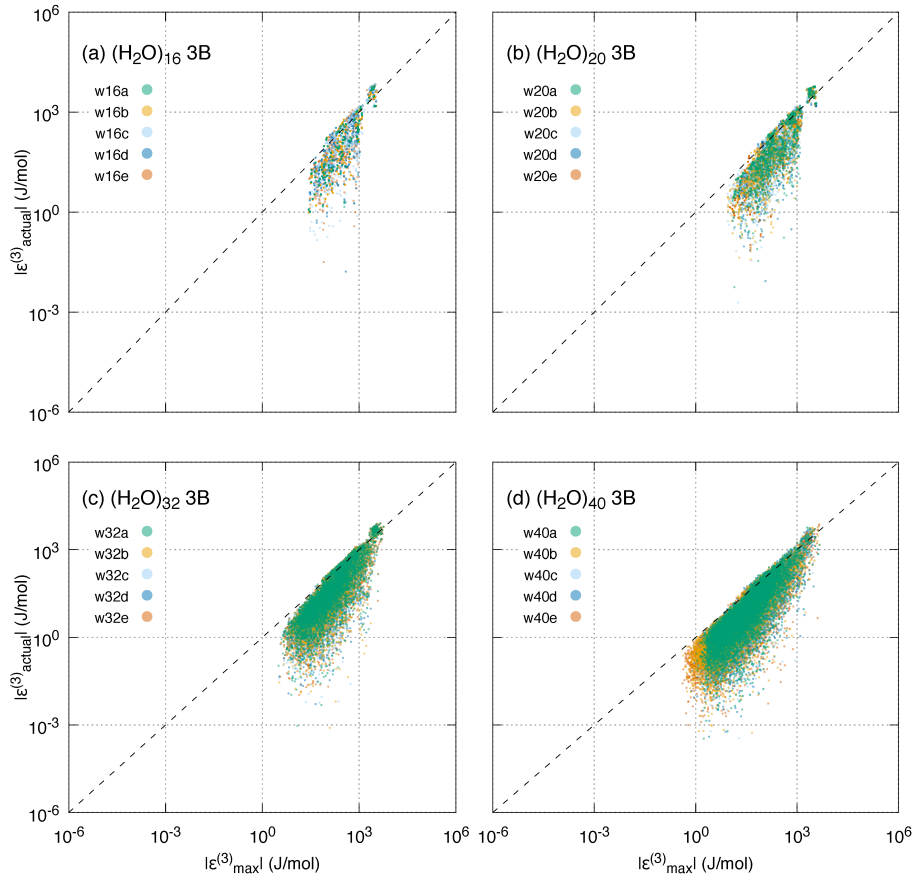


Figure A.5: Comparison of the magnitude of the actual three-body interactions, $|\varepsilon_{\text{actual}}|$, with that of the predicted maximum three-body interactions, $|\varepsilon_{\text{max}}|$ for water clusters of different size, including the (a) $(H_2O)_{16}$, (b) $(H_2O)_{20}$, (c) $(H_2O)_{32}$ and (d) $(H_2O)_{40}$ clusters.

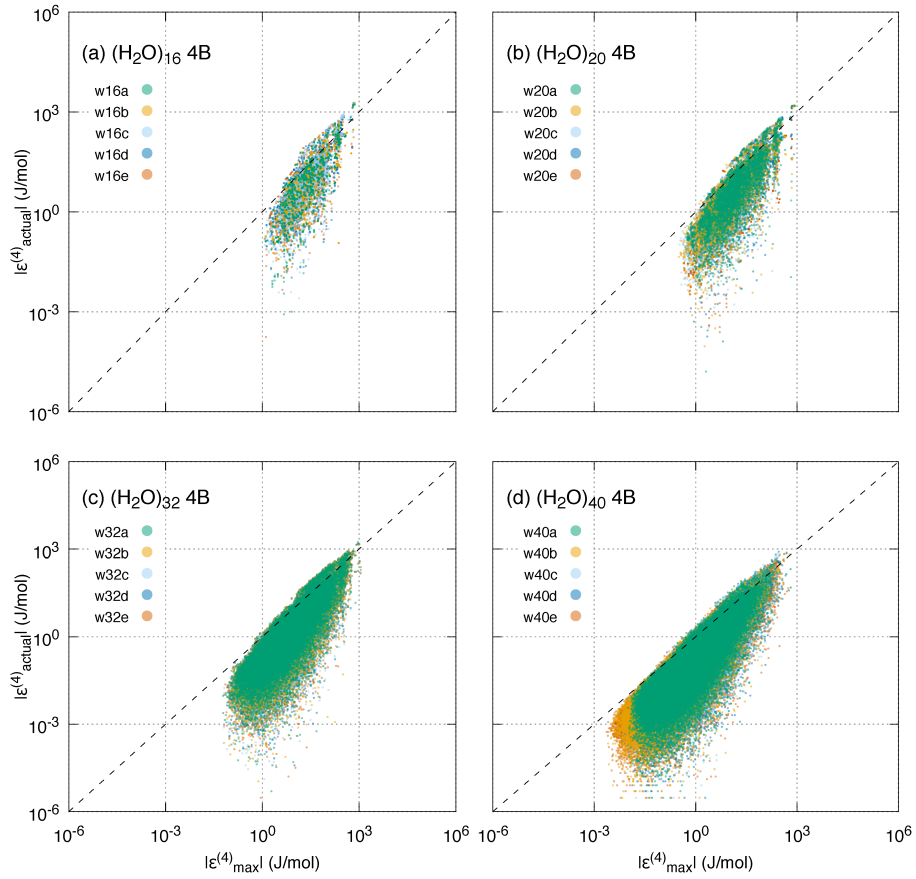


Figure A.6: Comparison of the magnitude of the actual four-body interactions, $|\varepsilon_{\text{actual}}|$, with that of the predicted maximum four-body interactions, $|\varepsilon_{\text{max}}|$ for water clusters of different size, including the (a) $(H_2O)_{16}$, (b) $(H_2O)_{20}$, (c) $(H_2O)_{32}$ and (d) $(H_2O)_{40}$ clusters.

A.2 Additional Tables

A.2.1 Fitted Coefficients for the Orientational Components of Four-body Interactions

For the four-body interactions, the orientational components obtained from GA are fitted to a sum of product of cosine functions of the following form

$$\sum_{1\leq i,j,k\leq 3} c_{i,j,k} \cos^i \theta_1 \cos^j \theta_2 \cos^k \phi \tag{A.1}$$

Since i, j, k runs from 1 to 3, there are a total of $3^3 = 27$ $c_{i,j,k}$ coefficients. The fitted coefficients are given in Table A.1.

Table A.1: Fitted $c_{i,j,k}$ coefficients for eq (A.1).

i	j	k	$c_{i,j,k}$	i	j	k	$c_{i,j,k}$	i	j	k	$c_{i,j,k}$
0	0	0	2.408239	0	0	1	0.405611	0	0	2	0.457653
1	0	0	0.223393	1	0	1	0.574079	1	0	2	-0.071959
2	0	0	1.433820	2	0	1	-0.109810	2	0	2	-0.509253
0	1	0	0.223393	0	1	1	0.574079	0	1	2	-0.071959
1	1	0	2.106957	1	1	1	3.473203	1	1	2	0.626254
2	1	0	-0.084414	2	1	1	1.774932	2	1	2	0.862841
0	2	0	1.433820	0	2	1	-0.109810	0	2	2	-0.509254
1	2	0	-0.084413	1	2	1	1.774932	1	2	2	0.862841
2	2	0	0.962129	2	2	1	1.675623	2	2	2	1.555943

B | MATHEMATICAL DERIVATIONS

This part of the Appendix includes long and tedious mathematical derivations of the key equations presented in the thesis.

B.1 Many-ghost Many-body Expansion

Before proceeding to derive the working equations for the MBE and MGMBE, we need to introduce a "general notation" in place of the specific notation used in the chapter, which we shall call the "chapter notation" (Table B.1).

Table B.1: List of selected quantities in the chapter (in chapter notation) and their corresponding general notation for the derivation of working equations.

Chapter	General	Definition
$E_{\mathrm{A}\cdots\mathrm{K}}$	$E_j^{(k)}$	Total energy of the j-th k-mer subcluster $\mathrm{A}\cdots\mathrm{K}$ calculated in its own basis
$E_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$	$E_{j,\beta}^{(k,m)}$	Total energy of the j-th k-mer subcluster $\mathrm{A}\cdots\mathrm{K}$ calculated in the presence of ghost functions centred on the β-th set of m-ghost-bodies $\mathrm{L}\cdots\mathrm{M}$
$\varepsilon'_{\mathrm{A}\cdots\mathrm{K}}$	$\varepsilon_i'^{(k)}$	k-body interaction of the i-th k-mer subcluster $\mathrm{A}\cdots\mathrm{K}$ computed in the nuclei-centred basis
$\varepsilon_{\mathrm{A}\cdots\mathrm{K}}$	$\varepsilon_i^{(k)}$	k-body interaction of the i-th k-mer subcluster $\mathrm{A}\cdots\mathrm{K}$ computed in the subcluster basis
$\varepsilon_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$	$\varepsilon_{i,\alpha}^{(k,m)}$	k-body interaction of the i-th k-mer subcluster $\mathrm{A}\cdots\mathrm{K}$ computed using total energies calculated with basis functions centred on both $\mathrm{A}\cdots\mathrm{K}$ and the α-th set of m-ghost-bodies $\mathrm{L}\cdots\mathrm{M}$
$\xi_{\mathrm{A}\cdots\mathrm{K},\mathrm{L}\cdots\mathrm{M}}$	$\xi_{i,\alpha}^{(k,m)}$	Basis Set Extension Effect (BSEE) from the α-th set of m-ghost-bodies $\mathrm{L}\cdots\mathrm{M}$ in the k-body interaction of the i-th k-mer subcluster $\mathrm{A}\cdots\mathrm{K}$

The general notation focusses on the number of monomer bodies and ghost-bodies involved—denoted by the superscript—allowing for a more general derivation up to any arbitrary number of bodies. Furthermore, the identity of the monomer bodies and ghost-bodies are not given explicitly in the general notation but represented by an integer variable denoted in the subscript. Also, the subscript is dropped when all the terms of the same type are summed, e.g, $\varepsilon^{(k)}$ gives the total k-body interaction computed in the subcluster basis.

B.1.1 Preliminaries: Many-body Interactions

Firstly, we need to establish compact equations for the total many-body interactions computed in the various basis (nuclei-centred, subcluster and cluster)to establish the cancellation of terms in the MGMBE. In the general notation (Table B.1), the k-body interaction of the i-th k-mer subcluster computed in the nuclei-centred basis, $\varepsilon_i'^{(k)}$, can be expressed in terms of total energies

$$\varepsilon_i'^{(k)} = \sum_{g=0}^{k-1} (-1)^g \sum_{j=1}^{\binom{k}{k-g}} E_j^{(k-g)} \tag{B.1}$$

where $E_j^{(k-g)}$ is the total energy of the j-th $(k-g)$-mer with monomers taken from the i-th k-mer subcluster. For example, in computing the three-body interaction of timer ABC, $g = 1$ would correspond to dimers $(k-g = 2)$ taken from ABC, which could be either AB, AC or BC. The first term, where $g = 0$ and $j = 1$, would uniquely correspond to the total energy of the subcluster, i.e., $E_1^{(k)} = E_i^{(k)}$. To obtain the total k-body interaction computed in the nuclei-centred basis, $\varepsilon'^{(k)}$, we need to sum all $\binom{n}{k}$individual k-body interactions

$$\varepsilon'^{(k)} = \sum_{i=1}^{\binom{n}{k}} \varepsilon_i'^{(k)} \tag{B.2}$$

When summing the individual many-body interactions, there are total energy terms repeated for subclusters with overlapping monomers. For example, E_A is involved in computing $\varepsilon'_{\mathrm{AB}}$, $\varepsilon'_{\mathrm{AC}}$ and in fact, all two-body interactions of the form $\varepsilon'_{\mathrm{AX}}$. Instead of performing tedious—and computationally inefficient—bookkeeping of the total energies involved for individual many-body interactions, we exploit the symmetry that each total energy term appears an equal number of times when the terms are collected to give $\varepsilon'^{(k)}$.

For $\varepsilon'^{(k)}$, there comprises $\binom{n}{k}$ individual $\varepsilon_i'^{(k)}$ given by eq (B.2) which in turn contains $\binom{k}{k-g}$ total energy terms of the form $E_j^{(k-g)}$ given by eq (B.1). This implies that in computing $\varepsilon'^{(k)}$, there are a total of $\binom{n}{k} \cdot \binom{k}{k-g}$ total energy terms of the form $E_j^{(k-g)}$. Since such terms involve $(k-g)$ monomers, there are only $\binom{n}{k-g}$ of such unique $E_j^{(k-g)}$ terms. Thus, each of these total energy terms would

be repeated

$$\frac{\binom{n}{k}\cdot\binom{k}{k-g}}{\binom{n}{k-g}} = \frac{(n-k+g)!}{(n-k)!g!} = \binom{n-k+g}{g}$$

times. In the general notation (Table B.1), we can combine eq (B.1) and (B.2) to give a compact expression of the $\varepsilon'^{(k)}$

$$\varepsilon'^{(k)} = \sum_{g=0}^{k-1}(-1)^g\binom{n-k+g}{g}E^{(k-g)} \tag{B.3}$$

where $E^{(k-g)}$ refers to the sum of the total energies of each $(k-g)$-mer, which can be trivially collected. Note that eq (B.3) is similar to eq(3.3) where the lower-body energies in the latter equation is replaced with total energy terms. Furthermore, eq (B.3) also apply to many-body interactions computed in the cluster basis where the many-body interactions share the same total energy terms. From the general expression in eq (B.3), we can also write out the total k-body interaction explicitly up to the four-body terms

$$\varepsilon'^{(1)} = E^{(1)} \tag{B.4a}$$

$$\varepsilon'^{(2)} = E^{(2)} - (n-1)E^{(1)} \tag{B.4b}$$

$$\varepsilon'^{(3)} = E^{(3)} - (n-2)E^{(2)} + \frac{(n-1)(n-2)}{2}E^{(1)} \tag{B.4c}$$

$$\varepsilon'^{(4)} = E^{(4)} - (n-3)E^{(3)} + \frac{(n-2)(n-3)}{2}E^{(2)} - \frac{(n-1)(n-2)(n-3)}{6}E^{(1)} \tag{B.4d}$$

When the subcluster basis is used to compute many-body interactions, the total energies cannot be reused. For example, $E_{\mathrm{A}\overline{\mathrm{B}}} \neq E_{\mathrm{A}\overline{\mathrm{C}}}$ in computing $\varepsilon_{\mathrm{AB}}$ and $\varepsilon_{\mathrm{AC}}$. Thus, the compact expression for the total k-body interaction computed in the subcluster basis, $\varepsilon^{(k)}$, would be different from the nuclei-centred counterpart derived earlier. In computing the $\varepsilon^{(k)}$, a total of $\binom{n}{k}\cdot\binom{k}{k-g}$ total energy terms are involved. For the many-body interaction computed in the subcluster basis, total energy terms of the form $E_{j,\beta}^{(k-g,g)}$ will be calculated where there are $(k-g)$-bodies and g-ghost-bodies. Interestingly, the number of unique terms is also $\binom{n}{k}\cdot\binom{k}{k-g}$. We obtain this by first considering that both the actual bodies and ghost-bodies contribute to the subcluster basis and there are $\binom{n}{k}$ ways to pick the subcluster basis of a k-mer subcluster. Subsequently, there are $\binom{k}{k-g}$ ways to pick out the $(k-g)$ actual bodies for placing the nuclei. Thus, each $E_{j,\beta}^{(k-g,g)}$ total energy term only appears once when collected. In the general notation (Table B.1), the compact expression of the $\varepsilon^{(k)}$ is

$$\varepsilon^{(k)} = \sum_{g=0}^{k-1}(-1)^g E^{(k-g,g)} \tag{B.5}$$

where $E^{(k-g,g)}$ refers to the sum of the total energies of each $(k-g)$-mer calculated in the presence of g-ghost-bodies. From eq (B.5), we can also write out explicitly the total k-body interaction computed in the subcluster basis up to the four-body terms

$$\varepsilon^{(1)} = E^{(1,0)} \tag{B.6a}$$
$$\varepsilon^{(2)} = E^{(2,0)} - E^{(1,1)} \tag{B.6b}$$
$$\varepsilon^{(3)} = E^{(3,0)} - E^{(2,1)} + E^{(1,2)} \tag{B.6c}$$
$$\varepsilon^{(4)} = E^{(4,0)} - E^{(3,1)} + E^{(2,1)} - E^{(1,3)} \tag{B.6d}$$

B.1.2 Working Equations for the MGMBE

In the MGMBE, many total energy terms are repeated when the different BSEE terms are summed. Here, we present the working equations for the total BSEE from m-ghost-bodies in the total k-body interaction, $\xi^{(k,m)}$. Furthermore, we describe the cancellation of total energy terms involving ghost functions in the MGMBE.

The MGMBE performs a two-dimensional many-body decomposition of the total energy of a cluster as shown in eq (4.9) in the chapter where the terms are arranged in a two-dimensional array. In the general notation (Table B.1), eq (4.9) in the chapter can be compactly written as

$$E_{\text{tot}} = \sum_{k=1}^{n} \sum_{\lambda=0}^{n-k} \sum_{i,\alpha}^{\binom{n}{k}\cdot\binom{n-k}{m}} \xi_{i,\alpha}^{(k,m)} \tag{B.7}$$

where the first two sums represent the two-dimensional decomposition while the last sum are essentially the BSEE terms arranged in a two-dimensional array in eq (4.9) in the chapter, i.e. $\sum_{i,\alpha}^{\binom{n}{k}\cdot\binom{n-k}{m}} \xi_{i,\alpha}^{(k,m)} = \sum_{\mathrm{A}<\cdots<\mathrm{K},\mathrm{L}<\cdots<\mathrm{M}}^{\binom{n}{k}\cdot\binom{n-k}{m}} \xi_{\mathrm{A}\cdots\mathrm{K}\overline{\mathrm{L}\cdots\mathrm{M}}}$. In fact, these terms are the total BSEE from the m-ghost-bodies in the total k-body interaction

$$\xi^{(k,m)} = \sum_{i,\alpha}^{\binom{n}{k}\cdot\binom{n-k}{m}} \xi_{i,\alpha}^{(k,m)} \tag{B.8}$$

In eq (4.10) in the chapter, we expressed the individual BSEE terms in terms of total energies for which $k+m \leq 4$. In the general notation (Table B.1), the BSEE from the α-th set of m-ghost-bodies in the k-body interaction of the i-th k-mer subcluster, $\xi_{i,\alpha}^{(k,m)}$, can be expressed as

$$\xi_{i,\alpha}^{(k,m)} = \sum_{\gamma=0}^{m} (-1)^{\gamma} \sum_{\beta=1}^{\binom{m}{m-\gamma}} \varepsilon_{i,\beta}^{(k,m-\gamma)} \tag{B.9}$$

where $\varepsilon_{i,\beta}^{(k,m-\gamma)}$ refers to the k-body interaction of the i-th k-mer subcluster computed in the presence of the β-th combination of $(m-\gamma)$ ghost-bodies taken from the α-th set of m-ghost-bodies. Since $\varepsilon_{i,\beta}^{(k,m-\gamma)}$ is a many-body interaction, it follows a form similar to eq (B.1) and thus can be further expressed in terms of total energies

$$\varepsilon_{i,\beta}^{(k,m-\gamma)} = \sum_{g=0}^{k-1} (-1)^g \sum_{j=1}^{\binom{k}{k-g}} E_{j,\beta}^{(k-g,g+m-\gamma)} \tag{B.10}$$

By inspecting eq (B.8), (B.9) and (B.10), we observe that in $\xi^{(k,m)}$, there is a total of $\binom{n}{k} \cdot \binom{n-k}{m} \cdot \binom{m}{m-\gamma} \cdot \binom{k}{k-g}$ total energy terms of the form $E_{j,\beta}^{(k-g,g+m-\gamma)}$. The first two terms in the product, $\binom{n}{k} \cdot \binom{n-k}{m}$ originates from eq (B.8) while the last two terms, $\binom{m}{m-\gamma}$ and $\binom{k}{k-g}$, arises from eq (B.9) and (B.10) respectively. Next, we consider the number of unique total energy terms. We first note that there is a total of $(k-g)+(g+m-\gamma) = (k+m-\gamma)$ actual bodies and ghost-bodies. Thus, the basis functions are centred on $(k+m-\gamma)$ monomers while nuclei are only placed on $(k-g)$ monomers. Thus, there are $\binom{n}{k+m-\gamma}$ ways to place the basis functions, followed by $\binom{k+m-\gamma}{k-g}$ to choose the nuclei, giving a total of $\binom{n}{k+m-\gamma} \cdot \binom{k+m-\gamma}{k-g}$ unique total energy terms. Thus, each of these terms would be repeated

$$\begin{aligned}
\frac{\binom{n}{k} \cdot \binom{n-k}{m} \cdot \binom{m}{m-\gamma} \cdot \binom{k}{k-g}}{\binom{n}{k+m-\gamma} \cdot \binom{k+m-\gamma}{k-g}} &= \frac{(n-k-m+\gamma)!}{(n-k-m)!\gamma!} \cdot \frac{(m-\gamma+g)!}{(m-\gamma)!g!} \\
&= \binom{n-k-m+\gamma}{\gamma} \cdot \binom{m-\gamma+g}{g}
\end{aligned}$$

times. Now, we can write a compact expression for $\xi^{(k,m)}$

$$\xi^{(k,m)} = \sum_{\gamma=0}^{m} (-1)^\gamma \sum_{g=0}^{k-1} (-1)^g \binom{n-k-m+\gamma}{\gamma} \cdot \binom{m-\gamma+g}{g} E^{(k-g,g+m-\gamma)} \tag{B.11}$$

where $E^{(k-g,g+m-\gamma)}$ refers to the sum of total energies of each $(k-g)$-mer calculated in the presence of $(g+m-\gamma)$-ghost-bodies. Note that when $m=0$, there is no BSEE from any ghost-body and we instead obtain the total k-body interaction computed in the subcluster basis, i.e., $\xi^{(k,0)} = \varepsilon^{(k)}$ and the working equations reduces to eq (B.5).

From eq (B.11), we write down explicitly the working equations for the total BSEE from m-ghost-bodies in the total k-body interaction, $\xi^{(k,m)}$, for the various combinations of k and m used in the chapter, namely $k+m \leq 4$. We first

begin with $k = 1$. When $k = 1$, we have

$$\xi^{(1,m)} = \sum_{\gamma=0}^{m} (-1)^{\gamma} \binom{n-1-m+\gamma}{\gamma} E^{(1,m-\gamma)} \tag{B.12}$$

in which, we have the subcases where $m = 0, 1, 2, 3$

$$\xi^{(1,0)} = E^{(1,0)} \tag{B.13a}$$

$$\xi^{(1,1)} = E^{(1,1)} - \binom{n-1}{1} E^{(1,0)} \tag{B.13b}$$

$$\xi^{(1,2)} = E^{(1,2)} - \binom{n-2}{1} E^{(1,1)} + \binom{n-1}{2} E^{(1,0)} \tag{B.13c}$$

$$\xi^{(1,3)} = E^{(1,3)} - \binom{n-3}{1} E^{(1,2)} + \binom{n-2}{2} E^{(1,1)} - \binom{n-1}{3} E^{(1,0)} \tag{B.13d}$$

When $k = 2$, we have (with the summation over g fully written out)

$$\begin{aligned} \xi^{(2,m)} = & \sum_{\gamma=0}^{m} (-1)^{\gamma} \binom{n-2-m+\gamma}{\gamma} E^{(2,m-\gamma)} \\ & - \sum_{\gamma=0}^{m} (-1)^{\gamma} \binom{n-2-m+\gamma}{\gamma} \cdot \binom{m-\gamma+1}{1} E^{(1,1+m-\gamma)} \end{aligned} \tag{B.14}$$

in which, we have the subcases where $m = 0, 1, 2$

$$\xi^{(2,0)} = E^{(2,0)} - E^{(1,1)} \tag{B.15a}$$

$$\begin{aligned} \xi^{(2,1)} = & \left\{ E^{(2,1)} - \binom{n-2}{1} E^{(2,0)} \right\} \\ & - \left\{ 2E^{(1,2)} - \binom{n-2}{1} E^{(1,1)} \right\} \end{aligned} \tag{B.15b}$$

$$\begin{aligned} \xi^{(2,2)} = & \left\{ E^{(2,2)} - \binom{n-3}{1} E^{(2,1)} + \binom{n-2}{2} E^{(2,0)} \right\} \\ & - \left\{ 3E^{(1,3)} - \binom{n-3}{1} \cdot 2E^{(1,2)} + \binom{n-2}{2} E^{(1,1)} \right\} \end{aligned} \tag{B.15c}$$

When $k = 3$, we have (with the summation over m fully written out)

$$\begin{aligned}\xi^{(3,m)} = &\sum_{\gamma=0}^{m}(-1)^{\gamma}\binom{n-3-m+\gamma}{\gamma}E^{(3,m-\gamma)} \\ &- \sum_{\gamma=0}^{m}(-1)^{\gamma}\binom{n-3-m+\gamma}{\gamma}\cdot\binom{m-\gamma+1}{1}E^{(2,1+m-\gamma)} \\ &+ \sum_{\gamma=0}^{m}(-1)^{\gamma}\binom{n-3-m+\gamma}{\gamma}\cdot\binom{m-\gamma+2}{2}E^{(1,2+m-\gamma)} \end{aligned} \tag{B.16}$$

in which, we have the subcases where $m = 0, 1$

$$\xi^{(3,0)} = E^{(3,0)} - E^{(2,1)} + E^{(1,2)} \tag{B.17a}$$

$$\begin{aligned}\xi^{(3,1)} = &\left\{E^{(3,1)} - \binom{n-3}{1}E^{(3,0)}\right\} \\ &- \left\{2E^{(2,2)} - \binom{n-3}{1}E^{(2,1)}\right\} \\ &+ \left\{3E^{(1,3)} - \binom{n-3}{1}E^{(1,2)}\right\}\end{aligned} \tag{B.17b}$$

When $k = 4$ and $\lambda = 0$, we have

$$\xi^{(4,0)} = E^{(4,0)} - E^{(3,1)} + E^{(2,2)} - E^{(1,3)} \tag{B.18}$$

B.1.3 Cancellation of Terms the MGMBE

Careful analysis of the working equations of the $\xi^{(k,m)}$ in eq (B.11) revealed that all the total energies involving any ghost functions vanishes when we sum the $\xi^{(k,m)}$ terms with $k+m = \alpha$, where α is a constant. We show this cancellation of terms explicitly for cases where $\alpha = 2, 3, 4$

$$\xi^{(2,0)} + \xi^{(1,1)} = E^{(2,0)} - \binom{n-1}{1}E^{(1,0)} \tag{B.19a}$$

$$\xi^{(3,0)} + \xi^{(2,1)} + \xi^{(1,2)} = E^{(3,0)} - \binom{n-2}{1}E^{(2,0)} + \binom{n-1}{2}E^{(1,0)} \tag{B.19b}$$

$$\xi^{(4,0)} + \cdots + \xi^{(1,3)} = E^{(4,0)} - \binom{n-3}{1}E^{(3,0)} + \binom{n-2}{2}E^{(2,0)} - \binom{n-1}{3}E^{(1,0)} \tag{B.19c}$$

From eq (B.19a–c), we observe that the second digit in the superscript in all the total energy terms are zero, of the form $E^{(k,0)}$, indicating that there are no ghost-bodies involved. In fact, the resulting expressions in each of the equation is equivalent to the k-body interaction computed in the nuclei-centred basis, $\varepsilon'^{(k)}$, given in eq (B.3). This is a surprising and important result and the implications are discussed in the chapter.

B.2 Derivation of Leading Many-body Terms

Using perturbation theory,[48] we derive the first-order and second-order induction energy between k bodies, $k = 2-4$. The leading terms from these energy expressions will reveal the most important contribution to the k-body interaction. Furthermore, we will identify trends that determine the coupling between bodies. Here, we consider only the dipole-dipole induction, which constitutes the most important many-body effect in highly polar molecules. Furthermore, we assume that the polarizability is isotropic and thus there is no change in the direction of the induced dipole. These equations can be generalized for any arbitrary-rank multipoles, for example monopole or quadrupole, by modifying the interaction tensor, T_{tu}^{AB}, which mainly affects the exponent of the inter-body distances in the denominator, R^{-n}. Throughout the entire derivation section, Einstein summation convention will be employed that implies summation over terms with the same indices.

The leading three-body and four-body terms are found in eq (B.27) and eq (B.36) respectively.

B.2.1 Prelude: The Two-body Terms

Consider two bodies A and B. The potential gradient at A due to B, $V_i^{\mathrm{A}(0)}(\mathbf{B})$, is given by

$$V_t^{\mathrm{A}(0)}(\mathbf{B}) = T_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}} = \frac{\mathscr{T}_{tu}^{\mathrm{AB}}}{{R_{\mathrm{AB}}}^3}\mu_u^{\mathrm{B}} \tag{B.20}$$

where $t,u \in \{x,y,z\}$ is the direction of the potential gradient in the molecule fixed, i.e., local axes of A and B respectively. T_{tu}^{AB} is the dipole-dipole interaction tensor while $\mathscr{T}_{tu}^{\mathrm{AB}}$ gives the orientational part. The (0) in the superscript of $V_t^{\mathrm{A}(0)}(\mathbf{B})$ indicates that this is the zero-order potential gradient due to the permanent dipole of B. The order here indicates the number of polarizations applied. Also, note that the potential gradient is the negative of the more familiar electric field. A similar expression exist for $V_u^{\mathrm{B}(0)}(\mathbf{A})$.

This potential gradient, $V_t^{\mathrm{A}(0)}(\mathbf{B})$, induces a dipole in A

$$\Delta\mu_t^{\mathrm{A}(1)}(\mathbf{B}) = -\alpha^{\mathrm{A}}V_t^{\mathrm{A}(0)}(\mathbf{B}) \tag{B.21}$$

where α^{A} is the isotropic polarizability of A and the (1) in the superscript of $\Delta\mu_t^{\mathrm{A}(1)}(\mathbf{B})$ indicates that this induced dipole at A is obtained from the first-order induction. As we have assumed an isotropic polarizability, the induced dipole is in the same direction as the potential gradient, which need not be so in general.

