



UNIWERSYTET ŚLĄSKI
W KATOWICACH

Institute of Chemistry
Faculty of Science and Technology
University of Silesia

mgr Urszula Góra
Ph.D Thesis

**INTERMOLECULAR INTERACTIONS IN WATER
CLUSTERS**

Badanie oddziaływań międzymolekularnych w klastrach wody

Doctoral supervisor:
dr hab Rafał Podeszwa, prof. UŚ

Katowice, 2024

ACKNOWLEDGEMENTS

I am very grateful to the coauthors of the presented papers, without their support none of this work would be possible.

For constant encouragement to finalize the thesis I would like to thank Ania, Mirek, Rafał and my mom.

To my best friends: Mira, Henryk and Hektor.

TABLE OF CONTENTS

ABSTRACT	vii
Chapter	
1 INTRODUCTION	1
1.1 Outline of the dissertation	1
1.2 Overview	1
1.3 SAMBA approach	12
1.4 CCpol23+ potential	15
1.4.1 CCpol23+ Errata	19
1.5 Current status of the water potentials	20
2 INTERACTION ENERGIES OF LARGE CLUSTERS FROM MANY-BODY EXPANSION	30
2.1 Introduction	31
2.2 Details of calculations	36
2.3 Complete basis set limits for water hexamer	40
2.3.1 CCSD(T) method only	41
2.3.2 Hybrid MP2 plus CCSD(T) approach	41
2.3.3 Non-truncated many-body expansion approach	43
2.4 Comparison with literature	46
2.5 Convergence of many-body expansion for water hexamers	50
2.6 Computer timings	56
2.7 Applications of effective many-body strategies	58
2.8 Large water clusters	61
2.9 Conclusions	65
2.10 Acknowledgments	68

3	PREDICTIONS FOR WATER CLUSTERS FROM A FIRST-PRINCIPLES TWO- AND THREE-BODY FORCE FIELD	93
3.1	Introduction	93
3.2	Choice of trimer configurations	100
3.3	<i>Ab initio</i> calculations	103
3.4	Two-body fit	107
3.5	Many-body induction energy model	110
3.6	Nonadditive three-body fit	114
	3.6.1 Functional form of fit	114
	3.6.2 Fitting of three-body potential	116
3.7	Application to clusters	118
	3.7.1 Water trimer	118
	3.7.2 Water hexamer	126
	3.7.3 24-mer	129
3.8	Summary and Conclusions	133
3.9	Acknowledgments	136
4	APPENDIX	156

ABSTRACT

This dissertation addresses the challenges of modeling intermolecular interactions in water. The primary motivation of the presented work was to find a protocol that would be accurate enough to predict the most stable isomer among water hexamers.

Traditionally, the calculation of energies for molecular clusters relies on the supermolecular method, which involves calculating energies for the entire cluster. As the system size expands, the computational cost of this approach increases rapidly. In this work, we demonstrate that the many-body expansion can be employed as an efficient alternative for calculating interaction energies. We take advantage of the rapid convergence of the many-body expansion to reproduce the results of canonical calculations with greater accuracy and reduced computational effort. By breaking down large clusters into smaller subclusters and using tailored computational methods based on the fragments' importance, we developed a protocol called the "stratified approximation" many-body approach (SAMBA). Using this approach, we achieved highly accurate benchmarks for the isomers of water hexamers and larger clusters, including 16-mers and 24-mers. Although we applied this approach only to water clusters, it can also be extended to other systems.

To model properties beyond single-point energies, an accurate intermolecular potential is required. Numerous potential energy surfaces (PESs) or force fields have been developed for water, but, as it turned out, their characterization of non-additive interaction energies, including three-body interactions, proved insufficient for accurately modeling larger water clusters. To address this issue, we developed a new three-body water potential that also incorporated improvements in the two-body part and in contributions from higher-order interactions. We conducted extensive calculations

for water trimer, with more than 70 thousand trimer interaction energies computed using state-of-the-art level of theory. The resulting potential was sufficiently accurate to model the stability of water hexamers, achieving the main objective of the dissertation, and provided benchmark-quality results for larger systems.

Modeling of water is a very competitive field of research. Other groups have also proposed alternative protocols and potentials to solve the same problem. Our approaches remained competitive and gained a significant number of citations.

Chapter 1

INTRODUCTION

1.1 Outline of the dissertation

The dissertation is structured as follows: The first chapter offers a general description of the problem, covering the background, key concepts, and a brief literature overview. This is followed by a summary of the publications that form the core of the dissertation. A concise overview of the developments that occurred after the main papers were published is also included. Chapter 2 presents the text of the publication entitled “Interaction energies of large clusters from many-body expansion” [1] with the statement of authors’ contributions, while Chapter 3 contains the text of the publication entitled “Predictions for water clusters from a first-principles two- and three-body force field” [2] with the statements of authors’ contributions. Finally, the Appendix includes a scientific Curriculum Vitae.

1.2 Overview

Water has been central to scientific research for centuries due to its importance and peculiar properties, such as its high boiling point and anomalous temperature-density relationship. These unusual characteristics, which differ significantly from other well-known substances, played a crucial role in the emergence and evolution of life on Earth. Theoretical chemists, in addition to experimentalists, have extensively studied water. Due to its relatively small size, the water molecule allows for high-level theoretical calculations, making it a benchmark system for various types of *ab initio* calculations [3–9]. Despite modest size, the theoretical modeling of water is still a challenging task requiring extremely high accuracy for correct description.

This dissertation focuses on the theoretical description of the physical chemistry of water. In this case, the hydrogen-oxygen bonds are not broken, and most of the physical phenomena occur due to intermolecular interactions between water molecules. These interactions are much weaker than the chemical bonds but are still responsible for the aggregation of water molecules in clusters, liquid and solid phases. These interactions, also known as van der Waals interactions or forces, arise from Coulomb attractions and repulsions between the electrons and nuclei of the interacting water molecules and can be subdivided into several categories (electrostatic, exchange, induction, and dispersion). Processes that occur within the individual molecules (intramolecular effects), like changes in bond lengths and angles or quantum effects from zero-point energies, are less significant in water aggregates but need to be considered for high accuracy.

To connect theoretical data with experiments and predict all observable physical properties of water, one needs the complete potential energy surface (PES) of interacting water molecules. PESs, also referred to as force fields, map the interaction energies to the geometries of the interacting molecules. Such maps can then be used in modeling the motions of the nuclei, either in classical molecular dynamics (MD) or classical Monte-Carlo (MC), or in more demanding quantum-chemical simulations of the nuclear motion. The motion of the nuclei, determined by the intermolecular potential energy, is responsible for the liquid water phase diagram, its density, diffusion coefficients and other physical chemistry phenomena, as well as the spectra of water clusters. An ideal water potential should be capable of modeling all these phenomena, both in the condensed phase and in clusters. It turns out that such a goal is extremely difficult, and despite decades of effort, water potentials that can reasonably predict all phenomena have been developed fairly recently [10] and are still not perfect. To understand why this is the case, it would be instructive to explain how the PESs are created.

There are two general paths one can take: using experimental data or theoretical calculations (or a combination of these two). The first approach results in a large

class of empirical potentials, where the force field parameters are fitted to ensure that MD simulations using these potentials reproduce some experimental properties, such as density of water at various temperatures. These potentials have been developed a long time ago, are widely used and include, for example, TIP4P [11] and SPCE [12] potentials. They are very simple, containing a small number of parameters and using basic site-site functions to describe the intermolecular potential such as Coulomb charge-charge interaction and the Lennard-Jones function:

$$\frac{A}{r^{12}} - \frac{B}{r^6}, \quad (1.1)$$

where A and B are parameters describing the repulsive and attractive part of the interaction, respectively, and r is the distance between sites (usually nuclei but off-atomic sites are also used). The simplicity of these potentials stems from two factors: the computational cost of the MD simulations depends on the complexity of the potentials, so simpler potentials are faster, and from the difficulty in fitting a large number of parameters into experimental properties. Although such potentials can predict the properties that were fitted fairly accurately, they lack universality and do not work well for small water clusters [13]. This problem fundamentally stems from their site-site character and distance-based dependence. Such site-site functions can physically describe only two-body effects. Since the bulk properties of water depend on more than two-body effects and result from cooperative three- and higher-body interactions, hence the site-site function are incapable to incorporate such effects.

This simple empirical-only approach can be augmented by incorporating *ab initio* data in the fitting procedure [14, 15] and by including an explicit many-body contribution via a polarization term. This is achieved by providing a polarizable site that creates an induced dipole moment from interactions with other monomers. This method mimics the physical induction effect, which is indeed the major component in the many-body expansion of water clusters. Therefore, such empirical potentials work better than the simple ones for small water clusters. However, a polarization term only

partly accounts for pairwise nonadditive interactions [16] in liquid water and larger water clusters and such potentials are still incapable of a universal representation of all water properties. Another type of empirical potential has been derived from water dimers [17–20], with parameters fitted to reproduce the spectroscopic data. While these potentials describe water dimers accurately, they overlook the nonadditive part of the interaction that is absolutely vital for liquid water or ice and is important for water clusters [21].

An alternative approach to potentials fitted to experimental data is to focus solely on the *ab initio* calculations of interaction energies. This approach, when combined with a physically-relevant functional form and accurate *ab initio* interaction energies can, in principle, yield a potential that can model both small clusters and large aggregates. This method was first developed by Clementi and coworkers [3, 22, 23]. However, these early attempts were very limited by computational resources. The number of grid points and the level of theory were insufficient to obtain meaningful results. The approach was subsequently expanded, and the *ab initio* water potentials were gradually improved [24–31]. Using symmetry-adapted perturbation theory (SAPT) [32] and utilizing 2,510 points for the two-body part of the potential [28], and later 7,533 grid points for the three-body part [31], the resulting SAPT-5s two-, three-body, and polarization-based many-body potential yielded fairly accurate dimer energies and various properties of liquid water. This potential was further refined in the CCPOL family of potentials, which successfully predicted water properties from the dimer to the liquid phase [10, 33, 34] for the first time. *Ab initio* potentials can sometimes predict properties before they are accurately measured. A notable example is the dissociation energy of the water dimer that was first predicted in 2000 [27], later refined [34, 35] and eventually confirmed by experimental measurements [36]. Despite these successes in *ab initio* water potentials, there is still room and a need for improvement. Certain water properties, such as anomalous density-temperature dependence or high dielectric constant, remain difficult to accurately model with the current potentials. During the work on publication provided in Chapter 2, it also became evident

that the existing potentials, (including the CCPOL family), were not accurate enough to predict subtle differences between isomers of the water hexamer clusters. There are two aspects where the CCPOL potentials fall short. One is monomer flexibility, as the original potential assumed fixed monomer geometries. Another limitation is the three-body part of the potentials, which was state-of-the art in 2003 [31], but the grid density was insufficient, and the level of theory was inadequate for modern applications. A two-body water potential with rigid monomers is six-dimensional, and after considering the intramonomer degrees of freedom for triatomic monomers the potential becomes 12-dimensional—challenging but feasible. The complexity increases rapidly with three-body flexible-monomer potentials, which are 21-dimensional. To provide a grid for fitting with just 3 points per dimension would require 10 billion data points, an impossible task without some serious approximations. There is a trade off: whether to peruse flexible-monomer potential that would be significantly less accurate due to the limits in the number of grid points or the level of theory, or a more accurate but limited in scope rigid-monomer potential. There are many problems where the latter potentials work well. Also, the lack of flexibility can be mitigated by carefully choosing the rigid-monomer geometry. It turns out that the geometry averaged over the lowest rovibrational state provides a more accurate representation of the potential than the equilibrium geometry of the monomers [6, 37]. Therefore, we decided to work on a rigid three-body potential and create the new state of the art three-center polarization model.

The polarization model used in the publication of Chapter 3 (similar models are also used in the literature) is detailed in Section 3.5. To further explain the concept, a simplified physical illustration is provided in Figure 1.1. This model also explains the nonadditivity of induction interactions. As shown in the figure, the negative partial charge of the oxygen atom of the molecule A induces a dipole moment on water molecule C. Although the figure illustrates the actual polarization of the electron cloud, the model simplifies this process by treating the induced dipole as a point dipole or several dipoles (the model of Section 3.5 uses three polarization centers). This dipole then

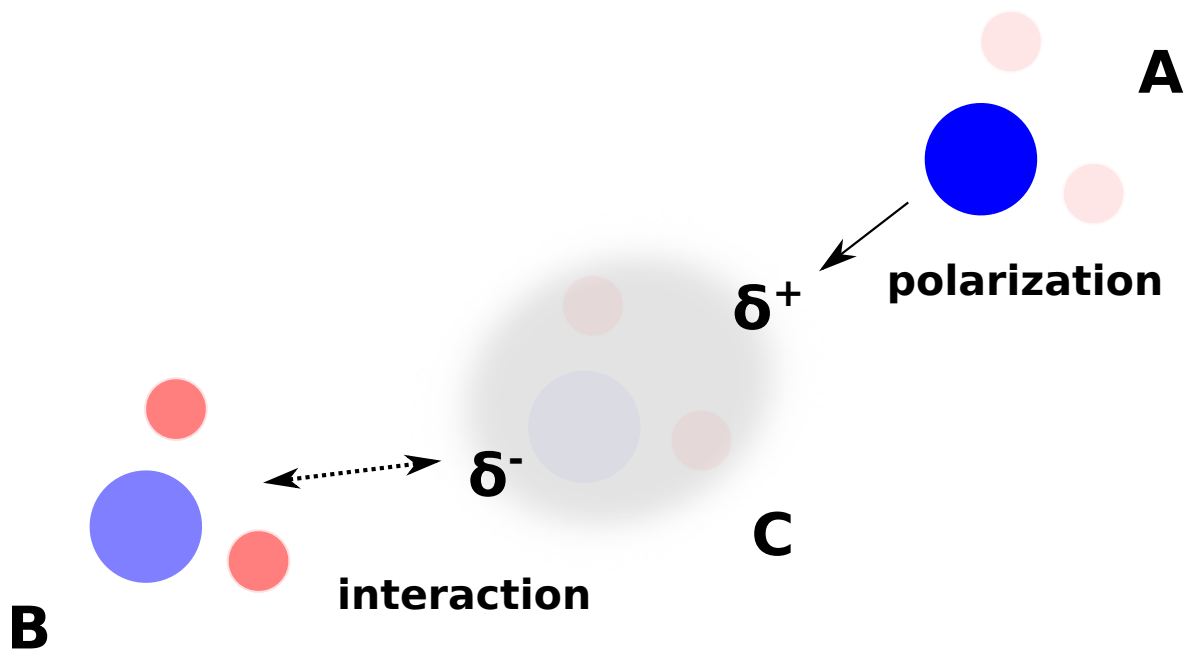


Figure 1.1: Polarization model. Charges from molecules A (one of the charges highlighted) polarizes molecule C. The induced dipole interacts with molecule B.

interacts with the partial charges of the molecule B. Molecule A and B also engage in direct interactions (not depicted), which are pure two-body in nature. Molecule C similarly interacts with both A and B through pairwise interactions (not depicted). The interaction shown in the figure occurs only in the presence of all molecules A, B, and C, and can be interpreted as an additional interaction between A and B due to the presence of molecule C. This extra interaction is included in the non-additive part and will be defined formally in Eq. (1.6). The polarization process is repeated for all the partial charges across all the molecules. The model can also be extended to include a larger number of molecules, representing the three-body contribution in larger aggregates. Alternatively, it can involve the polarization of multiple molecules and the simultaneous participation of more than three molecules. In this scenario, the model incorporates four-body or higher terms. As interactions involving more molecules become increasingly weaker, the model disregards those below a certain

threshold, thus keeping the computational cost of the procedure down. One significant advantage of the model is that only polarizability, number, and location of the centers need to be determined, making the model easily adaptable to any order of the many-body expansion. Nevertheless, the iterative process of polarizing molecules makes this model significantly more computationally expensive than simple pair interactions, and polarizable force fields are considerably slower than the purely two-body non-polarizable potentials. However, it remains less complicated than explicit many-body potentials.

The orientation of the water molecules in Figure 1.1 results in an attractive three-body contribution, making the interaction cooperative, i.e., the attraction between A and B is stronger in the presence of C. In water clusters with strongly attractive arrangements and in liquid water, this is the most frequent situation, making the three-body induction energy usually negative. However, unlike the two-body case, non-additive energies can vary in sign, and a proper three-body potential must account for all arrangements (cf. Figure 3.2).

Despite its success in capturing the major aspects of non-additive interactions in water and other polar systems, the polarization model has several limitations. Firstly, the physical process represented by this model is accurate only at large intermonomer separations, as the multipole expansion diverges at short distances [32]. This issue is mitigated by introducing damping [cf., Eq. (3.17)], which effectively suppresses the interaction at short distances, avoiding divergence but, simultaneously, preventing accurate representation of short-range effects. As a result, other parts of the potential must be adjusted to compensate for this missing part, reducing the overall generality of the potential. Secondly, these models are usually low order, involving isotropic centers and dipole-dipole polarizability, whereas actual molecules are anisotropic and involve higher-order polarizabilities. This issue can be address by including these higher-order terms, although this comes at the cost of increased computational time and stronger divergence at short ranges. Lastly, the polarization model only accounts for induction effects, completely omitting the exchange or dispersion components. These effects

are less important in water than induction non-additivity, they are nonetheless non-negligible [38]. Physically-correct potentials should incorporate explicit three-body effects to capture all possible interactions. This requirement motivated the development of three-body potential in Ref. 31 and its significant enhancement in publication of Chapter 3. Despite this, the polarization model remains useful. A potential function that captures the difference between the three-body nonadditivity and that of the polarization model is more accurate than one that attempts to model the total three-body non-additive energy. Additionally, developing explicit potentials beyond three-body interactions is extremely expensive both in development and usage, making the polarization model valuable for modeling four-body and higher-order contributions. This strategy was employed in the publication of Chapter 3.

Related only tangentially to the developments in the dissertation, it is also worth briefly discussing the physical picture of the interactions in the water dimer (and larger water clusters) based on the SAPT decomposition of the interaction energy into physically meaningful components. In the dissertation, the methods used involved a supermolecular approach, where the interaction energy is calculated as the difference between the energy of the cluster and that of the monomers:

$$E_{\text{int}} = E_{\text{cluster}} - \sum_i E_i, \quad (1.2)$$

where E_i is the energy of the i -th monomer. The simplest case of this approach is the interaction energy of a dimer:

$$E_{\text{int}} = E_{AB} - E_A - E_B, \quad (1.3)$$

where E_{AB} is the energy of the dimer and E_A and E_B are the energies of monomers A and B, respectively. In the SAPT approach, the interaction energy is calculated directly as a sum of physically-interpretable contributions [32]: electrostatic (resulting from interactions of permanent multipole moments of the interacting monomers),

exchange (resulting from the Pauli exclusion principle), induction (resulting from interaction between the permanent multipole moments on one monomers with the induced moments of the other), dispersion (resulting from interaction between multipole moments generated from instantaneous electron fluctuations), exchange-induction and exchange-dispersion (corrections to induction and dispersion due to antisymmetry). For the dimer, induction and dispersion contributions are always attractive, while exchange (including exchange-induction and exchange dispersion) is always repulsive, and electrostatic energy may be either and is usually the dominant factor in deciding whether a particular orientation is favorable or not. Dispersion energy originates from electronic correlation and in the supermolecular approach it requires a post-Hartree Fock treatment [39]. While the leading terms of electrostatic, exchange and induction energies are captured by the supermolecular Hartree-Fock calculations, accurate descriptions necessitate post-Hartree Fock correlated methods. Dispersion energy in water dimer is secondary to electrostatic and induction energies, but accurate treatment of the interaction energy requires a correlated method, such as those used in the publications provided in Chapters 2 and 3.

The supermolecular recipe given by Eq. (1.2) is very simple and straightforward in calculating the interaction energy of a cluster, but an alternative approach using the many-body expansion offers several benefits. As demonstrated in the publication of Chapter 2, this approach is more effective than Eq. (1.2) for calculating interaction energy. Moreover, the many body expansion provides a framework upon which the intermolecular potentials can be constructed, such as the potential developed in publication of Chapter 3. Describing interaction energy of Eq. (1.2) using an analytic function of atomic coordinates is practically impossible, unless the cluster is very small (in the case of water, up to three molecules). This difficulty arises from the number of variables that are necessary to describe the whole large cluster, or what is often referred to as the “dimensionality curse”. The complete description of the whole cluster of N monomers requires $3NL - 6$ relative coordinates, where L is the number of atoms in the monomer ($L = 3$ for water). This number quickly becomes overwhelming, and,

therefore, the prudent approach is to divide the interaction into contributions coming from pairs, triples, etc. For a cluster that has N monomers, the many-body expansion is as follows [40]:

$$E_{\text{int}} = E_{\text{int}}[2, N] + E_{\text{int}}[3, N] + \dots + E_{\text{int}}[N, N], \quad (1.4)$$

where $E_{\text{int}}[2, N]$ represents the sum of pair interactions, and $E_{\text{int}}[K, N]$ with $K > 2$ includes non-additive, K -body contributions. The pair (two-body) contribution is typically the largest and includes all possible dimer contributions within the cluster:

$$E_{\text{int}}[2, N] = \sum_{i < j} E_{\text{int}}(i, j)[2, 2], \quad (1.5)$$

where $E_{\text{int}}(i, j)[2, 2]$ is the interaction energy of a dimer consisting of i th and j th monomer in the cluster. Many intermolecular potentials ignore non-additive contributions, or approximate them with polarization models. This leads to significant inaccuracies, as non-additive contributions are crucial for systems like liquid water or large clusters, contributing up to 23% of the interaction energy for larger clusters, as shown in the publication of Chapter 2.

The next term in Eq.(1.4), the three-body energy, represents the sum of the three-body energies of all trimers within the cluster,

$$E_{\text{int}}[3, N] = \sum_{i < j < k} E_{\text{int}}(i, j, k)[3, 3], \quad (1.6)$$

where $E_{\text{int}}(i, j, k)[3, 3]$ is the three-body nonadditive interaction energy of a trimer consisting of monomers i , j , and k . The non-additive energy contains the part of the trimer energy that is not pairwise additive, i.e., it is the difference between the interaction energy of the trimer with monomers i , j , and k , $E_{\text{int}}(i, j, k)$, and all the

pairwise interaction energies within a trimer:

$$E_{\text{int}}(i, j, k)[3, 3] = E_{\text{int}}(i, j, k) - E_{\text{int}}(i, j)[2, 2] - E_{\text{int}}(j, k)[2, 2] - E_{\text{int}}(i, k)[2, 2]. \quad (1.7)$$

The methodology can be extended to four and higher-body terms [39, 40]. From Eq. (1.6), it is evident that the number of the three-body terms is significantly larger than two-body terms, but, as presented in publication of Chapter 2, this issue is manageable.

The stratified approximation strategy [“stratified approximation” many-body approach (SAMBA)] employs a truncated many-body expansion with K -dependent basis set and theory levels, as presented in publication of Chapter 2. This strategy serves as an alternative to canonical supermolecular calculations of cluster interactions energies as described in Eq. (1.2). While more calculations are required for each sub-cluster compared to the canonical supermolecular approach, these calculations can be performed at progressively lower levels of theory and with smaller basis sets, leading to significant savings in computational resources without compromising accuracy. The procedure can also yield benchmark-quality results that are not feasible with the canonical approach at such precision.

The benchmarks offer insight into the required level of accuracy for the calculations. With this understanding, along with a polarization model and a set of grid points, one can finally start developing an *ab initio* force field. Fitting of intermolecular potentials might seem straightforward: parameters of the potential’s functional form are adjusted to minimize errors with respect to benchmark values on a grid. However, such a simplistic application is unlikely to produce a good potential at first or even after several attempts. There are several challenges: the functional form may lack flexibility, the grid points may not represent the whole space adequately, and the potential’s accuracy can vary depending on the distance between the molecules. Finally, the space between the grid points may be poorly captured by the potential. While these issues arise in two-body potentials, they are even more pronounced in three-body ones,

due to the significantly larger dimensionality and hence much sparser representation of its physical complexity. Consequently, fitting becomes a heuristic process of trial and error, involving development of multiple potential versions and fine-tuning them based on performance. The experience with the development of the potential in the publication of Chapter 3 (and also with other potentials) led later to the development of AutoPES [41], where the process is more automated.

1.3 SAMBA approach

In “stratified approximation” many-body approach (SAMBA) presented in detail in the publication of Chapter 2, we aimed to develop a more efficient method to compute the total interaction energy of a molecular cluster with reduced computational resources, while maintaining or improving accuracy compared to the traditional methods. The main tool for achieving this goal is the many body expansion of Eq. (1.4), which utilizes the fast convergence of this expansion, allowing the calculation of subsequent terms with progressively lower levels of theory. While neither the use of the many-body expansion nor the concept of “embedding”—where less important energy fragments are calculated with lower accuracy methods—is new, the combination of these approaches at the level of simultaneous modification of both the basis set and the level of theory is novel. This combination enables benchmark-quality results. To validate our approach, we conducted a series of extended computational tests across various levels of theory, employing numerous methods and basis sets. Our focus was on water clusters, particularly the water hexamer, where the two isomers, cage and prism, are the lowest-energy structures with very slight energy differences. Determining the relative stability of cage and prism is challenging and depends on the level of theory. Additionally, due to water’s polar nature, many-body effects contribute significantly, causing the many-body expansion to converge slowly, making this system close to the worst-case scenario. Despite this specific focus, the presented scheme is universal and applicable to other systems.

In our calculations, we used the aug-cc-pVXZ basis sets of Dunning and coworkers [42] with X ranging from 2 to 6, and applied complete basis set (CBS) extrapolations. The methods employed included Hartree-Fock (HF), second order Møller-Plesset perturbation theory (MP2), and coupled cluster method with singles, doubles, and non-iterative triples [CCSD(T)]. The latter method is considered the “gold standard” in benchmarking calculations, offering very high accuracy with significant but manageable computational costs. To ensure that the calculations did not accumulate numerical errors that could compromise their accuracy, the integral and convergence thresholds typically used in the software packages were set to tighter values than the defaults.

We explored various levels of theory and basis sets to map the errors across all stages of the many-body expansion. The key finding, supported by data for the water hexamer in Table 2.10, is that the SAMBA approach indeed offers advantages over the canonical supermolecular calculation. For the water hexamer, the SAMBA approach either achieved similar accuracy to the canonical representation with approximately 24-fold reduction in compute time or delivered better accuracy by ensuring more thorough convergence of the two-body energies with only about a 25% increase in computational time (and a 15-fold reduction in computational cost compared to the canonical calculations).