This induced dipole, $\Delta\mu_t^{\mathrm{A}(1)}(\mathbf{B})$, then interacts with the potential gradient to give the first-order induction energy at A due to B

$$\begin{aligned} E_{\text{ind}}^{\text{A}(1)}(\mathbf{B}) &= \frac{1}{2}\Delta\mu_t^{\text{A}(1)}(\mathbf{B})V_t^{\text{A}(0)}(\mathbf{B}) \\ &= -\frac{1}{2}\alpha^{\text{A}}\left\{V_t^{\text{A}(0)}(\mathbf{B})\right\}^2 \end{aligned} \tag{B.22}$$

B.2.2 The Leading Three-body Contributions

Next, consider how the first-order induction energy at A changes when a third body C is present. Now there is an additional field at A due to C which has a similar form as eq (B.20) and the new total field at A due to both B and C is simply the vectorial sum

$$V_t^{\text{A}(0)}(\mathbf{B},\mathbf{C}) = V_t^{\text{A}(0)}(\mathbf{B}) + V_t^{\text{A}(0)}(\mathbf{C}) \tag{B.23}$$

In eq (B.22), we have purposefully written $E_{\text{ind}}^{\text{A}(1)}(\mathbf{B})$ in terms of α^{A} and $V_t^{\text{A}(0)}(\mathbf{B})$. When a third body C is present, $E_{\text{ind}}^{\text{A}(1)}(\mathbf{B},\mathbf{C})$ can be written in a similar fashion by replacing the potential gradient $V_t^{\text{A}(0)}(\mathbf{B})$ with $V_t^{\text{A}(0)}(\mathbf{B},\mathbf{C})$. Thus, the first-order induction at A due to B and C is

$$\begin{aligned} E_{\text{ind}}^{\text{A}(1)}(\mathbf{B},\mathbf{C}) &= -\frac{1}{2}\alpha^{\text{A}}\left\{V_t^{\text{A}(0)}(\mathbf{B}) + V_t^{\text{A}(0)}(\mathbf{C})\right\}^2 \\ &= E_{\text{ind}}^{\text{A}(1)}(\mathbf{B}) + E_{\text{ind}}^{\text{A}(1)}(\mathbf{C}) - V_t^{\text{A}(0)}(\mathbf{C})\alpha^{\text{A}}V_t^{\text{A}(0)}(\mathbf{B}) \end{aligned} \tag{B.24}$$

We observe that there are additional terms which are not present if A interacted with B in the absence of C. These are the three-body contributions to the induction energy at A

$$\varepsilon_{\text{3B}}^{\text{A}(1)}(\mathbf{B},\mathbf{C}) = -V_t^{\text{A}(0)}(\mathbf{C})\alpha^{\text{A}}V_t^{\text{A}(0)}(\mathbf{B}) \tag{B.25}$$

Analogous expressions exist for the induction energy at B and C as well. To obtain the three-body first-order induction energy, we sum the contributions from A, B and C, which is given by

$$\begin{aligned} \varepsilon_{\text{3B}}^{\text{tot}(1)} &= -V_t^{\text{A}(0)}(\mathbf{C})\alpha^{\text{A}}V_t^{\text{A}(0)}(\mathbf{B}) - V_u^{\text{B}(0)}(\mathbf{A})\alpha^{\text{B}}V_u^{\text{B}(0)}(\mathbf{C}) - V_v^{\text{C}(0)}(\mathbf{B})\alpha^{\text{C}}V_v^{\text{C}(0)}(\mathbf{A}) \\ &= -\frac{\mu_v^{\text{C}}\,\mathscr{T}_{vt}^{\text{CA}}\alpha^{\text{A}}\,\mathscr{T}_{tu}^{\text{AB}}\mu_u^{\text{B}}}{{R_{\text{CA}}}^3{R_{\text{AB}}}^3} - \frac{\mu_t^{\text{A}}\,\mathscr{T}_{tu}^{\text{AB}}\alpha^{\text{B}}\,\mathscr{T}_{uv}^{\text{BC}}\mu_v^{\text{C}}}{{R_{\text{AB}}}^3{R_{\text{BC}}}^3} - \frac{\mu_u^{\text{B}}\,\mathscr{T}_{uv}^{\text{BC}}\alpha^{\text{C}}\,\mathscr{T}_{vt}^{\text{CA}}\mu_t^{\text{A}}}{{R_{\text{BC}}}^3{R_{\text{CA}}}^3} \end{aligned} \tag{B.26}$$

where $v \in \{x,y,z\}$ refers to the direction of the potential gradient in the local axes of C.

In the special case where A, B and C are identical molecules, the dipoles

and polarizabilities would be the same and eq (B.26) simplifies to

$$\varepsilon_{3\mathrm{B}}^{\mathrm{tot}(1)} = -\alpha\mu^2 \left(\frac{\mathscr{T}_{z't}^{\mathrm{CA}} \mathscr{T}_{tz}^{\mathrm{AB}}}{{R_{\mathrm{CA}}}^3 {R_{\mathrm{AB}}}^3} + \frac{\mathscr{T}_{z'u}^{\mathrm{AB}} \mathscr{T}_{uz}^{\mathrm{BC}}}{{R_{\mathrm{AB}}}^3 {R_{\mathrm{BC}}}^3} + \frac{\mathscr{T}_{z'v}^{\mathrm{BC}} \mathscr{T}_{vz}^{\mathrm{CA}}}{{R_{\mathrm{BC}}}^3 {R_{\mathrm{CA}}}^3} \right) \tag{B.27}$$

which are the leading three-body contributions to the many-body induction. As we will see later, higher-order induction also introduces three-body contributions but the additional terms contains additional R^{-3} terms. Thus, eq (B.27) contains the three-body contributions that decay the slowest with inter-body distance. Furthermore, the three separate terms in eq (B.27) represent the three different non-branching paths in the leading three-body contributions.

Now, let us understand the origin of the many-body contributions. When the sum of the potential gradients are squared in eq (B.24), there is a loss of linearity which gave rise to cross terms as shown in eq (B.25). These cross terms are the origins of the three-body induction energy. Careful inspection revealed that each of the cross terms have the potential gradients terms centred at particular body. For example, both $V_t^{\mathrm{A}(0)}(\mathbf{C})$ and $V_t^{\mathrm{A}(0)}(\mathbf{B})$ in eq (B.25) are potential gradients terms centred at A but due to the multipoles of different bodies, namely C and B respectively. Consequently, the powers of intermolecular separation in the denominator of these cross terms, i.e., the $R_{\mathrm{CA}}^{-3} R_{\mathrm{AB}}^{-3}$ terms, have to share a common centre, specifically the body at which the potential gradients are centred. We will make use of this observation in the next subsection.

B.2.3 Four-body Contributions

No Four-body Contributions in First-order Induction

Now we consider how first-order induction at A changes when a fourth body D is present. A fourth body alters the three-body treatment by having an additional field at A due to D. Following the same arguments as eq (B.24), the first-order induction energy at A is

$$\begin{aligned} E_{\mathrm{ind}}^{\mathrm{A}(1)}(\mathbf{B},\mathbf{C},\mathbf{D}) = E_{\mathrm{ind}}^{\mathrm{A}(1)}(\mathbf{B}) + E_{\mathrm{ind}}^{\mathrm{A}(1)}(\mathbf{C}) + E_{\mathrm{ind}}^{\mathrm{A}(1)}(\mathbf{D}) \\ - V_t^{\mathrm{A}(0)}(\mathbf{D}) \alpha^{\mathrm{A}} V_t^{\mathrm{A}(0)}(\mathbf{B}) \\ - V_t^{\mathrm{A}(0)}(\mathbf{D}) \alpha^{\mathrm{A}} V_t^{\mathrm{A}(0)}(\mathbf{C}) \\ - V_t^{\mathrm{A}(0)}(\mathbf{C}) \alpha^{\mathrm{A}} V_t^{\mathrm{A}(0)}(\mathbf{B}) \end{aligned} \tag{B.28}$$

Similar to the three-body case, we observe cross terms that are of the form $V_t^{\mathrm{A}(0)}(\mathbf{C}) \alpha^{\mathrm{A}} V_t^{\mathrm{A}(0)}(\mathbf{B})$ from the loss of linearity. However, these cross terms only involve three bodies simultaneously, i.e, they are strictly three-body. Furthermore, recall that the powers of intermolecular separation in the denominator share a common centre. Thus, only a maximum of three centres can be involved

and the four-body contribution to the first-order induction energy is zero.

$$\varepsilon_{4\mathrm{B}}^{\mathrm{tot}(1)} = 0 \tag{B.29}$$

Eq (B.28) can be extended to more bodies and it is clear that there are no n-body interactions, $n \geq 4$, in the first-order induction energy.

Leading Four-body Terms in Second-order Induction

To obtain the leading four-body contribution in the induction energy, we look at the second order induction energy. Upon first-order induction, the first-order total dipole of B is being modified with an additional induced dipole due to A, C and D

$$\mu_u^{\mathrm{B}} + \Delta\mu_u^{\mathrm{B}(1)}(\mathbf{A},\mathbf{C},\mathbf{D}) = \mu_u^{\mathrm{B}} - \alpha^{\mathrm{B}} \left\{ V_u^{\mathrm{B}(0)}(\mathbf{A}) + V_u^{\mathrm{B}(0)}(\mathbf{C}) + V_u^{\mathrm{B}(0)}(\mathbf{D}) \right\} \tag{B.30}$$

and analogous expressions exist for the first-order total dipole of A, C and D. Consequently, the potential gradient at A due to B, which is of the same form as eq (B.20) gets modified to

$$\begin{aligned} V_t^{\mathrm{A}(1)}(\mathbf{B}) &= \frac{\mathscr{T}_{tu}^{\mathrm{AB}}}{{R_{\mathrm{AB}}}^3} \left\{ \mu_u^{\mathrm{B}} + \Delta\mu_u^{\mathrm{B}(1)}(\mathbf{A},\mathbf{C},\mathbf{D}) \right\} \\ &= \frac{\mathscr{T}_{tu}^{\mathrm{AB}}}{{R_{\mathrm{AB}}}^3} \left\{ \mu_u^{\mathrm{B}} - \alpha^{\mathrm{B}} V_u^{\mathrm{B}(0)}(\mathbf{A}) - \alpha^{\mathrm{B}} V_u^{\mathrm{B}(0)}(\mathbf{C}) - \alpha^{\mathrm{B}} V_u^{\mathrm{B}(0)}(\mathbf{D}) \right\} \\ &= \frac{\mathscr{T}_{tu}^{\mathrm{AB}} \mu_u^{\mathrm{B}}}{{R_{\mathrm{AB}}}^3} - \frac{\mathscr{T}_{tu}^{\mathrm{AB}} \alpha^{\mathrm{B}} \mathscr{T}_{ut'}^{\mathrm{BA}} \mu_{t'}^{\mathrm{A}}}{{R_{\mathrm{AB}}}^6} - \frac{\mathscr{T}_{tu}^{\mathrm{AB}} \alpha^{\mathrm{B}} \mathscr{T}_{uv}^{\mathrm{BC}} \mu_v^{\mathrm{C}}}{{R_{\mathrm{AB}}}^3 {R_{\mathrm{BC}}}^3} - \frac{\mathscr{T}_{tu}^{\mathrm{AB}} \alpha^{\mathrm{B}} \mathscr{T}_{uw}^{\mathrm{BD}} \mu_w^{\mathrm{D}}}{{R_{\mathrm{AB}}}^3 {R_{\mathrm{BD}}}^3} \end{aligned} \tag{B.31}$$

where $w \in \{x,y,z\}$ refers to the direction of the potential gradient in the local axes of D and the superscript (1) in $V_i^{\mathrm{A}(1)}(\mathbf{B})$ indicates that this is the first-order potential gradient at A due to B, i.e., the potential gradient at A due to the first-order total dipole of B. Let us examine the final expression of eq (B.31) in detail. The first term is due to the permanent dipole of B and is already present in the zero-order potential gradient, $V_t^{\mathrm{A}(0)}(\mathbf{B})$, given in eq (B.20). The remaining terms originate from the induced dipole of B and thus have in common the isotropic polarizability of B, α^{B}.

Now, let us first write the second-order induction energy at A due to B, C and D in terms of potential gradients

$$
\begin{aligned}
E_{\mathrm{ind}}^{\mathrm{A}(2)}(\mathbf{B},\mathbf{C},\mathbf{D}) &= \frac{1}{2}\Delta\mu_t^{\mathrm{A}(2)}(\mathbf{B},\mathbf{C},\mathbf{D})V_t^{\mathrm{A}(0)}(\mathbf{B},\mathbf{C},\mathbf{D}) \\
&= -\frac{1}{2}V_t^{\mathrm{A}(1)}(\mathbf{B},\mathbf{C},\mathbf{D})\alpha^{\mathrm{A}}V_t^{\mathrm{A}(0)}(\mathbf{B},\mathbf{C},\mathbf{D}) \\
&= -\frac{1}{2}\alpha^{\mathrm{A}} \\
&\quad \Big\{V_t^{\mathrm{A}(1)}(\mathbf{B})V_t^{\mathrm{A}(0)}(\mathbf{B}) + V_t^{\mathrm{A}(1)}(\mathbf{C})V_t^{\mathrm{A}(0)}(\mathbf{C}) + V_t^{\mathrm{A}(1)}(\mathbf{D})V_t^{\mathrm{A}(0)}(\mathbf{D}) \\
&\quad + V_t^{\mathrm{A}(1)}(\mathbf{C})V_t^{\mathrm{A}(0)}(\mathbf{B}) + V_t^{\mathrm{A}(1)}(\mathbf{D})V_t^{\mathrm{A}(0)}(\mathbf{B}) + V_t^{\mathrm{A}(1)}(\mathbf{D})V_t^{\mathrm{A}(0)}(\mathbf{C}) \\
&\quad + V_t^{\mathrm{A}(1)}(\mathbf{B})V_t^{\mathrm{A}(0)}(\mathbf{C}) + V_t^{\mathrm{A}(1)}(\mathbf{B})V_t^{\mathrm{A}(0)}(\mathbf{D}) + V_t^{\mathrm{A}(1)}(\mathbf{C})V_t^{\mathrm{A}(0)}(\mathbf{D})\Big\}
\end{aligned}
\tag{B.32}
$$

Note that to correctly obtain the second-order induction energy, the second-order induced dipole has to be applied onto the zero-order/permanent potential gradient and not a potential gradient of any other order. Unlike eq (B.28), we cannot easily separate out the two-body components from the repeated-centre terms, i.e., the $V_t^{\mathrm{A}(1)}(\mathbf{B})\alpha^{\mathrm{A}}V_t^{\mathrm{A}(0)}(\mathbf{B})$ terms. This is because the first-order potential gradient $V_t^{\mathrm{A}(1)}(\mathbf{B})$ contains contributions from both two and three bodies as shown in eq (B.31). Nonetheless, we can analyse separately the repeated-centre and cross-centre terms to find the leading four-body contributions to the induction energy.

Let us first expand one of the repeated-centre terms, say

$$
\begin{aligned}
V_t^{\mathrm{A}(1)}(\mathbf{B})\alpha^{\mathrm{A}}V_t^{\mathrm{A}(0)}(\mathbf{B}) &= \left\{\frac{\mu_{u'}^{\mathrm{B}}\mathscr{T}_{u't}^{\mathrm{BA}}}{{R_{\mathrm{BA}}}^3} - \frac{\mu_{t'}^{\mathrm{A}}\mathscr{T}_{t'u'}^{\mathrm{AB}}\alpha^{\mathrm{B}}\mathscr{T}_{u't}^{\mathrm{BA}}}{{R_{\mathrm{BA}}}^6} - \frac{\mu_v^{\mathrm{C}}\mathscr{T}_{vu'}^{\mathrm{CB}}\alpha^{\mathrm{B}}\mathscr{T}_{u't}^{\mathrm{BA}}}{{R_{\mathrm{CB}}}^3{R_{\mathrm{BA}}}^3}\right. \\
&\quad \left. - \frac{\mu_w^{\mathrm{D}}\mathscr{T}_{wu'}^{\mathrm{DB}}\alpha^{\mathrm{B}}\mathscr{T}_{u't}^{\mathrm{BA}}}{{R_{\mathrm{DB}}}^3{R_{\mathrm{BA}}}^3}\right\}\alpha^{\mathrm{A}}\left\{\frac{\mathscr{T}_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}}}{{R_{\mathrm{AB}}}^3}\right\} \\
&= \frac{\mu_{u'}^{\mathrm{B}}\mathscr{T}_{u't}^{\mathrm{BA}}\alpha^{\mathrm{A}}\mathscr{T}_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}}}{{R_{\mathrm{AB}}}^6} - \frac{\mu_{t'}^{\mathrm{A}}\mathscr{T}_{t'u'}^{\mathrm{AB}}\alpha^{\mathrm{B}}\mathscr{T}_{u't}^{\mathrm{BA}}\alpha^{\mathrm{A}}\mathscr{T}_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}}}{{R_{\mathrm{AB}}}^9} \\
&\quad - \frac{\mu_v^{\mathrm{C}}\mathscr{T}_{vu'}^{\mathrm{CB}}\alpha^{\mathrm{B}}\mathscr{T}_{u't}^{\mathrm{BA}}\alpha^{\mathrm{A}}\mathscr{T}_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}}}{{R_{\mathrm{CB}}}^3{R_{\mathrm{BA}}}^6} \\
&\quad - \frac{\mu_w^{\mathrm{D}}\mathscr{T}_{wu'}^{\mathrm{DB}}\alpha^{\mathrm{B}}\mathscr{T}_{u't}^{\mathrm{BA}}\alpha^{\mathrm{A}}\mathscr{T}_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}}}{{R_{\mathrm{DB}}}^3{R_{\mathrm{BA}}}^6}
\end{aligned}
\tag{B.33}
$$

In the final expression of eq (B.33), the first two terms involves only the bodies A and B, equating to $E_{\mathrm{ind}}^{\mathrm{A}(2)}(\mathbf{B})$ with the terms being the first-order induction energy and second-order correction respectively. By symmetry, the $E_{\mathrm{ind}}^{\mathrm{A}(2)}(\mathbf{C})$ and $E_{\mathrm{ind}}^{\mathrm{A}(2)}(\mathbf{D})$ are found in the $V_t^{\mathrm{A}(1)}(\mathbf{C})\alpha^{\mathrm{A}}V_t^{\mathrm{A}(0)}(\mathbf{C})$ and $V_t^{\mathrm{A}(1)}(\mathbf{D})\alpha^{\mathrm{A}}V_t^{\mathrm{A}(0)}(\mathbf{D})$ terms respectively. This implies that the two-body components to the induction energy are only found in the repeated-centre terms in eq (B.32). The latter two

terms in the final expression of eq (B.33) represent part of the second-order correction to the three-body induction energy.

Next, we expand one of the cross-centre terms, say

$$\begin{aligned} V_t^{\mathrm{A}(1)}(\mathbf{C})\alpha^{\mathrm{A}}V_t^{\mathrm{A}(0)}(\mathbf{B}) &= \left\{ \frac{\mu_v^{\mathrm{C}}\mathscr{T}_{vt}^{\mathrm{CA}}}{{R_{\mathrm{CA}}}^3} - \frac{\mu_{t'}^{\mathrm{A}}\mathscr{T}_{t'v}^{\mathrm{AC}}\alpha^{\mathrm{C}}\mathscr{T}_{vt}^{\mathrm{CA}}}{{R_{\mathrm{CA}}}^6} - \frac{\mu_{u'}^{\mathrm{B}}\mathscr{T}_{u'v}^{\mathrm{BC}}\alpha^{\mathrm{C}}\mathscr{T}_{vt}^{\mathrm{CA}}}{{R_{\mathrm{BC}}}^3{R_{\mathrm{CA}}}^3} \right. \\ &\quad \left. - \frac{\mu_w^{\mathrm{D}}\mathscr{T}_{wv}^{\mathrm{DC}}\alpha^{\mathrm{C}}\mathscr{T}_{vt}^{\mathrm{CA}}}{{R_{\mathrm{DC}}}^3{R_{\mathrm{CA}}}^3} \right\} \alpha^{\mathrm{A}} \left\{ \frac{\mathscr{T}_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}}}{{R_{\mathrm{AB}}}^3} \right\} \\ &= \frac{\mu_v^{\mathrm{C}}\mathscr{T}_{vt}^{\mathrm{CA}}\alpha^{\mathrm{A}}\mathscr{T}_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}}}{{R_{\mathrm{CA}}}^3{R_{\mathrm{AB}}}^3} - \frac{\mu_{t'}^{\mathrm{A}}\mathscr{T}_{t'v}^{\mathrm{AC}}\alpha^{\mathrm{C}}\mathscr{T}_{vt}^{\mathrm{CA}}\alpha^{\mathrm{A}}\mathscr{T}_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}}}{{R_{\mathrm{CA}}}^6{R_{\mathrm{AB}}}^3} \\ &\quad - \frac{\mu_{u'}^{\mathrm{B}}\mathscr{T}_{u'v}^{\mathrm{BC}}\alpha^{\mathrm{C}}\mathscr{T}_{vt}^{\mathrm{CA}}\alpha^{\mathrm{A}}\mathscr{T}_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}}}{{R_{\mathrm{BC}}}^3{R_{\mathrm{CA}}}^3{R_{\mathrm{AB}}}^3} \\ &\quad - \frac{\mu_w^{\mathrm{D}}\mathscr{T}_{wv}^{\mathrm{DC}}\alpha^{\mathrm{C}}\mathscr{T}_{vt}^{\mathrm{CA}}\alpha^{\mathrm{A}}\mathscr{T}_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}}}{{R_{\mathrm{DC}}}^3{R_{\mathrm{CA}}}^3{R_{\mathrm{AB}}}^3} \end{aligned} \tag{B.34}$$

In the final expression of eq (B.34), the first term is part of the three-body first-order induction, corresponding to the first term in eq (B.26). The next two terms then constitute part of the second-order correction to the three-body induction energy. Notably, the third term represents a cyclic coupling between three bodies, explicitly the coupling $B \to C \to A \to B$, which is unlike the previous three-body contributions. The last term is of significance: the first four-body contribution to the induction energy. All in all, the repeated-centre terms such as eq (B.33) contains the two-body and three-body contributions while the cross-centre terms such as eq (B.34) contains the three-body and four-body contributions.

Let us consolidate all the four-body contributions in the second-order induction, which are also the leading four-body contributions in many-body induction. By symmetry, we expect six of these four-body contributions in $E_{\mathrm{ind}}^{\mathrm{A}(2)}(\mathbf{B},\mathbf{C},\mathbf{D})$, one from each of the cross-centre terms, i.e, the last six terms in eq (B.32). We also need to consider $E_{\mathrm{ind}}^{\mathrm{B}(2)}(\mathbf{A},\mathbf{C},\mathbf{D})$, $E_{\mathrm{ind}}^{\mathrm{C}(2)}(\mathbf{A},\mathbf{B},\mathbf{D})$ and $E_{\mathrm{ind}}^{\mathrm{D}(2)}(\mathbf{A},\mathbf{B},\mathbf{C})$. Thus, there will be a total of $6 \times 4 = 24$ of such four-body contributions in the second-order induction energy. However, since the following expressions are equivalent,

$$\frac{\mu_w^{\mathrm{D}}\mathscr{T}_{wv}^{\mathrm{DC}}\alpha^{\mathrm{C}}\mathscr{T}_{vt}^{\mathrm{CA}}\alpha^{\mathrm{A}}\mathscr{T}_{tu}^{\mathrm{AB}}\mu_u^{\mathrm{B}}}{{R_{\mathrm{DC}}}^3{R_{\mathrm{CA}}}^3{R_{\mathrm{AB}}}^3} = \frac{\mu_u^{\mathrm{B}}\mathscr{T}_{ut}^{\mathrm{BA}}\alpha^{\mathrm{A}}\mathscr{T}_{tv}^{\mathrm{AC}}\alpha^{\mathrm{C}}\mathscr{T}_{vw}^{\mathrm{CD}}\mu_w^{\mathrm{D}}}{{R_{\mathrm{BA}}}^3{R_{\mathrm{AC}}}^3{R_{\mathrm{CD}}}^3} \tag{B.35}$$

half of the 24 four-body contributions are equivalent to the other half, giving only 12 unique terms, corresponding to the 12 non-branching paths. In the special case where all four bodies are identical, the dipoles and polarizabilities would be the same and the leading four-body contributions to the many-body induction is given by

$$\varepsilon_{4\mathrm{B}}^{\mathrm{tot}(2)} = \alpha^2\mu^2 \left(\frac{\mathscr{T}_{z'u}^{\mathrm{AB}}\mathscr{T}_{uv}^{\mathrm{BC}}\mathscr{T}_{vz}^{\mathrm{CD}}}{R_{\mathrm{AB}}{}^3 R_{\mathrm{BC}}{}^3 R_{\mathrm{CD}}{}^3} + \frac{\mathscr{T}_{z'u}^{\mathrm{AB}}\mathscr{T}_{uw}^{\mathrm{BD}}\mathscr{T}_{wz}^{\mathrm{DC}}}{R_{\mathrm{AB}}{}^3 R_{\mathrm{BD}}{}^3 R_{\mathrm{DC}}{}^3} + \frac{\mathscr{T}_{z'v}^{\mathrm{AC}}\mathscr{T}_{vu}^{\mathrm{CB}}\mathscr{T}_{uz}^{\mathrm{BD}}}{R_{\mathrm{AC}}{}^3 R_{\mathrm{CB}}{}^3 R_{\mathrm{BD}}{}^3} + \right.$$
$$\frac{\mathscr{T}_{z'v}^{\mathrm{AC}}\mathscr{T}_{vw}^{\mathrm{CD}}\mathscr{T}_{wz}^{\mathrm{DB}}}{R_{\mathrm{AC}}{}^3 R_{\mathrm{CD}}{}^3 R_{\mathrm{DB}}{}^3} + \frac{\mathscr{T}_{z'w}^{\mathrm{AD}}\mathscr{T}_{wu}^{\mathrm{DB}}\mathscr{T}_{uz}^{\mathrm{BC}}}{R_{\mathrm{AD}}{}^3 R_{\mathrm{DB}}{}^3 R_{\mathrm{BC}}{}^3} + \frac{\mathscr{T}_{z'w}^{\mathrm{AD}}\mathscr{T}_{wv}^{\mathrm{DC}}\mathscr{T}_{vz}^{\mathrm{CB}}}{R_{\mathrm{AD}}{}^3 R_{\mathrm{DC}}{}^3 R_{\mathrm{CB}}{}^3} +$$
$$\frac{\mathscr{T}_{z't}^{\mathrm{BA}}\mathscr{T}_{tv}^{\mathrm{AC}}\mathscr{T}_{vz}^{\mathrm{CD}}}{R_{\mathrm{BA}}{}^3 R_{\mathrm{AC}}{}^3 R_{\mathrm{CD}}{}^3} + \frac{\mathscr{T}_{z't}^{\mathrm{BA}}\mathscr{T}_{tw}^{\mathrm{AD}}\mathscr{T}_{wz}^{\mathrm{DC}}}{R_{\mathrm{BA}}{}^3 R_{\mathrm{AD}}{}^3 R_{\mathrm{DC}}{}^3} + \frac{\mathscr{T}_{z'v}^{\mathrm{BC}}\mathscr{T}_{vt}^{\mathrm{CA}}\mathscr{T}_{tz}^{\mathrm{AD}}}{R_{\mathrm{BC}}{}^3 R_{\mathrm{CA}}{}^3 R_{\mathrm{AD}}{}^3} +$$
$$\left. \frac{\mathscr{T}_{z'w}^{\mathrm{BD}}\mathscr{T}_{wt}^{\mathrm{DA}}\mathscr{T}_{tz}^{\mathrm{AC}}}{R_{\mathrm{BD}}{}^3 R_{\mathrm{DA}}{}^3 R_{\mathrm{AC}}{}^3} + \frac{\mathscr{T}_{z't}^{\mathrm{CA}}\mathscr{T}_{tu}^{\mathrm{AB}}\mathscr{T}_{uz}^{\mathrm{BD}}}{R_{\mathrm{CA}}{}^3 R_{\mathrm{AB}}{}^3 R_{\mathrm{BD}}{}^3} + \frac{\mathscr{T}_{z'u}^{\mathrm{CB}}\mathscr{T}_{ut}^{\mathrm{BA}}\mathscr{T}_{tz}^{\mathrm{AD}}}{R_{\mathrm{CB}}{}^3 R_{\mathrm{BA}}{}^3 R_{\mathrm{AD}}{}^3} \right) \tag{B.36}$$

B.2.4 Summary

Behaviour of Induction

A careful analysis revealed that the terms generated from the k-order correction to the induction energy, or simply the k-order induction, contain $(k+1)$ dipole-dipole interaction tensors, corresponding to a $L^{-3(k+1)}$ overall distance dependence, where L is one of the inter-body distances. This originate from there being k interaction tensors in the k-order induced dipole and one additional interaction tensor when the induced dipole is applied onto the permanent potential gradient. Since each interaction tensor connects only two bodies, $(k+1)$ interaction tensors can provide coupling between at most $(k+2)$ bodies. Thus, the k-order induction can involve at most $(k+2)$ bodies. Indeed, this is what we observe with the leading three-body terms arising from the first-order induction and having a L^{-6} overall distance dependence while the four-body counterparts emerge from second-order induction and have a L^{-9} overall dependence.

We also observe that two adjacent interaction tensors have to share a common centre, specifically at the body that is polarized. Mathematically, the polarizability of the common centre, say A, is sandwiched between two adjacent interaction tensors, in the form $T_{vt}^{\mathrm{CA}}\alpha^{\mathrm{A}}T_{tu}^{\mathrm{AB}}$. This results in a path linking all the bodies. More importantly, only the terminal of this path is allowed to interact with another body. This is evidenced in both eq (B.33) and eq (B.34) where the expanded $V_t^{\mathrm{A}(1)}(\mathbf{X})$ can only be interacted with at the terminal A. This implies that the path joining all the bodies contains no branching. This is how induction propagates.

It should be noted that the the terminal of the induction non-branching path can be connected to a previously induced body, giving rise to repeated instances of a particular interaction tensor, for example, the last term in eq (B.33). These terms constitutes the higher-order corrections.