A general strategy for the water hexamer is summarized in Table 2.8. We listed the levels of theory and basis sets for each step of the many-body expansion to achieve different levels of accuracy, ranging from a broad 10 kcal/mol criterion to a very fine 0.02 kcal/mol. While achieving the latter would likely require post-CCSD(T) calculations, the 0.1 kcal/mol accuracy is more feasible with just up to the CCSD(T) level of theory. The results indicate rapid convergence in the many-body expansion, making problematic five- and higher-body calculations unnecessary in most cases. Four-body terms can be handled at a lower level of theory and with smaller basis sets, with MP2/aug-cc-pVDZ sufficiently representing these calculations. However, for extremely high accuracy, CCSD(T)/aug-cc-pVDZ is recommended. In contrast, it is more effective to focus on the two-body terms since they are both very

large and relatively inexpensive to compute. The best results for these can be achieved using CCSD(T), large basis sets, and CBS extrapolations. Three-body terms are also significant, and for higher accuracy goals, the CCSD(T) level of theory is needed, but the basis set size can be reduced compared to the two-body calculations.

Table 2.9 illustrates the main source of the SAMBA efficiency. In SAMBA, each K -body term is computed with its respective basis set rather than the entire cluster's basis set. This approach substantially reduces costs while introducing only minimal residual basis set superposition error (BSSE). This holds true to a lesser extent for calculations that are not counterpoise corrected, such as the literature benchmark of the 24-mer [43]. In SAMBA, since higher-body terms can be ignored beyond a certain point, there is no need for extremely expensive full-cluster calculations. Following the SAMBA approach and with the recommended strategy of including up to four-body terms, the larger system considered is the tetramer, regardless of the overall system size. This approach is particularly advantageous for the studied 24-mer (and to the 16-mer), where the literature benchmark [43] was computed using a modest basis set, lacking diffuse functions and with some polarization functions removed, without counterpoise corrections. Those calculations were performed on a massive 223,200-core machine and took 76 years of combined processor time. Our calculations required 200 times less aggregate processor time, did not require such a large compute hardware, and provided significantly better-converged energies with respect to the basis set. Detailed results are shown in Tables 2.11, 2.12, and Figure 2.6. Results for the 16-mer are included in the Supporting Information of the paper.

It should be noted that the SAMBA approach does have some disadvantages. The primary benefit of the canonical supermolecular approach is the significantly smaller number of calculations required to obtain the interaction energy, making the process less prone to human and computational errors. In contrast, SAMBA requires a series of complex scripts to prepare inputs and gather results, which introduces more opportunities for mistakes. Crashes or incorrect results are easier to detect and avoid in canonical calculations than in the numerous calculations needed for SAMBA.

Additionally, the large number of calculations involved in SAMBA imposes stricter requirements to prevent error accumulation or computational artifacts. Lastly, geometry optimizations are much more difficult to perform with the SAMBA approach. While it might be possible to combine the gradients from individual calculations, this is a non-trivial task compared to canonical calculations, and such an extension to SAMBA has not yet been implemented. As a result, the SAMBA approach is currently limited to single-point calculations.

In summary, the SAMBA approach is recommended for high-accuracy benchmark calculations for smaller water clusters as well as larger systems like the 24-mer. For even larger systems, the strategy may need to be combined with more simplified approaches to handle very high-body terms, such as our polarization model developed in the publication of Chapter 3, where its application to large clusters is also discussed. This scheme should be applicable to other structures as well. Water is a highly polar system, and despite this, the many-body expansion converges relatively quickly. For non-polar systems, the convergence is expected to be even faster, although non-additive dispersion effects, which dominate in non-polar systems, require higher levels of theory, making MP2 unsuitable for these cases.

1.4 CCpol23+ potential

During the work on the publication of Chapter 2, it turned out that the existing three-body potential [31] is not accurate enough to properly model interactions in water clusters. Therefore, the primary goal of the subsequent work was to develop a new three-body potential from first principles. Three-body non-additive effects account for approximately 20% of the interaction energy for water clusters larger than a dimer [31]. The CCpol potential's two-body component was already quite accurate, and the main source of the error in modeling water clusters stemmed from the three-body part of the potential.

We began by selecting a large pool of grid points to cover a wide range of relevant trimer geometries. The initial set of 7,533 grid points from Ref. 31 served

as a starting point. Additional geometries were taken from trimers extracted from clusters, ranging from tetramer to 21-mers, which included configurations with large intermonomer separations. Another set of points was generated randomly, ensuring that they were far from the existing points. Additional points included snapshots from MD simulations and quantum Monte-Carlo simulations for water hexamers, which aimed to sample the rovibrational motions of the cluster. Further points were added to represent regions near the minima of water hexamers. Finally, more points were included to properly model regions with short intermolecular separations. Overall, the grid set included 71,456 configurations, providing comprehensive coverage for possible applications of the potential in liquid simulations and cluster modeling.

Next, we conducted a series of tests to identify a combination of the level of theory and basis sets that would offer the best balance of accuracy and computational efficiency. Based on our previous work on the publication of Chapter 2, we knew that CCSD(T) was necessary for sufficient accuracy, but its relatively high cost required us to search for an optimal basis set. The main findings are summarized in Table 3.1. The optimal approach turned out to be a hybrid scheme, where interaction energies are calculating with MP2 in the aug-cc-pVTZ basis set and CCSD(T) in the aug-cc-pVDZ set. This method resulted in negligible loss of accuracy compared to full CCSD(T)/aug-cc-pVTZ, while achieving a fourfold reduction in compute time. Although a single geometry calculation takes around an hour, the total number of grid points makes these time savings significant. All points used a rigid-monomer geometry averaged over the ground rovibrational state of the water molecule.

The two-body CC-pol-8s [44] potential served as the two-body component of our new complete potential, but it required slight modifications due to the introduction of a new polarization model. Since this polarization model is applied across all levels of the many-body expansion, adjustments were necessary for the two-body part. Despite these modifications, both the functional form and performance of the two-body part remained very similar to the CC-pol-8s potential, with only minor differences. The new

polarization model, which utilized three polarizable sites instead of one, was specifically designed to accurately reproduce four-body and higher terms in water hexamer. Interestingly, it also performed better in recovering three-body interaction energies in a sample of trimers from an MD simulation (cf. Table 3.5). The average error in the four-body interaction energies for a set of water hexamers was just 5.8%, with the largest error being 12.2%. This is a noteworthy outcome of the paper since simple polarization models tend to have errors of about 50% for trimer energies [16], suggesting that polarization models aimed at higher-order terms can also accurately model these interactions.

The three-body functional form followed the general structure of the potential described in Ref 45. However, instead of using a Legendre polynomial expansion, we opted for exponential site-site terms. The 6 symmetry-unique sites used in the fit resulted in 63 unique nonlinear parameters and 364 linear adjustable parameters. Given the symmetry of the water molecule and the need to respect permutational symmetry, the fitting process was carefully designed to ensure that the resulting fit was invariant under all symmetry operations, including permutations of symmetry-equivalent atoms and monomers.

The parameters were fitted through a multi-stage procedure, with the initial values of the nonlinear parameters chosen randomly. For the fits with the lowest root mean square error (RMSE), an additional step was taken to reduce divergence at short distances, and further points were added to address short-range issues. All grid points were assigned equal weights, except for this additional short-range set. The RMSE for the complete non-short range set of 70,268 points was 0.0184 kcal/mol, while for trimers extracted from hexamers, it was 0.0145 kcal/mol. This represents a significant reduction in error compared to the potential in Ref. 31, where the RMSE was 0.07 kcal/mol.

Additional comparisons were made using 600 points extracted from MD simulations, as shown in Table 3.5 (cf. also Table 1.1 with additional results from the MB-pol potential [46]). The RMSE of 0.0184 kcal/mol indicates that the three-body fit has

accuracy comparable to the two-body fit, which has an RMSE of about 0.01 kcal/mol. With three dimers in a trimer, one would expect an error larger by $\sqrt{3} \approx 1.73$ if the errors were random. A more elaborate fitting form could have reduced the three-body fit error, but a highly complex form would have increased the computational time required for simulations using the fit.

The result of the work is a family of potentials: CCpol2 represents the two-body potential, CCpol3 is the three-body potential (the main focus of the work), CCpol23 combines the two- and three-body potentials, and CCpol23+ includes two-body, three-body, and higher-body terms from the polarization model. These potentials have been applied to water clusters. The first test involved trimers extracted from water hexamers, with results shown in Table 3.2. Overall, CCpol3 performed very well, with errors only 3 to 5 times larger than the underlying *ab initio* calculations used to train the potential, and better than the potentials available in the literature at the time. It correctly identified the cage structure as having a lower three-body energy than the prism. Another test focused on characteristic points of the water trimer. Good performance in this case would indicate possible good results in modeling the trimer spectra. The details are provided in Table 3.1 and Figure 3.1.

The results calculated with rigid-monomer geometries of the water trimer optimized with CCpol23 are excellent, with an RMSE for the barriers between stationary points of only 0.06 kcal/mol compared to the CBS CCSD(T) results, and the largest error being 0.126 kcal/mol for the highest barriers. In the case of optimization with flexible monomers, the barrier errors increase by about 2.5 times, mainly due to differences in monomer geometries, as even the rigid CCSD(T) CBS results show significant errors. By applying a “monomer flexibility correction,” the barrier height errors are reduced to the same level as in the rigid-monomer case. The accuracy is consistent with the potential’s uncertainties and demonstrates that the intermolecular geometries optimized with the CCpol23 potential are very close to those optimized with flexible monomers.

For larger systems, it is possible to test the full CCpol23+ potential that includes

the description of the higher than three-body effect via the polarization model. A test of the relative stability of various water hexamer isomers is presented in Figure 3.3. Despite being compared to methods involving flexible monomers, CCpol23+ performs best among all available potentials in recovering the relative stability of the isomers. It turns out that the flexible-monomer effects are of similar magnitude across the isomers, and the accuracy of the intermolecular potentials plays a more significant role than flexibility.

1.4.1 CCpol23+ Errata

The applications of CCpol23+, both by the original authors and other researchers, revealed several issues. They are detailed in an erratum published alongside the website where the potential can be downloaded [47]. The first two issues involve the two-body part. The first one relates only to the description in the paper of how the potential was damped in Eq. (3.10) since the computer code was unaffected and worked as intended. Only a part of the $u_{ab}(r_{ab})$ was damped, instead of the whole contribution as stated in the paper. Another issue is a programming error where certain terms in the short-range part of the potential were damped multiple times instead of once. This error affected the results, but only at very short distances, which are physically irrelevant in most applications.

Ref. 48 pointed out another problem. When applying the CCpol3 potential to calculate the third virial coefficient, some exchange terms that should quickly decay at long distances diverged, leading to incorrect behavior at very large separations. The solution is to cut off these terms at 10 Å in intermonomer separation. After applying this fix, the three-body potential could then be used to properly calculate the third virial coefficient. Additional issues were identified in the two-body part at very short distances, where the potential exhibits unphysical behavior. The fix involved setting the potential to a very high value below a certain cutoff distance. It is important to note that these issues are not unique to CCpol; similar fixes were also required for the MB-pol potential.

Table 1.1: RMSEs (in kcal/mol) of nonadditive three-body energies on a sample of 600 configurations from an MD simulation. ‘MB-pol’ is the potential of 46. Refer to the caption of Table 3.5 for further details.

polarization model (old CCPOL)	0.1070
polarization model (present work)	0.0734
SAPT-3B/old CCPOL	0.0418
WHBB6	0.0642
HBB2-pol	0.0374
MB-pol	0.0203
CCpol3	0.0154

1.5 Current status of the water potentials

At the time the manuscript of Chapter 3 was submitted, the MB-pol [46] three-body potential had just been published, following an earlier two-body potential [49] and improving the earlier the HBB2-pol [50] potential. We learned about the MB-pol potential after our manuscript had been reviewed and no comparison of the MB-pol to our potential could be made, only a short citation was given as a note added to the proof.

The MB-pol potential used 12,000 grid points calculated at the CCSD(T)/aug-cc-pVTZ level, applying permutationally invariant polynomials, monomer flexibility, and a polarization model for four-body and higher-order terms. Compared to our CCpol23+ potential of Chapter 3, the MB-pol potential employed a slightly higher-level *ab initio* calculations [CCSD(T)/aTZ compared to MP2/aTZ/CCSD(T)/aDZ; however, as shown in our publication this change in the level of theory has a negligible impact compared to other possible inaccuracies in the fit], with about six times fewer grid points. Additionally, the MB-pol had to also model monomer flexibility with this reduced set of points, possibly impacting accuracy.

To evaluate the MB-pol potential, I have updated Figures 3.1, 3.2, and Table 3.5 to include MB-pol results. The new data are presented in Figures 1.2, 1.3 and Table 1.1. Overall, these results support the claim from Ref. 46 that MB-pol is a

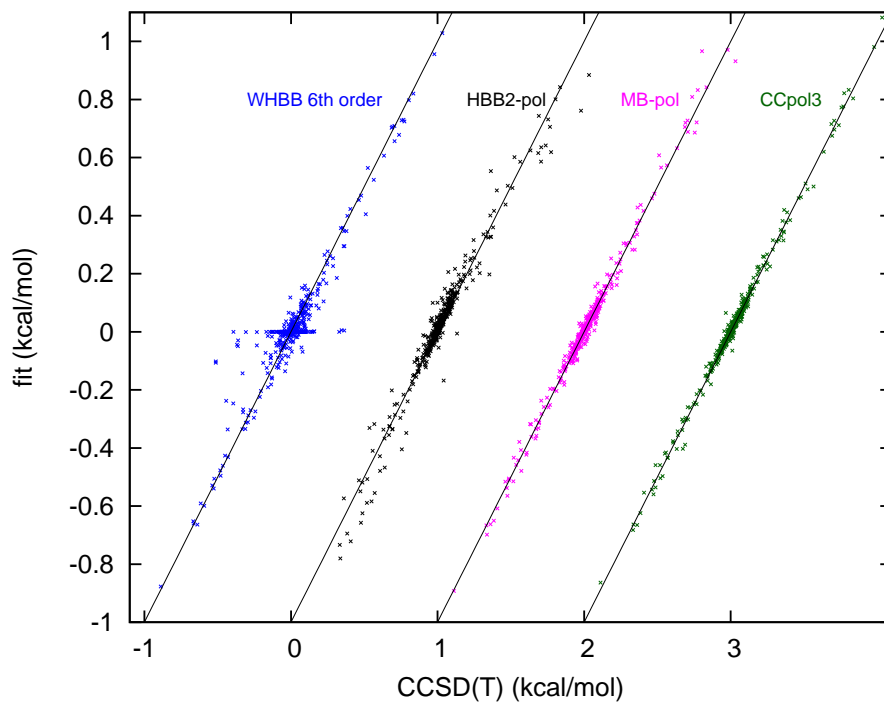


Figure 1.2: Accuracy of the three-body potentials on a sample of 600 MD points. MB-pol refers to the potential of Ref. 46. For other details, see the caption of Figure 3.2.