Five-and-higher-body Interactions

Following the arguments in Section B.2.4, the five-body interaction would first appear in the third-order induction, having a L^{-12} overall distance dependence. The terms would be of the form

$$\frac{\mu_s^{\mathrm{E}} \mathscr{T}_{sw}^{\mathrm{ED}} \alpha^{\mathrm{D}} \mathscr{T}_{wv}^{\mathrm{DC}} \alpha^{\mathrm{C}} \mathscr{T}_{vt}^{\mathrm{CA}} \alpha^{\mathrm{A}} \mathscr{T}_{tu}^{\mathrm{AB}} \mu_u^{\mathrm{B}}}{R_{\mathrm{ED}}{}^3 R_{\mathrm{DC}}{}^3 R_{\mathrm{CA}}{}^3 R_{\mathrm{AB}}{}^3} \tag{B.37}$$

Due to the distance dependence, the five-body interaction is often very weak and negligible in most if not all systems. For example, in water, we would have $\mu = 0.8$ a.u. and $\alpha = 10$ a.u. while the $\mathscr{T}^{\mathrm{AB}}$ can only take a maximum value of 2. Assuming the typical O–O distance in hydrogen bonded water molecules ($R = 5.2$ a.u. ≈ 2.75 Å), eq (B.37) evaluates to be 0.026 m-E_{h} or 0.068 kJ mol^{-1}. Due to the R^{-12} overall dependence, this five-body interaction diminishes rapidly with distance. By increase the distance to be $R = 6$ a.u. (about 3.18 Å), the energy drops to 0.004 m-E_{h} or 0.010 kJ mol^{-1}, making the five-body interaction almost non-existent! We can extend this analysis further to six-body interactions. The L^{-15} overall distance dependence would render the six-body interactions to be insignificant in all systems.

Thus, we conclude that many-body inductive effects propagate in non-branching paths. The k-order induction terms contain $(k+1)$ dipole-dipole interaction tensors, corresponding to a $L^{-3(k+1)}$ overall distance dependence. From this, we also deduced that five-body induction interaction are very likely to be negligible while higher-body effects can be completely ignored.

C | CD CONTENTS AND SUPPORTING PUBLICATIONS

C.1 CD Contents

The following files were attached in the CD accompanying this thesis.

Filename	Description
thesis.pdf	PDF containing a digital copy of this thesis
chp3xyz.pdf	PDF containing the Cartesian coordinates of the molecules studied in Chapter 3
chp4xyz.pdf	PDF containing the Cartesian coordinates of the molecules studied in Chapter 4
chp5xyz.pdf	PDF containing the Cartesian coordinates of the molecules studied in Chapter 5

C.2 Supporting Publications

Here, we attach the following publications which has been covered in this thesis.

1) Ouyang, J.F.; Cvitkovic, M.W.; Bettens, R.P.A. *J. Chem. Theory Comput.* **2014,** *10,* 3699–3707.

2) Ouyang, J.F.; Bettens, R.P.A. *Chimia* **2015,** *69,* 104–111.

3) Ouyang, J.F.; Bettens, R.P.A. *J. Chem. Theory Comput.* **2015,** *11,* 5132–5143.

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

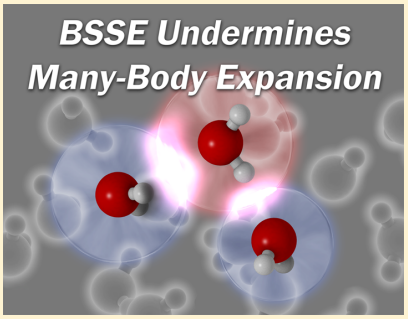
Trouble with the Many-Body Expansion

John F. Ouyang, Milan W. Cvitkovic, and Ryan P. A. Bettens*

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

Supporting Information

ABSTRACT: Longstanding conventional wisdom dictates that the widely used Many-Body Expansion (MBE) converges rapidly by the four-body term when applied to large chemical systems. We have found, however, that this is not true for calculations using many common, moderate-sized basis sets such as 6-311++G** and aug-cc-pVDZ. Energy calculations performed on water clusters using these basis sets showed a deceptively small error when the MBE was truncated at the three-body level, while inclusion of four- and five-body contributions drastically increased the error. Moreover, the error per monomer increases with system size, showing that the MBE is unsuitable to apply to large chemical systems when using these basis sets. Through a systematic study, we identified the cause of the poor MBE convergence to be a many-body basis set superposition effect exacerbated by diffuse functions. This was verified by analysis of MO coefficients and the behavior of the MBE with increasing monomer−monomer separation. We also found poor convergence of the MBE when applied to valence-bonded systems, which has implications for molecular fragmentation methods. The findings in this work suggest that calculations involving the MBE must be performed using the full-cluster basis set, using basis sets without diffuse functions, or using a basis set of at least aug-cc-pVTZ quality.



1. INTRODUCTION

The many-body expansion (MBE) is a useful and ubiquitous formalism in the theoretical study of large chemical systems.[1−12] The MBE expresses the total energy, E_{tot}, of an n-body system as the sum of one-body, two-body, etc., up to n-body energy contributions (see next section for a detailed description). Calculating E_{tot} directly for large systems is often computationally unaffordable. The benefit of the MBE is that for many systems E_{tot} can be well approximated by truncating the expansion to just the first few terms. Truncated MBEs have found especially widespread use in the study of water clusters, in which most intermolecular interactions are assumed to be pairwise additive (i.e., completely captured in an MBE truncated after the two-body term). The remaining (mostly inductive) interaction energy is accounted for by the rest of the terms in the MBE.

A longstanding and crucial question in modeling aqueous systems is how many terms of the MBE are necessary to adequately approximate the total energy. The earliest studies addressing this question were performed on water dimers, trimers, and tetramers. They found that three-body effects accounted for about 10% of the interaction energy, and that four- and higher-body effects were negligible.[13,14] Subsequent work on slightly larger clusters agreed that four- and higher-body energy contributions were minute.[15−20] The most thorough examination of many-body effects was performed on water hexamers by Xantheas in 1994, in which he found "the contribution from four-body and higher terms to be negligible for these systems."[21]

The results from these studies eventually coalesced into an oft-cited piece of conventional wisdom: that the many-body expansion for water converges rapidly by the four-body term and in a well behaved manner.[22−24] Indeed, most current ab-initio-based simulation models use MBEs truncated at three or occasionally four bodies.[25−28]

Despite this, in the course of refining our group's Combined Fragmentation Method (CFM)[11,12] for use with noncovalent systems, we recently decided to verify the rapid convergence of the MBE for a few water clusters. We calculated all the terms in the MBEs of four $(H_2O)_{16}$ clusters with HF/6-31++G**, expecting, per conventional wisdom, to observe convergence to the true cluster energy by at most the five-body term (convergence herein defined as consistently having an error less than 1 m-E_h). Instead, not only did these MBEs not converge by anywhere near the five-body term, the convergence was notably erratic (Figure 1). Particularly worrying was that while truncating the many-body expansion at the four-body term led to a decent result (3.2−4.2 m-E_h error for 4444-a,c1b,cie), inclusion of the five-body term *increased* the error (4.6−4.7 m-E_h error for 4444-a,c1b,cie) rather than further converging the MBE toward the true cluster energy.

We are by no means the first to observe problems with the MBE.[19,31−33] To our knowledge, however, there has been no thorough examination of under what circumstances the MBE fails to converge rapidly. In the following sections, we will demonstrate the wide extent of the MBE convergence problem

Received: May 7, 2014
Published: June 12, 2014

Journal of Chemical Theory and Computation Article

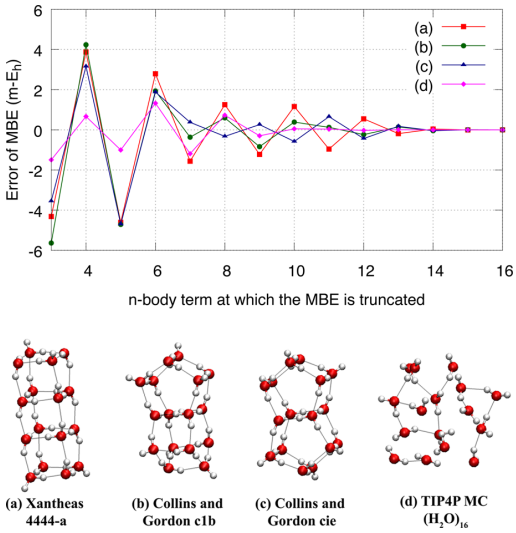


Figure 1. Slow, erratic convergence of the MBE toward the full-cluster energy for four $(H_2O)_{16}$ clusters calculated at HF/6-31++G**. The linear fused cube 4444-a (a) was obtained from Yoo et al.,[29] while the fused pentameric structures c1b (b) and cie (c) were obtained from Pruitt et al.[30] The TIP4P MC structure (d) was obtained by taking a random fragment of 16 water molecules from a TIP4P Monte Carlo simulation of 400 water molecules.

and show that a many-body Basis Set Superposition Effect (BSSE) is its cause. We stress that we do not question the accuracy or validity of the many methods that use many-body expansions, or question previous studies of many-body effects in water. Our purpose in this work is simply to refine the conventional wisdom about the MBE.

2. COMPUTATIONAL DETAILS

For a chemical system comprising n monomers, the MBE of the total energy of the system, E_{tot}, is the finite sum

$$E_{tot} = \sum_{k=1}^{n} \varepsilon^{(n,k)} \tag{1}$$

where $\varepsilon^{(n,k)}$ is the total k-body energy of the system. The total k-body energy of the system is the component of the total energy due to all k-body effects, so $\varepsilon^{(n,1)}$ is the sum of the energies of all isolated monomers; $\varepsilon^{(n,2)}$ is the sum of the energies of all dimers minus the energies of the monomers they comprise, i.e., it is the sum of all the pairwise interaction energies. Thus, the total interaction energy of the system is given by

$$\varepsilon_{tot\backslash 1} = E_{tot} - \varepsilon^{(n,1)} \tag{2}$$

For example, consider a system of three molecules, A, B, and C. The one-body energy is

$$\begin{aligned}\varepsilon^{(3,1)} &= \varepsilon_A^{(3,1)} + \varepsilon_B^{(3,1)} + \varepsilon_C^{(3,1)} \\ &= E_A + E_B + E_C\end{aligned} \tag{3}$$

the two-body energy is

$$\begin{aligned}\varepsilon^{(3,2)} &= \varepsilon_{AB}^{(3,2)} + \varepsilon_{AC}^{(3,2)} + \varepsilon_{BC}^{(3,2)} \\ &= E_{AB} - (E_A + E_B) + E_{AC} - (E_A + E_C) + E_{BC} \\ &\quad - (E_B + E_C)\end{aligned} \tag{4}$$

and the three-body energy is

$$\begin{aligned}\varepsilon^{(3,3)} &= \varepsilon_{ABC}^{(3,3)} \\ &= E_{tot} - (E_{AB} - (E_A + E_B) + E_{AC} - (E_A + E_C) \\ &\quad + E_{BC} - (E_B + E_C)) - (E_A + E_B + E_C)\end{aligned} \tag{5}$$

where in all the above, $\varepsilon_\alpha^{(3,k)}$ and E_α are the k-body energy and the total energy respectively of the subsystem α of the trimer. For a more detailed explanation of the MBE, see Xantheas.[21]

$\varepsilon^{(n,k)}$ can also be expressed recursively in terms of the total energy and lower-body energies[34]

$$\varepsilon^{(n,k)} = \sum_\alpha E_\alpha - \sum_{i=1}^{k-1} \left[\frac{(n-i)!}{(n-k)!(k-i)!} \right] \varepsilon^{(n,i)} \tag{6}$$

where E_α is the total energy of the k-mer subsystem α of which there are $\binom{n}{k}$.

We also wish to clarify our use of the term basis set superposition effect. When small basis sets are used in ab initio calculations of molecular clusters, basis functions from one molecule can be utilized by other molecules to compensate for the incompleteness of their basis set. This results in an improved description of the wave function of all molecules in the cluster, which leads to a lowering of the total energy known as the BSSE. One way of quantifying the BSSE is the counterpoise (CP) method.[35] In the CP method, the familiar expression for the BSSE in the interaction energy of two molecules, A and B, is given as

$$E_{BSSE} = (E_A(\mathbf{ab}) - E_A(\mathbf{a})) + (E_B(\mathbf{ab}) - E_B(\mathbf{b})) \tag{7}$$

where **a**, **b**, and **ab** are the basis sets of molecule A, molecule B, and the cluster AB, respectively. Applying the CP method to the interaction energy of A and B, the BSSE-free interaction energy is

$$\begin{aligned}\varepsilon_{tot\backslash 1}^{CP} &= \varepsilon_{tot\backslash 1} - E_{BSSE} \\ &= E_{AB}(\mathbf{ab}) - (E_A(\mathbf{a}) + E_B(\mathbf{b})) - E_{BSSE} \\ &= E_{AB}(\mathbf{ab}) - (E_A(\mathbf{ab}) + E_B(\mathbf{ab}))\end{aligned} \tag{8}$$

Now, all the quantities are calculated consistently in the same basis, namely the basis set of the cluster AB. The brilliance of the CP method lies in that it does not try to remove the lowering of energy in the total energy of the cluster AB due to sharing of basis functions, which is a natural consequence of the variational principle. Instead, it does the opposite where the constituents A and B are calculated in the basis set of the cluster AB so as to achieve a similar lowering of energy.

In the spirit of the CP method, we define the BSSE for the sum of the total energies of all $\binom{n}{k}$ k-mers in the cluster containing n monomers as

$$E_{BSSE}^{(n,k)} = E^{(n,k)}(\boldsymbol{n}\textbf{-mer}) - E^{(n,k)}(\boldsymbol{k}\textbf{-mer}) \tag{9}$$

where ***n*-mer** and ***k*-mer** are the basis set of the full molecular cluster (cluster basis) and the basis set of the k monomers considered (k-mer basis), respectively. Using our definition, the

removal of the BSSE in the total energies when performing a many-body energy decomposition (such as eqs 4 and 5) will result in all the quantities being calculated consistently in the cluster basis, ensuring that the MBE remains formally exact. Indeed, the use of a consistent cluster basis has been employed previously[21,36,37] to obtain BSSE-free many-body energies. From eq 9, BSSE can be seen as a lowering of the total energy of the k-mer in the cluster due to the sharing of basis functions from the remaining $n - k$ monomers. Thus, when $k = n$, there is no BSSE, i.e., $E_{\mathrm{BSSE}}^{(n,n)} = 0$. Notably, this definition of BSSE reduces to the familiar expression in eq 7 in the context of the interaction energy of a cluster where $E_{\mathrm{BSSE}} = E_{\mathrm{BSSE}}^{(n,1)}$.

All quantum chemical calculations were performed using the Gaussian 09 package[38] or the MOLPRO suite of programs[39] at the Hartree−Fock (HF) or second-order Møller−Plesset perturbation (MP2) level of theory. A variety of Pople split-valence basis sets were used, along with the series of Dunning correlation-consistent cc-pVXZ basis sets, X = 2−4, labeled VDZ, VTZ, and VQZ. An "A" or "dA" prepended to these basis sets indicate they are augmented or doubly augmented, respectively, with diffuse functions.

3. RESULTS AND DISCUSSION

3.1. Extent of the Poor Convergence of the MBE. Prompted by our initial results (Figure 1), we attempted to ascertain the extent of the MBE convergence problem. We calculated MBEs up to the five-body term for a variety of $(H_2O)_n$, n between 6 and 57, geometries (Table 1). Some geometries are optimized structures from the literature; others were taken from TIP4P Monte Carlo simulations. These latter structures were included as they are representative of geometries encountered in simulations using truncated-MBE-based water models. All calculations were performed using the AVDZ basis set, which is a better yet still computationally manageable basis set compared to the 6-31++G** basis used in Figure 1. The MBEs in Table 1 were only calculated to at most the five-body term due to the steep computational cost of calculating high-body terms for large clusters: the number of additional calculations required to obtain the k-body energy of an n-body system is $(n!)/(k!(n - k)!)$.

Table 1 shows that the MBEs of very small clusters (n = 6−8) indeed converge by the three-body term, as shown in previous studies.[21] But for larger clusters, while MBEs truncated at the three-body term appear converged, inclusion of four- and five-body energies *increases* the error in the MBE. A notable oscillatory behavior also occurs wherein the error changes sign when three- and higher-body energies are included.

More alarmingly, Figure 2 shows that for the clusters studied in Table 1, the four- and five-body MBE truncation errors per monomer increase with system size. This was also noticed previously in a smaller sample of water clusters by Kulkarni et al., who called for further examination.[19] This is concerning as the error-per-monomer should be an intensive, not extensive, property. Otherwise, the scalability of truncated-MBE-based computational methods, such as fragmentation methods and bulk material simulations, becomes questionable.

Table 1. Error (m-E_h) in the Total Energy of Water Clusters $(H_2O)_n$, n = 6−57, As Approximated by an MBE Truncated after the Two- through Five-Body Term[a]

$(H_2O)_n$	two-body	three-body	four-body	five-body
6[b]	−9.769	−0.488	−0.006	0.017
8[b]	−16.905	−0.426	0.491	−0.518
10[b]	−23.227	−1.444	1.092	−1.099
12[b]	−26.745	−0.399	1.519	−2.541
14[b]	−27.613	−1.302	2.478	−3.563
16[b]	−37.045	−0.631	2.940	−5.200
18[b]	−37.060	−0.563	3.227	−5.848
20[b]	−46.844	−0.741	4.219	−8.101
24[c]	−106.289	−10.762	3.252	
32[d]	−130.794	−6.612	16.895	
57[e]	−97.126	−5.401	27.334	
6[f]	−1.674	−0.024	0.022	−0.009
8[f]	−2.591	−0.072	0.072	−0.023
10[f]	−1.704	0.041	0.157	−0.098
12[f]	0.189	0.266	0.274	−0.344
14[f]	−6.870	−0.003	0.391	−0.495
16[f]	−11.364	−0.378	1.153	−1.555
18[f]	−6.524	−0.191	1.382	−2.009
20[f]	−14.251	−0.721	1.692	−2.910
45[f]	−36.931	−2.117	8.373	

[a]All calculations done at HF/AVDZ level of theory. [b]Optimized water clusters obtained from Maheshwary et al.[40] [c]Optimized water clusters obtained from Gora et al.[24] [d]Optimized water clusters obtained from Pruitt et al.[30] [e]Optimized water clusters obtained from Richard and Herbert.[41] [f]Disordered random fragments of $(H_2O)_n$ obtained from a TIP4P Monte Carlo simulations. For these disordered fragments, the MBE truncation errors were averaged over four different random fragments for n = 6 − 20.

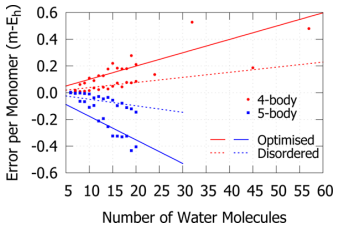


Figure 2. Error-per-monomer of the MBE truncated at the four-body level (red solid circles) and at the five-body level (blue solid squares) for optimized water clusters (solid lines) and disordered water clusters (dashed lines) from Table 1.

3.2. Cause of the Poor Convergence of the MBE. Having found the MBE convergence problem to be widespread, we sought to determine its cause. Initially, we thought the poor convergence was due to the MBE not properly capturing the many-body induction energy. Since induction energy is closely related to the polarizability of the molecules in a system, we performed MBE calculations for the optimized clusters studied in Table 1 using a series of basis sets of increasing polarizability (Table 2). If the poor MBE convergence were due to induction, the error for truncating the MBE should worsen with increasing polarizability. It is clear from Table 2 that this is not the case.

Instead, poor MBE convergence only occurred when using small, incomplete basis sets augmented with diffuse functions, namely 6-311++G** (P3) and AVDZ (D). This led us to suspect that the convergence problem was due to BSSE. As a preliminary test of this, the MBEs in Table 2 were recalculated using the cluster basis, as opposed to the usual k-mer basis, in all k-mer calculations. This eliminated the BSSE in the MBE calculations as explained in section 2. As shown in Table 3, the poor MBE convergence observed in Table 2 disappeared when

Table 2. Error in Approximating the Total Energy of Optimized Water Clusters $(H_2O)_n$, n = 8−20, with an MBE Truncated at the Four-Body (4B) and Five-Body (5B) Term Using Basis Sets of Increasing Isotropic Dipole−Dipole Polarizability $\bar{\alpha}$

		basis set					
		P1[a]	P2[a]	P3[a]	D[b]	T[b]	Q[b]
$\bar{\alpha}$ (a.u.)		3.74	4.87	6.50	7.97	8.23	8.30
$(H_2O)_n$		error (m-E_h)					
8	4B	0.01	0.07	1.24	0.49	0.05	0.01
	5B	0.02	0.01	−0.67	−0.52	−0.04	0.00
10	4B	−0.06	−0.01	0.91	1.09	0.04	−0.05
	5B	0.04	0.01	−0.65	−1.10	−0.09	−0.02
12	4B	0.42	0.21	1.13	1.52	0.01	
	5B	0.01	−0.02	−0.55	−2.54	−0.36	
14	4B	0.30	0.25	1.57	2.48	0.19	
	5B	0.06	0.02	−1.12	−3.56	−0.47	
16	4B	0.47	0.32	2.82	2.94	0.00	
	5B	0.11	0.02	−1.75	−5.20	−0.66	
18	4B	0.82	0.38	1.32	3.23		
	5B	0.03	−0.02	−0.66	−5.85		
20	4B	0.98	0.48	2.00	4.22		
	5B	0.03	−0.04	−1.15	−8.10		

[a]All calculations performed at the HF level. Water geometries are from Maheshwary et al.[40] (the same geometries as used in Table 1). Pople basis set P1: 3-21G, P2: 6-31G**, and P3: 6-311+G(2d,p). [b]Dunning basis set AVXZ where X = D or T or Q.

Table 3. Error for Truncating the MBE at the Four-Body (4B) and Five-Body (5B) Term Using the Cluster Basis, As Opposed to the k-mer Basis Used in Table 2, in All k-Body Calculations[a]

		basis set				
		P1[b]	P2[b]	P3[b]	D[c]	T[c]
$\bar{\alpha}$ (a.u.)		3.74	4.87	6.50	7.97	8.23
$(H_2O)_n$		error (m-E_h)				
8	4B	−0.07	−0.02	0.01	0.04	0.04
	5B	0.01	0.01	0.00	0.00	0.00
10	4B	−0.05	−0.03	−0.02	−0.02	−0.01
	5B	0.00	0.00	−0.01	−0.01	−0.01
12	4B	−0.01	0.00	−0.01	0.00	
	5B	0.00	−0.01	−0.02	−0.03	
14	4B	0.01	0.02	0.03	0.05	
	5B	0.00	0.00	−0.01	−0.01	
16	4B	−0.10	−0.05	−0.02	0.02	
	5B	0.03	0.02	0.00	−0.01	
18	4B	−0.03				
	5B	0.00				
20	4B	−0.03				
	5B	0.00				

[a]Results shown for a series of optimized water clusters $(H_2O)_n$, n = 8−20, at the HF level using various basis set of increasing isotropic dipole−dipole polarizability $\bar{\alpha}$. [b]Pople basis set P1: 3-21G, P2: 6-31G**, and P3: 6-311+G(2d,p). [c]Dunning basis set AVXZ where X = D or T.

BSSE was removed. (Note that not all terms were recalculated due to the computational cost of using the full-cluster basis.) In fact, the BSSE present in the many-body energies can be easily computed as the difference between the errors in both tables.

To verify that BSSE was the cause of the poor MBE convergence, we calculated the full MBEs for two $(H_2O)_{10}$ clusters, **10PP** and **10OB** (Figure 3), using basis sets of

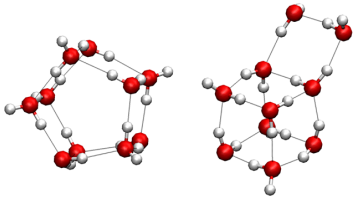


Figure 3. $(H_2O)_{10}$ clusters chosen for a more detailed study on the cause of poor MBE convergence. Both the pentagonal prism (**10PP**) and open box (**10OB**) were obtained from Maheshwary et al.[40]

increasing quality and diffusiveness and using both k-mer bases and cluster bases. MP2 calculations were also performed for the 6-31G** and the 6-311G** series to investigate the effects of electron correlation on the MBE convergence. As the results were similar for both **10PP** and **10OB** clusters, only the errors of the MBE for **10PP** are presented in Figure 4. Results for **10OB** are in the Supporting Information.

It is clear from Figure 4 that when the k-mer basis is used (solid lines), the more diffuse functions that are present, the worse the MBE convergence (red and orange solid lines), with the exception of AVTZ and AVQZ. This is all precisely what one would expect if BSSE were the cause of the poor MBE convergence: more diffuse functions lead to more overlap of basis functions between water molecules, increasing BSSE, except for basis sets like AVTZ and AVQZ which are so complete that water monomers need not rely on diffuse functions from their neighbors to describe their wave functions. Indeed, Truhlar and co-workers have made a similar observation by examining the effects of increasing augmentation in the Dunning basis sets.[42−44] It should be noted that there are still tiny oscillations in the MBE truncation errors for AVTZ and AVQZ in the range of 10−50 μ-E_h, which are hard to see in the figure. Moreover, when the cluster basis is used (dashed lines) and BSSE is eliminated, the MBE converges by the four-body term regardless of the presence of diffuse functions. Figure 4c−f further show that when electron correlation is included, the MBE errors are amplified. This can be attributed to additional BSSE associated with electron correlation—it is known that correlation energy converges more slowly toward the complete basis set limit than the SCF energy.[45]

We also examined the MO coefficients in these calculations to specifically check whether the BSSE originated from the diffuse functions. HF calculations were performed using VDZ, AVDZ, and AVTZ basis sets on an arbitrarily chosen monomer from **10PP** with the ghost basis functions of all other waters in the cluster present. (The choice of monomer does not significantly affect the results due to the symmetry of the cluster.) The distribution of MO coefficients for the occupied MOs is shown in Figure 5. By performing calculations on a single monomer in the cluster basis, all observations are solely due to BSSE and not physical interaction between molecules. If diffuse functions were causing the BSSE—that is, if water molecules were using diffuse basis functions centered on other molecules to improve the description of their own wave

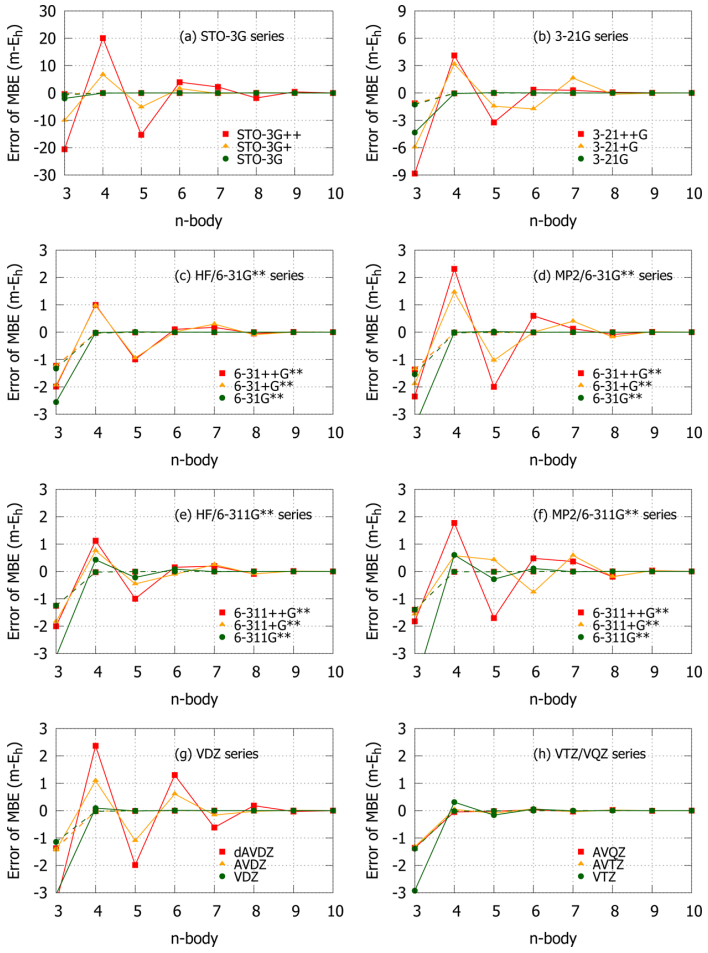


Figure 4. Convergence of the MBE for **10PP** using various basis sets: (a) STO-3G series, (b) 3-21G series, (c) HF/6-31G** series, (d) MP2/6-31G** series, (e) HF/6-311G** series, (f) MP2/6-311G** series, (g) VDZ series, and (h) VTZ/VQZ series. Solid lines represent MBE calculated using the *k*-mer basis, while dashed lines represent MBEs calculated using the cluster basis. It should be noted that the diffuse functions of the 6-31++G** basis set were used as the diffuse functions for the STO-3G and 3-21G basis sets, as these basis sets have no defined diffuse functions.

function—then there should be significant MO coefficients for basis functions centered on the ghost molecules. Similarly, a many-body BSSE effect can be inferred if there are significant nonzero MO coefficients arising from *many* of these ghost molecules simultaneously.

For VDZ (Figure 5a), significant nonzero MO coefficients, represented by red or blue colored regions, are only found for a few water molecules' basis functions. In contrast, the AVDZ basis set (Figure 5b) has significant nonzero MO coefficients on all the water molecules' basis functions. As the ghost water molecules in Figure 5 are ordered by their proximity to the monomer under study, the colored regions become fainter across the horizontal axis due to decreasing overlap of the basis functions from more distant ghosts. Nonetheless, the nonzero coefficients imply that the BSSE is many-body in nature, with contributions from all monomers in the system. The contributions come primarily from the diffuse functions (denoted D in the figure) of both oxygen and hydrogen, again implicating diffuse functions in causing the BSSE. The MO coefficient distribution for AVTZ (Figure 5c) also shows contributions from diffuse functions, but less so than those for AVDZ. Again, this is due to AVTZ being a more complete basis set: the wave function of the monomer can be described using its own core, valence, and diffuse basis functions without the need for the basis functions of its neighbors. In fact, AVTZ's BSSE contribution to the total energy of the monomer is low at 2.7 ppm of the total energy of the monomer (−0.203 m-E_h) in contrast to the higher contribution from both VDZ (68 ppm, −5.197 m-E_h) and AVDZ (10 ppm, −0.768 m-E_h).