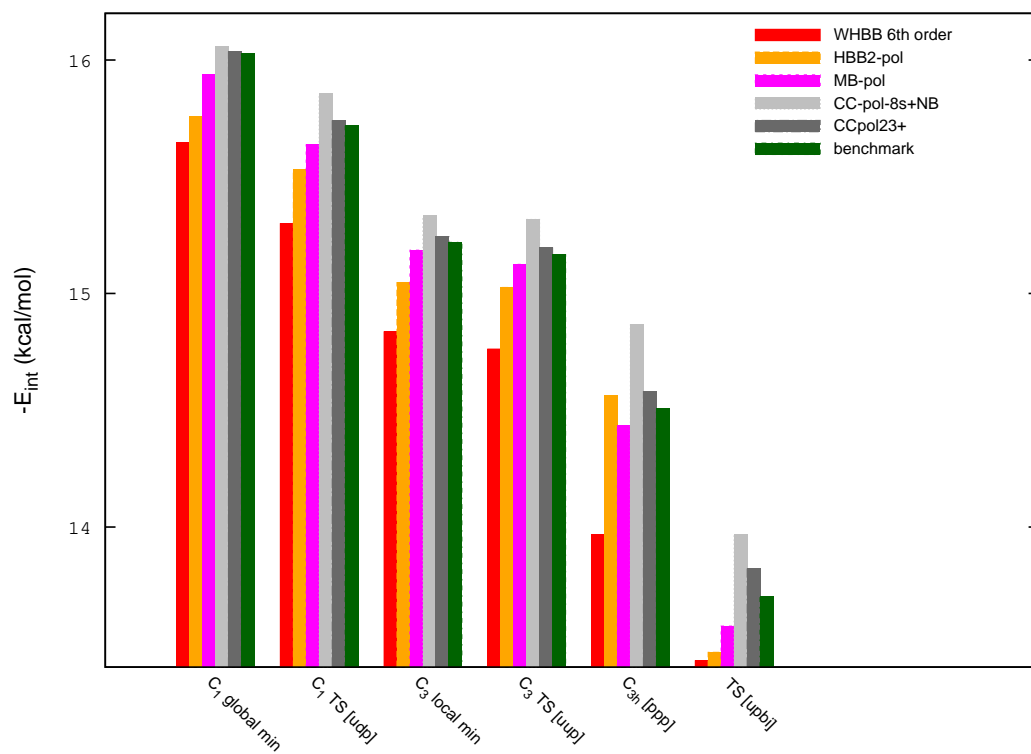


Figure 1.3: Interaction energies for water trimer characteristic points. MB-pol is the potential of Ref. 46. Other symbols are explained in the caption of Figure 3.1.

significant improvement over earlier literature results. However, CCpol3 outperforms MB-pol, both for the benchmark of sample points from an MD simulation and for the water trimer characteristic points. It is worth noting that these benchmarks use rigid monomer geometries, which favor CCpol3. As mentioned in publication of Chapter 3, inaccuracies in the three-body potentials usually have a larger impact on results than neglecting flexibility effects. Therefore, CCpol3 (and the full CCpol23+) should outperform MB-pol in these cases. This is further supported by an application of the CCpol3 potential to the calculations of the third virial coefficient for water [48], where CCpol3 results showed better agreement with experimental benchmarks than MB-pol.

Given that MB-pol performed very well for liquid water and for other benchmarks [51], it remained a strong competitor to CCpol23+. Flexible effects has been included in the two-body part of the CCpol potential [52] but not in the three-body one. Nevertheless, CCpol3 have been recommended for cases when the accuracy of the three-body potential was critical and the flexible effects are less significant.

CCpol23+ and other potentials that were developed by the time of publication of Ref. 2 have remained among the leading *ab initio* water potentials for nearly a decade. However, with recent advancements in computing power, a new generation of improved potentials has emerged. Recently, several potentials from the family of permutationally invariant polynomials [53] received updates, featuring more grid points, a higher level of theory, and improvements in potential representation. The first of these was q-AQUA [54], which utilized 71,892, 45,332, and 3,692 grid points for the two-, three-, and four-body components, respectively. This potential incorporated the first explicit four-body potential, developed earlier [55]. Initial tests of the latter potential showed generally good accuracy, particularly in comparison to approximate polarization models for four-body interactions. However, the relatively small number of grid points, which is insufficient for such a high-dimensional problem, raises concerns about its performance on configurations that differ significantly from the training data. Despite this, the full q-AQUA potential, with its relatively large set of the two- and three-body grid points and improved theory (CCSD(T)-F12a/aVTZ for the three-body), enhances the

modeling of clusters and liquid water. The potential was further expanded with a polarization model for interactions beyond four body terms [56], making it a complete potential. Another significant development was the update to the MB-pol potential [46], now called MB-pol(2023) [57], which was trained on the same dataset as q-AQUA and also incorporates explicit four-body terms, though again on a relatively small dataset. With over 2,000 parameters for the three-body component, this potential offers improved accuracy over MB-pol and likely surpassed CCpol23+, even for rigid-monomer systems. Unfortunately, the code implementing MB-pol(2023) has not yet been released.

Recently, several potentials based on neural networks have also been developed. A set of potentials released by Paesani and coworkers [58] aimed to address the primary disadvantage of sophisticated potentials like MB-pol: their high computational cost. Once trained, neural network potentials are relatively fast to compute, especially when using specialized hardware such as graphical processing units (GPU) or tensor processing units (TPU). The potential from Ref. 58 was trained on MB-pol data, and while it could not surpass MB-pol in accuracy, the authors aimed to achieve comparable accuracy at a lower computational cost. This goal was only partially realized, and for certain applications, such as phase transitions, the neural network potentials performed significantly worse. Another noteworthy development by Zhu *et al.* [59] involved training a neural network potential on a vast dataset of 220,000 grid points for two-body and an even more extensive 430,000 points for three-body interactions at the CCSD(T) level. This potential showed excellent performance for water trimer spectra, although its performance on larger water clusters remains to be tested. With the recent development of such high-quality potentials, the goal of finding a universal potential that performs equally well for both small and large systems is much closer to being achieved.

Bibliography

- [1] U. Góra, R. Podeszwa, W. Cencek, and K. Szalewicz, *J. Chem. Phys.* **135**, 224102 (2011).
- [2] U. Góra, W. Cencek, R. Podeszwa, A. van der Avoird, and K. Szalewicz, *J. Chem. Phys.* **140**, 1941011 (2014).
- [3] O. Matsuoka, E. Clementi, and M. Yoshimine, *J. Chem. Phys.* **64**, 1351 (1976).
- [4] B. Jeziorski and M. van Hemert, *Mol. Phys.* **31**, 713 (1976).
- [5] K. Szalewicz, S. J. Cole, W. Kolos, and R. J. Bartlett, *J. Chem. Phys.* **89**, 3662 (1988).
- [6] E. M. Mas and K. Szalewicz, *J. Chem. Phys.* **104**, 7606 (1996).
- [7] W. Klopper, J. G. C. M. van Duijneveldt-van de Rijdt, and F. B. van Duijneveldt, *Phys. Chem. Chem. Phys.* **2**, 2227 (2000).
- [8] G. S. Tschumper, M. L. Leininger, B. C. Hoffman, E. F. Valeev, H. F. Schaefer III, and M. Quack, *J. Chem. Phys.* **116**, 690 (2002).
- [9] J. R. Lane, *J. Chem. Theory Comput.* **9**, 316 (2012).
- [10] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *Science* **315**, 1249 (2007).
- [11] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).
- [12] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, *J. Phys. Chem.* **91**, 6269 (1987).
- [13] K. Szalewicz, C. Leforestier, and A. van der Avoird, *Chem. Phys. Lett.* **482**, 1 (2009).

- [14] L.-P. Wang, T. Head-Gordon, J. W. Ponder, P. Ren, J. D. Chodera, P. K. Eastman, T. J. Martinez, and V. S. Pande, *J. Phys. Chem. B* **117**, 9956 (2013).
- [15] K. T. Wikfeldt, E. R. Batista, F. D. Vilaça, and H. Jonsson, *Phys. Chem. Chem. Phys.* **15**, 16542 (2013).
- [16] O. Akin-Ojo and K. Szalewicz, *J. Chem. Phys.* **138**, 024316 (2013).
- [17] R. S. Fellers, C. Leforestier, L. B. Braly, M. G. Brown, and R. J. Saykally, *Science* **284**, 945 (1999).
- [18] R. S. Fellers, L. B. Braly, R. J. Saykally, and C. Leforestier, *J. Chem. Phys.* **110**, 6306 (1999).
- [19] N. Goldman, R. S. Fellers, M. G. Brown, L. B. Braly, C. J. Keoshian, C. Leforestier, and R. J. Saykally, *J. Chem. Phys.* **116**, 10148 (2002).
- [20] N. Goldman, C. Leforestier, and R. J. Saykally, *Phil. Trans. Royal Soc. (London), Ser. A* **363**, 493 (2005).
- [21] K. Szalewicz, R. Bukowski, and B. Jeziorski, in *Theory and Applications of Computational Chemistry: The First 40 Years. A Volume of Technical and Historical Perspectives*, edited by C. E. Dykstra, G. Frenking, K. S. Kim, and G. E. Scuseria (Elsevier, Amsterdam, 2005) Chap. 33, pp. 919–962.
- [22] E. Clementi and P. Habitz, *J. Phys. Chem.* **87**, 2815 (1983).
- [23] U. Niesar, G. Corongiu, E. Clementi, G. R. Kneller, and D. K. Bhattacharya, *J. Phys. Chem.* **94**, 7949 (1990).
- [24] C. Millot and A. J. Stone, *Mol. Phys.* **77**, 439 (1992).
- [25] E. M. Mas, K. Szalewicz, R. Bukowski, and B. Jeziorski, *J. Chem. Phys.* **107**, 4207 (1997).

- [26] C. Millot, J. C. Soetens, M. T. C. M. Costa, M. P. Hodges, and A. J. Stone, *J. Phys. Chem.* **102**, 754 (1998).
- [27] G. C. Groenenboom, E. M. Mas, R. Bukowski, K. Szalewicz, P. E. S. Wormer, and A. van der Avoird, *Phys. Rev. Lett.* **84**, 4072 (2000).
- [28] E. M. Mas, R. Bukowski, K. Szalewicz, G. C. Groenenboom, P. E. S. Wormer, and A. van der Avoird, *J. Chem. Phys.* **113**, 6687 (2000).
- [29] G. C. Groenenboom, P. E. S. Wormer, A. van der Avoird, E. M. Mas, R. Bukowski, and K. Szalewicz, *J. Chem. Phys.* **113**, 6702 (2000).
- [30] M. J. Smit, G. C. Groenenboom, P. E. S. Wormer, A. van der Avoird, R. Bukowski, and K. Szalewicz, *J. Phys. Chem. A* **105**, 6212 (2001).
- [31] E. M. Mas, R. Bukowski, and K. Szalewicz, *J. Chem. Phys.* **118**, 4386 (2003).
- [32] B. Jeziorski, R. Moszyński, and K. Szalewicz, *Chem. Rev.* **94**, 1887 (1994).
- [33] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *J. Chem. Phys.* **128**, 094313 (2008).
- [34] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *J. Chem. Phys.* **128**, 094314 (2008).
- [35] A. Shank, Y. Wang, A. Kaledin, B. J. Braams, and J. M. Bowman, *J. Chem. Phys.* **130**, 144314 (2009).
- [36] B. E. Rocher-Casterline, L. C. Ch'ng, A. K. Mollner, and H. Reisler, *J. Chem. Phys.* **134**, 211101 (2011).
- [37] M. Jeziorska, P. Jankowski, K. Szalewicz, and B. Jeziorski, *J. Chem. Phys.* **113**, 2957 (2000).
- [38] R. Podeszwa and K. Szalewicz, *J. Chem. Phys.* **126**, 194101 (2007).