As a final test of our hypothesis, we investigated how MBE convergence is affected by the average nearest-neighbor distances of water molecules in a cluster. HF/AVDZ MBE calculations using the *k*-mer basis were performed on a series of progressively expanded structures derived from **10PP** (Figure 6). The expanded structures were constructed by scaling the distance between the center-of-mass of each water and the center-of-mass of the entire **10PP** cluster. This ensures that the nearest-neighbor distances of all water molecules are increased

Journal of Chemical Theory and Computation Article

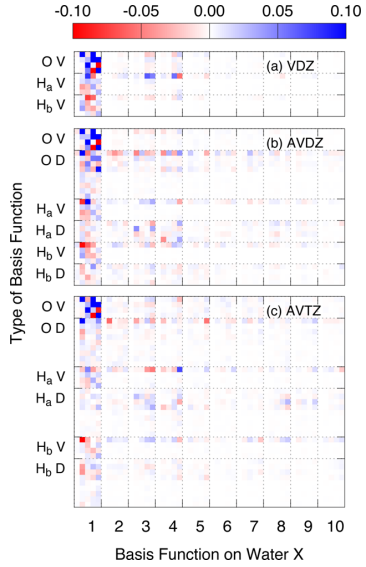


Figure 5. Distribution of MO coefficients of an arbitrarily chosen monomer of **10PP** calculated with the cluster basis using (a) VDZ, (b) AVDZ, and (c) AVTZ basis sets. The vertical axis shows the basis functions arranged according to the nuclei (O or H). Only valence functions (denoted V), functions with the smallest exponent, and diffuse functions (denoted D) are shown. The horizontal axis shows which water molecule the basis functions are centered on: Water 1 is the monomer under study, and the rest are ghost molecules.

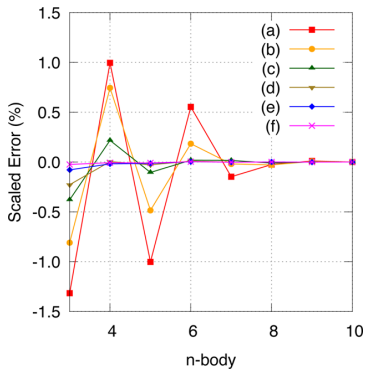


Figure 6. Convergence of the MBE for expanded structures derived from **10PP** with the following mean nearest-neighbor distance: (a) 5.25 a_0, (b) 5.67 a_0, (c) 6.61 a_0, (d) 7.56 a_0, (e) 8.50 a_0, and (f) 9.45 a_0. The MBE truncation error has been normalized to the one-body error (i.e., the interaction energy of the cluster).

by the same factor. As the mean nearest-neighbor distance increases, the oscillations in the MBE error gradually disappear. This is because when the waters are farther apart, the overlap between diffuse functions on different waters decreases exponentially and so does the BSSE. This can be seen explicitly in Figure 7, where an exponential fit captures the decay of BSSE with increasing interwater-molecule distance.

The curious reader may wonder why the error oscillates from positive to negative in nearly all the poorly convergent MBEs we have shown. Our best explanation is that this behavior is related to the inclusion/exclusion principle inherent in the

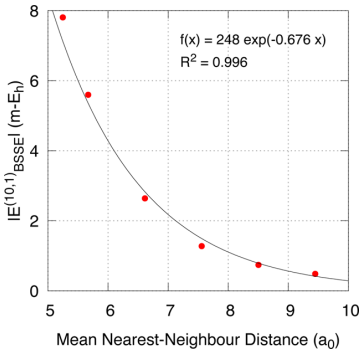


Figure 7. Magnitude of the BSSE in the interaction energy $|E^{(10,1)}_{\mathrm{BSSE}}|$ of the expanded structures derived from **10PP** shown in Figure 6 shows an exponential decay with increasing mean nearest-neighbor distance.

MBE. To obtain a system's k-body energy, $\varepsilon^{(k)}$, the total energy of each k-mer in the system has the total energy of all its constituent $(k-1)$-mers subtracted from it. But this results in oversubtraction of $(k-2)$-body energies, so the $(k-2)$-mer total energies have to be added back, and so on (see again eq 4). When subsequently obtaining the $(k+1)$-body energy, the signs of the terms in the expression switch: k-body energies are subtracted where they were previously added etc. (in addition to there being many more terms in the calculation). So if a particular k-body energy is underestimated—perhaps due to an inadequate basis set—it will contribute to the overestimation of the $(k+1)$-body energy and then the underestimation of the $(k+2)$-body energy and so on, leading to oscillation in the error. Of course, the MBE by definition converges by the final term. And each subsequent term in the MBE contains less BSSE-derived error, $E^{(n,k+1)}_{\mathrm{BSSE}} < E^{(n,k)}_{\mathrm{BSSE}}$, since the more water monomers that are in a calculation, the closer the basis set is to the correct, full-cluster basis. But in the early MBE terms, there is much utilization of neighboring waters' diffuse basis functions, and thus more error in each calculated energy, and thus oscillations which diminish as more terms are added.

3.3. Methods to Improve MBE Convergence. We have shown that rapid convergence of the MBE can be guaranteed by performing calculations in the cluster basis or with a high-quality basis set. These are, of course, fairly dispiriting solutions as they greatly increase computational cost, so we have examined several alternatives.

Since basis function overlap is distance-dependent, we tested a distance-cutoff basis (d-c basis) which includes ghost functions only from waters within a specified cutoff distance of the water molecules in a calculation. Testing this d-c basis on the $(H_2O)_{16}$ 4444-a cluster using various cutoff distances gave poor results (Table 4), however. This is likely due to the small MBE truncation errors involved in the m-E_h range. With slight changes in cutoff distances, one more or one fewer water's ghost functions might be included in a calculation, which could lead to significant changes in calculated energies, drastically affecting the MBE truncation errors. That said, we think the distance-cutoff basis might work with a larger cutoff distance, but in those cases it would be more economical to use a high-quality basis set instead.

Another workaround that has been proposed, albeit for a different problem, is the k-mer-centerd basis set (kCBS) approach of Gora et al.[24] The kCBS approach attempts to remove BSSE in an MBE calculation by calculating each k-mer

dx.doi.org/10.1021/ct500396b | J. Chem. Theory Comput. 2014, 10, 3699−3707

Table 4. Error for Truncating the MBE up to the Five-Body Term, Performed Using Various Methods to Improve MBE Convergence[a]

	error (m-E_h)			
method	two-body	three-body	four-body	five-body
k-mer basis	−42.418	−4.763	4.588	−5.563
cluster basis	−56.170	−1.217	−0.043	−0.056
d-c basis, 3 Å	−35.325	−28.202	10.527	9.549
d-c basis, 4 Å	−41.468	−46.804	85.769	−108.259
d-c basis, 5 Å	−48.275	−34.555	87.929	−161.926
*k*CBS	−50.464	4.779	6.268	6.279
charge field	−12.466	−0.449	2.073	−2.054

[a]Calculations were performed at HF/6-31++G** on the $(H_2O)_{16}$ 4444-a cluster shown in Figure 1. *k*-mer basis and cluster basis are shown for reference.

and all its subcalculations using the *k*-mer's basis set. That is, a dimer's two-body contribution would be computed as total energy of the dimer minus the total energy of its constituent monomers, all calculated with the dimer basis set. This results in substantially more calculations to compute the MBE since calculations from previous terms cannot be reused, but it does mean each calculation has no BSSE. We applied the *k*CBS approach to the $(H_2O)_{16}$ 4444-a cluster. From Table 4, we see that the MBE does converge rapidly, but to an incorrect value. This is likely because the *k*CBS approximation is not formally exact: the terms in the MBE do not cancel due to the different numbers of basis functions used in each term's calculations. The *k*CBS approach certainly does converge correctly when a high-quality basis set is used, as has been demonstrated in the literature, but this seems to be due to the high quality of the basis set, not the *k*CBS method.

Strategies unrelated to BSSE for improving MBE convergence are widely used. Many truncated-MBE-based computational methods incorporate a charge field to approximate higher-order many-body effects, typically by interacting the one- or two-body fragments with a charge field representing the rest of the system.[2−4,6,10,46,47] While we have not done a thorough analysis, preliminary results using embedded charges from Stone's distributed multipole analysis[48] indicate that embedded charges dampen, but do not remove, the oscillatory MBE behavior (Table 4). This is not surprising as the embedded charges only serve to approximate the *physical* higher-order many-body effects arising from induction and thus do not remove the many-body BSSE.

Other methods include high-order many-body effects by performing a low-level ab initio calculation on the full system.[3,4,49,50] Such methods capture many-body effects far better than methods using only a truncated MBE.[51] The full-system calculations in these methods are not susceptible to BSSE-based MBE convergence issues since they use full-system basis, but lower-body calculations performed using only the *k*-mer basis are still susceptible.

Thus, unhappily, we have found no alternative for avoiding poor MBE convergence that is more efficient than using the full-cluster basis or a high-quality basis set. As the use of the cluster basis is computationally prohibitive, our recommendation is to use a high-quality basis set for MBE calculations; our results indicate that at least AVTZ-quality is prudent.

3.4. Extension to Valence-Bonded Systems. So far we have only presented data for noncovalent water clusters, but MBE convergence problems also arise in valence-bonded systems. This has great implications for fragmentation methods, which in most cases use a truncated MBE, or something analogous to it, to approximate the total energies of large chemical systems.[41,52−54]

In fragmentation methods, small groups of adjacent atoms are treated as bodies. Using our CFM algorithm[11] to define groups/bodies, we calculated the MBEs for a 22-carbon $C_{22}H_{24}$ conjugated alkene and α-cyclodextrin (Figure 8). Slow MBE convergence is observed in both systems when incomplete basis sets with diffuse functions are used, as seen in the case of AVDZ for $C_{22}H_{24}$ and 6-31++G** for α-cyclodextrin. This is no surprise as the same borrowing of basis functions from adjacent groups that causes poor convergence in water clusters occurs in these valence-bonded systems. The errors in the $C_{22}H_{24}$ MBE are small (even negligible) because the molecule's linear shape minimizes basis function overlap. Compare this to the MBE of the more compact α-cyclodextrin, where the errors are beyond chemical accuracy until the inclusion of the nine-body term. Since fragmentation methods rarely include five-or-higher-body effects, it seems likely that fragmentation calculations using BSSE-prone basis sets are liable to, and have in the past been afflicted by, preventable, BSSE-based errors. On an interesting related point, due to the above-mentioned affects, we expect that any calculation that is performed in order to predict bond-breaking energies would be overestimated.

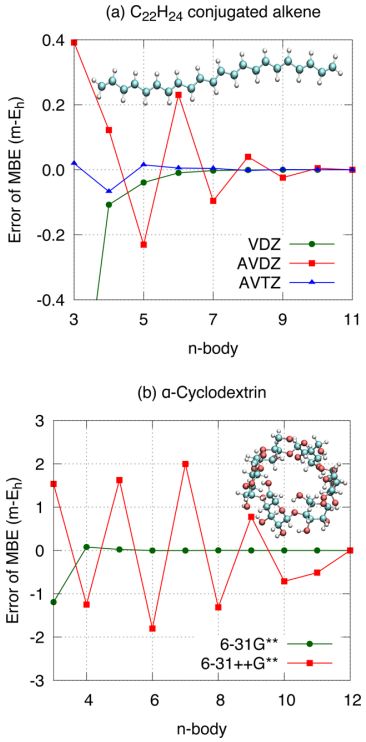


Figure 8. Convergence of the MBE for the total energy of (a) $C_{22}H_{24}$ conjugated alkene and (b) α-cyclodextrin at HF level of theory for various basis sets. Here, the CFM algorithm was used to define the bodies in the MBE. The inset shows the structures of the molecules.

dx.doi.org/10.1021/ct500396b | J. Chem. Theory Comput. 2014, 10, 3699−3707

It should be noted, though, that poor MBE convergence in a valence-bonded system depends on the definition of "body". Another type of MBE that our group has examined is to treat distortions in the internal degrees-of-freedom of a molecule as bodies. Using the equilibrium geometry as a reference, an MBE can be used to calculate the distortion energy of a molecule. We demonstrate a proof-of-concept using the methanol molecule (Figure 9). We distorted the molecule randomly in all 12

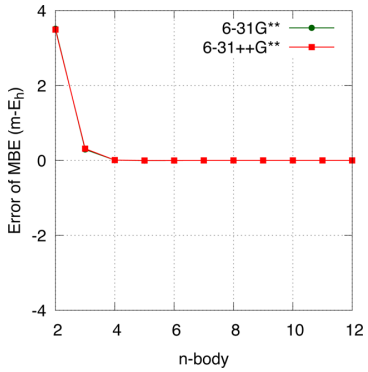


Figure 9. Convergence of the MBE for the distortion energy of a methanol molecule at the HF level of theory for various basis sets. Here, distortions in the 12 internal degrees of freedom are treated as bodies in the MBE. Convergence is clearly independent of the choice of basis set—the results from both basis sets overlap almost perfectly.

degrees of freedom, yielding a total distortion energy of about 140 m-E_h. The degree-of-freedom MBE converges by the four-body term, even when a BSSE-prone basis set is used. This is expected as a consistent basis set is used in all calculations, essentially equivalent to the use of a full-cluster basis. Apart from intramolecular degrees-of-freedom, intermolecular degrees-of-freedom or a combination of both could be treated in the same manner. The utility of such an approach is obvious. A high-dimensional system is broken down into to numerous, completely independent (and thus highly parallelizable) much lower-dimensional function evaluations. Future work will explore how degree-of-freedom MBEs can be used to construct accurate, high-dimensional potential energy surfaces from many lower-dimensional surfaces.

4. CONCLUSIONS

There is no question that the many-body expansion is a theoretically sound and extremely useful formalism in the study of large molecular systems. But it is likewise clear from our observations that care must be taken in its implementation. Rapid convergence at the four-body term of the MBE cannot be assumed, even when convergence appears to have occurred. Incautious use of MBEs with systems and levels of theory susceptible to BSSE is liable to yield errors well beyond chemical accuracy. Moreover, the error per monomer worsens extensively with system size. Such concerns are relevant in valence-bonded and noncovalent systems alike. We conclude that the use of a consistent basis set, either in the form of the full-cluster basis or a high-quality basis set (at least AVTZ quality), is necessary to avoid poor MBE convergence due to BSSE.

■ ASSOCIATED CONTENT

Supporting Information

The graphs for the convergence of the MBE for **10OB** using the basis set in Figure 4 are given in section S1. The Cartesian coordinates for the water clusters studied in this work are given in section S2. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +65 6516 2846. Fax: +65 6779 1691. E-mail: ryan.pa.bettens@nus.edu.sg.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the National University of Singapore's support from the Academic Research Fund, grant number R-143-000-549-112, and the Fulbright U.S. Student Program. The authors also thank the Centre for Computational Science and Engineering for the use of their computers.

■ REFERENCES

(1) Dahlke, E. E.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 10595−10601.
(2) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 46−53.
(3) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 1342−1348.
(4) Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 33−41.
(5) Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904−6914.
(6) Addicoat, M. A.; Collins, M. A. *J. Chem. Phys.* **2009**, *131*, 104103.
(7) Řezác, J.; Salahub, D. R. *J. Chem. Theory Comput.* **2009**, *6*, 91−99.
(8) Weiss, S. N.; Huang, L.; Massa, L. *J. Comput. Chem.* **2010**, *31*, 2889.
(9) Beran, G. J. O.; Nanda, K. *J. Phys. Chem. Lett.* **2010**, *1*, 3480−3487.
(10) Wang, X.; Liu, J.; Zhang, J. Z. H.; He, X. *J. Phys. Chem. A* **2013**, *117*, 7149−7161.
(11) Le, H.-A.; Tan, H.-J.; Ouyang, J. F.; Bettens, R. P. A. *J. Chem. Theory Comput.* **2011**, *8*, 469−478.
(12) Tan, H.-J.; Bettens, R. P. A. *Phys. Chem. Chem. Phys.* **2013**, *15*, 7541−7547.
(13) Elrod, M. J.; Saykally, R. J. *Chem. Rev.* **1994**, *94*, 1975−1997.
(14) Hermansson, K. *J. Chem. Phys.* **1988**, *89*, 2149.
(15) Xantheas, S. S. *Philos. Mag. B* **1996**, *73*, 107−115.
(16) Chen, W.; Gordon, M. S. *J. Phys. Chem.* **1996**, *100*, 14316−14328.
(17) Hodges, M. P.; Stone, A. J.; Xantheas, S. S. *J. Phys. Chem. A* **1997**, *101*, 9163−9168.
(18) Xantheas, S. S. *Chem. Phys.* **2000**, *258*, 225−231.
(19) Kulkarni, A. D.; Ganesh, V.; Gadre, S. R. *J. Chem. Phys.* **2004**, *121*, 5043−5050.
(20) Xantheas, S. S. *Struct. Bonding (Berlin, Ger.)* **2005**, *116*, 119−148.
(21) Xantheas, S. S. *J. Chem. Phys.* **1994**, *100*, 7523−7534.
(22) Christie, R. A.; Jordan, K. D. *Struct. Bonding (Berlin, Ger.)* **2005**, *116*, 27−41.
(23) Szalewicz, K.; Leforestier, C.; van der Avoird, A. *Chem. Phys. Lett.* **2009**, *482*, 1−14.
(24) Gora, U.; Podeszwa, F.; Cencek, W.; Szalewicz, K. *J. Chem. Phys.* **2011**, *135*, 224102.
(25) Wang, Y.; Huang, X.; Shepler, B. C.; Braams, B. J.; Bowman, J. M. *J. Chem. Phys.* **2011**, *134*, 094509.

(26) Kumar, R.; Wang, F.-F.; Jenness, G. R.; Jordan, K. D. *J. Chem. Phys.* **2010**, *132*, 014309.
(27) Medders, G. R.; Babin, V.; Paesani, F. *J. Chem. Theory. Comput.* **2013**, *9*, 1103−1114.
(28) Kumar, R.; Keyes, T. *Theor. Chem. Acc.* **2012**, *131*, 1197.
(29) Yoo, S.; Aprá, E.; Zeng, X. C.; Xantheas, S. S. *J. Phys. Chem. Lett.* **2010**, *1*, 3122−3127.
(30) Pruitt, S. R.; Addicoat, M. A.; Collins, M. A.; Gordon, M. S. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7752−7764.
(31) Hermann, A.; Krawczyk, R. P.; Lein, M.; Schwerdtfeger, P. *Phys. Rev. A* **2007**, *76*, 013202.
(32) Riega, H.; Almeida, R.; Rincon, L. *Rev. Mex. Fis. S* **2006**, *52*, 204−207.
(33) Cui, J.; Liu, H.; Jordan, K. D. *J. Phys. Chem. B* **2006**, *110*, 18872−18878.
(34) Kaplan, I. G.; Santamaria, R.; Novaro, O. *Mol. Phys.* **1995**, *84*, 105−114.
(35) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553−566.
(36) Valiron, P.; Mayer, I. *Chem. Phys. Lett.* **1997**, *275*, 46−55.
(37) Milet, A.; Moszynski, R.; Wormer, P. E. S.; van der Avoird, A. J. *Phys. Chem. A* **1999**, *103*, 6811−6819.
(38) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision C.01; Gaussian Inc.: Wallingford, CT, 2009.
(39) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Lindh, R.; Mitrushenkov, A.; Rauhut, G.; Shamasundar, K. R.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hesselmann, A.; Hetzer, G.; Hrenar, T.; Jansen, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; O'Neill, D. P.; Palmieri, P.; Peng, D.; Pflüger, K.; Pitzer, R.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M. *MOLPRO*, version 2012.1; Cardiff University: Cardiff, U. K.; Universität Stuttgart: Stuttgart, Germany, 2012. See http://www.molpro.net.
(40) Maheshwary, S.; Patel, N.; Sathyamurthy, N.; Kulkarni, A. D.; Gadre, S. R. *J. Phys. Chem. A* **2001**, *105*, 10525−10537.
(41) Richard, R. M.; Herbert, J. M. *J. Chem. Phys.* **2012**, *137*, 064113.
(42) Papajak, E.; Leverentz, H. R.; Zheng, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1197−1202.
(43) Papajak, E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2011**, *7*, 10−18.
(44) Papajak, E.; Zheng, J.; Xu, X.; Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2011**, *7*, 3027−3034.
(45) Halkier, A.; Koch, H.; Jørgensen, P.; Christiansen, O.; Nielsen, I. M. B.; Helgaker, T. *Theor. Chem. Acc.* **1997**, *97*, 150−157.
(46) Beran, G. J. O. *J. Chem. Phys.* **2009**, *130*, 164115.
(47) Le, H.-A.; Lee, A. M.; Bettens, R. P. A. *J. Phys. Chem. A* **2009**, *113*, 10527−10533.
(48) Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128−1132.
(49) Bates, D. M.; Smith, J. R.; Janowski, T.; Tschumper, G. S. *J. Chem. Phys.* **2011**, *135*, 044123.
(50) Richard, R. M.; Lao, K. U.; Herbert, J. M. *J. Phys. Chem. Lett.* **2013**, *4*, 2674−2680.
(51) Mata, R.; Stoll, H. *Chem. Phys. Lett.* **2008**, *465*, 136−141.
(52) Hua, D.; Leverentz, H. R.; Amin, E. A.; Truhlar, D. G. *J. Chem. Theory Comput.* **2011**, *7*, 251−255.
(53) Kurbanov, E. K.; Leverentz, H. R.; Truhlar, D. G.; Amin, E. A. *J. Chem. Theory Comput.* **2012**, *8*, 1−5.
(54) Kurbanov, E. K.; Leverentz, H. R.; Truhlar, D. G.; Amin, E. A. *J. Chem. Theory Comput.* **2013**, *9*, 2617−2628.

doi:10.2533/chimia.2015.104 Chimia 69 (2015) 104–111

Modelling Water: A Lifetime Enigma

John F. Ouyang and Ryan P. A. Bettens*

Abstract: The first attempt to describe water dates back to 1933 with the Bernal–Fowler model and it would take another forty years before the first computer simulation of liquid water by Barker and Watts in 1969. Since then, over a hundred different water models have been proposed. Despite being widely studied, water remains poorly understood. Examining the evolution of water models, we identified three distinct philosophies in water modelling, namely the employment of effective point charges in pioneering empirical models, the incorporation of polarization to describe many-body inductive effects and the extensive use of *ab initio* calculations to describe short-range effects. In doing so, we can appraise the current understanding of water and identify attributes that a water model should possess to capture the intricate interactions between water molecules.

Keywords: Force field · Molecular dynamics · Polarizable · Potential energy surface · Water models

1. Introduction

Considering the rich history of water modelling, it would be prudent to ask why scientists across different disciplines are enthralled by water. An obvious motivation would be its abundance which suggests that water is undeniably important in the grand scheme of nature. The strange properties associated with water also spur academic curiosity to unravel the mysteries behind this small molecule. Most importantly, deciphering the interactions between water molecules would lead to basic understanding of intermolecular forces, which govern many dynamic processes in nature.

Given its ubiquity in nature, water has been the subject of extensive research. On Earth, water is the central solvent for naturally occurring chemical processes. In particular, water is the medium for biochemical interactions, widely recognized as the 'matrix of life'.[1] Its place in biology goes beyond a passive solvent, having many active roles in molecular biology.[2–4] Water-mediated hydrogen bonding provides exchangeable and extensible linkages to manoeuvre the peptide backbone during protein folding, allowing proteins to achieve their active conformation rapidly.[5] Hydration changes can induce modification in DNA conformation and interfacial water possesses a unique sequence-dependent hydration structure, acting as a 'hydration fingerprint' for the recognition of the DNA sequence.[6] On a cosmic scale, detection of water vapour in the atmosphere of an extrasolar gas-giant planet suggests that the presence of water is common in gas-giants.[7] Closer to home, studies on the isotopic composition of water in meteorites help us gain insights about the origins of the early solar system.[8,9] Interestingly, most water in the universe exists as different forms of amorphous ice and their transitions in cold dense interstellar molecular clouds causes radical recombination, resulting in the synthesis of complex organic molecules.[10] The role of water in many chemical and biological processes that are responsible for sustaining life, is the driving force behind understanding its behaviour under different conditions, and in various environments.

Being one of the most studied substances, many physical properties of water are accepted as international standards such as its triple point and density.[11] Even so, many of these physical properties are considered anomalous as they contradict the general theories of the liquid state of matter. The most widely known property would be the maximum density of water at 4 °C, making water the only liquid to expand upon cooling. Other anomalies include the non-monotonic behaviour of its isothermal compressibility and specific heat.[12,13] Furthermore, water exhibits a very high boiling point and dielectric constant for a simple liquid. Although the aforementioned anomalies were known for some time, new anomalous behaviours are constantly uncovered. It was found that supercooled water becomes more diffusive as pressure is increased to about 200 MPa at room temperature.[14] Also, the discovery of another supercooled liquid water state at 150 K challenges the notion of a single supercooled regime at ambient pressure[15] and this newly discovered supercooled state may lead to the identification of a possible second critical point in supercooled confined water.[16] If the liquid state is strange, the solid state would be bizarre with water having fifteen known forms of ice, many of which were only recently discovered.[17,18] It is ironic that while better technology has allowed us to probe the properties of water further, these observed phenomena can exacerbate confusion as they remain unexplained.

The wealth of knowledge on water, many of which deemed anomalous, imposes severe tests on any newly proposed water model. Despite being a chemically simple molecule, water is notoriously hard to model. First, water can give rise to extensive hydrogen bonding networks.[19] As early as 1920, hydrogen bond is first suggested to occur in water[20] and it is commonly agreed that these fleeting hydrogen bonds makes water unique from most other liquids. Dimer interactions are dominated by a deep minimum at the hydrogen-bonded configuration,[21–23] implying that certain configurations are preferred in water clusters and bulk water. The strong directionality of hydrogen bonding is the reason for the inclusion of explicit water molecules in simulating water-mediated processes such as protein folding.[3] However, the hydrogen bond minimum is not overly stabilising, making dynamic hydrogen bonding rearrangements possible in bulk water.[19] Second, the description of water is complicated by strong non-additive inductive effects that manifest in water due to the large dipole and polarizability of water. Such inductive effects can enhance the dipole moment of water molecules by more than 60% in the condensed phase.[24] This is further complicated by the fact that the introduction of polarizability can be rather deceptive,[25] compounded by

*Correspondence: Dr. R. P. A. Bettens
Department of Chemistry
National University of Singapore
3 Science Drive 3, Singapore 117543
E-mail: chmbrpa@nus.edu.sg

reasons which will be covered in Section 3.1. All in all, water is especially sensitive to how the forces between molecules are described and thus demand a thorough and basic understanding of intermolecular forces.

2. Water Models

The Bernal–Fowler (BF) model can be considered the first realistic water model, describing water as a collection of point charges and a repulsion-dispersion term.[26] A similar representation would be used later in the first Monte Carlo simulation of water by Barker and Watts[27] and the first Molecular Dynamics (MD) simulation of water by Rahman and Stillinger.[28] Since the first computer simulation of water, a myriad of water models, exceeding a hundred to date, have been proposed. While there already exist several excellent reviews on the progress of modelling water,[29–33] we still wish to survey the water modelling scene with the aim of highlighting the qualities of a good water model.

In the aforementioned reviews, water models are categorized based on (i) the interaction between water monomers and (ii) the treatment of water monomers. Polarizable models treat many-body inductive effects explicitly using point polarizabilities whereas non-polarizable models describe this polarization in an averaged manner in the pairwise interactions. Rigid water models constrain the intramolecular degrees of freedom, typically to that of the vibrational averaged geometry while flexible counterparts relax all degrees of freedom. Due to *ab initio* calculations approaching experimental accuracy, water models can also be classified based on the nature of the data (*ab initio* or experimental or both) used to parameterise the model.

Instead of following these traditional and possibly restricting classifications, we analysed the evolution of water models and broadly identified three distinct philosophies in the saga of water modelling, namely the employment of enhanced point charges in pioneering models to effectively describe induction in a pairwise potential, the incorporation of polarization in later models to describe explicitly the many-body inductive effects and the extensive use of *ab initio* data in state-of-the-art models to accurately describe water–water interaction at all ranges (Fig. 1). Water models are not necessarily grouped based on chronological order as these demarcations represent distinct principles of water modelling rather than actual time periods. In doing so, we have alluded to the long history of water modelling and its coming of age.

2.1 *Pioneering Empirical Water Models*

This class of water models has its origins in legacy water models, aimed at describing water with a low computational cost and thus often utilise a rigid water monomer. Similar to the BF model, these models are empirical and non-polarizable, using point charges to represent electrostatics and a Lennard-Jones term for dispersion and repulsion. Induction effects are effectively described by increasing the point charges to simulate an enhanced dipole moment found in the condensed phase. Parameters are fitted to reproduce macroscopic experimental data such as the liquid density and heat of vaporization. The reliance on experimental data can be reconciled by noting that these models flourished in the 1980s while highly accurate *ab initio* tools such as the Coupled-Cluster Single and Double, and perturbative treatment of Triple excitations [CCSD(T)] level of theory[34] and Dunning correlation basis sets[35] were only developed in 1989 and became computationally feasible many years later. Consequently, these models only work well at reproducing macroscopic properties of the condensed phase near the conditions under which they are parameterized, typically ambient pressure and temperature, and are targeted towards applications such as biomolecular simulations rather than basic scientific enquiry about the anomalous properties of water. The low computational cost associated with these models would make them remain the preferred choice for the most computationally demanding applications. For example, the TIP3P model is the default water model used in the CHARMM force field for biomolecular simulations.[36] One of the earliest water models in this class is the MCY model,[37] well-known for being constructed entirely from *ab initio* Hartree-Fock (HF) calculations. We will also look further into two families of these pioneering water models, namely the TIP*n*P and SPC water models.

2.1.1 *TIPnP Family*

First developed by Jorgensen in 1981 as the Transferable Intermolecular Potential functionS (TIPS),[38] it was later refined into the TIP3P and TIP4P model[39] which most water scientists are familiar

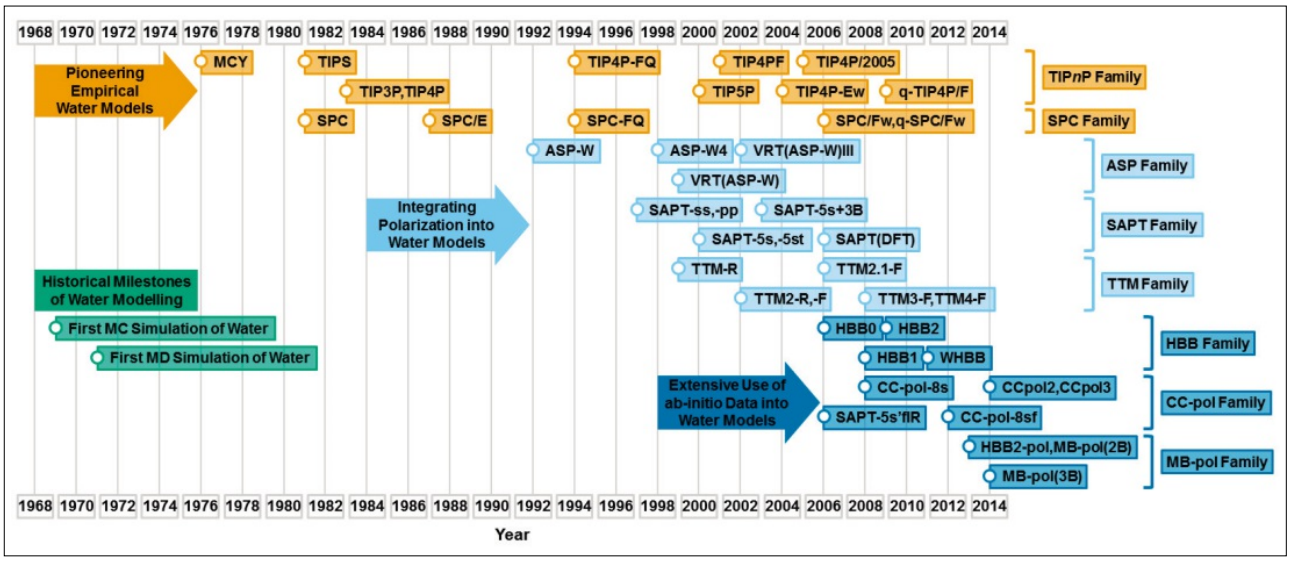


Fig. 1. Timeline showing the year of implementation of various water models reviewed in this paper. Water models are grouped (using different colour schemes) according to the three distinct philosophies of modelling water identified. Within each class of water models, the models are further subdivided into different families of water model that share similar traits.

with. Here, the *n*P refers to the number of point sites in the model where point charges and/or Lennard-Jones terms are placed. In the simplest case, the atomic sites were used as seen in TIP3P. An additional M-site along the HOH angle bisector is introduced in TIP4P to displace the negative charge towards the hydrogens as placing the negative charge on the oxygen would lead to an excessively high dipole moment.[27] In an attempt to describe inductive effects, TIP4P-FQ (FQ for fluctuating charge) was introduced where the point charges fluctuate in response to the environment to equalize the electronegativities of the sites.[40] Later, Mahoney and Jorgensen would introduce more TIP*n*P variants, namely TIP5P[41] and TIP4PF.[42] TIP5P replaced the M-site with two tetrahedral negative charges to mimic the lone pairs on water but this resulted in a overly structured water in simulations. TIP4PF is a flexible version of TIP4P where intramolecular stretching and bending are described by quadratic terms and the same study showed that the inclusion of quantum effects improve the predictions made by this flexible water model.

Surprisingly, the models mentioned thus far truncate long-range electrostatics at a certain cut-off distance. The TIP4P-Ew model is designed for use with Ewald techniques to account for long-range electrostatics, commonly employed in biomolecular simulations.[43] Numerous other parameterization attempts were made, such as TIP4P/2005[44] and q-TIP4P/F,[45] which are optimised to better reproduce the thermodynamic properties of water and to account for quantum effects respectively.

2.1.2 *SPC Family*

Apart from the TIP*n*P family, another family of water models is the Single Point Charge (SPC) model, which only uses the three atomic sites to place point charges and/or Lennard-Jones terms.[46] Simple values were used for its parameters such as 1.0 Å for the O–H bond length and an ideal tetrahedral angle of 109.5° instead of the experimental gas-phase values used in TIP4P. Shortly, the improved SPC/E model was proposed to account for polarization self-energy.[47] Similar to TIP4P-FQ, SPC-FQ was introduced to incorporate induction effects.[40] Likewise, flexible monomer versions such as SPC/Fw[48] and variants parameterized to account for quantum effects, such as the q-SPC/Fw[49] model, have been introduced.

2.2 *Integrating Polarization into Water Models*

The increase in computational power saw a transition towards increasingly complicated water models with an emphasis on the non-additivity of water–water interactions, in particular induction/polarization effects. Polarization is often incorporated explicitly *via* central or distributed point dipole–dipole polarizabilities, derived from the use of perturbation theory to treat intermolecular forces.[50] Despite the rigorous theoretical background, such an implementation may lead to deceptive results as we shall see in Section 3.1. Furthermore, higher-order multipoles, typically up to quadrupoles, are employed to represent electrostatics instead of point charges in recognition of the anisotropic nature of the electron distribution. This led to more elaborate analytic potentials that required more parameters that would come from a mix of *ab initio* and experimental spectroscopic data. This class of water models flourished in the 1990s and 2000s when accurate *ab initio* second-order Møller–Plesset perturbation theory (MP2) and later CCSD(T) calculations become amendable. As the majority of the parameters are monomer properties such as the dipole moment and polarizability, highly accurate *ab initio* calculations can be performed on the small water monomer system. In some cases, the Vibration-Rotation-Tunnelling (VRT) spectroscopic data was used in the parameterization as they represent information at the atomistic level as opposed to bulk water properties. Using these water models, there would be more studies devoted towards water clusters, underscoring the importance of microscopic understanding of water. As the functional form of these water models grew more complex, it would naturally encompass a larger variety of models and some of the notable water models include the ASP, SAPT and TTM family of water models.

2.2.1 *ASP Family*

The Anisotropic Site Potential with Wormer's dispersion (ASP-W) model, based on Hayes–Stone intermolecular perturbation theory (IMPT),[51,52] is one of the earliest rigid water models to adopt higher-order multipoles.[53] For electrostatics, distributed multipoles are present on both the oxygen and hydrogen atomic sites, up to quadrupole and dipole respectively whereas induction is computed at first order (instead of full iteration) using point polarizabilites on oxygen up to quadrupole. Site anisotropy was likewise incorporated into the dispersion and repulsion terms. Further refinements by the same group led to the inclusion of a new charge transfer term, creating the ASP-W2 and ASP-W4 models, used to study the stationary structures of the water dimer.[54] The difference between both models lies in the order of multipoles used with the original multipoles being retained in ASP-W2 and hexadecapole present on each atom in ASP-W4.

The ASP-W functional form was also fitted to the $(D2O)_2$ VRT spectra, giving rise to the VRT(ASP-W) model.[55] VRT(ASP-W) is the first water model to achieve spectroscopic accuracy, able to reproduce most of the tunnelling barriers in the water dimer. While this is not surprising given the use of same experimental data in constructing the model, it is worthwhile to note that the use of the rigid monomer approximation can still lead to accurate predictions at the atomistic level. Later improvements would give rise to the VRT(ASP-W)II and VRT(ASP-W) III model, where induction is computed to full iteration.[56]

2.2.2 *SAPT Family*

Both SAPT-ss and SAPT-pp water models,[57] employing rigid water monomers, were developed based on Symmetry-Adapted Perturbation Theory (SAPT).[58] SAPT-ss comprises a site–site form, with a similar placement of sites as TIP4P but instead uses the functional form of the MCY model. Point charges and exponential terms are fitted to 1056 SAPT energies. The SAPT-pp is more complicated, describing the intermolecular interactions using expansions of functions in interatomic vectors and Euler angles, again fitted to the same 1056 SAPT energies.

Due to its complexity, SAPT-pp fell into disuse and the site–site form was evolved to the SAPT-5s model.[59] To reflect the anisotropy of electron distribution, two new symmetry distinct sites representing lone pairs and out-of-plane charges were added, giving a total of five symmetry distinct sites (eight sites in total). An elaborate functional form was adopted using a polynomial-exponential terms to represent exchange-repulsion and an inverse power (6-8-10) series to describe induction and dispersion. Consequently, no iteration of the induced dipole is required in calculating the induction term as it is represented by fitted coefficients. The model's exchange-repulsion parameters are also tuned to better reproduce the water dimer's acceptor tunnelling splitting, giving the revised SAPT-5st.[60]

All the SAPT models mentioned above only contain a pair potential. Thus, three-body SAPT(HF) energies were incorporated into SAPT-5s, giving the SAPT-5s+3B model.[61] This new three-body potential is the first to include functional forms to model three-body exchange effects using a combination of exponential and Legendre polynomial terms. Long-range effects are described using a damped induced dipole model. Later, the SAPT-5s functional form is refitted using SAPT(DFT) energies and this new SDFT-5s model[62] gives more accurate results, attributed to the faster basis set convergence with DFT.

2.2.3 *TTM Family*

TTM-R,[63] the first of Thole-Type Model (TTM) water models, is based on Thole's idea of using smeared out multipoles to mirror the diffuse picture of electron distribution.[64] TTM-R utilises TIP4P-style point charges for electrostatic interactions and an inverse power (6-10-12) series to represent dispersion and repulsion. Smeared charges and dipole are present on all atomic sites for induction and intramolecular polarization can occur, accounting for charge transfer. As the TTM-R model consistently over-binds small water clusters, the TTM2-R model was proposed by refitting the inverse power (6-10-12) series to minimum energy pathways connecting the global minimum and other stationary points of the water dimer.[65]

Monomer flexibility was then incorporated using the Partridge–Schwenke intramolecular Potential Energy Surface (PES) and Dipole Moment Surface (DMS)[66] resulting in the TTM2-F model, the first water model to properly reproduce an increase in the monomer bending angle in water clusters.[67] A revised TTM2.1-F model,[68] intended for simulations, was proposed by modifying the inverse power (6-10-12) series that decreases unphysically below 2.5 Å as such repulsive regions may be sampled during condensed-phase simulations.

Two unrelated updates, the TTM3-F[69] and TTM4-F model[70] were also reported. Aimed at describing the vibrational spectra of water clusters and bulk water, TTM3-F has modified partial charges to reflect the behaviour that water dissociates to H^+/OH^- in liquid as opposed to radical formation in the gas phase. On the other hand, TTM4-F is reparameterized to better reproduce polarizability surface. Notably, the popular AMOEBA water model uses a Thole-type induction model.[71]

2.3 *Extensive Use of ab initio Data in Water Models*

As *ab initio* methods matured into reliable tools rivalling experimental accuracy, we ushered in an era of water models empowered by *ab initio* data. This class of water models relies on high-quality large datasets (in the order of 10^5 data points) of CCSD(T) energies, the gold standard of quantum chemistry. The water models are deeply rooted in the Many-Body Expansion (MBE) where the total energy of a system can be decomposed into one-body (monomer contribution), two-body (pairwise interactions), three-body contributions and so on. Separate PES are constructed for each of these *k*-body terms by fitting large energy datasets which sample important configuration space encountered in water clusters and during condensed phase simulations. The extensive amount of high-quality data required can only be fulfilled by large volumes of accurate *ab initio* calculations which only became amendable in recent years. The shift towards large datasets and complicated PES construction techniques stems from the realization that short-range effects such as charge transfer and exchange cannot be accurately described by simple analytic forms. Thus, sufficiently flexible functional forms are required to map the accurate *ab initio* dataset into high-quality PES for on-the-fly evaluation of energies. Water monomer flexibility is a another common feature in these models although a rigid monomer constraint is often imposed in demanding calculations such as condensed phase simulations and VRT spectra prediction. As a result, these models are mainly focussed on studies of water clusters with few examples of condensed phase simulations. As the construction of these water models is laborious, there were only three families of such *ab initio* water models, namely the HBB, CC-pol and MB-pol family of water models.

2.3.1 *HBB Family*

The HBB water models describe the PES for each of these *k*-body terms using permutationally invariant polynomials involving interatomic distances, incorporating the permutation symmetry of identical atoms, *i.e.* the hydrogen and oxygen atoms. This alleviates the steep computational cost in evaluating high-dimensional PES and drastically reduces the number of data points required for fitting the PES. The first HBB0 model uses polynomials of Morse-type exponential functions, fitted to 19805 CCSD(T)/AVTZ energies.[72] Like all HBB models, all $N(N-1)/2$ interatomic distances were used to preserve the permutational symmetry, more than the actual $3N-6$ degrees-of-freedom present in the system. In the next revision HBB1, the same functional form is refitted to an additional 10227 CCSD(T)/AVTZ energies to better describe the low-energy configuration space below 10000 cm^{-1}.[73] This refitting led the RMS fitting error to drop by a factor of two, suggesting that the quality of the functional form was previously not maximized in HBB0.

A hybrid pair potential was developed in the new HBB2 model, comprising long-range and short-range components.[74] The short-range component remains to be described by permutationally invariant polynomials while the long-range component is described using the TTM3-F model. This led to slight improvements in accuracy and large computational savings as the TTM3-F potential is much faster to compute. The HBB*n* models only contain a pair potential and cannot be used to describe water clusters where higher-body effects have to be considered. Thus, the WHBB model is introduced where a three-body potential is again constructed using permutationally invariant polynomials, fitted to 40000 MP2/AVTZ energies.[75] Interestingly, it was mentioned that the three-body potential is shorter range than the two-body counterpart and a cutoff was implemented when the maximum O–O distance is greater than 8 Å. Four-and-higher-body effects are described by induction using the TTM3-F model. For all the water models in the HBB family, the one-body potential is provided by the Partridge–Schwenke intramolecular PES.

2.3.2 *CC-pol Family*

The CC-pol family of water models is the successor of the SAPT family, utilising *ab initio* energies computed at CCSD(T) instead of SAPT energies. The first CC-pol model[76] is similar to the SAPT-5s model except that induction is now explicitly iterated instead of using a fitted inverse power series. CC-pol is able to reproduce the water dimer VRT spectra except for the interchange splitting transition, attributed to the rigid monomer approximation.

The CC-pol-8s model revamped the placement of the interaction sites, having eight symmetry distinct sites (25 sites in total).[77] The three-dimensional Cartesian space was scanned in regular intervals, followed by finer subgrids to ensure that the most optimal positions were chosen. As only point charges were used (as opposed to higher-order multipoles), the presence of more interaction sites better represents the anisotropy of the electron distribution and led to a four-fold decrease in the fitting errors. A flexible variant, CC-pol-8sf,[78] was developed where monomer contribution to the interaction energy is obtained from an earlier flexible SAPT-5s'fIR water model.[79]

Feeling that the order of 10^5 data points is inadequate to build an accurate full 21-dimensional flexible-monomer three-body PES, the authors reverted to a rigid monomer model, consisting of the pair potential CCpol2 and three-body potential CCpol3.[80] CCpol2 is essentially the same as CC-pol-8s, except that short-range damping is included to improve the description at very small intermolecular distances as these regions may be sampled during condensed phase simulations. The CCpol3 model, fitted to 71456 CCSD(T) energies, gives improved polarization from the use of three atomic polarization centres, instead of one. Four-and-higher-body interactions are described using a simple polarization model. Surprisingly, the polarization model gives accurate four-body energies to within a few percent, whereas such models are known to have significant errors for three-body interactions.

2.3.3 *MB-pol Family*

The MB-pol family incorporates many features from the HBB family of *ab initio* based water models. The prototype HBB2-pol model[81] borrows from the HBB2 model using a hybrid pair potential and the Partridge–Schwenke intramolecular PES. The same HBB2 PES was used for the short-range component of the pair potential while the long-range component was replaced with the TTM4-F model. Furthermore, a three-body hybrid potential is included where the short-range component again incorporates the permutational symmetry, fitted to 8019 CCSD(T) trimer energies, while the long-range counterpart, as well as four-and-higher-body effects, are described by induction in the TTM4-F model. The TTM4-F component greatly reduced the order of the permutationally invariant polynomials and the associated computational cost, making HBB2-pol amendable to condensed phase simulations. TTM4-F was chosen after careful comparison with two other polarizable flexible water models, namely TTM3-F and AMOEBA.

The eventual MB-pol model is described in two papers, detailing the hybrid pair potential[82] and higher-body effects separately.[83] The hybrid pair potential MB-pol(2B) was improved with the addition of two new sites to represent the lone pairs of water, which greatly improved the flexibility of the functional form in the short-range component. Thus, the permutationally invariant polynomials now involves intersite distances between the atomic sites and/or the lone pair sites, fitted to 42508 CCSD(T) dimer energies. The three-body potential MB-pol(3B) is described in a similar fashion as in HBB2-pol but fitted to a larger dataset of 12347 energies. All long range effects are handled by induction using the TTM4-F model. It was noted that short-range corrections are not required at the four-and-higher-body level, in agreement with CCpol3 authors' observation that a simple polarization model is sufficient.

On a final note, both HBB2-pol and MB-pol are the first water models constructed from extensive CCSD(T) energies dataset to be employed in classical and quantum simulations of liquid water.[84,85] In both instances, many structural and dynamic properties of liquid water under ambient conditions were reproduced, such as the radial distribution functions, bulk water density and diffusion coefficient.

3. Qualities of a Good Water Model

After reviewing the plethora of water models shaped by different philosophies, we identified several key features for the proper description of water. They are namely (i) the inclusion of polarizability to account for non-additive effects, (ii) fitting or interpolating energies to account for short-range effects, (iii) incorporation of monomer flexibility, (iv) accounting for quantum effects in simulations and (v) transferability and dissociable water model.

3.1 *Inclusion of Polarizability*

As we witness from the integration of polarization into water models, (Section 2.2) the inclusion of polarizability is crucial in describing the significant many-body inductive effects that arise from the high dipole and polarizability of water. Neglecting polarization effects in empirical point charge water models (such as TIN*n*P and SPC models in Section 2.1) prevents an accurate description of virial coefficients, vapour pressures, critical pressure and dielectric constant.[86] The first three quantities involve gas phase properties which are very sensitive to changes in the environment. Clearly, the degree of polarization in the gas phase would differ greatly from that in the condensed phase for which the empirical models are calibrated. Likewise, polarization is required to reproduce the enhanced dipole moment in condensed phase to properly reproduce the dielectric constant.

There are several excellent reviews[30,87–89] on the implementation of polarization as it found importance not only in water models but also in ion solvation, other small molecules and protein simulations. Three methods for incorporating polarization exist, namely fluctuating charge, Drude oscillator and induced point dipole models. While the first two methods have been implemented in water models, (*e. g.* TIP4P-FQ, SPC-FQ[40] for fluctuating charge and SWM4-DP[90] for Drude oscillator) the induced point dipole model remains the most implemented for water models. In fact, the ASP, SAPT and TTM families of water models in Section 2.2 all use some kind of induced point dipole model. In principle, higher-order multipoles such as the quadrupole can also be induced as seen in the ASP water models but they see little action elsewhere (SAPT and TTM families only involve inducible dipole) perhaps due to the laborious theoretical expressions involved. While the introduction of inducible dipole models is increasingly prevalent, Guillot cautions that poor implementation can lead to deceptive results.[29] The induced dipole moment is given as the product of the polarizability with the electric field. The electric field is often represented by the point charges/multipoles present in the model and this may be inadequate if higher-order multipoles are not considered.[91] Furthermore, there is also dipole–quadrupole and quadrupole–quadrupole polarizabilities which are often neglected and these inductive effects can be significant given that water has a strong quadrupole moment.

Finally, Thole[40] and Applequist *et al.*[92] have pointed out that the point induced dipole moment may become infinite at small distances, which is commonly known as the 'polarization catastrophe'. This can be avoided by screening the dipole–dipole interaction at short distances, either using a Tang–Toennies damping function[93] as seen in the ASP and SAPT models or using smeared out charges and dipoles in TTM models. This screening is an indication that point multipoles cannot properly describe the electronic distribution at small distances, underscoring the importance of accounting for short-range effects.

3.2 *Short-range Effects*

At short intermolecular distances R, the power series expansion of inverse R which defines the point multipole diverges, causing the failure of point multipoles at short-range. Furthermore, there is a charge penetration effect as the electrons are 'not fully felt' within the electron cloud. Physically, this can be interpreted as the unrealistic representation of the electronic distribution as if it was concentrated at a point. Possible remedies include the use of damping functions or smeared out multipoles as seen in Section 2.2 as well as partitioning the electronic distribution using distributed multipoles.[50] Despite these corrections, other short-range interactions such as exchange-repulsion and charge transfer have to be explicitly accounted for. The distinction between short-range and long-range interactions (electrostatic, induction and dispersion) is rooted in their different physical character where short-range effects vary exponentially with intermolecular distance while long-range effects behave as some inverse power of intermolecular distance.[50] Thus, it would be prudent to separate the total interaction energy into short-range and long-range components due to their intrinsically different nature as seen in the HBB2, WHBB, HBB2-pol and MB-pol water models.

Unfortunately, unlike long-range interactions which have well-defined formulae based on IMPT, no exact analytic form exists for short-range interactions. Otherwise, high quality *ab initio* methods which can describe these subtle short-range effects up to any desired numerical precision would have been developed in vain. For the ASP, SAPT and TTM families of models, short-range exchange-repulsion effects were modelled by simple exponential and/or polynomial-exponential terms.

As these approaches proved inadequate, large *ab initio* data sets are fitted to more complicated functional forms to accurately describe these exchange-repulsion effects (Section 2.3). Currently, two such functional forms have been implemented. The permutationally invariant polynomials in HBB and MB-pol families of models incorporate the permutational symmetry of identical nuclei into exponential terms involving interatomic distances. On the other hand, CC-pol models uses simple polynomial-exponential terms but applied between a large number of symmetry-distinct sites, greatly increasing the flexibility of the functional form. Inevitably, both methods incorporate some form of symmetry which serves to alleviate the high computational cost. Furthermore, both methods involve fitting of the coefficients of the terms from *ab initio* data. An alternative to fitting methods would be interpolation methods. Examples include Shepard interpolation[94,95] and Interpolating Moving Least Square[96,97] as well as simpler interpolating methods such as cubic splines. While interpolation methods ensure that the PES passes exactly through the dataset, care has to be taken that the asymptotic behaviour of the PES is enforced in interpolating models which are otherwise naturally incorporated into the functional forms used in fitting models. Nonetheless, it would be interesting to see new *ab initio* based water models based on interpolation methods and compare their accuracy with existing models.

Another essential formalism employed to describe short-range effects would be the Many-Body Expansion (MBE). Without the use of MBE, the dimensionality of the system would be too large for any fitting or interpolation method to be feasible. Instead, using the MBE, large water clusters or even bulk water can be decomposed into many-body contributions, truncated at the four-body level. However, basis set superposition effects causes poor convergence of the MBE when diffuse basis functions are involved[98] and these diffuse functions are crucial in accurately describing the hydrogen bonding between water molecules.

3.3 *Monomer Flexibility*

In the MBE formalism, the one-body contribution would correspond to intramolecular distortions of water monomer. Due to computational limitations, pioneering empirical water models often employ rigid monomers. While later models would comprise of flexible monomers, a rigid monomer approximation is still preferred for computationally demanding calculations such as spectra prediction and condensed phase simulations. Also, a large dataset is required to fit flexible monomer potentials which can disfavour their use as seen in the CCpol2 and CCpol3 water models. It is recommended that the vibrational averaged geometry be used over the equilibrium geometry when a rigid monomer approximation is necessary.

Monomer flexibility is integral in the atomistic understanding of water as subtle changes in bond lengths and angles can affect the predicted energetics and VRT spectra of water clusters. The first water models to include flexible monomers use quadratic terms to describe the stretching and bending motions, modelling the vibrational modes as harmonic oscillators. This is overly simplistic in dealing with the quantum mechanical effects that arises when the electron clouds of the two hydrogens overlap during the bending motion. Thus, more sophisticated intramolecular PES were constructed, the most popular being the Partridge–Schwenke intramolecular PES, which is used in the TTM, HBB and MB-pol families of water models. The Partridge–Schwenke PES is also accompanied with an intramolecular DMS which supplies the dipole moment required in the calculation of long-range interactions. This could be the reason why higher-order multipoles are not involved in the long-range components of these models as an accurate quadrupole moment surface do not exist yet.

It is important to realize that these intramolecular vibrations are quantum mechanical in nature and their treatment within classical simulations may not yield satisfactory results.[99–101] The representative example would be the harmonic oscillator where the classical probability would be greatest away from the equilibrium while the quantum counterpart has the maximum probability at the equilibrium position. Thus, flexible water models should be simulated using methods that incorporate quantum effects.

3.4 *Nuclear Quantum Effects*

Nuclear quantum effects and monomer flexibility are intertwined since the motions of the nuclei obey the laws of quantum mechanics rather than the classical counterpart. This is especially so for water due to the presence of the light hydrogen nuclei and extensive hydrogen bonding, both of which exhibit strong nuclear quantum effects. Thus, processes involving the hydrogen nuclei such as Grotthuss proton shuttling[102] require nuclear quantum effects to be accounted for.[103]

Furthermore, disregarding nuclear quantum effects can lead to poor description of the heat capacity of both liquid and solid water[104,105] and low-temperature properties such as the densities of ice polymorphs.[99] In addition, when nuclear quantum effects are neglected, isotopic effects cannot be probed, which can have a significant influence in bulk properties. For example, the enthalpy of vaporization is a measure of the strength of the hydrogen bonding within liquid water. Classically, there should be no isotopic effects present. However, it has been shown experimentally that the isotopic effects on the vaporization enthalpy is important, increasing by 0.4 kcal mol^{-1} from water to tritiated water.[86]

A variety of quantum simulation methods exist and some of the computational methodologies have been reviewed.[106] The most commonly employed method would be Path Integral Molecular Dynamics (PIMD),[107–109] which exploits the isomorphism between the quantum partition function expressed in path integral formalism and the classical partition function of a ring-polymer. This isomorphism provides a way to sample the quantum nuclear configuration through modifications of the classical MD technique. Other quantum simulation methods would include Path Integral Monte Carlo (PIMC),[110,111] Path Integral Hybrid Monte Carlo (PIHMC),[107,112,113] Centroid Molecular Dynamics (CMD)[114–118] and Ring Polymer Molecular Dynamics (RPMD).[119,120]

While PIMD simulations have been performed for the HBB2-pol and MB-pol *ab initio* based models at ambient conditions,[84,85] extreme conditions (low temperatures, critical point) have not been explored to elucidate the anomalous behaviour of water. On a side note, studies on the quantum effects of water performed on empirical water models such as TIP4P should be interpreted with caution. As such water models are parameterized to reproduce experimental values using classical simulations, quantum effects are included in these models in an effective manner. Thus, performing quantum simulations on these water models to investigate quantum effects seems counterproductive unless the model has been reparameterized for such purposes.

3.5 *Transferability and Ability to Dissociate*

While less discussed in literature, it is ideal to develop a water model to be used outside pure water systems for applications such as explicit solvation of proteins. The empirical and polarizable models (Section 2.1 and 2.2) are highly transferable due to the use of point multipoles which share the same functional form regardless of the molecular species. This is not the case for *ab initio* based water models (Section 2.3) that rely on the MBE as new PES have to be constructed for new combinations of *k*-body interactions.

Finally, very few models in literature

are able to dissociate into H^+/OH^- ions. Water dissociation is difficult to handle as the products (charged ions) are very different from the reactant (neutral molecules). This is complicated by the fact that water dissociates homolytically into radicals in the gas phase. It would be optimal to use on-the-fly *ab initio* simulation techniques such as Car-Parrinello Molecular Dynamics (CPMD)[121] to study water dissociation as these *ab initio* methods do not make any distinction between H^+/OH^- ions and neutral water molecules.

4. Concluding Remarks and Outlook

The scene of water modelling remains a vibrant one, especially in the last 15 years where countless water models of distinct modelling philosophies have been developed with the sole aim to better understand this mysterious liquid. The strengths and (more often) inadequacies of these water models have provided useful information on the essential ingredients for the making of a universal water model.

It is only very recently, with the extensive use of *ab initio* data and availability of quantum simulations, that water models possess the right qualities to accurately describe water at both the microscopic and macroscopic level. Yet, there still leaves room for development, in seeking new ways to describe short-range effects using interpolation techniques and employing higher-order multipoles in long-range interactions so that more of the configuration space can be described cheaply.

Nonetheless, it is due time to put these state-of-the-art water models to more rigorous tests to reproduce experimental results at extreme conditions. If these water models were to succeed at these trials, then perhaps it would be possible to explain the many anomalies of water, fulfilling the role of computations in assisting experiments to dispel confusion and eventually pushing the boundaries of science.

Acknowledgements

The authors thank the National University of Singapore's support from the Academic Research Fund, grant number R-143-000-549-112.