- [39] G. Chałasiński and M. M. Szczęśniak, *Chem. Rev.* **94**, 1723 (1994).
- [40] C. E. Dykstra, G. Frenking, K. S. Kim, and G. E. Scuseria, eds., *Theory and Applications of Computational Chemistry: The First 40 Years. A Volume of Technical and Historical Perspectives* (Elsevier, Amsterdam, 2005).
- [41] M. Metz, K. Piszczatowski, and K. Szalewicz, *J. Chem. Theory Comput.* **12**, 5895 (2016).
- [42] R. A. Kendall, T. H. Dunning, Jr., and R. J. Harrison, *J. Chem. Phys.* **96**, 6796 (1992).
- [43] E. Apra, R. J. Harrison, W. A. deJong, A. P. Rendell, V. Tipparaju, and S. S. Xantheas, in *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis* (ACM, New York, 2010) article number 66.
- [44] W. Cencek, K. Szalewicz, C. Leforestier, R. van Harrevelt, and A. van der Avoird, *Phys. Chem. Chem. Phys.* **10**, 4716 (2008).
- [45] E. M. Mas, R. Bukowski, and K. Szalewicz, *J. Chem. Phys.* **118**, 4386 (2003).
- [46] V. Babin, G. R. Medders, and F. Paesani, *J. Chem. Phys.* **10**, 2906 (2014).
- [47] “CCpol23+ water potential,” <https://www.physics.udel.edu/~szalewic/ccpol23/>.
- [48] G. Garberoglio, P. Jankowski, K. Szalewicz, and A. Harvey, *Faraday Discuss.* **212**, 467 (2018).
- [49] V. Babin, C. Leforestier, and F. Paesani, *J. Chem. Theory Comput.* **9**, 5395 (2013).
- [50] V. Babin and F. Paesani, *Chem. Phys. Lett.* **580**, 1 (2013).
- [51] R. Medders, V. Babin, and F. Paesani, *J. Chem. Theory Comput.* **10**, 2906 (2014).

- [52] P. Jankowski, G. Murdachaew, R. Bukowski, O. Akin-Ojo, C. Leforestier, and K. Szalewicz, *J. Phys. Chem. A* **119**, 2940 (2015).
- [53] B. J. Braams and J. M. Bowman, *Int. Rev. Phys. Chem.* **28**, 577 (2009).
- [54] Q. Yu, C. Qu, P. L. Houston, R. Conte, A. Nandi, and J. M. Bowman, *J. Phys. Chem. Lett.* **13**, 5068 (2022).
- [55] A. Nandi, C. Qu, P. L. Houston, R. Conte, Q. Yu, and J. M. Bowman, *J. Phys. Chem. Lett.* **12**, 10318 (2021).
- [56] C. Qu, Q. Yu, P. L. Houston, R. Conte, A. Nandi, and J. M. Bowman, *J. Chem. Theory Comput.* **19**, 3446 (2023).
- [57] X. Zhu, M. Riera, E. F. Bull-Vulpe, and F. Paesani, *J. Chem. Theory Comput.* **19**, 3551 (2023).
- [58] Y. Zhai, R. Rashmi, E. Palos, and F. Paesani, *J. Phys. Chem.* **160**, 144501 (2024).
- [59] Y.-C. Zhu, S. Yang, J.-X. Zeng, W. Fang, L. Jiang, D. H. Zhang, and X.-Z. Li, *J. Am. Chem. Soc.* **144**, 21356 (2022).

Chapter 2

INTERACTION ENERGIES OF LARGE CLUSTERS FROM MANY-BODY EXPANSION ¹

Abstract

In the canonical supermolecular approach, calculations of interaction energies for molecular clusters involve a calculation of the whole cluster, which becomes expensive as the cluster size increases. We propose a novel approach to this task by demonstrating that interaction energies of such clusters can be constructed from those of small subclusters with a much lower computational cost by applying progressively lower-level methods for subsequent terms in the many-body expansion. The efficiency of such “stratified approximation” many-body approach (SAMBA) is due to the rapid convergence of the many-body expansion for typical molecular clusters. The method has been applied to water clusters $(\text{H}_2\text{O})_n$, $n = 6, 16, 24$. For the hexamer, the best results that can be obtained with current computational resources in the canonical supermolecular method were reproduced to within about one tenth of the uncertainty of the canonical approach while using 24 times less computer time in the many-body expansion calculations. For $(\text{H}_2\text{O})_{24}$, SAMBA is particularly beneficial and we report interaction energies with accuracy that is currently impossible to obtain with the canonical supermolecular approach. Moreover, our results were computed using two orders of magnitude smaller computer resources than used in the previous best calculations for this system. We also show that the basis-set superposition errors should be removed in calculations for large clusters.

¹ The text appeared in U. Góra, R. Podeszwa, W. Cencek, and K. Szalewicz, *J. Chem. Phys.* **135**, 224102 (2011).

2.1 Introduction

Calculations of cluster interaction energies are the subject of numerous papers, for a few examples see Refs. 1–6. Such calculations are usually performed using the *canonical supermolecular approach*, i.e., subtracting from the cluster total energy the energies of monomers. As the cluster size increases, the calculations eventually become prohibitively time consuming. For example, for water clusters the largest calculation performed on $(\text{H}_2\text{O})_{24}$ using the coupled-cluster method with single, double, and noniterative triple excitations [CCSD(T)] and a modified cc-pVTZ basis set containing 1224 orbitals required 76 years of central-processor unit (CPU) time [5, 7]. The calculations performed for smaller clusters can also be very time consuming if a high accuracy is required, like for example in the work of Bates and Tschumper (BT) [6], aimed at a precise determination of the energy difference between the prism and cage structures of the water hexamer. This difference is only about 0.25 kcal/mol or 0.5% of the total hexamer interaction energy.

An alternative way of performing such calculations is to build the N -body interaction energy from the sum of 2-body, 3-body, etc. contributions to this energy. The total interaction energy of an N -monomer cluster (an N -mer) can be described by the following expression:

$$E_{\text{int}} = E_{\text{int}}[2, N] + E_{\text{int}}[3, N] + \cdots + E_{\text{int}}[N, N], \quad (2.1)$$

where $E_{\text{int}}[K, N]$ is the K -body contribution to the interaction energy of the N -mer. The leading two-body contribution, $E_{\text{int}}[2, N]$, consists of all the pair interaction energies within the N -mer

$$E_{\text{int}}[2, N] = \sum_{i < j}^N E_{\text{int}}(i, j)[2, 2], \quad (2.2)$$

where i, j denote the monomers in the cluster. Similarly, the three-body term is a sum

of three-body nonadditive contributions from all non-equivalent trimers:

$$E_{\text{int}}[3, N] = \sum_{i < j < k} E_{\text{int}}(i, j, k)[3, 3], \quad (2.3)$$

where the three-body nonadditive term for a given trimer, $E_{\text{int}}(i, j, k)[3, 3]$, is defined as the difference between the total interaction energy of this trimer, $E_{\text{int}}(i, j, k)$, and the sum of pair interaction energies:

$$E_{\text{int}}(i, j, k)[3, 3] = E_{\text{int}}(i, j, k) - E_{\text{int}}(i, j)[2, 2] - E_{\text{int}}(j, k)[2, 2] - E_{\text{int}}(i, k)[2, 2]. \quad (2.4)$$

Analogous definitions hold for four- and higher-body nonadditive terms [8, 9]. The nonadditive K -body contribution can also be directly expressed via the total energies of all subclusters of a given cluster containing up to K monomers [9]

$$E_{\text{int}}[K, N] = \sum_{i=1}^K (-1)^{K-i} \binom{N-i}{K-i} S_{\text{tot}}[i, N] \quad (2.5)$$

where $S_{\text{tot}}[i, N]$ is the sum of the total energies of all the i -monomer subclusters of the N -mer cluster. For the three-body component (and similarly for other components),

$$S_{\text{tot}}[3, N] = \sum_{i < j < k} E_{\text{tot}}(i, j, k),$$

where $E_{\text{tot}}(i, j, k)$ is the total energy of a trimer consisting of monomers i , j , and k .

Equation (2.1) defines interaction energies relative to the sum of the energies of isolated monomers, each monomer at an identical geometry as in the cluster. Such definition is most often used in studies of cluster dynamics on potential energy surfaces formed by rigid monomers. In studies of the energetics of clusters, one often defines interaction energies relative to the sum of energies of isolated monomers at their respective equilibrium geometries. We will call such quantities the *relaxed* interaction

energies and the many-body expansion for such energies

$$E_{\text{int}}^{\text{rel}} = E_{\text{int}}[1, N] + E_{\text{int}}[2, N] + E_{\text{int}}[3, N] + \cdots + E_{\text{int}}[N, N], \quad (2.6)$$

acquires an additional, one-body term, $E_{\text{int}}[1, N]$, accounting for the energetic difference between the total energies of the monomers in the cluster geometries and their equilibrium geometries. Note that one can also analyze the many-body expansion in terms of the total cluster energies which removes the dependence on the reference point.

Truncated many-body expansions of Eq. (2.1) are widely used in physics since the beginnings of quantum mechanics. The first explicit treatment of the nonadditive interaction energies is apparently the 1943 work of Axilrod-Teller [10] and Muto [11]. Starting from 1970, the individual terms in many-body expansion have been particularly often computed for small water clusters, see, e.g., Refs. 12–14.

As one can see from Eq. (2.5), calculations of nonadditive contributions require a large number of calculations of the total energies for subclusters. Thus, a straightforward many-body approach—calculations of all the terms in Eq. (2.1) using a fixed basis set and level of theory—would be much more time consuming than the corresponding canonical supermolecular calculation. However, as we will show, the former approach becomes much more effective than the latter one if a proper strategy is used. It should be mentioned here that in the symmetry-adapted perturbation theory (SAPT) approach [9, 15, 16] to intermolecular interactions, the nonadditive contributions are calculated directly [9, 17], avoiding the problem discussed here. A general nonadditive SAPT approach has been developed so far for three-body interactions [17–19].

The main reason that the many-body expansion can be useful is the fact that it converges very fast and therefore can be truncated at a low K . The value of K depends on the accuracy required. For example, as will be shown later, if for larger water clusters one aims at an uncertainty of about one kcal/mol, one needs to include only two- and three-body contributions. In fact, for such clusters one can always restrict the expansion to up to four-body terms since the higher-body contributions

are smaller than the uncertainties of the best possible calculations of the lower-body terms. Although calculations of cluster interaction energies from a truncated many-body expansion still require many more calculations than the canonical supermolecular approach even if the maximum K is low, the calculations for $K \ll N$ are orders of magnitude less time consuming than the calculations for N . Furthermore, we will show that as K increases, the calculations can be performed at decreasing levels of theory and sizes of basis sets. In consequence, with such a strategy very significant savings of computer time can be achieved using truncated many-body expansions without sacrificing the accuracy of results.

The strategy of using a truncated many-body expansion with K -dependent basis set and theory levels—which can be called an “stratified approximation” many-body approach (SAMBA)—as an alternative to the canonical supermolecular calculations of cluster interactions energies is the main novel idea of the present work. This strategy and the demonstration that it results in much more accurate interaction energies than obtainable by the canonical supermolecular approach within given computational resources has not to our knowledge been published before. There have been many papers which presented calculations of many-body contributions, but in almost all cases the goal was to get an insight into the physical decomposition and not to perform an actual calculation of the total interaction energy. In some cases, e.g., in Ref. 20, the level of the method and the size of the basis set was decreased with K , but this was done only to make such calculations affordable. Also, the many-body expansion restricted to two-body and three-body terms has been often applied to develop intermolecular interaction potentials, see for example the work for water reviewed in Ref. 21. Another application of the many-body expansion is the “electrostatically embedded” method of Dahlke and Truhlar [22]. The closest precursor of our proposed approach is the work of Christie and Jordan [23]. However, these authors restrict the theory level to second-order many-body perturbation theory with Møller-Plesset partition of the Hamiltonian (MP2) only, so that they do not include a hierarchy of theory levels. They do propose to use smaller basis sets for higher K , but in a way more limited than in our approach.

As an illustration of our method, we have applied SAMBA in calculations for two water clusters: hexamer and icosikaitetramer, $(\text{H}_2\text{O})_{24}$. Additionally, results for the $(\text{H}_2\text{O})_{16}$ cluster have been included in the Supplementary Information [24]. Let us stress that the proposed approach is by no means restricted to water clusters, however, such clusters are particularly suitable for testing it. First, water clusters contain large many-body effects since the polar character of the water molecule implies large nonadditive induction interactions [25]. Therefore, the many-body expansion is more slowly convergent than for non-polar molecules, where the presented approach should work better. Second, due to the importance of water, there are highly-accurate benchmarks available. Water hexamer is distinguished among water clusters since it is a transitional structure [26, 27] between simple cyclic structures of $(\text{H}_2\text{O})_n$, $n = 3-5$ and larger clusters that have a clear three dimensional structure, resembling the hydrogen-bond organization of condensed phases. Water hexamer is also of interest due to the question concerning its minimum structure. As already mentioned, this cluster has in particular two energetically very close minimum structures, called cage and prism, and the difference

$$\Delta_{\text{p-c}} = E_{\text{int}}(\text{prism}) - E_{\text{int}}(\text{cage})$$

equals to only -0.25 kcal/mol (Ref. 6). Although the prism structure is the global minimum on the potential energy surface, if the zero-point vibrational energy (ZPVE) is included, the cage becomes the lowest-energy hexamer [28], in agreement with molecular-beam observations of rotational spectra [28]. However, later work by Steinbach *et al.* [29] assigned their observed infra-red intramonomer spectra to the book isomer. More recently, the cage paradigm has been put into doubt by BT [6] who argued that the difference of ZPVE between the two structures should be smaller than 0.25 kcal/mol, so that the observed hexamer should be the prism. This work may also indicate that the book isomer is not likely to be observed as its interaction energy is 0.72 kcal/mol above the prism structure. Of course, higher-energy isomers can possibly be observed at suitably high temperature ranges [30, 31]. Note that although we will discuss $\Delta_{\text{p-c}}$

in several places, the main emphasis of this paper is on cluster interaction energies E_{int} . As a byproduct of the $(\text{H}_2\text{O})_n$, $n = 6, 16$, and 24 applications, we have provided the values of both the total interaction energies and of the many-body contributions for the investigated clusters that are more accurate than published values. As stated recently by Wang *et al.* [32]: “definitive analysis of n -body contributions . . . remains to be done”. Thus, our calculations answer this request. Our many-body expansion for $(\text{H}_2\text{O})_{24}$ represents to our knowledge the largest cluster for which such an expansion has been performed at a high level of theory. The published cases are mostly limited to N of about eight and the MP2 method [33, 34], with the exception the MP2 calculations of Cui *et al.* [35] for $(\text{H}_2\text{O})_{21}$.