Received: January 2, 2015

[1] F. Franks, 'Water: A Matrix of Life', 2nd ed., Royal Society Of Chemistry, Cambridge, England, **2000**.
[2] M. J. Tait, F. Franks, *Nature* **1971**, 230, 91.
[3] M. Chaplin, *Nature Rev.* **2006**, *7*, 861.
[4] P. Ball, *Chem. Rev.* **2008**, *108*, 74.
[5] M. S. Cheung, A. E. Garcia, J. N. Onuchic, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 685.
[6] M. Fuxreiter, M. Mezei, I. Simon, R. Osman, *Biophys. J.* **2005**, *89*, 903.
[7] G. Tinetti, A. Vidal-Madjar, M.-C. Liang, J.-P. Beaulieu, Y. Yung, S. Carey, R. J. Barber, J. Tennyson, I. Ribas, N. Allard, G. E. Ballester, D. K. Sing, F. Selsis, *Nature* **2007**, *448*, 169.
[8] F. Robert, *Science* **2001**, *293*, 1056.
[9] N. Sakamoto, Y. Seto, S. Itoh, K. Kuramoto, K. Fujino, K. Nagashima, A. N. Krot, H. Yurimoto, *Science* **2007**, *317*, 231.
[10] P. Jenniskens, D. F. Blake, *Science* **1994**, *265*, 753.
[11] F. Franks, in 'Water: A Comprehensive Treatise Vol 1 Physics and Physical Chemistry of Water', Ed. F. Franks, Plenum Press, New York, USA, **1972**.
[12] P. H. Poole, U. E. F. Sciortino, H. E. Stanley, *Nature* **1992**, *360*, 324.
[13] H. E. Stanley, S. V. Buldyrev, M. Canpolat, S. Havlin, O. Mishima, M. R. Sadr-Lahijany, A. Scala, F. Starr, *Physica D* **1999**, *133*, 453.
[14] F. X. Prielmeier, E. W. Lang, R. J. Speedy, H.-D. Lüdemann, *Phys. Rev. Lett.* **1987**, *59*, 1128.
[15] R. S. Smith, B. D. Kay, *Nature* **1999**, *398*, 788.
[16] L. Liu, S.-H. Chen, A. Faraone, C.-W. Yen, C.-Y. Mou, *Phys. Rev. Lett.* **2005**, *95*, 117802.
[17] C. G. Salzmann, P. G. Radaelli, A. Hallbrucker, E. Mayer, J. L. Finney, *Science* **2006**, *311*, 1758.
[18] C. G. Salzmann, P. G. Radaelli, E. Mayer, J. L. Finney, *Phys. Rev. Lett.* **2009**, *103*, 105701.
[19] I. Ohmine, H. Tanaka, *Chem. Rev.* **1993**, *93*, 2545.
[20] W. M. Latimer, W. H. Rodebush, *J. Am. Chem. Soc.* **1920**, *42*, 1419.
[21] B. J. Smith, D. J. Swanton, J. A. Pople, III, H. F. S.; L. Radom, *J. Chem. Phys.* **1990**, *92*, 1240.
[22] G. S. Tschumper, M. L. Leininger, B. C. Hoffman, E. F. Valeev, III, H. F. S.; M. Quack, *J. Chem. Phys.* **2002**, *116*, 690.
[23] J. R. Lane, *J. Chem. Theory Comput.* **2013**, *9*, 316.
[24] E. R. Batista, S. S. Xantheas, *J. Chem. Phys.* **1998**, *109*, 4546.
[25] B. Guillot, *J. Mol. Liq.* **2002**, *101*, 219.
[26] J. D. Bernal, R. H. Fowler, *J. Chem. Phys.* **1933**, *1*, 515.
[27] J. A. Barker, R. O. Watts, *Chem. Phys. Lett.* **1969**, *3*, 144.
[28] A. Rahman, F. H. Stillinger, *J. Chem. Phys.* **1971**, *55*, 3336.
[29] B. Guillot, *J. Mol. Liq.* **2002**, *101*, 219.
[30] T. A. Halgren, W. Damm, *Curr. Opin. Struct. Biol.* **2001**, *11*, 236.
[31] J. L. Finney, *J. Mol. Liq.* **2001**, *90*, 303.
[32] K. Szalewicz, C. Leforestier, A. van der Avoird, *Chem. Phys. Lett.* **2009**, *482*, 1.
[33] O. Demerdash, E.-H. Yap, T. Head-Gordon, *Annu. Rev. Phys. Chem.* **2014**, *65*, 149.
[34] K. Raghavachari, G. W. Trucks, J. A. Pople, M. Head-Gordon, *Chem. Phys. Lett.* **1989**, *157*, 479.
[35] T. H. Dunning, Jr., *J. Chem. Phys.* **1989**, *90*, 1007.
[36] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiurkiewicz-Kuczera, D. Yin, M. Karplus, *J. Phys. Chem. B* **1998**, *102*, 3586.
[37] H. Popkie, H. Kistenmacher, E. Clementi, *J. Chem. Phys.* **1973**, *59*, 1325.
[38] W. L. Jorgensen, *J. Am. Chem. Soc.* **1981**, *103*, 335.
[39] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926.
[40] S. W. Rick, S. J. Stuart, B. J. Berne, *J. Chem. Phys.* **1994**, *101*, 6141.
[41] M. W. Mahoney, W. L. Jorgensen, *J. Chem. Phys.* **2000**, *112*, 8910.
[42] M. W. Mahoney, W. L. Jorgensen, *J. Chem. Phys.* **2001**, *115*, 10758.
[43] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, *J. Chem. Phys.* **2004**, *120*, 9665.
[44] J. L. F. Abascal, C. Vega, *J. Chem. Phys.* **2005**, *123*, 234505.
[45] S. Habershon, T. E. Markland, D. E. Manolopoulos, *J. Chem. Phys.* **2009**, *131*, 024501.
[46] H. J. C. Berendsen, J. P. M. Postma, W. F. von Gunstaren, J. Hermans, in 'Intermolecular Forces', Ed. B. Pullman, Reidel, Dordrecht, Holland, **1981**, pp 331–342.
[47] H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, *J. Phys. Chem.* **1987**, *91*, 6269.
[48] Y. Wu, H. L. Tepper, G. A. Voth, *J. Chem. Phys.* **2006**, *124*, 024503.
[49] F. Paesani, W. Zhang, D. A. Case, T. E. Cheatham III, G. A. Voth, *J. Chem. Phys.* **2006**, *125*, 184507.
[50] A. J. Stone, 'The Theory of Intermolecular Forces', 2nd ed., Oxford University Press, Oxford, England, **2013**.
[51] I. C. Hayes, A. J. Stone, *Mol. Phys.* **1984**, *53*, 69.
[52] I. C. Hayes, A. J. Stone, *Mol. Phys.* **1984**, *53*, 83.
[53] C. Millot, A. J. Stone, *Mol. Phys.* **1992**, *77*, 439.
[54] C. Millot, J.-C. Soetens, M. T. C. M. Costa, M. P. Hodges, A. J. Stone, *J. Phys. Chem. A* **1998**, *102*, 754.
[55] R. S. Fellers, C. Leforestier, L. B. Braly, M. G. Brown, R. J. Saykally, *Science* **1999**, *284*, 945.
[56] N. Goldman, R. S. Fellers, M. G. Brown, L. B. Braly, C. J. Keoshian, C. Leforestier, R. J. Saykally, *J. Chem. Phys.* **2002**, *116*, 10148.
[57] E. M. Mas, K. Szalewicz, R. Bukowski, B. Jeziorski, *J. Chem. Phys.* **1997**, *107*, 4207.
[58] B. Jeziorski, R. Moszynski, K. Szalewicz, *Chem. Rev.* **1994**, *94*, 1887.
[59] E. M. Mas, R. Bukowski, K. Szalewicz, G. C. Groenenboom, P. E. S. Wormer, A. van der Avoird, *J. Chem. Phys.* **2000**, *113*, 6687.
[60] G. C. Groenenboom, P. E. S. Wormer, A. van der Avoird, E. M. Mas, R. Bukowski, K. Szalewicz, *J. Chem. Phys.* **2000**, *113*, 6702.
[61] E. M. Mas, R. Bukowski, K. Szalewicz, *J. Chem. Phys.* **2003**, *118*, 4386.
[62] R. Bukowski, K. Szalewicz, G. Groenenboom, A. van der Avoird, *J. Chem. Phys.* **2006**, *125*, 044301.
[63] C. J. Burnham, J. Li, S. S. Xantheas, M. Leslie, *J. Chem. Phys.* **1999**, *110*, 4566.
[64] B. T. Thole, *Chem. Phys.* **1981**, *59*, 341.
[65] C. J. Burnham, S. S. Xantheas, *J. Chem. Phys.* **2002**, *116*, 1500.
[66] H. Partridge, D. W. Schwenke, *J. Chem. Phys.* **1997**, *106*, 4618.
[67] C. J. Burnham, S. S. Xantheas, *J. Chem. Phys.* **2002**, *116*, 5115.
[68] G. S. Fanourgakis, S. S. Xantheas, *J. Phys. Chem. A* **2006**, *110*, 4100.
[69] G. S. Fanourgakis, S. S. Xantheas, *J. Chem. Phys.* **2008**, *128*, 074506.
[70] C. J. Burnham, D. J. Anick, P. K. Mankoo, G. F. Reiter, *J. Chem. Phys.* **2008**, *128*, 154519.
[71] P. Ren, J. W. Ponder, *J. Phys. Chem. B* **2003**, *107*, 5933.
[72] X. Huang, B. J. Braams, J. M. Bowman, *J. Phys. Chem. A* **2006**, *110*, 445.
[73] X. Huang, B. J. Braams, J. M. Bowman, R. E A. Kelly, J. Tennyson, G. C. Groenenboom, A. van der Avoird, *J. Chem. Phys.* **2008**, *128*, 034312.
[74] A. Shank, Y. Wang, A. Kaledin, B. J. Braams, J. M. Bowman, *J. Chem. Phys.* **2009**, *130*, 144314.
[75] Y. Wang, X. Huang, B. C. Shepler, B. J. Braams, J. M. Bowman, *J. Chem. Phys.* **2011**, *134*, 094509.
[76] R. Bukowski, K. Szalewicz, G. C. Groenenboom, A. van der Avoird, *J. Chem. Phys.* **2008**, *128*, 094313.

[77] W. Cencek, K. Szalewicz, C. Leforestier, R. van Harrevelt, A. van der Avoird, *Phys. Chem. Chem. Phys.* **2008**, *10*, 4716.
[78] C. Leforestier, K. Szalewicz, A. van der Avoird, *J. Chem. Phys*. **2012**, *137*, 014305.
[79] K. Szalewicz, G. Murdachaew, R. Bukowski, O. Akin-Ojo, C. Leforestier, in 'Lecture Series on Computer and Computational Science: ICCMSE 2006', Eds. G. Maroulis, T. Simos, Brill Academic, Leiden, **2006**, pp 482–491.
[80] U. Góra, W. Cencek, R. Podeszwa, A. van der Avoird, K. Szalewicz, *J. Chem. Phys.* **2014**, *140*, 194101.
[81] G. R. Medders, V. Babin, F. A. Paesani, *J. Chem. Theory Comput.* **2013**, *9*, 1103.
[82] V. Babin, C. Leforestier, F. Paesani, *J. Chem. Theory Comput.* **2013**, *9*, 5395.
[83] V. Babin, G. R. Medders, F. Paesani, *J. Chem. Theory Comput.* **2014**, *10*, 1599.
[84] V. Babin, G. R. Medders, F. Paesani, *J. Phys. Chem. Lett.* **2012**, *3*, 3765.
[85] G. R. Medders, V. Babin, F. Paesani, *J. Chem. Theory Comput*. **2014**, *10*, 2906.
[86] C. Vega, J. L. F. Abascal, *Phys. Chem. Chem. Phys.* **2011**, *13*, 19663.
[87] S. W. Rick, S. J. Stuart, *Rev. Comp. Chem.* **2002**, *18*, 89.
[88] J. W. Ponder, D. A. Case, *Adv. Prot. Chem.* **2003**, *66*, 27.
[89] P. E. M. Lopes, B. Roux, A. D. MacKerell Jr., *Theor. Chem. Acc.* **2009**, *124*, 11.
[90] G. Lamoureux, A. D. MacKerell Jr., B. Roux, *J. Chem. Phys.* **2003**, *119*, 5185.
[91] S. Cardamone, T. J. Hughes, P. L. A. Popelier, *Phys. Chem. Chem. Phys.* **2014**, *16*, 10367.
[92] J. Applequist, J. R. Carl, K.-K. Fung, *J. Am. Chem. Soc.* **1972**, *94*, 2952.
[93] K. T. Tang, J. P. Toennies, *J. Chem. Phys.* **1984**, *80*, 3726.
[94] J. Ischtwan, M. A. Collins, *J. Chem. Phys.* **1994**, *100*, 8080.
[95] R. P. A. Bettens, M. A. Collins, *J. Chem. Phys.* **1999**, *111*, 816.
[96] G. G. Maisuradze, D. L. Thompson, A. F. Wagner, M. Minkoff, *J. Chem. Phys.* **2003**, *119*, 10002.
[97] Y. Guo, A. Kawano, D. L. Thompson, A. F. Wagner, M. Minkoff, *J. Chem. Phys.* **2004**, *121*, 5091.
[98] J. F. Ouyang, M. W. Cvitkovic, R. P. A. Bettens, *J. Chem. Theory Comput.* **2014**, *10*, 3699.
[99] C. McBride, C. Vega, E. G. Noya, R. Ramírez, L. M. Sesé, *J. Chem. Phys.* **2009**, *131*, 024506.
[100] E. G. Noya, C. Vega, L. M. Sesé, R. Ramírez, *J. Chem. Phys.* **2009**, *131*, 124518.
[101] M. M. Conde, C. Vega, C. McBride, E. G. Noya, R. Ramírez, L. M. Sesé, *J. Chem. Phys.* **2010**, *132*, 114503.
[102] N. Agmon, *Chem. Phys. Lett.* **1995**, *244*, 456.
[103] C. Knight, G. A. Voth, *Acc. Chem. Res.* **2012**, *45*, 101.
[104] W. Shinoda, M. Shiga, *Phys. Rev. E* **2005**, 71, 041204.
[105] M. Shiga, W. Shinoda, *J. Chem. Phys.* **2005**, *123*, 134502.
[106] F. Paesani, G. A. Voth, *J. Phys. Chem. B* **2009**, *113*, 5702.
[107] M. E. Tuckerman, B. J. Berne, G. J. Martyna, M. L. Klein, *J. Chem. Phys.* **1993**, *99*, 2796.
[108] D. Marx, M. Parrinello, *J. Chem. Phys.* **1995**, *104*, 4077.
[109] M. Shiga, M. Tachikawa, S. Miura, *J. Chem. Phys.* **2001**, *115*, 9149.
[110] J. A. Barker, *J. Chem. Phys.* **1979**, *70*, 2914.
[111] R. A. Kuharski, P. J. Rossky, *J. Chem. Phys.* **1985**, *82*, 5164.
[112] S. Miura, *J. Chem. Phys.* **2007**, *126*, 114308.
[113] K. Suzuki, M. Tachikawa, M. Shiga, *J. Chem. Phys.* **2010**, *132*, 144108.
[114] J. Cao, G. A. Voth, *J. Chem. Phys.* **1994**, *100*, 5093.
[115] J. Cao, G. A. Voth, *J. Chem. Phys.* **1994**, *100*, 5106.
[116] J. Cao, G. A. Voth, *J. Chem. Phys.* **1994**, *101*, 6157.
[117] J. Cao, G. A. Voth, *J. Chem. Phys.* **1994**, *101*, 6168.
[118] J. Cao, G. A. Voth, *J. Chem. Phys.* **1994**, *101*, 6184.
[119] I. R. Criag, D. E. Manolopoulos, *J. Chem. Phys.* **2004**, *121*, 3368.
[120] B. J. Braams, D. E. Manolopoulos, *J. Chem. Phys.* **2006**, *125*, 124105.
[121] K. Laasonen, M. Sprik, M. Parrinello, R. Car, *J. Chem. Phys.* **1993**, *99*, 9080.

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

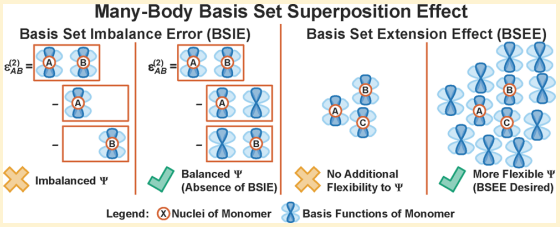
Many-Body Basis Set Superposition Effect

John F. Ouyang and Ryan P. A. Bettens*

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

Supporting Information

ABSTRACT: The basis set superposition effect (BSSE) arises in electronic structure calculations of molecular clusters when questions relating to interactions between monomers within the larger cluster are asked. The binding energy, or total energy, of the cluster may be broken down into many smaller subcluster calculations and the energies of these subsystems linearly combined to, hopefully, produce the desired quantity of interest. Unfortunately, BSSE can plague these smaller fragment calculations. In this work, we carefully examine the major sources of error associated with reproducing the binding energy and total energy of a molecular cluster. In order to do so, we decompose these energies in terms of a many-body expansion (MBE), where a "body" here refers to the monomers that make up the cluster. In our analysis, we found it necessary to introduce something we designate here as a many-ghost many-body expansion (MGMBE). The work presented here produces some surprising results, but perhaps the most significant of all is that BSSE effects up to the order of truncation in a MBE of the total energy cancel exactly. In the case of the binding energy, the only BSSE correction terms remaining arise from the removal of the one-body monomer total energies. Nevertheless, our earlier work indicated that BSSE effects continued to remain in the total energy of the cluster up to very high truncation order in the MBE. We show in this work that the vast majority of these high-order many-body effects arise from BSSE associated with the one-body monomer total energies. Also, we found that, remarkably, the complete basis set limit values for the three-body and four-body interactions differed very little from that at the MP2/aug-cc-pVDZ level for the respective subclusters embedded within a larger cluster.



1. INTRODUCTION

A major barrier to the theoretical study of large chemical systems is the fact that the computational effort of electronic structure methods increases drastically with system size. To circumvent this, one can look to fragmentation methods[1−4] where a large chemical system is broken up into numerous small subsystems. From there, only certain important interactions between these subsystems are considered for electronic structure calculations so as to recover the total energy of the system. Beyond a specified distance cutoff, each of the subsystems can be treated as multipoles (point charges, dipoles, quadrupoles, and so on) and interacted according to perturbation theory[5] to further reduce computational cost. Fundamentally, fragmentation methods are rooted in the many-body expansion (MBE), which decomposes the total energy of a cluster, E_{tot}, as the sum of one-body total energies, two-body pairwise interactions, three-body interactions, and so on, up to n-body interactions.[6] The effects of additional bodies are expected to diminish quickly with increasing bodies, allowing for a good estimate of E_{tot} by truncating the MBE at a low order, typically at or before the four-body interaction.[7−9]

However, the basis set superposition effect[10] (BSSE) comes into play as energy differences are involved in computing the many-body interactions. BSSE arises when monomers within a molecular cluster borrow basis functions from other monomers to compensate for their basis set incompleteness. The same applies to any subcluster within the molecular cluster. Thus, when the total energies of interacting monomers and their isolated counterparts are compared, there is an imbalance in the computed many-body interactions. To eliminate this basis set imbalance error (BSIE) for a dimer system, Boys and Bernardi proposed the counterpoise (CP) method to compute the binding energy where the monomer energies are calculated in the dimer basis.[11] To clarify our use of terminology, we use the term "*location* basis" to describe the placement of basis functions at the specified *location* in the cluster. The CP method was extended for many-monomer molecular clusters to give the Site−Site Function Counterpoise (SSFC) method by calculating the monomer energies in the cluster basis.[12] The many-body counterpoise (MBCP) method[13,14] was proposed later to approximate the expensive calculation of the monomer energies in the cluster basis by performing a MBE-like decomposition of the effects of the ghost functions present. Nonetheless, the consistent use of the cluster basis in the SSFC method allows for a meaningful decomposition of the binding energy into its many-body contributions. The SSFC method is not a unique extension of the CP method.[12,15] Valiron and Mayer proposed that the many-body interaction of a subcluster can be instead computed using the set of basis functions centered on the subcluster of interest, i.e., the subcluster basis.[16] These many-body interactions can then be summed to give the

Received: April 13, 2015
Published: September 23, 2015

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132−5143

Valiron−Mayer Function Counterpoise (VMFC) corrected binding energy. While both the SSFC and VMFC methods eliminate BSIE through the use of a consistent basis, the use of the cluster basis in the former incorporates an additional basis set extension effect (BSEE) where the monomers surrounding the subcluster of interest can extend their basis functions—functions present in the cluster basis but not the subcluster basis—to improve the quality of the computed many-body interactions.

The crucial point from the aforementioned counterpoise methods is that many-body BSSE can be divided into two components, namely the BSIE and BSEE (see section 2.2 for a detailed description). The BSIE (where the E stands for error) is undesirable, causing pairwise interactions and consequently the binding energy of large clusters to be overstabilizing. On the other hand, the BSEE (where the E stands for effect) is necessary to reproduce the binding energy and total energy of molecular clusters because all the monomers are better electronically described with the additional external basis functions. Earlier, we showed that the use of the cluster basis leads to rapid convergence of the MBE,[17] indicating that the BSEE is indeed present in the total energy. When the subcluster basis is used, we observed that the MBE converged rapidly, but to an "incorrect value." There is a significant difference between this "incorrect value" and the total energy, which is essentially the BSEE. More importantly, the rapid convergence associated with the subcluster basis suggested that the BSEE diminishes as rapidly as the many-body interactions. This is of relevance as we noticed that the many-body interactions computed using the subcluster basis are commonly employed in the construction of ab initio based potential energy surface (PES) in the literature.[18−21] Furthermore, perturbation theory is only compatible with many-body interactions computed using the subcluster basis as the former is BSSE-free by definition regardless of the type of BSSE. Thus, it is important to determine if BSEE is significant in these many-body interactions.

In this work, we examine the amount of BSSE, in particular BSEE, present in many-body interactions in order to identify the major sources of error associated with reproducing the binding energy and total energy of a molecular cluster via an MBE. First, we investigate whether the BSEE is significant in the many-body interactions up to the four-body level. Second, we introduce the many-ghost many-body expansion (MGMBE) to precisely and quantitatively account for both the BSIE and BSEE. Remarkably, we found that the oscillatory behavior of the MBE when diffuse functions are involved can be traced to the BSEE in the one-body interactions, i.e., the monomer total energies. Third, with the removal of the monomer total energies and associated BSEE, the MGMBE is able to accurately reproduce the binding energies of molecular clusters using the energies of numerous subclusters that are no larger than four monomers. Notably the utilization of embedded charges, or a coulomb field, is entirely unnecessary to accomplish this.

2. THEORY

Before discussing the theory behind the MBE, many-body BSSE, and MGMBE, we need to define the following terms and quantities which will be constantly used throughout this work. From here on, we denote the molecular cluster of interest simply as the "cluster" while a "subcluster" refers to a collection of monomers taken from the cluster. In the counterpoise methods, additional basis functions are placed on the locations of nuclei in the cluster, but without the nuclei being present in the electronic structure calculation, and these functions are called "ghost functions." We also use the term "*location* basis" to describe the *location* at which basis functions are placed in the calculation. For example, the cluster basis refers to the placement of basis functions at the locations of all nuclei present in the cluster. Each of the "bodies" in the many-body interactions refers to a monomer from the cluster, which is taken to be an individual water molecule in this work. When discussing the BSEE, we denote a "ghost-body" as the set of ghost functions centered on a monomer surrounding the subcluster of interest. Table 1 summarizes the relevant quantities described in this work.

Table 1. List of Important Quantities Presented in This Work, Followed by a Brief Definition and the Equation in Which It First Appeared

quantity	definition	eq
E_{tot}	total energy of a cluster.	1
$E_{\mathrm{A\cdots K\overline{L\cdots M}}}$	total energy of k-mer subcluster A···K calculated in the presence of ghost functions centered on L···M	4[a]
$\varepsilon^{\mathrm{C}}_{\mathrm{tot}}$	binding energy computed using the cluster basis	7
$E^{(k)}_{\mathrm{ext.}}$	basis set extension effect (BSEE) in the total k-body interaction	8
$\xi_{\mathrm{A\cdots K,L\cdots M}}$	BSEE from m-ghost-body L···M in the k-body interaction of A···K	9
$\varepsilon_{\mathrm{A\cdots K\overline{L\cdots M}}}$	k-body interaction of A···K computed using total energies calculated with basis functions centered on A···KL···M	10[b]

[a] $E_{\mathrm{A\cdots K\overline{L\cdots M}}}$ is mentioned much earlier in text at the beginning of section 2.2. [b] $\varepsilon_{\mathrm{A\cdots K\overline{L\cdots M}}}$ is defined and explained much earlier in text at the second paragraph of section 2.3.

2.1. Many-Body Expansion. For a cluster containing n monomers, the MBE allows us to decompose the total energy of the cluster, E_{tot}, into its many-body contributions

$$E_{\mathrm{tot}} = \sum_{\mathrm{A}}^{\binom{n}{1}} \varepsilon'_{\mathrm{A}} + \sum_{\mathrm{A<B}}^{\binom{n}{2}} \varepsilon'_{\mathrm{AB}} + \sum_{\mathrm{A<B<C}}^{\binom{n}{3}} \varepsilon'_{\mathrm{ABC}} + \sum_{\mathrm{A<B<C<D}}^{\binom{n}{4}} \varepsilon'_{\mathrm{ABCD}} + \ldots + \varepsilon'_{\mathrm{A\cdots N}} \quad (1)$$

where $\varepsilon'_{\mathrm{A\cdots K}}$ is the k-body interaction of the k-mer subcluster A···K, of which there are $\binom{n}{k}$ of such terms (Figure 1). The prime symbol in $\varepsilon'_{\mathrm{A\cdots K}}$ indicates that the basis functions are placed exclusively at the location of the nuclei; i.e., no ghost functions are involved. In this work, we truncate the MBE at the four-body level and thus only provide the relevant equations up to the four-body interaction.

$\varepsilon'_{\mathrm{A\cdots K}}$ is not directly obtainable from electronic structure calculations, which only gives the total energy, $E_{\mathrm{A\cdots K}}$, of the k-mer subcluster of interest. Thus, we need to write the many-body interactions in terms of the total energies. $\varepsilon'_{\mathrm{A\cdots K}}$ is defined recursively using lower-body interactions[2,6,7,22] and can then be expressed in terms of total energies

$$\varepsilon'_{\mathrm{A}} = E_{\mathrm{A}} \quad (2a)$$

$$\varepsilon'_{\mathrm{AB}} = E_{\mathrm{AB}} - (E_{\mathrm{A}} + E_{\mathrm{B}}) \quad (2b)$$

$$\varepsilon'_{\mathrm{ABC}} = E_{\mathrm{ABC}} - (E_{\mathrm{AB}} + E_{\mathrm{AC}} + E_{\mathrm{BC}}) + (E_{\mathrm{A}} + E_{\mathrm{B}} + E_{\mathrm{C}}) \quad (2c)$$

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132−5143

Journal of Chemical Theory and Computation Article

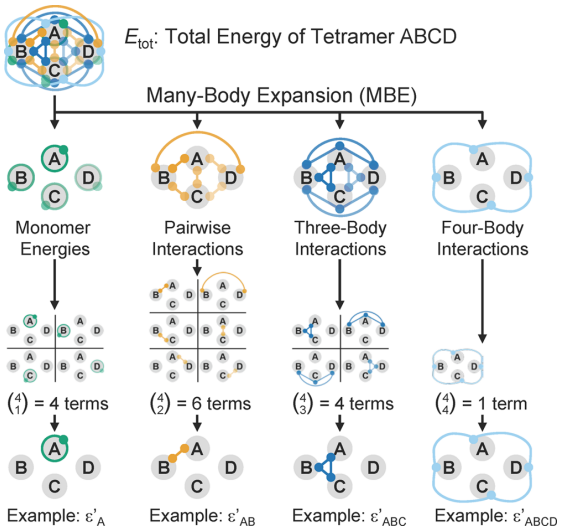


Figure 1. MBE allows us to easily identify the numerous interactions between monomers that is encompassed within the total energy of the cluster. The MBE is illustrated here for a tetramer ABCD (n = 4) where the total energy, E_{tot}, is decomposed into the total k-body interactions, k = 1−4, which comprises $\binom{n}{k}$ individual terms. The explicit formulas for calculating each individual k-body interaction are given in eqs 2a−2d in the text.

$$\varepsilon'_{ABCD} = E_{ABCD} - (E_{ABC} + E_{ABD} + E_{ACD} + E_{BCD}) + (E_{AB} + E_{AC} + E_{AD} + E_{BC} + E_{BD} + E_{CD}) - (E_A + E_B + E_C + E_D) \quad (2d)$$

Here, the one-body interaction, ε'_A, is the total energy of isolated monomer A while the two-body interaction, ε'_{AB}, gives the pairwise interaction between monomers A and B. The three-body interaction, ε'_{ABC}, can be understood as the effect of a third monomer C on the interaction between the other two monomers A and B, and the higher-body interactions can be interpreted similarly. In order to employ the MBE, the k-body interaction of individual subclusters presented in eqs 2a−2d has to be collected to give the total k-body interaction. Many total energy terms are repeated during this collection process, and a compact expression of the total k-body interaction is derived and presented in the Supporting Information.

Other than expressing the total energy of the cluster using the MBE, another quantity of interest is the binding energy of the cluster

$$\varepsilon_{tot} = E_{tot} - \sum_{A}^{\binom{n}{1}} E_A = \sum_{A<B}^{\binom{n}{2}} \varepsilon'_{AB} + \sum_{A<B<C}^{\binom{n}{3}} \varepsilon'_{ABC} + \sum_{A<B<C<D}^{\binom{n}{4}} \varepsilon'_{ABCD} + \ldots + \varepsilon'_{A\cdots N} \quad (3)$$

2.2. Many-Body Basis Set Superposition Effect. In the electronic structure calculation of a cluster, basis functions from one monomer can be exploited by other monomers to compensate for their basis set incompleteness, and this is known as BSSE. In many-body systems, this BSSE can be divided into two components, namely the basis set imbalance error (BSIE) and basis set extension effect (BSEE). The distinction between these two components becomes clear when we compare the various placements of basis functions, i.e., the *location* basis, in computing the many-body interactions (Table 2). So far, the total energy, $E_{A\cdots K}$, is written such that it is determined by the identity of the k-mer subcluster or, more specifically, the location of the nuclei constituting the subcluster. In the context of molecular orbital based electronic structure calculations, the total energy also depends on the placement of basis functions. For example, the total energy of monomer A calculated in the nuclei-centered basis centered on A alone would be different from that using the set of basis functions centered on the cluster A···N. From here on, we use the more general notation, $E_{A\cdots K\overline{L\cdots M}}$. The overline in the subscript indicates the presence of ghost functions centered on L···M in the electronic structure calculation.

Table 2. Comparison of the Choice of Basis in Computing the Many-Body Interactions and Binding Energy, Together with the Name of the Method Reported in the Literature[a]

location basis	nuclei-centered	subcluster	cluster
many-body interactions	$\varepsilon'_{A\cdots K}$	$\varepsilon_{A\cdots K}$	$\varepsilon_{A\cdots K\overline{L\cdots N}}$
eq for many-body interactions	eqs 2a−2d	eqs 4a−4d	eqs 6a−6d
binding energy	ε_{tot}	ε^{S}_{tot}	ε^{C}_{tot}
eq for binding energy	eq 3	eq 5	eq 7
name in literature	uncorrected[b]	VMFC/ *k*CBS	SSFC/CP
absence of BSIE?[c]	no	yes	yes
presence of BSEE?[c]	no	no	yes
does MBE converge to E_{tot}?	yes	no	yes
MBE convergence	slow, oscillatory	rapid	rapid
further remarks	default basis	agree with PT	very expensive

[a]Furthermore, the absence of the BSIE and presence of BSEE in the many-body interactions are compared with their effects on the convergence of the MBE of the total energy. [b]There is no formal name for the nuclei-centered basis as it is the default basis in electronic structure calculations. [c]As summarized toward the end of section 2.2, the *absence* of BSIE and the *presence* of BSEE is desirable, as in the case of the cluster basis.

The most straightforward method to compute the many-body interactions is to place basis functions exclusively at the locations of nuclei in the electronic structure calculations. This is the usual way of calculating an electronic energy of a molecule. Thus, the k-body interaction computed using the nuclei-centered basis follows eqs 2a−2d presented earlier. We emphasize again that the prime symbol in the many-body interaction, $\varepsilon'_{A\cdots K}$, indicates the nuclei-centered basis where the number of basis functions are different across the different total energy terms. This is in contrast to the "consistent" subcluster and cluster basis which will be introduced shortly. Similarly, the binding energy computed using the nuclei-centered basis, ε_{tot}, follows eq 3. The ε_{tot} is often called the uncorrected binding energy for reasons that will be obvious in the following discussion. In computing ε'_{AB} using eq 2b, it is clear that E_{AB} is calculated using more basis functions as compared to E_A and E_B. In calculating E_{AB}, monomer A can utilize basis functions centered on monomer B to improve the description of its wave function and vice versa. This is obviously absent in the

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132−5143

calculation of E_A and E_B. This imbalance in the number of basis functions used in the three different calculations of the total energy is the origin of the BSIE. The same BSIE manifests in higher-body interactions in eqs 2c and 2d and the binding energy in eq 3. For the two-body interactions, the BSIE leads to the interactions being overstabilizing. Previously, we also found that MBEs using $\varepsilon'_{A\cdots K}$ exhibit slow and oscillatory convergence, especially when diffuse basis functions are present.[17]

To remove the BSIE in many-body interactions, we need to ensure that there is a common set of basis functions employed in each of the total energy calculation. The smallest common set is one that is centered on the subcluster for which the many-body interaction is computed. We denote this as the subcluster basis and the k-body interaction can be written as

$$\varepsilon_A = E_A \tag{4a}$$

$$\varepsilon_{AB} = E_{AB} - (E_{A\overline{B}} + E_{B\overline{A}}) \tag{4b}$$

$$\varepsilon_{ABC} = E_{ABC} - (E_{AB\overline{C}} + E_{AC\overline{B}} + E_{BC\overline{A}}) + (E_{A\overline{BC}} + E_{B\overline{AC}} + E_{C\overline{AB}}) \tag{4c}$$

$$\varepsilon_{ABCD} = E_{ABCD} - (E_{ABC\overline{D}} + E_{ABD\overline{C}} + E_{ACD\overline{B}} + E_{BCD\overline{A}}) + (E_{AB\overline{CD}} + E_{AC\overline{BD}} + E_{AD\overline{BC}} + E_{BC\overline{AD}} + E_{BD\overline{AC}} + E_{CD\overline{AB}}) - (E_{A\overline{BCD}} + E_{B\overline{ACD}} + E_{C\overline{ABD}} + E_{D\overline{ABC}}) \tag{4d}$$

where the total energy terms containing the same number of monomers are grouped together to reduce clutter. For each of the many-body interactions, the same set of basis functions centered on the subcluster of interest is employed for all the total energy calculations and this introduces ghost functions denoted by the overline in the subscript of total energy terms. Note that, unlike the nuclei-centered basis, the prime symbol is not present here as there is a consistent number of basis functions in each total energy calculation. The binding energy computed using the subcluster basis is

$$\varepsilon^{S}_{tot} = \sum_{A<B}^{\binom{n}{2}} \varepsilon_{AB} + \sum_{A<B<C}^{\binom{n}{3}} \varepsilon_{ABC} + \sum_{A<B<C<D}^{\binom{n}{4}} \varepsilon_{ABCD} + \ldots + \varepsilon_{A\cdots N} \tag{5}$$

The ε^{S}_{tot} is known as the Valiron−Mayer function counterpoise (VMFC) corrected binding energy[16] in the literature. In our previous study,[17] we referred to the subcluster basis as the kCBS method as named by Góra et al.[23] The subcluster basis is the standard way of predicting many-body interactions in the construction of ab intito based PES as it is free of BSIE.[18−21,24] Furthermore, these many-body interactions are reproduced with high accuracy using multipoles and perturbation theory[5]—which are BSSE-free by definition—at intermediate to long intermolecular separations. Unlike the ε_{tot} in eq 3, we cannot express ε^{S}_{tot} as the difference between the total energy and the monomer total energies. This is because the sum of the $\varepsilon_{A\cdots K}$ does not add up to the total energy; i.e., eq 1 does not hold true here. This is related to the fact that total energies between different $\varepsilon_{A\cdots K}$'s cannot be reused. For example, the different E_A and $E_{A\overline{B}}$ are involved in computing ε_A and ε_{AB}, respectively, whereas the nuclei-centered counterpart would only require the same E_A in both cases. Thus, when the $\varepsilon_{A\cdots K}$'s are summed in a MBE according to eq 1, the total energy terms do not cancel to give the exact total energy eventually. This implies that there are some effects present in the total energy that are not accounted for in the subcluster basis. In fact, this is due to the second component of BSSE—the BSEE.

Apart from the subcluster basis, another common set of basis functions that remove BSIE is one that is centered on the cluster. We denote this as the cluster basis and the k-body interaction can be written as

$$\varepsilon_A = E_{A\overline{B\cdots N}} \tag{6a}$$

$$\varepsilon_{AB} = E_{AB\overline{C\cdots N}} - (E_{A\overline{BC\cdots N}} + E_{B\overline{AC\cdots N}}) \tag{6b}$$

$$\varepsilon_{ABC} = E_{ABC\overline{D\cdots N}} - (E_{AB\overline{CD\cdots N}} + E_{AC\overline{BD\cdots N}} + E_{BC\overline{AD\cdots N}}) + (E_{A\overline{BCD\cdots N}} + E_{B\overline{ACD\cdots N}} + E_{C\overline{ABD\cdots N}}) \tag{6c}$$

$$\varepsilon_{ABCD} = E_{ABCD\overline{E\cdots N}} - (E_{ABC\overline{DE\cdots N}} + E_{ABD\overline{CE\cdots N}} + E_{ACD\overline{BE\cdots N}} + E_{BCD\overline{AE\cdots N}}) + (E_{AB\overline{CDE\cdots N}} + E_{AC\overline{BDE\cdots N}} + E_{AD\overline{BCE\cdots N}} + E_{BC\overline{ADE\cdots N}} + E_{BD\overline{ACE\cdots N}} + E_{CD\overline{ABE\cdots N}}) - (E_{A\overline{BCDE\cdots N}} + E_{B\overline{ACDE\cdots N}} + E_{C\overline{ABDE\cdots N}} + E_{D\overline{ABCE\cdots N}}) \tag{6d}$$

For all the total energy calculations, the set of basis functions centered on the entire cluster is employed. Thus, basis functions centered on other monomers surrounding the subcluster of interest are involved, indicated by the overline in the many-body interaction. For example, computing $\varepsilon_{AB\overline{C\cdots N}}$ in eq 6b requires total energies involving basis functions centered on C···N surrounding the subcluster AB. The binding energy computed using the cluster basis is

$$\varepsilon^{C}_{tot} = E_{tot} - \sum_{A}^{\binom{n}{1}} E_{A\overline{B\cdots N}} = \sum_{A<B}^{\binom{n}{2}} \varepsilon'_{AB\overline{C\cdots N}} + \sum_{A<B<C}^{\binom{n}{3}} \varepsilon'_{ABC\overline{D\cdots N}} + \sum_{A<B<C<D}^{\binom{n}{4}} \varepsilon'_{ABCD\overline{E\cdots N}} + \ldots + \varepsilon'_{A\cdots N} \tag{7}$$

The ε^{C}_{tot} is named the site−site function counterpoise (SSFC) corrected binding energy.[12] It is commonly referred to simply as the counterpoise (CP) method as the binding energy is a direct generalization of the CP method for a dimer system.[11] The cluster basis ensures that a common set of basis functions is employed in each of the total energy calculations, removing the undesirable BSIE. Furthermore, all the total energy terms employ the same basis, allowing for the sum of these many-body interactions to add up to the total energy according to eq 1. Comparing the subcluster basis and cluster basis, there is an additional effect in the latter where the ghost functions surrounding the subcluster, e.g. functions centered on C···N in the case of $\varepsilon_{AB\overline{C\cdots N}}$, improve the many-body interaction associated with the subcluster. This is the BSEE. Mathematically, we define the BSEE in the total k-body interaction, $E^{(k)}_{ext}$, as the difference between the total k-body interaction computed using the cluster basis and subcluster basis

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132−5143

Journal of Chemical Theory and Computation Article

$$E^{(1)}_{\text{ext.}} = \sum_{\text{A}}^{\binom{n}{1}} (\varepsilon_{\text{A}\overline{\text{B}\cdots\text{N}}} - \varepsilon_{\text{A}}) \tag{8a}$$

$$E^{(2)}_{\text{ext.}} = \sum_{\text{A<B}}^{\binom{n}{2}} (\varepsilon_{\text{AB}\overline{\text{C}\cdots\text{N}}} - \varepsilon_{\text{AB}}) \tag{8b}$$

$$E^{(3)}_{\text{ext.}} = \sum_{\text{A<B<C}}^{\binom{n}{3}} (\varepsilon_{\text{ABC}\overline{\text{D}\cdots\text{N}}} - \varepsilon_{\text{ABC}}) \tag{8c}$$

$$E^{(4)}_{\text{ext.}} = \sum_{\text{A<B<C<D}}^{\binom{n}{4}} (\varepsilon_{\text{ABCD}\overline{\text{E}\cdots\text{N}}} - \varepsilon_{\text{ABCD}}) \tag{8d}$$

Unlike the BSIE, the BSEE is important in reproducing the total energy of a cluster. We have previously shown that the MBEs using $\varepsilon_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{N}}}$ exhibit rapid convergence to the total energy by the four-body interaction.[17] This indicates that the total energy contains the BSEE as part of the variational optimization and/or the perturbative treatment of electron correlation in the electronic structure calculation of the total energy. The borrowing of basis functions from other monomers surrounding the subcluster does improve the flexibility of the wave function of the subcluster and consequently the quality of the many-body interaction computed. On a side note, this is likewise true in valence-bonded systems where the bonding between atoms can be improved by the basis functions from other surrounding atoms. This importance of BSEE also applies to the binding energy where the BSEE should be incorporated into the many-body interactions used to compute the binding energy. Therefore, we consider $\varepsilon^{\text{C}}_{\text{tot}}$, and *not* $\varepsilon^{\text{S}}_{\text{tot}}$, to be the best estimate of the binding energy at a given level of theory and basis set.

To summarize many-body BSSE, there are two components, namely the BSIE and BSEE. The first component is undesirable, arising from an imbalance in the number of basis functions when computing energy differences in the many-body interactions and the binding energy. This BSIE can be removed by using a common set of basis functions in each of the total energy calculations which can be fulfilled by the use of the subcluster basis or the cluster basis. The second component originates from the extension of the subcluster basis due to the presence of monomers surrounding the subcluster in the cluster. This BSEE is necessary to reproduce the binding energy and total energy of the cluster and can be accounted for using the cluster basis. However, computing the many-body interactions in the cluster basis is very expensive and defeats the usefulness of the MBE in decomposing a large many-body system into manageable few-body subsystems. Thus, we wish to analyze the amount of many-body BSSE present so as to accurately yet cheaply reproduce the binding energy and total energy.

2.3. Many-Ghost Many-Body Expansion. To account for both the BSIE and BSEE (section 2.2), we introduce the many-ghost many-body expansion (MGMBE). The MGMBE defines these two components of many-body BSSE up to the order of truncation of the many-body interactions, allowing us to establish the amount of many-body BSSE present in the many-body interactions. The MGMBE performs a two-dimensional many-body decomposition with each decomposition accounting for one component of many-body BSSE (Figure 2). The

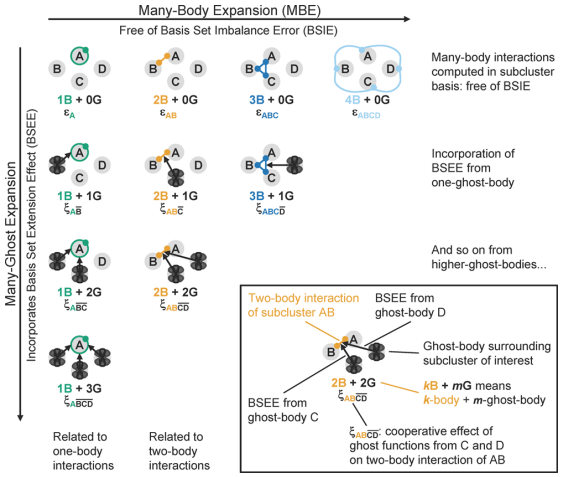


Figure 2. MGMBE performs a two-dimensional many-body decomposition with the first being an MBE (left to right) up to the k-body interaction computed using the subcluster basis, $\varepsilon_{\text{A}\cdots\text{K}}$, using eqs 4a−4d. It is important to note that the BSIE is removed by performing this calculation. A second many-ghost expansion (top to bottom) then decomposes the BSEE from the cluster basis into contributions from m-ghost-body, represented by the black lobes. These BSEE terms, $\xi_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{M}}}$, are shown here for $k + m \leq 4$, which can be computed using eqs 10a−10f. The inset explains the various symbols in the figure while the comments at the right and bottom edge summarizes the components of the MGMBE along each row and column, respectively. We also note that along the diagonal where $k + m$ is constant, the BSIE and BSEE cancel as these terms share the same basis functions.

first decomposition involves the MBE (section 2.1) using the many-body interactions computed using the subcluster basis, $\varepsilon_{\text{A}\cdots\text{K}}$, ensuring that these interactions are free of the BSIE. The second decomposition, denoted the many-ghost expansion, breaks down the BSEE present in the cluster basis into contributions from one ghost body, two ghost bodies, and so on, up to $(n - k)$ ghost bodies. To reiterate, a ghost body refers to the set of ghost functions centered on a monomer surrounding the subcluster of interest. Both decompositions can be truncated at a low order to hopefully reproduce the binding energy and total energy of the cluster at a low computational cost. We note that the MGMBE is a logical extension of the earlier many-body counterpoise (MBCP) method.[13,14] The MBCP method seeks to cheaply approximate the $\varepsilon^{\text{C}}_{\text{tot}}$ in eq 7 by performing two separate many-body decompositions on both the E_{tot} and $E_{\text{AB}\cdots\overline{\text{N}}}$. The former decomposition is an MBE computed using the nuclei-centered basis while the latter decomposition is essentially the many-ghost expansion performed on the monomer total energies. In the MGMBE, we extend the many-ghost expansion for any arbitrary k-body interaction to identify the BSEE present.

Previously, we mentioned that the total energy of a subcluster does not depend solely on the identity of the subcluster of interest but also on the placement of basis functions. To recap, we denoted $E_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{M}}}$ as the total energy of subcluster A···K calculated in the presence of additional ghost functions centered on L···M. The same situation applies to the many-body interactions, evident from the discussion on the nuclei-centered, subcluster and cluster basis in section 2.2. Here, we denote $\varepsilon_{\text{A}\cdots\text{K}\overline{\text{L}\cdots\text{M}}}$ as the k-body interaction of the k-

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132−5143

mer subcluster A···K computed using total energies calculated with the set of basis functions centered on A···KL···M. The overline in the subscript denotes the ghost bodies, namely the set of ghost functions centered on monomers L···M surrounding the subcluster. For example, $\varepsilon_{AB\overline{CD}} = E_{AB\overline{CD}} - (E_{A\overline{BCD}} + E_{B\overline{ACD}})$. In particular, the many-body interaction computed using the subcluster basis, $\varepsilon_{A\cdots K}$, and the cluster basis, $\varepsilon_{A\cdots K\overline{L\cdots N}}$, are specific cases of this general notation. In the former, there are no ghost bodies involved, while the entire cluster (excluding the subcluster of interest) constitutes all the ghost bodies in the latter case. Now, we can write the MGMBE of the total energy as

$$E_{tot} = \sum_{A}^{\binom{n}{1}} \varepsilon_A + \sum_{A<B}^{\binom{n}{2}} \varepsilon_{AB} + \sum_{A<B<C}^{\binom{n}{3}} \varepsilon_{ABC} + \sum_{A<B<C<D}^{\binom{n}{4}} \varepsilon_{ABCD} + \ldots + \sum_{A,B}^{\binom{n}{1}\cdot\binom{n-1}{1}} \xi_{A\overline{B}} + \sum_{A<B,C}^{\binom{n}{2}\cdot\binom{n-2}{1}} \xi_{AB\overline{C}} + \sum_{A<B<C,D}^{\binom{n}{3}\cdot\binom{n-3}{1}} \xi_{ABC\overline{D}} + \ldots + \sum_{A,B<C}^{\binom{n}{1}\cdot\binom{n-1}{1}} \xi_{A\overline{BC}} + \sum_{A<B,C<D}^{\binom{n}{2}\cdot\binom{n-2}{2}} \xi_{AB\overline{CD}} + \ldots + \sum_{A,B<C<D}^{\binom{n}{1}\cdot\binom{n-1}{3}} \xi_{A\overline{BCD}} + \ldots + \xi_{higher} \quad (9)$$

where $\xi_{A\cdots K\overline{L\cdots M}}$ is the BSEE from m-ghost-body L···M in the k-body interaction of k-mer subcluster A···K, of which there are $\binom{n}{k}\cdot\binom{n-k}{m}$ of such terms. The first line in eq 9 gives the MBE using many-body interactions computed using the subcluster basis using eqs 4a–4d). While these many-body interactions are free of BSIE, they lack the important BSEE. These missing BSEE terms are added in the following lines with each line introducing contributions from a different number of ghost bodies. For cases where $k + m \leq 4$, $\xi_{A\cdots K\overline{L\cdots M}}$ can be expressed as

$$\xi_{A\overline{B}} = \varepsilon_{A\overline{B}} - \varepsilon_A \quad (10a)$$

$$\xi_{AB\overline{C}} = \varepsilon_{AB\overline{C}} - \varepsilon_{AB} \quad (10b)$$

$$\xi_{ABC\overline{D}} = \varepsilon_{ABC\overline{D}} - \varepsilon_{ABC} \quad (10c)$$

$$\xi_{A\overline{BC}} = \varepsilon_{A\overline{BC}} - (\varepsilon_{A\overline{B}} + \varepsilon_{A\overline{C}}) + \varepsilon_A \quad (10d)$$

$$\xi_{AB\overline{CD}} = \varepsilon_{AB\overline{CD}} - (\varepsilon_{AB\overline{C}} + \varepsilon_{AB\overline{D}}) + \varepsilon_{AB} \quad (10e)$$

$$\xi_{A\overline{BCD}} = \varepsilon_{A\overline{BCD}} - (\varepsilon_{A\overline{BC}} + \varepsilon_{A\overline{BD}} + \varepsilon_{A\overline{CD}}) + (\varepsilon_{A\overline{B}} + \varepsilon_{A\overline{C}} + \varepsilon_{A\overline{D}}) - \varepsilon_A \quad (10f)$$

The meaning of these terms can be better understood by looking at specific examples. For example, $\xi_{ABC\overline{D}}$ in eq 10c quantifies the amount by which the ghost functions centered on D affect the three-body interaction of ABC, i.e., the BSEE from D on ε_{ABC}. Likewise, $\xi_{AB\overline{CD}}$ in eq 10e gives the cooperative effect of the ghost functions centered on both C and D on the two-body interaction of AB, and higher-ghost-body BSEEs can be interpreted similarly. These terms represent the many-body decomposition of the BSEE present in the cluster basis. As such, eqs 10a–10c, eqs 10d and 10e, and eq 10f resemble eq 2a, eq 2b, and eq 2c, respectively. Upon comparison between the two sets of equations, there is an additional $\varepsilon_{A\cdots C}$ term (last term in each equation) in eq 10. This is the 0-ghost-body term where there is no BSEE and the equivalent in a MBE corresponds to a 0-body interaction which is zero and thus omitted in the many-body interaction expressions. In order to compute the $\xi_{A\cdots K\overline{L\cdots M}}$ terms, all the $\varepsilon_{A\cdots K\overline{L\cdots M}}$ terms have to be expressed in terms of total energies that can be readily obtained from electronic structure calculations. Here, we give an example where we express $\xi_{AB\overline{CD}}$ in terms of total energies

$$\begin{aligned}\xi_{AB\overline{CD}} &= \varepsilon_{AB\overline{CD}} - \varepsilon_{AB\overline{C}} - \varepsilon_{AB\overline{D}} + \varepsilon_{AB} \\ &= (E_{AB\overline{CD}} - E_{A\overline{BCD}} - E_{B\overline{ACD}}) \\ &\quad - (E_{AB\overline{C}} - E_{A\overline{BC}} - E_{B\overline{AC}}) \\ &\quad - (E_{AB\overline{D}} - E_{A\overline{BD}} - E_{B\overline{AD}}) + (E_{AB} - E_{A\overline{B}} - E_{B\overline{A}})\end{aligned} \quad (11)$$

From eq 11, we observe that the maximum number of basis functions is limited to that of four monomers in computing $\xi_{AB\overline{CD}}$. In fact, the maximum number of basis functions is limited to $(k + m)$ monomers in computing $\xi_{A\cdots K\overline{L\cdots M}}$. It is also obvious that rewriting the $\xi_{A\cdots K\overline{L\cdots M}}$ terms in terms of total energies can lead to cumbersome expressions. Fortunately, many total energy terms are repeated across different $\xi_{A\cdots K\overline{L\cdots M}}$'s which can be collected to give a more compact expression when all the $\xi_{A\cdots K\overline{L\cdots M}}$ terms are summed. The derivation of these working equations is presented in the Supporting Information.

The two many-body decompositions in the MGMBE can be truncated at a low order to hopefully reproduce the binding energy and total energy of a cluster. Given that the two decompositions are independent, the BSEE present in each of the k-body interactions can be truncated at a different m ghost body. A prudent choice would be to truncate at order (k, m) such that $k + m = \alpha$, keeping the maximum number of basis functions in each electronic structure calculation to that of α monomers. For example, truncating the MGMBE at $\alpha = 2$ would include the ε_A, ε_{AB}, and $\xi_{A\overline{B}}$ terms while truncation at $\alpha = 3$ includes the previously mentioned terms as well as ε_{ABC}, $\xi_{AB\overline{C}}$, and $\xi_{A\overline{BC}}$ terms.

A surprising result surfaced when the truncation order of the MGMBE is such that $k + m = \alpha$. Careful analysis of the working equations in the Supporting Information revealed that all the total energies involving any ghost functions vanish when we sum the $\varepsilon_{A\cdots K}$ and $\xi_{A\cdots K\overline{L\cdots M}}$ terms with $k + m = \alpha$, where α is a constant. Consequently, we obtain the many-body interactions computed using the nuclei-centered basis from this summation. This point is worth emphasizing, which is the inclusion of BSEE terms in the MGMBE cancels exactly all of the total energies involving any ghost functions in eq 9. This implies that an MBE using the nuclei-centered basis truncated at α bodies incorporates some BSEE, in particular contributions from up to $m = (\alpha - k)$ ghost bodies in each of the k-body interactions. We stress that this surprising result only occurs when the MGMBE terms are summed *across* different k number of interacting bodies to obtain either the binding energy or total energy. To illustrate this cancellation, let us consider the sum of $\xi_{A\overline{B}}$, $\xi_{B\overline{A}}$, and ε_{AB}. The first two terms would be $\xi_{A\overline{B}} = E_{A\overline{B}} - E_A$ and $\xi_{B\overline{A}} = E_{B\overline{A}} - E_B$, respectively, and the total energies involving ghost functions would be eliminated when we include the $\varepsilon_{AB} = E_{AB} - E_{A\overline{B}} - E_{B\overline{A}}$. Thus, we are left with the $\varepsilon'_{AB} = E_{AB} - E_A - E_B$. In essence, the BSEE in $\xi_{A\cdots K\overline{L\cdots M}}$ replaces the total energy terms involving ghost functions in the $\varepsilon_{A\cdots K}$ with corresponding ghost-free terms, "transforming" it into the nuclei-centered counterpart, $\varepsilon'_{A\cdots K}$. Expressed alternatively—the

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132–5143

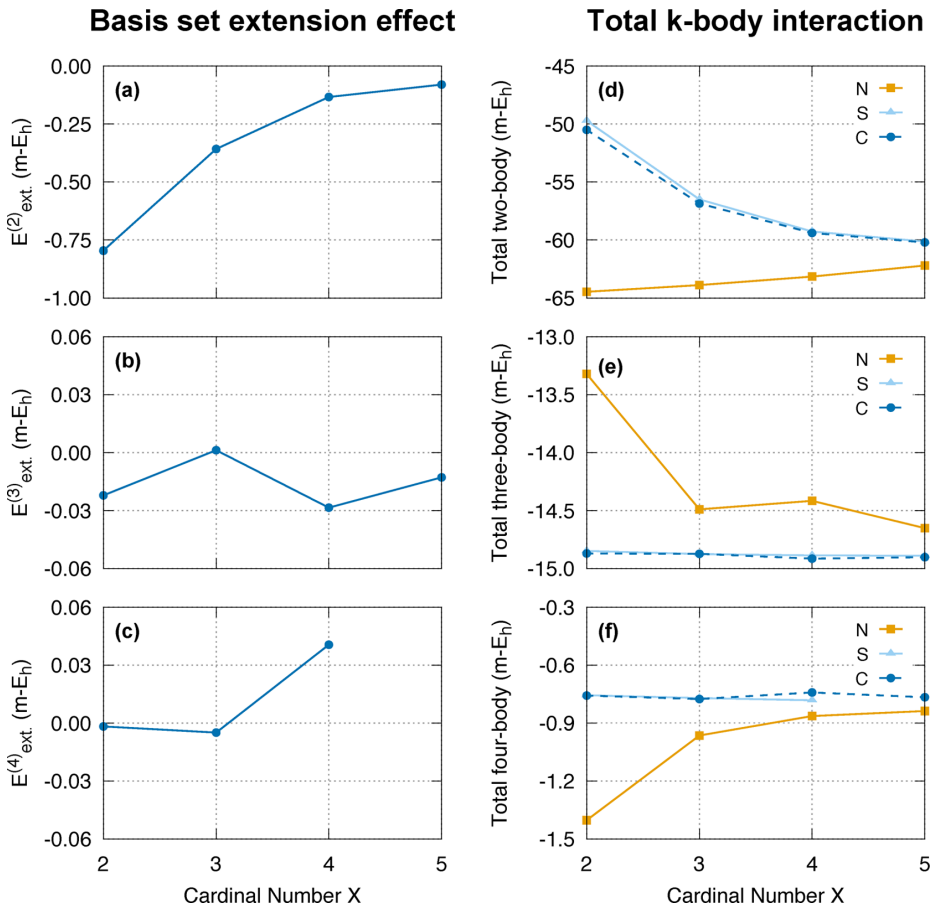


Figure 3. (a–c) BSEE in the total k-body interaction, $E^{(k)}_{\text{ext.}}$, as defined in eq 8 and the (d–f) total k-body interaction for the cage isomer of $(H_2O)_6$ with increasing basis set quality at MP2/AVXZ. The total k-body interactions are computed using various *location* bases, namely the nuclei-centered (N), subcluster (S), and cluster (C) basis described in section 2.2 to determine the effects of many-body BSSE on the many-body interactions. In particular, the lines for the cluster basis are dashed to show clearly the similarities between that and the subcluster basis results. The $E^{(4)}_{\text{ext.}}$ and total four-body interaction computed using the subcluster basis at MP2/AVSZ are omitted due to steep computational cost.

BSIE for a higher-body interaction (something that must be subtracted) is an BSEE for a lower-body interaction (something that must be added)—and the two effects cancel each other exactly!

3. RESULTS AND DISCUSSION

All quantum chemical calculations were performed using the MOLPRO suite of programs.[25] Calculations were carried out at the second-order Møller–Plesset perturbation (MP2) level of theory using the aug-cc-pVnZ, labeled AVnZ, n = D, T, Q, 5, Dunning correlation-consistent basis sets augmented with diffuse functions.[26,27] The explicitly correlated MP2 (MP2-F12) level of theory[28] was employed with the AVDZ basis set.

3.1. Basis Set Extension Effect in Many-Body Interactions. In our previous work, we observed rapid convergence in the MBEs using either the subcluster or cluster basis. This indirectly suggests that the difference between these two MBEs—the BSEE—should converge rapidly with the number of bodies. We computed the $E^{(k)}_{\text{ext.}}$ and total k-body interaction for the $(H_2O)_6$ cage and prism isomers up to the four-body term with increasing basis set quality. Both isomers are taken from Richard et al.[14] and showed similar trends. Thus, the data for the cage isomer are shown in Figure 3, while the prism isomer counterparts are in Figure S1 of the Supporting Information. Similar studies exist in the literature but are performed on small trimer and tetramer clusters[29,30] or focused on the binding energy.[14,31] Instead, we choose to separately examine the BSEE in each of the total k-body interactions, especially between the two-body and the three-and-higher-body interactions, because they are dominated by different intermolecular interactions.[5]

At the two-body level (Figure 3a), the $E^{(2)}_{\text{ext.}}$ is always negative, indicating that the additional ghost functions in the cluster basis help to lower the two-body interactions. As expected, increasing the quality of the basis set decreases this borrowing of basis functions to improve the two-body interactions. These BSEEs are generally small, below 1 m-E_h, because the additional basis functions in the cluster basis are not centered on the nuclei or on regions between nuclei where the interaction occurs. This is in contrast to the use of midbond functions where the placement of basis functions at regions between interacting molecules improves the description of the interaction.[32] We

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132–5143

point out that the $E^{(k)}_{ext.}$ also serves as an error indicator of how well the many-body interactions computed using the subcluster basis can be used in place of the cluster basis counterpart to reproduce the binding energy or total energy. Thus, the $E^{(2)}_{ext.}$ can still be substantial if very high accuracy is demanded. For the higher-body interactions (Figure 3b,c), the $E^{(3)}_{ext.}$ and $E^{(4)}_{ext.}$ are minuscule—smaller than 0.045 m-E_h—and we can treat the many-body interactions computed in both the subcluster and cluster basis to be practically the same. This is of comfort as the use of the subcluster basis renders the construction of MBE-based ab initio water potentials[24,33] possible. The reduction in dimensionality from applying the MBE is preserved unlike the cluster basis which depends on the geometry of the cluster. Indeed, the many-body interactions computed using the subcluster basis were used to construct ab initio based PES to study large water clusters and bulk water.[18−21]

The tiny $E^{(3)}_{ext.}$ and $E^{(4)}_{ext.}$ bring us to an unrelated but important result. At the complete basis set (CBS) limit, there is no BSSE, i.e., $E^{(k)}_{ext.} = 0$. While the converse is not necessarily true, it is worthwhile to investigate if the CBS limit can be approximated using moderate-sized basis sets. Clearly, this is true for the three-body (Figure 3e) and four-body interactions (Figure 3f). Both the total three-body and four-body interaction computed using the subcluster or cluster basis (light and dark blue lines) appear to have converged, presumably to the CBS limit, varying by 0.005−0.015 m-E_H. This was mentioned in passing recently in the construction of an ab initio water PES where the three-body interactions computed using the subcluster basis at CCSD(T)/AVTZ are very similar to the CBS limit values.[20] With the removal of BSIE, we only require an AVDZ basis set to obtain CBS limit three-body and four-body interactions. This result implies that primarily the convergence of the total energies with increasing basis set quality comes from changes in the one-body and two-body interactions. Thus, we can obtain the total energies of water clusters with increasing basis set quality by recalculating the one-body and two-body interactions at the respective basis sets. The fact that three-and-higher-body interactions do not require a large basis set to achieve the CBS limit eliminates the need for extrapolation or ad hoc measures as commonly employed for two-body interactions. The ad hoc methods involve taking a fraction of the two-body interaction computed using the subcluster and nuclei-centered basis,[34−37] motivated by the well-documented trend[34,36−38] that these two quantities converge to the CBS limit from above and below, respectively (Figure 3d).