2.2 Details of calculations

In order to enable precise comparisons of the proposed approach with accurate benchmark values, we have performed very large basis set calculations for the water hexamer, computing both the total interaction energy and the many-body contributions. Then, to show the power of our approach, we have applied it to the water icosikaitetramer. We have used hexamer’s prism and cage structures since these are the two lowest-energy minima on the potential energy surface of the system, but also checked the proposed approach on several other hexamer structures. The geometries were optimized using the CC-pol-8s+NB (coupled-cluster based, polarizable, 8-site plus nonadditive effects) potential [36, 37]. The two-body part of this potential is an improved refit of the data used for the CC-pol-5s potential, the first *ab initio* potential that correctly predicted both spectroscopic properties of water dimers and bulk properties of liquid water [38]. The nonadditive part is from Refs. 25, 39. The potential was fitted to a set of interaction energies obtained from complete basis-set (CBS) extrapolations of CCSD(T) calculations. The monomers in the CC-pol-8s+NB (abbreviated later on to CC-pol) potential are frozen in the r_0 geometry, the average geometry of the ground state vibration. The CC-pol hexamer geometries are given in the Supporting Information [24]. One reason for choosing these geometries is that we plan to perform

diffusion quantum Monte Carlo calculations (DQMC) for the hexamer using the CC-pol-8s+NB potential. The other reason is that literature geometries of water cluster usually include distorted monomers, which would be an unnecessary complication in developing the computational strategy proposed by us. We emphasize that the choice of the test systems is not relevant for the development of our method since the convergence of many-body expansion is similar for all water clusters investigated to date. For the water icosikaitetramer, we have used the structures from Refs. 5, 7, 40.

In most of our calculations, we have used the series of augmented correlation-consistent aug-cc-pVXZ basis sets [41], $X = 2-6$. We will label the consecutive basis sets aDZ, aTZ, aQZ, a5Z, and a6Z. We will also occasionally use non-augmented bases, denoted as XZ. The MOLPRO suite of programs [42] was used for all the calculations except for the CCSD(T)/aQZ calculations for the water hexamer where the PQS code [43] was used. All MP2 and CCSD(T) calculations were performed using the frozen-core approximation. In MOLPRO calculations, we used energy convergence thresholds of 10^{-10} and 10^{-11} hartree for the CC and Hartree-Fock (HF) iterations, respectively. Additionally, the density threshold for HF was 10^{-7} and the threshold for the sum of squares of the changes in CC amplitudes was 10^{-14} . For PQS, the energy threshold was 10^{-9} hartree. All integrals were calculated with 10^{-15} and 10^{-14} precision in MOLPRO and PQS, respectively. These thresholds are a few orders of magnitude tighter than MOLPRO defaults. Since the thresholds should produce interaction energies accurate to at least 10^{-7} kcal/mol, such thresholds should make our results practically independent of possible accumulations of numerical errors in additions involved in Eq. (2.5).

To perform CBS extrapolations, we used algorithms of Halkier *et al.* [44, 45] For HF interaction energies we assumed:

$$E_{\text{int}}^{\text{HF}}(X) = E_{\text{int}}^{\text{HF}}(\text{CBS}) + Ae^{-\alpha X}, \quad (2.7)$$

whereas for MP2 or CCSD(T) interaction energies the following formula was used:

$$E_{\text{int}}^{\text{MET}}(X) = E_{\text{int}}^{\text{MET}}(\text{CBS}) + BX^{-3}, \quad (2.8)$$

where MET stands for the method used and $E_{\text{int}}^{\text{MET}}(X)$ are the interaction energies obtained using the method MET and the aXZ basis set. We have taken $\alpha = 1.63$ in Eq. (2.7), as recommended in Ref. 45. The remaining parameters were found by solving two linear equations with X and $X - 1$. We will also use symbols such as $E_{\text{int}}^{\text{MET}}(\text{Q5})$, which will denote $E_{\text{int}}^{\text{MET}}(\text{CBS})$ extrapolated from basis sets with $X = 4$ and 5. We have used the same extrapolations for basis sets with and without midbond functions. In the former case we used the same midbond functions for calculations in X and $X - 1$ bases. Appropriateness of such extrapolations was shown in Refs. 46, 47.

Some of our interaction energies were computed using a hybrid approach combining MP2 calculations in larger bases with CCSD(T) calculations in smaller ones. To this end, let us define the quantity:

$$\delta E_{\text{int}}^{\text{CCSD(T)}} = E_{\text{int}}^{\text{CCSD(T)}} - E_{\text{int}}^{\text{MP2}}, \quad (2.9)$$

where both the CCSD(T) and the MP2 energies have to be calculated in the same basis set. The hybrid interaction energy is then defined as

$$E_{\text{int}}^{\text{CCSD(T)}}(X, Y) = E_{\text{int}}^{\text{MP2}}(X) + \delta E_{\text{int}}^{\text{CCSD(T)}}(Y), \quad (2.10)$$

with $X > Y$. In a variant of this approach, the $E_{\text{int}}^{\text{MP2}}$ component is CBS extrapolated, and in another variant both components are extrapolated. This will be shown by an appropriate notation, for example, $E_{\text{int}}^{\text{CCSD(T)}}(\text{Q5}, \text{T})$ will mean that the $E_{\text{int}}^{\text{MP2}}$ part was extrapolated using the aQZ and a5Z results and the $\delta E_{\text{int}}^{\text{CCSD(T)}}$ part was computed in the aTZ basis set.

All our calculations use the counterpoise (CP) correction for the basis-set superposition error (BSSE) [48, 49]. This correction has been applied in two versions.

Unless stated otherwise, we will use the “orthodox” CP approach which consists in performing calculations of all the energies entering Eq. (2.5) in the basis set of the whole N -monomer cluster. This means that, for example in calculations for a hexamer, all dimer interaction energies are computed using the hexamer-centered basis set (HCBS). In some calculations utilizing many-body expansions, we will use a “simplified” CP correction such that calculations of the K -body nonadditivity for a given K -mer are performed in the K -mer-centered basis set (K CBS). The CP correction requires additional computer resources compared to noCP calculations since in the canonical supermolecular approach one has to perform N calculations for monomers in the complete cluster basis set. The time of such calculations is non-negligible, for example, for the water hexamer in the aTZ basis set the time required increases by about 40%. Thus, the CP correction is often neglected in calculations for larger clusters. Our results will shed light on uncertainties resulting from this approximation.

Most of the hexamer calculations described here will use rigid monomer geometries, but to compare with literature, we will present also calculations with distorted monomers, which requires a specification of how these calculations were performed. Following the notation of Ref. 49, we can write the CP-corrected N -mer interaction energy as:

$$E_{\text{int}}(\mathbf{Q}_1, \dots, \mathbf{Q}_N) = E_{\text{tot}}(\mathbf{Q}_1, \dots, \mathbf{Q}_N) - \sum_{i=1}^N E_{\text{tot}}^{\text{NCBS}}(\mathbf{Q}_i), \quad (2.11)$$

where $\mathbf{Q}_i = (\mathbf{R}_i, \boldsymbol{\omega}_i, \boldsymbol{\xi}_i)$ stands for the set of all coordinates needed to specify the spatial position \mathbf{R}_i , orientation $\boldsymbol{\omega}_i$, and the internal geometry $\boldsymbol{\xi}_i$ of the i th monomer. Note that although the total energy of the i th monomer, $E_{\text{tot}}^{\text{NCBS}}(\mathbf{Q}_i)$, is shown to depend only on \mathbf{Q}_i , it in fact depends on the complete geometry of the N -mer due to the location of the “ghost” functions in the N -mer-centered basis set (NCBS) calculations. Thus, even in the rigid-monomer approximation where all $\boldsymbol{\xi}_i$ are the same, all the monomer energies are generally different since each monomer has differently distributed ghost functions. In calculations with flexible monomers, the interaction energies defined in this way are sometimes called the “vertical” interaction energies.

In calculations with flexible monomers, one usually compares the so-called "relaxed" interaction energies, i.e., energies relative to the equilibrium geometry r_e of the isolated monomer. Such an energy differs from the vertical interaction energy $E_{\text{int}}(\mathbf{Q}_1, \dots, \mathbf{Q}_N)$ by the one-body terms defined earlier, also called "distortion" corrections, describing the increase of monomer energies due to their departure from the r_e geometry

$$E_{\text{int}}[1, N] \equiv E_{\text{dist}} = \sum_{i=1}^N E_{\text{tot}}(\boldsymbol{\xi}_i) - N E_{\text{tot}}(\boldsymbol{\xi}_{r_e}) \quad (2.12)$$

where $\boldsymbol{\xi}_{r_e}$ is the monomer equilibrium geometry. There is no superscript *NCBS* since E_{dist} is an energetic effect of small monomer distortions and it can be computed in the monomer-centered only basis set. The relaxed interaction energy can now be defined as

$$E_{\text{int}}^{\text{rel}}(\mathbf{Q}_1, \dots, \mathbf{Q}_N) = E_{\text{int}}(\mathbf{Q}_1, \dots, \mathbf{Q}_N) + E_{\text{dist}} \quad (2.13)$$

2.3 Complete basis set limits for water hexamer

In calculations of the interaction energy of a cluster, one can choose one of the two approaches discussed above and there are several strategies possible within each approach. We will illustrate these choices and their performance on the example of the water hexamer. We will push each method to the limits of the current computational resources. The simplest approach (Sec. 2.3.1) is to perform canonical supermolecular calculations for the whole cluster using a single method, both in terms of the level of theory and of the basis set. Since for the water hexamer one aims at an error of below 1 kcal/mol in the interaction energies, it has to be at least the CCSD(T) method. The largest basis applied in the literature at this level was the aTZ basis [50, 51]. We were able to perform such calculations in the significantly larger aQZ basis. A possible improvement of this approach (Sec. 2.3.2) is to use a hybrid method, defined by Eq. (2.10). Since one can perform the MP2 calculation for the hexamer in bases up to a5Z, this may reduce the uncertainty of the final interaction energy despite the fact that the $\delta E_{\text{int}}^{\text{CCSD(T)}}$ component remains at the aQZ level. Finally, one can use

the non-truncated many-body expansion and compute each K -body contribution as accurately as possible (Sec. 2.3.3). For example, for the 2-body term one can use the a6Z bases at the MP2 level and the a5Z bases at the CCSD(T) level.

2.3.1 CCSD(T) method only

The results of the CCSD(T) calculations for the water hexamer using the simplest version of the canonical supermolecular approach are presented in Table 2.1. We list in this and other tables results with four decimal digits. This is done to allow a better analysis of the convergence patterns and to list some small contributions which would amount to zero with less digits. The largest basis that we were able to use was aQZ. The magnitude of the interaction energy computed in the aQZ basis set increases by 1.68 (1.72) kcal/mol for the cage (prism) structure relative to the aTZ value. CBS extrapolations further increase these values by 1.11 (1.14) kcal/mol for the cage (prism) structure. These large changes indicate that even the aQZ results are not well converged with respect to the basis set. The error of the extrapolated result should be of the order of the difference between the CBS and largest basis set values. We have arbitrarily assumed half of this value as our estimate of the uncertainty of the CBS results, i.e., the uncertainty is 0.6 kcal/mol for both structures.

The difference between the prism and cage interaction energies is very small and changes dramatically in relative terms. No reliable estimate of this difference can be made from these calculations.

2.3.2 Hybrid MP2 plus CCSD(T) approach

Remaining within the canonical supermolecular approach, one can hopefully obtain more accurate energies using the hybrid approach of Eq. (2.10). For the MP2 calculations, the largest basis set that we could apply was a5Z. For these calculations, we have used the density-fitting (also known as resolution of identity) approach [52] with the auxiliary basis set of Ref. 53. This approximation should lead to a negligible loss of accuracy. We have tested it on the aQZ basis set and found that it amounted to

less than 0.0001 kcal/mol. On the other hand, the use of the density fitting in Hartree-Fock calculations resulted in an unacceptable error of more than 0.01 kcal/mol, and, therefore, we have performed the Hartree-Fock calculations without the density fitting. For the CCSD(T) calculations, the largest basis set was, of course, the same as used in the previous subsection.

The results presented in Table 2.2 show that the MP2 (Q5)-extrapolated interaction energies differ from the ones computed in the a5Z basis by 0.51 (0.52) kcal/mol for the cage (prism) structure. The analogous value for the (TQ) extrapolation vs. the aQZ basis set is 1.03 (1.06) kcal/mol, very similar to the increments seen in the previous subsection. The values of $\delta E_{\text{int}}^{\text{CCSD(T)}}$ are very small, only about 1% of the total interaction energy. The difference between the (TQ) and aQZ values is 0.08 kcal/mol, for both structures. We will estimate the accuracy of the best result in Table 2.2 as half of the difference between $E_{\text{int}}^{\text{CCSD(T)}}(5, \text{Q})$ and $E_{\text{int}}^{\text{CCSD(T)}}((\text{Q5}), (\text{TQ}))$, which amounts to about 0.3 kcal/mol. The difference between the ((TQ),T) and ((Q5),(TQ)) hybrid results is smaller than this value: 0.15 kcal/mol for both structures. Although the current uncertainty is significantly reduced from that of the pure CCSD(T) approach, it is still large enough to try to improve it, as it will be described in the next section.

Whereas the convergence of the differences between the prism and cage interaction energies is very poor, these energies are very close to each other in the ((TQ),T) and ((Q5),(TQ)) hybrid approaches: only 0.0003 kcal/mol apart and furthermore the latter one is only 0.0012 kcal/mol from the result of Sec. 2.3.1. This might suggest that these values are close to the CBS limit. We will show later that this is actually not the case. In fact, the (5,Q) result which stands apart here, is closest to our best estimate established later.

One should mention here that whereas the hybrid method works very well for water clusters, where $\delta E_{\text{int}}^{\text{CCSD(T)}}$ is always small, it may perform poorly for systems with larger $\delta E_{\text{int}}^{\text{CCSD(T)}}$, for example for the argon dimer [54].

2.3.3 Non-truncated many-body expansion approach

We will now try to obtain as accurate as possible interaction energies of the hexamer clusters from the many-body expansion. The two-body energies can be calculated in much larger basis sets than used in Secs. 2.3.1 and 2.3.2, up to a6Z/a5Z at the MP2/CCSD(T) level and including midbond functions. This was possible since instead of using HCBSs, we used dimer-centered "plus" basis sets (DC⁺BS), where the plus indicates the presence of midbond functions. This approach removes the major part of BSSE, as it will be shown in later sections. Furthermore, the true CBS limits are, of course, the same in CP-corrected and non-corrected approaches, and our calculations are fairly close to such limits. We have applied the hybrid MP2/CCSD(T) approach. The results are shown in Table 2.3. The MP2 energies were computed in the a5Z and a6Z bases with the *3s3p2d2f1g* midbond basis set from Refs. 55, 56. One can see that there is still a significant change of 0.19 (0.20) kcal/mol for the cage (prism) structures between the a5Z and a6Z MP2 interaction energies. The CBS extrapolation leads to a further increment of 0.26 (0.27) kcal/mol relative to the a6Z results.

For the $\delta E_{\text{int}}^{\text{CCSD(T)}}$ contribution, we performed calculations with the aQZ and a5Z basis sets plus the midbond functions. The difference between the results in these two basis sets is 0.06 (0.07) kcal/mol for the cage (prism) structure and the CBS extrapolation decreases further the magnitude of this contribution by 0.07 kcal/mol for both structures. Thus, the errors of $\delta E_{\text{int}}^{\text{CCSD(T)}}$ are much smaller than those of $E_{\text{int}}^{\text{MP2}}$ and therefore the hybrid method works very well for the two-body contributions.

To obtain the estimate of accuracy of our best value of the two-body contribution, we took the difference between the $E_{\text{int}}^{\text{CCSD(T)}}(6,5)$ and $E_{\text{int}}^{\text{CCSD(T)}}((56),(Q5))$ results, equal to 0.197 kcal/mol for both structures, and assumed half of this difference, or about 0.1 kcal/mol as the uncertainty of our results. The value of $\Delta_{\text{p-c}}$ appears to be converged much better, to within 0.001 kcal/mol. Table 2.3 shows also results obtained in the DCBS and HCBS schemes, which will be discussed in more detail later on. The issue of using *K*-mer centered basis sets is not relevant at this point since the final results in Table 2.3 are very close to the CBS limits.

The three-body contributions, $E_{\text{int}}[3, 6]$, presented in Table 2.4, were computed using trimer-centered basis sets (TCBS) up to $X = 5$. We have also included the aTZ result obtained using the HCBS scheme. The comparison with the aTZ/TCBS result shows that the two schemes give practically identical results (differences are 0.003 and 0.004 kcal/mol for the cage and prism hexamers, respectively). The convergence with X is very fast. The CCSD(T)/(TQ) values and the hybrid ones obtained using MP2/(Q5) and CCSD(T)/(TQ) differ by only 0.003 (0.002) kcal/mol for the cage (prism) structure. The difference between $E_{\text{int}}^{\text{CCSD(T)}}(5, \text{Q})$ and $E_{\text{int}}^{\text{CCSD(T)}}((\text{Q}5), (\text{TQ}))$ is 0.012 kcal/mol for both structures and we will assume half of this difference, i.e., 0.006 kcal/mol, as our estimate of uncertainty of the three-body contribution. The accuracy achieved for the three-body contributions is much higher than for the two-body ones. Thus, we could have used smaller basis sets and skip the CBS extrapolations for the former terms. For example, the aTZ results differ from the best values by only 0.04 kcal/mol for both structures, much less than the uncertainty of the two-body term. In contrast, it would not be possible to lower the level of theory, as the MP2 approach results in errors of about 0.4 kcal/mol.

One may notice that the errors of $E_{\text{int}}^{\text{MP2}}$ and of $\delta E_{\text{int}}^{\text{CCSD(T)}}$ are of opposite sign and comparable magnitude, and therefore partly cancel, which makes the convergence of $E_{\text{int}}^{\text{CCSD(T)}}$ faster than that of its components. In such situations, the hybrid approach is less effective, as shown by the negligible difference between the $E_{\text{int}}^{\text{CCSD(T)}}(\text{TQ})$ and $E_{\text{int}}^{\text{CCSD(T)}}((\text{Q}5), (\text{TQ}))$ values discussed above.

The higher-body terms are presented in Table 2.5. For $K = 4, 5$, and 6, we have just used the results computed using aTZ/HCBS. In view of the basis set convergence for the three-body terms discussed above (see also a further discussion of the basis set convergence of nonadditive components in Sec. 2.5) and the smallness of the $K > 3$ contributions, the absolute errors of these terms are negligible. For the four-body contribution, the difference between the aTZ and aDZ results (see Table 1 in the Supporting Information [24]) is only 0.0035 (0.0075) kcal/mol for the cage (prism) structure. Thus, we assume 0.006 kcal/mol as the uncertainty of this term. Similar

estimates for the five- and six-body terms are 0.002 kcal/mol or less. Summing all the uncertainties discussed above gives 0.12 kcal/mol as the uncertainty of our hexamer interaction energy.

The 0.12 kcal/mol uncertainty estimated above is significantly lower than that of the results in Sec. 2.3.1 (0.6 kcal/mol) and 2.3.2 (0.3 kcal/mol). For convenience, the latter results are quoted in the last two rows of Table 2.5. The actual differences between these results and the result of this section are 0.26 (0.28) and 0.23 (0.24) kcal/mol for the cage (prism) configuration, respectively. This indicates that our estimates of uncertainties are reasonable despite being conservative and the close agreement between the results of Secs. 2.3.1 and 2.3.2 is accidental. The same is even more true in the case of Δ_{p-c} which is about 1.7 times smaller in magnitude than the values from Secs. 2.3.1 and 2.3.2. The result $\Delta_{p-c} = -0.024$ kcal/mol obtained from the many-body expansion may still have a fairly large relative error as its magnitude is five times smaller than the uncertainty of this approach. On the other hand, the results in Table 2.3 indicate convergence to within 0.001 kcal/mol. Thus, the convergence of Δ_{p-c} may be even more than an order of magnitude faster than that of total interaction energies. One would expect this to be the case to some extent since the cage and prism structures are fairly close to each other which should lead to some cancellation of errors. Indeed, even if we assume that the uncertainty of our Δ_{p-c} is of the same magnitude as the quantity itself, thus would mean an order of magnitude reduction of error. Therefore, although for the CC-pol-optimized rigid-monomer structures Δ_{p-c} is so small that these two structures have to be assumed to be isoenergetic, the flexible-monomer structures [57] differ by 0.25 kcal/mol (Ref. 6) and therefore such difference can be meaningfully determined from calculations as accurate as presented here.

The results from Table 2.5 are shown in Fig. 2.1 as percentage contributions to the total hexamer interaction energy. Compared to similar graphs in literature, the two-body contribution is much more significant. This is due to the fact that, apparently for the first time, the many-body expansion has been computed with fairly saturated two-body contributions. This makes these contributions larger relatively to

the higher-body terms.

The percentage contributions of nonadditive effects are nearly identical for the cage and prism structures due to their geometrical similarity. For the other structures, these effects will be different, but not substantially different, see Sec. 2.5.

In conclusion, we have shown in this section that the simplest strategy of computing interaction energies of a cluster, i.e., the use of a single method and basis set, can be improved by using the hybrid approach, and then further improved using the many-body expansion. The main reason for the latter improvement is the fact that the two-body component converges much slower in basis set than the higher-body ones. The many-body approach allows one to compute the former component much more accurately than it is possible in the canonical supermolecular approach.

2.4 Comparison with literature

Our calculations presented in the previous section used larger basis sets than any published paper on water hexamers. Therefore, it would be in order to compare to the previous largest calculations, performed by BT [6]. Since BT use different hexamer geometries (optimized in Ref. 57), we had to repeat our calculation for the BT geometries. The two sets of geometries are different due to (a) the use of rigid monomers in our calculations and the optimization of monomers' internal coordinates in the BT geometry and (b) the different potential energy surfaces in the two optimizations. Since BT used the hybrid method, we have performed calculations analogous to those described Sec. 2.3.2. We first obtained CP-corrected vertical interaction energies of Eq. (2.11). The results are presented in Table 2.6. Similarly to the calculations in Sec. 2.3.2, we calculated the MP2 energies with the a5Z basis set and the CCSD(T) ones with the aQZ basis. Note that the $\delta E_{\text{int}}^{\text{CCSD(T)}}$ component is not large but its basis set dependence is relatively significant and the (TQ)-extrapolated values differ from the aTZ ones by about 0.2 kcal/mol. Performing a similar analysis as in Sec. 2.3.2, one can obtain an uncertainty estimate of 0.30 kcal/mol for the vertical interaction energy.

To compare to the relaxed interaction energies of BT [6], we now have to compute the distortion energy of Eq. (2.12). We used just the CCSD(T) approach and monomer-centered basis sets (MCBS). The resulting distortion energies as well as the relaxed interaction energies are given in Table 2.7. The MP2 values are also given for the distortion energies, but we have not used the hybrid approach here. One may note that indeed the MCBS approach is entirely sufficient for calculations of the distortion energies since the (Q5)-extrapolated quantities have uncertainties of only 0.02 (0.03) kcal/mol for the cage (prism) structure.

Now we can compare with the results of BT [6] quoted in Table 2.7. These authors computed the MP2 energies using the MP2-R12 approach [58, 59] without any CP corrections. BT estimated the uncertainty of their MP2 energies to be 0.1 kcal/mol from comparisons to Xantheas, Burnham, and Harrison (XBH) [60] MP2/CBS(Q5) results. The latter authors used slightly different hexamer geometries. From comparisons with our results at the a5Z level, this difference leads to about 0.04 kcal/mol increase of interaction energies. Our extrapolated results for the cage are higher by 0.05 kcal/mol than the CP-corrected XBH results (for the prism structure, only the uncorrected results were computed in Ref. 60). The overall 0.09 kcal/mol difference is probably due to the different extrapolation algorithm used by XBH. The BT MP2-R12 interaction energy of the cage hexamer differs by 0.09 and 0.04 kcal/mol from XBH and our results, respectively (the latter difference is 0.03 kcal/mol for the prism hexamer). Thus, the 0.1 kcal/mol estimate of the uncertainty of the BT MP2-R12 results is reasonable, although our method of estimating uncertainties gives a somewhat larger value of 0.26 kcal/mol. Furthermore, the 0.10 kcal/mol difference between CP and noCP (Q5)-extrapolated results of XBH indicates that the discussed uncertainties may be larger than 0.1 kcal/mol. One may notice that the CBS extrapolations from bases with large X are quite competitive in accuracy with MP2-R12 calculations. Similar observations were made recently [61] in the case of CCSD(R)-F12 approach compared to orbital extrapolations for the argon dimer.

Although BT have not explicitly estimated the accuracy of their $\delta E_{\text{int}}^{\text{CCSD(T)}}$ component, computed in the haTZ basis set [which uses the aTZ (TZ) basis for oxygens (hydrogens)] without CP corrections, they did compute this value for the prism structure with the CP correction which changed this component from -0.06 to -0.07 kcal/mol only. BSSE of only 0.01 kcal/mol [although our calculations in the same basis set gave BSSE of 0.05 (0.03) kcal/mol for the cage (prism) structures] might have indicated that the uncertainty of $\delta E_{\text{int}}^{\text{CCSD(T)}}$ is negligible (although in general the size of BSSE cannot be used to estimate accuracy of interaction energies [62]). Our calculations show that this is not the case and in contrast to the excellent agreement at the MP2 level, our $E_{\text{int}}^{\text{CCSD(T),rel}}/\text{CBS}$ values differ from the BT final results—defined as $E_{\text{int}}^{\text{MP2,rel}}(\text{R12}) + \delta E_{\text{int}}^{\text{CCSD(T),rel}}(\text{haTZ})$ —by 0.39 and 0.38 kcal/mol for the cage and prism structures, respectively. Clearly, the reason for this discrepancy is the relatively small basis set and the lack of the CP correction in BT’s calculations of $\delta E_{\text{int}}^{\text{CCSD(T)}}$ whereas this contribution fairly significantly depends on the quality of basis set, as shown above. Notice also that for the cage structure, BT obtained a wrong sign for $\delta E_{\text{int}}^{\text{CCSD(T)}}$. The very small difference of 0.01 kcal/mol between the CP and noCP values of this quantity found by BT for the prism structure is accidental since for the significantly larger aQZ basis set the BSSE errors of $\delta E_{\text{int}}^{\text{CCSD(T)}}$ are 0.32 and 0.35 kcal/mol for the cage and prism structures, respectively. These comparisons show that the CP correction can be critical even in calculations of small contributions to interaction energies of larger clusters. Furthermore, it is advisable to test the convergence of results by performing calculations in a series of basis sets.