As with our previous work,[17] there is no guarantee that the observations made on the small $(H_2O)_6$ clusters still hold true when larger clusters are studied. To this end, we computed the BSEE up to the four-body interaction for a homologous series of optimized $(H_2O)_{8−16}$ clusters taken from Maheshwary et al.,[39] which is presented together with the hexamer results (Figure 4). Since various cluster sizes are involved, all the energies reported henceforth will be on a per monomer basis. Calculations were performed at MP2/AVDZ and MP2/AVTZ. The explicitly correlated MP2-F12 theory model[28] was also employed with the AVDZ basis set as this combination typically yielded results of MP2/AVQZ quality,[40] complementing the MP2/AVDZ and MP2/AVTZ results.

As mentioned earlier, $E^{(k)}_{ext.}$ serves as an error indicator of how well the cheaper subcluster basis can be used in place of the more expensive cluster basis. Here, we wish to clarify what we deem to be an acceptable value for $E^{(k)}_{ext.}$. Studies on atomization energies and reaction enthalpies often require calculations to

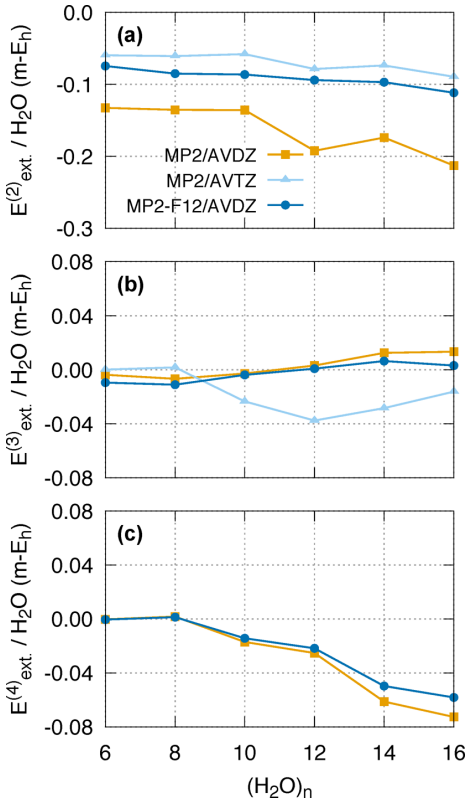


Figure 4. BSEE in the total k-body interaction, $E^{(k)}_{ext.}$, per (H_2O) monomer for water clusters of increasing size, $(H_2O)_{6−16}$, computed at MP2 levels of theory with various basis sets. The results for $E^{(4)}_{ext.}$ are not available at the MP2/AVTZ level due to the steep computational cost involved in computing the four-body interactions.

agree with experiments within chemical accuracy, which is 4.2 kJ mol^{-1} or 1.6 m-E_h.[41] However, the MBE is often used to study the dynamical evolution of large molecular clusters and do not involve bond breaking. As such, we introduce the "dynamical accuracy" where the error for large clusters is computed on a per monomer basis as the properties derived from dynamical simulations are intensive in nature. A suitable dynamical accuracy might be 10% of the average molecular kinetic energy at room temperature, $(3/2)kT$, which is about 0.14 m-E_h or 0.37 kJ mol^{-1}.

From Figure 4a, we again observe that the $E^{(2)}_{ext.}$ decreases with increasing basis set quality where the use of the higher quality MP2/AVTZ (light blue line) or the explicitly correlated MP2-F12/AVDZ (dark blue line) halved the small $E^{(2)}_{ext.}$ present in MP2/AVDZ (orange line). While the $E^{(2)}_{ext.}$ per monomer at MP2/AVTZ falls within dynamical accuracy, the BSEE exhibits a slow increase with increasing cluster size. Fortunately, due to the small system size, this small $E^{(2)}_{ext.}$ can be practically eliminated through the use of larger basis sets or CBS extrapolation. Indeed, CBS extrapolation is routinely applied to two-body interactions employed in ab initio two-body water potentials.[21,42,43] At the higher-body level, we confirmed that the $E^{(3)}_{ext.}$ and $E^{(4)}_{ext.}$ are, if not negligible, then acceptable. The $E^{(3)}_{ext.}$ is insignificant, always below 0.040 m-E_h per monomer (Figure 4b). The $E^{(4)}_{ext.}$ shows an increasing trend with increasing cluster

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132−5143

size (Figure 4c). Nonetheless, the value is quite small (<0.080 m-E_h per monomer) and would be even smaller if a larger basis set such as AVTZ is used. Furthermore, there would be some partial cancellation of the BSEE when the three-body and four-body interactions are summed. Therefore, we conclude that the cheaper subcluster basis can be employed in computing three-body and four-body interactions in place of their more expensive cluster basis counterpart.

Next, we overlaid the total k-body interaction computed using the subcluster basis at different basis set quality (Figure 5). The cluster basis counterpart shows identical trends and is

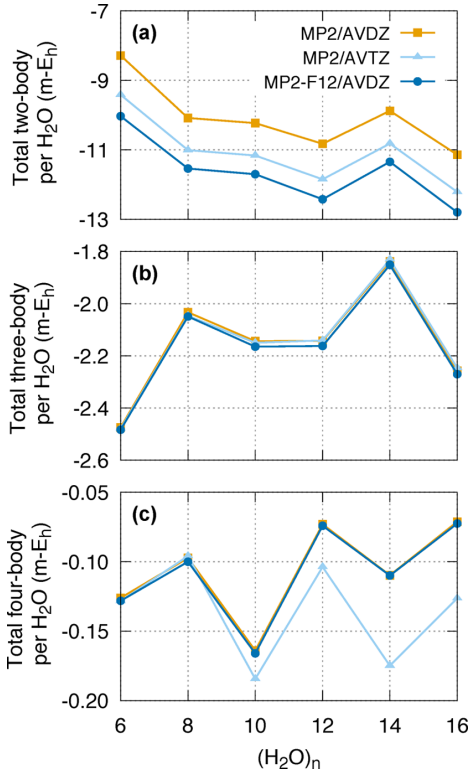


Figure 5. Comparison of the total k-body interaction computed using the subcluster basis per (H_2O) monomer for water clusters of increasing size, $(H_2O)_{6-16}$, computed at various levels of theory and basis sets.

presented in Figure S2 of the Supporting Information. It is clear that the total three-body and four-body interactions remain the same regardless of the basis set used (Figure 5b,c). The total four-body interaction at MP2/AVTZ (light blue line) appears to be different due to the scale of the energy axis, which exaggerates the small difference (<0.055 m-E_h) between the MP2/AVTZ and MP2/AVDZ values. The total two-body interaction (Figure 5a) becomes more stabilizing with increasing basis set quality, echoing the hexamer results.

In summary, we made three key observations: (i) the $E^{(2)}_{ext.}$ is small but significant and diminishes with increasing basis set quality, (ii) the $E^{(3)}_{ext.}$ and $E^{(4)}_{ext.}$ are much smaller, supporting the use of the cheaper subcluster basis to compute the three-body and four-body interactions, and (iii) the three-body and four-body interactions computed using the subcluster basis have converged to the CBS limit using an AVDZ basis set.

3.2. Many-Ghost Many-Body Expansion of the Total Energy. CBS extrapolation at the two-body level would eliminate the $E^{(2)}_{ext.}$, which can be coupled with AVDZ-quality three-body and four-body interactions to yield binding energies of CBS quality without using the expensive cluster basis. Apart from that, one may also be interested in reproducing the total energy at a particular basis set. This is useful in assessing the accuracy of fragmentation methods,[1−4] such as the Combined Fragmentation Method,[44−46] where small groups of adjacent atoms are treated as bodies and selected many-body interactions are computed to approximate the total energies of large chemical systems. We employed the MGMBE truncated at different order in an attempt to reproduce the total energy (Table 3). In this way, we can determine whether the omission of certain BSEEs affects the accuracy of the predicted total energy.

From Table 3, including the BSEE from a one-ghost-body into the one-body interaction decreased the error by 1 order of magnitude as seen in the entry {1, 0, 0, 0}. This suggests that the one-body interaction is very sensitive to the BSEE. This is not surprising as the one-body interaction constitutes the majority (≈99.98%) of the total energy. The error decreased again when more BSEE was incorporated (entry {2, 1, 0, 0}). However, from entry {2, 1, 0, 0} to {3, 2, 1, 0}, further inclusion of BSEE resulted in a larger error. Hypothesizing that this could be due to the BSEE in the one-body interaction, we varied the truncation order of the BSEE in the one-body interaction (entry {1, 2, 1, 0}, {2, 2, 1, 0}, and {3, 2, 1, 0}) and observed a fluctuation in the error. While the data are not shown here, the error actually oscillates wildly, changing in sign from positive (entry {1, 2, 1, 0}) to negative (entry {2, 2, 1, 0}) and back to positive again (entry {3, 2, 1, 0}). Recall the surprising result in section 2.3 that the MBE using the nuclei-centered basis truncated at the α-body term contains the BSEE from up to $m = (\alpha - k)$ ghost-bodies in each of the k-body interactions. This suggests that the similar oscillatory behavior reported previously[17] in the MBEs using the nuclei-centered basis could be due to the BSEE present in the one-body interaction. To determine if the two oscillatory behaviours are related, we compared the convergence of the MBE using the nuclei-centred basis to the total energy of the cluster with that of the MGMBE of the one-body interaction to the total one-body interaction in the cluster basis (Figure 6).

It is clear from Figure 6 that the two many-body decompositions are practically identical except for the first two data points. It appears to be the case that the poor convergence of the MBE using the nuclei-centered basis is almost completely caused by the BSEE in the one-body interaction. The differences in the first two data points is because the MBE (Figure 6a) includes the actual many-body interactions together with the BSEE. In the first two data points, there are additional errors in the MBE associated with neglecting these many-body interactions. From the four-body term onward, the majority of the many-body interactions are accounted for, and virtually all the remaining error is apparently due to BSEE in the one-body interaction.

The errors in the MBE associated with the BSEE not only applies to the brute force computation of all the $\binom{n}{k}$ individual k-body interactions but also to "internally consistent" selected many-body interactions virtually always employed in fragmen-

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132−5143

Journal of Chemical Theory and Computation Article

Table 3. Root Mean Square of the Error per (H_2O) Monomer (RMSE, m-E_h) and Maximum Absolute Error per (H_2O) Monomer (MxAE, m-E_h), in Reproducing the Total Energy for a Series of Optimized Water Clusters from Figure 4 Calculated at MP2/AVDZ, MP2/AVTZ and MP2-F12/AVDZ[a]

m-ghost-body in *k*-body[b]				MP2/AVDZ		MP2/AVTZ		MP2-F12/AVDZ	
k = 1	k = 2	k = 3	k = 4	RMSE	MxAE	RMSE	MxAE	RMSE	MxAE
0	0	0	0	2.505	2.779	1.286	1.402	0.811	0.868
1	0	0	0	0.122	0.162	0.065	0.110	0.431	0.518
2	1	0	0	0.101	0.150	0.032	0.060	0.267	0.343
1	2	1	0	0.280	0.324	0.091	0.105	0.510	0.595
2	2	1	0	0.145	0.197	0.066	0.095	0.338	0.461
3	2	1	0	0.320	0.463	0.088	0.141	0.298	0.462
all	all	all	all	0.044	0.075	—[c]	—[c]	0.037	0.062

[a]The MGMBE includes up to the four-body interaction (k = 1−4) of which the BSEEs are truncated at different *m*-ghost bodies. Error here is defined as the total energy of the cluster minus the MGMBE-predicted total energy. The error per (H_2O) monomer is first obtained before the RMS or maximum is taken. [b]The digits give the highest number of ghost bodies, *m*, that are incorporated into the *k*-body interaction using the MGMBE, and "all" refers to the cluster basis which includes all the BSEEs. For example, the second entry, {1, 0, 0, 0}, indicates that the BSEE from up to one ghost body is incorporated in the one-body interactions, and there are no BSEEs included for the two-to-four-body interactions. [c]As the four-body interaction computed using the cluster basis is computationally expensive at MP2/AVTZ, an estimate of the total energy is unavailable.

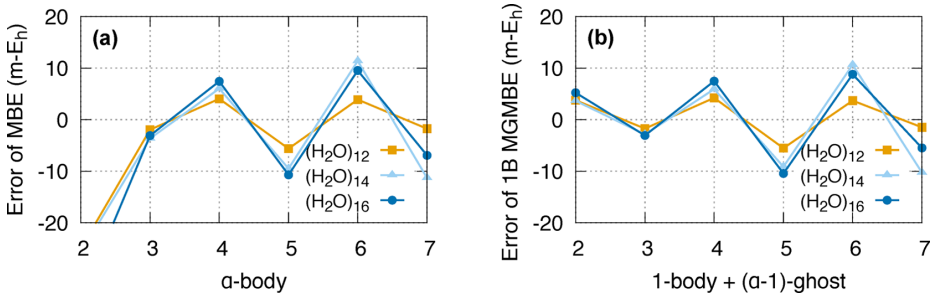


Figure 6. (a) The error of the MBE of the total energy using the nuclei-centered basis truncated at the α-body term follows an almost identical trend as (b) the error of the MGMBE of the total one-body interaction in the cluster basis truncated at the $(\alpha - 1)$-ghost-body term. The calculations were performed at MP2/AVDZ for $(H_2O)_{12}$, $(H_2O)_{14}$, and $(H_2O)_{16}$ from Figure 4. The error of the MBE is defined as the difference between the total energy, E_{tot}, and the sum of the total one-body interaction up to the total α-body interaction. Similarly, the error of the MGMBE is defined as the difference between the one-body interaction computed using the cluster basis, $E_{A\,\overline{B\cdots N}}$, and the sum of the BSEE from up to $(\alpha - 1)$ ghost bodies summed across all the monomers.

Table 4. Root Mean Square of the Error per (H_2O) Monomer (RMSE, m-E_h) and Maximum Absolute Error per (H_2O) Monomer (MxAE, m-E_h), in Reproducing the ε^C_{tot} for the Same Water Clusters in Table 3[a]

m-ghost-body in *k*-body[b]			MP2/AVDZ		MP2/AVTZ		MP2-F12/AVDZ	
k = 2	k = 3	k = 4	RMSE	MxAE	RMSE	MxAE	RMSE	MxAE
0	0	0	0.160	0.197	0.046	0.058	0.090	0.104
1	0	0	0.046	0.070	0.073	0.114	0.060	0.090
2	1	0	0.009	0.015	0.019	0.043	0.015	0.028
all	all	all	0.044	0.075	—[c]	—[c]	0.037	0.062

[a]The MGMBE includes up to the four-body interaction (k = 2−4) of which the BSEEs are truncated at different *m* ghost bodies. Error here is defined as the ε^C_{tot} of the cluster minus the MGMBE-predicted ε^C_{tot}. The error per (H_2O) monomer is first obtained before the RMS or maximum is taken. [b]The digits give the highest number of ghost bodies, *m*, that are incorporated into the *k*-body interaction using the MGMBE, and "all" refers to the cluster basis which includes all the BSEE. For example, the second entry, {1,0,0}, indicates that the BSEE from up to one ghost body is incorporated in the two-body interactions, and there are no BSEEs included for the three-body and four-body interactions. [c]As the four-body interaction computed using the cluster basis is computationally expensive at MP2/AVTZ, an estimate of the ε^C_{tot} is unavailable.

tation methods. These interactions are "internally consistent" in a sense that the many-body interactions of the selected fragments (interacting groups of atoms) and their constituent lower-body interactions are included and only included once. This allows for the BSIE and BSEE to cancel. The poor convergence of the MBE/MGMBE allows us to explain certain observations in fragmentation methods. "Grafting" is employed in some fragmentation methods[47−49] where the total energy of the system is calculated at a lower level of theory or basis set to serve as a correction to the predicted total energy. Such grafting approaches not only correct for missing important many-body interactions but also account for the BSEE to a large extent, explaining the low errors associated with these methods. Since the BSEE converges poorly with respect to the number of ghost bodies, the expensive one-body interactions computed using the cluster basis are required to accurately reproduce the total energy. Future investigations to develop cheaper alternatives to the cluster basis will be undertaken. One possibility includes the

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132−5143

omission of certain basis functions from the basis set, in particular the tight valence-type functions (i.e., not diffuse functions), on ghost bodies that are far away from the monomer of interest. Thus, only the contributing diffuse functions remain.

3.3. Many-Ghost Many-Body Expansion of the Binding Energy. We have shown that the poor convergence of the MBE using the nuclei-centered basis is caused by the BSEE in the one-body interactions. With the removal of the one-body interactions and its associated BSEE, we would expect the remaining energy to converge rapidly with the number of bodies. This remaining energy is the binding energy, ε^{C}_{tot}, and the accuracy of the MGMBE is evaluated in Table 4.

From Table 4, the incorporation of the BSEE greatly reduces the error in reproducing the ε^{C}_{tot}, eventually giving a tiny error per monomer of below 0.015−0.043 m-E_h (entry {2, 1, 0}), which is well within dynamical accuracy. In fact, entry {2, 1, 0} gives a lower error than entry {all, all, all} which incorporates all the BSEE in the two-to-four-body interactions. This can be attributed to a reversal in the sign of the error. In entry {0, 0, 0}, the absence of BSEE which stabilizes the binding energy results in negative errors. On the other hand, the errors from the incorporation of all the BSEE up to the four-body interaction in entry {all, all, all} are positive due to the neglect of higher-than-four-body interactions. Thus, there is some form of error cancellation between the two factors when the majority of the BSEE is accounted for in entry {2, 1, 0}. Furthermore, the maximum number of basis functions ever employed in any total energy calculation is limited to that of four monomers in entry {2, 1, 0}, originating from either the interacting bodies or ghost bodies. This allows for expensive theoretical models such as the Coupled Cluster Singles and Doubles with perturbative Triples [CCSD(T)] to be applied to obtain highly accurate ε^{C}_{tot} for large clusters or even bulk-water simulations. It should be emphasized that no charge embedding scheme[9,50,51] was used, although they are commonly applied to water clusters. The use of such schemes is prevalent in the literature due to the belief that the water−water interactions are highly many-body in nature. However, our results indicate that we only require up to the four-body interactions. It is likely that any apparent higher-than-four-body effects are caused by the BSEE in the one-body interactions which we have shown to be highly many-body in nature (Figure 6).

Notably, the calculations involved in entry {2, 1, 0} are equivalent to that in a MBCP(4) calculation.[13,14] A MBCP(4) calculation would involve a MBE using the nuclei-centered basis truncated at the four-body term minus the one-body interactions with the BSEE truncated at the $(4 - 1)$ = 3-ghost-body level. This is equivalent to a "{3, 2, 1, 0}" MGMBE calculation of the total energy minus the one-body interactions and its associated BSEE, i.e., entry {2, 1, 0} in Table 4. Thus, an "$\{\alpha - 2, \alpha - 1, ..., 0\}$" MGMBE calculation of the ε^{C}_{tot} is identical to an MBCP(α) calculation. Note that for the MBCP method, the MG1BE has to be truncated at one order less than that of the MBE. This is important to ensure that all the BSEE in the one-body interactions are properly removed, and this requirement only becomes obvious with the analysis of the BSEE using the MGMBE presented in this work.

4. CONCLUSION

Through a systematic study of water clusters with improving basis set and increasing cluster size, we concluded that one has to account for many-body BSSE in order to reproduce the many-body interactions computed using the cluster basis. There are two distinct components to the many-body BSSE. The first arises due to an imbalance in the number of basis functions used to compute a particular k-body interaction. In this case, the k-body total energy calculation utilizes many more basis functions than does the lower-body counterparts which are necessary to extract the k-body interaction. The second arises due to the fact that a k-body within a much larger cluster is further stabilized by the basis functions of the surrounding bodies denoted as the BSEE. If one wants to reproduce the binding energy and/or the total energy through a many-body approach, the first BSIE is undesirable as it leads to erroneous many-body interactions. However, the BSEE 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. Thus, the best estimate of the binding energy at a given basis set would be the total energy minus the one-body intramolecular interactions computed using the cluster basis.

We found that both components of the many-body BSSE are accounted for in the three-body and four-body interactions computed using the subcluster basis and that these interactions appear to have converged to the CBS limit at the AVDZ level. For the two-body interactions, and particularly for the one-body intramolecular interactions, important BSEEs are significant and have to be accounted for, thus making the use of the subcluster basis insufficient. To account for both the BSIE and the BSEE, we introduce the MGMBE in this work. The MGMBE performs a two-dimensional many-body decomposition with each decomposition accounting for one component of many-body BSSE. Through the MGMBE of the total energy, we found that the oscillatory behavior encountered in MBEs using diffuse functions is caused by the BSEE in the one-body interactions. With the adequate removal of the one-body interactions and the associated BSEE, the MGMBE successfully reproduces the binding energies of clusters using numerous small calculations that involve no more than four monomers.

Despite the utility of decomposing a large cluster into small subsystems, the MBE and the MGMBE come with a limitation. The number of four-body calculations increases quartically with the cluster size, substantially hindering the scalability of these methods. To circumvent this, a forthcoming publication will establish a rigorous criterion to select out all potentially significant many-body interactions.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00343.

The derivation of the working equations for the MBE (section 2.1) and the MGMBE (section 2.3); Figures S1 and S2, showing the results for the prism isomer counterpart to Figure 3 and the cluster basis counterpart to Figure 5 respectively; and Cartesian coordinates for the water clusters studied in this work (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +65 6516 2846. Fax: +65 6779 1691. E-mail: ryan.pa.bettens@nus.edu.sg.

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132−5143

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the National University of Singapore's support from the Academic Research Fund, grant number R-143-000-549-112. The authors also thank the Centre for Computational Science and Engineering for the use of their computers.

■ REFERENCES

(1) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. *Chem. Rev.* **2012**, *112*, 632−672.
(2) Richard, R. M.; Herbert, J. M. *J. Chem. Phys.* **2012**, *137*, 064113.
(3) Collins, M. A.; Bettens, R. P. A. *Chem. Rev.* **2015**, *115*, 5607−5642.
(4) Raghavachari, K.; Saha, A. *Chem. Rev.* **2015**, *115*, 5643−5677.
(5) Stone, A. J. *The Theory of Intermolecular Forces*, 2nd ed.; Oxford University Press: Oxford, England, 2013.
(6) Kaplan, I. G.; Santamaria, R.; Novaro, O. *Mol. Phys.* **1995**, *84*, 105−114.
(7) Xantheas, S. S. *J. Chem. Phys.* **1994**, *100*, 7523−7534.
(8) Pedulla, J.; Kim, K.; Jordan, K. *Chem. Phys. Lett.* **1998**, *291*, 78−84.
(9) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 46−53.
(10) Liu, B.; McLean, A. D. *J. Chem. Phys.* **1973**, *59*, 4557−4558.
(11) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553−566.
(12) Wells, B. H.; Wilson, S. *Chem. Phys. Lett.* **1983**, *101*, 429−434.
(13) Richard, R. M.; Lao, K. U.; Herbert, J. M. *J. Phys. Chem. Lett.* **2013**, *4*, 2674−2680.
(14) Richard, R. M.; Lao, K. U.; Herbert, J. M. *J. Chem. Phys.* **2013**, *139*, 224102.
(15) Turi, L.; Dannenberg, J. J. *J. Phys. Chem.* **1993**, *97*, 2488−2490.
(16) Valiron, P.; Mayer, I. *Chem. Phys. Lett.* **1997**, *275*, 46−55.
(17) Ouyang, J. F.; Cvitkovic, M. W.; Bettens, R. P. A. *J. Chem. Theory Comput.* **2014**, *10*, 3699−3707.
(18) Wang, Y.; Huang, X.; Shepler, B. C.; Braams, B. J.; Bowman, J. M. *J. Chem. Phys.* **2011**, *134*, 094509.
(19) Medders, G. R.; Babin, V.; Paesani, F. *J. Chem. Theory Comput.* **2013**, *9*, 1103−1114.
(20) Babin, V.; Medders, G. R.; Paesani, F. *J. Chem. Theory Comput.* **2014**, *10*, 1599−1607.
(21) Góra, U.; Cencek, W.; Podeszwa, R.; van der Avoird, A.; Szalewicz, K. *J. Chem. Phys.* **2014**, *140*, 194101.
(22) Hermann, A.; Krawczyk, R. P.; Lein, M.; Schwerdtfeger, P. *Phys. Rev. A: At., Mol., Opt. Phys.* **2007**, *76*, 013202.
(23) Góra, U.; Podeszwa, R.; Cencek, W.; Szalewicz, K. *J. Chem. Phys.* **2011**, *135*, 224102.
(24) Szalewicz, K.; Leforestier, C.; van der Avoird, A. *Chem. Phys. Lett.* **2009**, *482*, 1−14.
(25) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Lindh, R.; Mitrushenkov, A.; Rauhut, G.; Shamasundar, K. R.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hesselmann, A.; Hetzer, G.; Hrenar, T.; Jansen, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; O'Neill, D. P.; Palmieri, P.; Peng, D.; Pflüger, K.; Pitzer, R.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M. *MOLPRO*, version 2012.1; Cardiff University: Cardiff, U. K.; Universität Stuttgart: Stuttgart, Germany, 2012. See http://www.molpro.net.
(26) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(27) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796−6806.
(28) Werner, H.-J.; Adler, T. B.; Manby, F. R. *J. Chem. Phys.* **2007**, *126*, 164102.
(29) Lendvay, G.; Mayer, I. *Chem. Phys. Lett.* **1998**, *297*, 365−373.
(30) Salvador, P.; Szczęśniak, M. M. *J. Chem. Phys.* **2003**, *118*, 537−549.
(31) Kamiya, M.; Hirata, S.; Valiev, M. *J. Chem. Phys.* **2008**, *128*, 074103.
(32) Williams, H. L.; Mas, E. M.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **1995**, *103*, 7374−7391.
(33) Ouyang, J. F.; Bettens, R. P. A. *Chimia* **2015**, *69*, 104−111.
(34) Halkier, A.; Klopper, W.; Helgaker, T.; Jørgensen, P.; Taylor, P. R. *J. Chem. Phys.* **1999**, *111*, 9157−9167.
(35) Shields, R. M.; Temelso, B.; Archer, K. A.; Morrell, T. E.; Shields, G. C. *J. Phys. Chem. A* **2010**, *114*, 11725−11737.
(36) Burns, L. A.; Marshall, M. S.; Sherrill, C. D. *J. Chem. Theory Comput.* **2014**, *10*, 49−57.
(37) Brauer, B.; Kesharwani, M. K.; Martin, J. M. L. *J. Chem. Theory Comput.* **2014**, *10*, 3791−3799.
(38) Halkier, A.; Koch, H.; Jørgensen, P.; Christiansen, O.; Nielsen, I. M. B.; Helgaker, T. *Theor. Chem. Acc.* **1997**, *97*, 150−157.
(39) Maheshwary, S.; Patel, N.; Sathyamurthy, N.; Kulkarni, A. D.; Gadre, S. R. *J. Phys. Chem. A* **2001**, *105*, 10525−10537.
(40) Werner, H.-J.; Adler, T. B.; Knizia, G.; Manby, F. R. In *Recent Progress in Coupled Cluster Methods*; Cársky, P., Paldus, J., Pittner, J., Eds.; Springer: Dordrecht, Holland, 2010; pp 573−619.
(41) Helgaker, T.; Ruden, T. A.; Jórgensen, P.; Olsen, J.; Klopper, W. *J. Phys. Org. Chem.* **2004**, *17*, 913−933.
(42) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. *J. Chem. Phys.* **2008**, *128*, 094313.
(43) Babin, V.; Leforestier, C.; Paesani, F. *J. Chem. Theory Comput.* **2013**, *9*, 5395−5403.
(44) Le, H.-A.; Tan, H.-J.; Ouyang, J. F.; Bettens, R. P. A. *J. Chem. Theory Comput.* **2012**, *8*, 469−478.
(45) Tan, H.-J.; Bettens, R. P. A. *Phys. Chem. Chem. Phys.* **2013**, *15*, 7541−7547.
(46) Collins, M. A.; Cvitkovic, M. W.; Bettens, R. P. A. *Acc. Chem. Res.* **2014**, *47*, 2776−2785.
(47) Furtado, J. P.; Rahalkar, A. P.; Shanker, S.; Bandyopadhyay, P.; Gadre, S. R. *J. Phys. Chem. Lett.* **2012**, *3*, 2253−2258.
(48) Sahu, N.; Yeole, S. D.; Gadre, S. R. *J. Chem. Phys.* **2013**, *138*, 104101.
(49) Saha, A.; Raghavachari, K. *J. Chem. Theory Comput.* **2014**, *10*, 58−67.
(50) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.
(51) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 1342−1348.

DOI: 10.1021/acs.jctc.5b00343
J. Chem. Theory Comput. 2015, 11, 5132−5143