Despite the fairly large errors in the $\delta E_{\text{int}}^{\text{CCSD(T)}}$ contributions, the difference $\Delta_{\text{p-c}}$ computed by us agrees to two significant digits with the value computed by BT. This is consistent with our finding that $\Delta_{\text{p-c}}$ converges an order of magnitude faster than the interaction energies. One may also note that the monomer distortion energies make a very small contribution to $\Delta_{\text{p-c}}$.

Another recent high-level calculations for the water hexamer were published by Kumar *et al.* [51] These authors computed interaction energies using the hybrid

approach at the level $E_{\text{int}}^{\text{CCSD(T)}(5,\text{T})}$, i.e., without CBS extrapolations and also without CP corrections. They used r_e rigid-monomer geometries optimized by them at the MP2/aTZ level. Their interaction energy for the prism structure lies 1.11 kcal/mol above our energy from Table 2.5, but after adding estimated corrections for BSSE and CBS extrapolations, the discrepancy increases to 1.9 kcal/mol. The reason for such a large discrepancy is the monomer geometry used by Kumar *et al.* which for the water dimer gives interaction energy 0.12 kcal/mol smaller in magnitude than that obtained with r_0 geometry [63], translating to 1.8 kcal/mol for the 15 dimers in the hexamer.

The values of $\Delta_{\text{p-c}}$ in the CC-pol, Kumar *et al.*, and BT geometries: -0.02, -0.14, and -0.25 kcal/mol, respectively, show that this quantity is relatively sensitive to the monomer flexibility. A DQMC calculation for water hexamer, such as performed by Liu *et al.* [28] and planned by us, could probably be done with flexible-monomer potentials, but there are no such potentials currently available which would be accurate enough. The existing flexible-monomer potentials for the water dimer [64, 65] are significantly less accurate in their intermolecular part than the CC-pol-8s potential [36]. There exists also a flexible-monomer nonadditive water trimer potential developed by Wang *et al.* [66]. These authors computed the interaction energy for the prism hexamer and obtained the value of -45.8 kcal/mol, 0.5 kcal/mol above our energy from Table 2.7. Such a difference is probably too large for reliable distinguishing between the interaction energies of the prism and cage structures. Thus, work aimed at this goal will have to use the significantly more accurate rigid-monomer potentials, which brings the question whether the effects due to monomer flexibility can be neglected. The difference between the prism and cage ZPVEs was 0.79 kcal/mol in the calculations of Ref. 28. With a difference of this order, the rigid-monomer predictions should be reliable. However, if this difference is less than 0.2 kcal/mol, as argued by BT [6], the rigid-monomer predictions may not provide clear-cut answers. It is possible that the rigid-monomer value of $\Delta_{\text{p-c}}$ would be closer to the flexible-monomer one if a monomer geometry similar to some averaged monomer geometry from the cage and prism structures of Ref. 57 were used instead of the vibrationally averaged geometry from CC-pol-8s+NB.

However, the use of the former geometry in DQMC calculations would probably not be a good idea since in the full-dimensional hexamer vibrational motion the faster intramonomer motions are approximately averaged as in CC-pol-8s+NB.

2.5 Convergence of many-body expansion for water hexamers

In Sec. 2.3.3, we have shown that if various computational strategies are applied to the water hexamer at the limits of computational resources, the many-body expansion approach provides the hexamer interaction energy with smallest uncertainties. In this section, we will show that several many-body contributions whose best available values are presented in Table 2.5 can be either neglected or calculated less expensively by using lower levels of theory than CCSD(T), smaller basis sets than used in Table 2.5, and neglecting a small part of the CP correction. The simplest approximation is just to neglect higher-body terms. The fast convergence of the many-body expansion is clearly seen in Fig. 2.1. Since the uncertainty of our best hexamer interaction energy is 0.12 kcal/mol, results in Table 2.5 show that the five- and six-body contributions can be safely neglected. In less accurate calculations, where uncertainties from other sources are of the order of 1 kcal/mol, it is possible to neglect also the four-body effects, contributing about 0.5 kcal/mol.

We will next consider the convergence with respect to the level of theory. The results are shown in Figs. 2.2 and 2.3 for the cage and prism structures, respectively. Plotted are the errors of the HF and MP2 K -body contributions relative to the CCSD(T)/aTZ/HCBS results at a given K -body level. This reference level is the highest we could afford when calculating the complete many-body expansion. We will discuss here results obtained in the HCBS approach. The K CBS approach, consisting in using the given K -mer basis set in calculating the K -body contribution, will be discussed below. Numerical data are given in the Supporting Information [24].

The figures show that the HF method, often used in calculations for larger clusters, is not adequate at the two-body level (and therefore would not be adequate in the canonical supermolecular approach) as it produces errors larger than 10 kcal/mol.

This shows that the two-body correlation effects cannot be neglected for water clusters. The HF method can be used for the three-body terms if the accuracy goal is 1 kcal/mol. At this level of accuracy, the higher-body effects would be neglected. In fact, for the three-body terms the HF method performs relatively best as it is almost as accurate as the MP2 method. Furthermore, the dependence of this term on basis set size is weak and the aDZ basis is completely adequate. If we tighten the accuracy threshold to 0.1 kcal/mol and four-body effects have to be included, these effects can possibly be computed at the HF level, although the errors are just barely below the threshold. For the four-body effects, as well as for the five-body ones, the errors of the HF level are about three times larger than those of the MP2 level. Since, however, the error of HF/aDZ at the five-body level is only about 0.01 kcal/mol, this level is adequate here for any conceivable current calculations. One may note that for the cage five-body contribution the relative errors of both the HF and MP2 values are so large that its inclusion actually increases the overall error. This fact is clearly related to the accidental smallness of this term and the same problem does not appear for the prism structure.

Somewhat surprisingly, except for the two-body term, the MP2 method does not bring huge improvements in the many-body expansion over the HF approach despite the fact that it reproduces the total hexamer interaction energy with errors below 1%, cf. Table 2.6. This rather high accuracy is partly fortuitous and partly due to the good accuracy of the MP2 two-body results. Water dimer MP2 interaction energies have typically a couple percent error compared to the CCSD(T) results, therefore 0.5 (0.75) kcal/mol or 1.5% (2%) error for the cage (prism) hexamer two-body interaction energy is not accidental. The still better total interaction energy is mainly due to the fact that the error of the three-body term is of similar magnitude and of the opposite sign relative to the two-body error. Note, however, that in relative terms the three-body error is more substantial, as it constitutes about 5% of the three-body energy. In general, the sign of the MP2 (as well as HF) error alternates with K , leading to the total interaction energies more accurate than individual K -body terms. In any case,

the MP2 level of theory would be adequate for all terms in the many-body expansion of hexamer’s interaction energy with the uncertainty goal of 1 kcal/mol. If the threshold is lowered to 0.5 kcal/mol, MP2 cannot be used anymore at the two-body level. It is also not very useful for higher K since at the three-body level it is not much better than HF, at the four-body level it is better but HF is good enough, and higher-body levels are negligible. With 0.1 kcal/mol threshold, MP2 would be a good choice at the four-body level as it reduces the errors by about a factor of two compared to the HF approach, to about 0.05 kcal/mol. For five-body contributions, the HF level is adequate for all practical purposes, but the use of MP2 does give about a factor of three reduction of this error. The MP2 results are much more dependent on basis set size than the HF results and the use of non-augmented bases in MP2 calculations is clearly inadequate. However, the relatively small aDZ basis set is sufficient for MP2 calculations except for the two-body contributions.

The basis set convergence of the individual K -body contributions at the CCSD(T) level of theory is shown in Figs. 2.4 and 2.5 for the cage and prism structures, respectively. Note that in contrast to Figs. 2.2 and 2.3, the reference energies are now the benchmark results from Table 2.5. We have seen, cf. Tables 2.1 and 2.2, that the basis set convergence for the total interaction energy is slow. We can now see that this error is almost exclusively the result of the slow basis set convergence of the two-body contributions. For example, the aDZ two-body contributions have about 7 kcal/mol, nearly 20% error relative to the benchmark two-body CCSD(T) value. On the other hand, the errors of the three- and higher-body contributions in the aDZ basis set are below 0.02 kcal/mol, i.e., are completely negligible. For the two-body contributions, we have also included the results in the aXZ/DC⁺BS basis sets, $X = T, Q, 5$, showing the greatly reduced errors both due to the use of midbond functions and the increased cardinal number X . In particular, the addition of bond functions to the aTZ basis reduces the error by a factor of two upon only 24% increase of the basis set size. The non-augmented bases, even TZ, are completely inadequate at the two-body level, but give only about 1 kcal/mol error at the three-body level, so in particular the DZ basis

would be a reasonable choice at this level in less accurate calculations. This basis is also adequate in calculations of four- and five-body contributions, although the application of the TZ basis reduces the errors several times.

Figures 2.2–2.5 show both the results computed using the full hexamer-centered basis sets in each calculations of total energies, from monomer to hexamer, and using the K -mer centered basis sets (K CBS) for a given K -mer (for this K -mer and all its subclusters). To make sure that the latter approach is properly understood, let us give an example. In calculating the three-body nonadditive energy for the trimer consisting of monomers 2, 4, and 6, the calculations of monomer, dimer, and trimer total energies are all performed in the same 2-4-6 trimer-centered basis set. However, in the calculations for the tetramer 2-3-4-6, the energy of the trimer 2-4-6 is computed in the tetramer-centered basis set. Thus, the CP method is rigorously applied at each K -body level, but not for the whole hexamer. The part of the CP correction missed in this way is negligibly small, as shown in Figs. 2.2–2.5. The additional error is a substantial fraction of the HCBS error only for the two-body contribution computed in the smallest non-augmented basis sets which are anyway inadequate for any purposes. With the aDZ basis set, the additional error is about 1 kcal/mol compared to the HCBS error of 7 kcal/mol. The corresponding errors in the aTZ basis set are 0.3 and 3 kcal/mol, a still smaller relative difference. This decrease should be expected since as the basis set approaches completeness, the difference between the HCBS and K CBS approaches has to disappear. For the higher than two-body contributions and augmented basis sets, differences between the HCBS and K CBS approaches are completely negligible and relative changes are smaller than in the two-body case (for example, for the three-body contribution to the prism interaction energy, the difference between the two approaches is only 0.006 kcal/mol in the aTZ basis set). This is due to the fact that the difference between the sizes of basis sets in the HCBS and K CBS approaches becomes smaller with larger K . The neglected part of the CP correction is still smaller when the CBS extrapolations are applied. The HCBS/ K CBS differences are somewhat larger in relative terms for the three- and four-body components computed in non-augmented

basis sets, so some care should be taken if these bases are applied (which is generally not recommended).

The optimal strategies for several assumed levels of accuracy are summarized in Table 2.8. All the calculations should be performed in K -mer-centered basis sets and midbond functions should be used for the two-body contributions. At any accuracy level, CBS extrapolations are absolutely necessary for the two-body contributions and will likely be beneficial for the three-body ones. In general, we do not recommend the (DT) extrapolations, but for all higher X such extrapolations significantly improve accuracy at low extra computational costs. For the four-body and higher contributions, bases larger than aTZ are not needed, so extrapolations should not be performed.

In addition to the thresholds discussed above, we have included in Table 2.8 the threshold of 0.6 kcal/mol, equal to the uncertainty of the best value of the interaction energy computed using the approach of Sec. 2.3.1 at the CCSD(T)/(TQ) level and we will now discuss this case. For the two-body contribution, one has to still use the CCSD(T)/(TQ) approach but now with midbond functions. According to Table 2.3, this procedure should be accurate to about 0.3 kcal/mol (the increase of accuracy is due to the use of midbond functions). Since the MP2 method, which was adequate with the 1 kcal/mol threshold, gives a nearly 0.4 kcal/mol error for the three-body term, we have to move to the CCSD(T) approach for this term. However, the aDZ basis will still be sufficiently accurate here. We now have to include the four-body contribution, which could actually be computed even at the HF/TZ level. However, to be on the safe side, we recommend MP2/aDZ.

The strategies discussed above were based only on the results for the cage and prism structures. However, other hexamer structures may exhibit a slower convergence of the many-body expansion. This may be indicated by the recent results of Kumar *et al.* [51]. These authors performed calculations for four hexamer structures without any CP corrections using the hybrid MP2/a5Z plus CCSD(T)/aTZ approach and computed also the two- and three-body interaction energies at the same level (using

KCBS strategy and still no CP corrections). For the ring hexamer, the difference between the total interaction energy and the sum of two- and three-body contributions is -1.41 kcal/mol, somewhat larger than the similar quantity for the cage and prism hexamers. To examine this issue, we have performed calculations for the bag, book, boat, and ring structures at the geometries of Ref. 57. The resulting many-body expansions at the CCSD(T)/aDZ level are presented in the Supporting Information [24]. For the book and ring structures, the CCSD(T)/aTZ results are also given. Despite somewhat larger many-body contributions for the bag and book structures, the accuracy of the methods based on the cage and prism ones is also sufficient for the former cases. However, for the boat and ring structures, a non-negligible (at the 0.1 kcal/mol accuracy level) five-body contribution should be included using MP2/aDZ. Moreover, a slightly stronger three-body basis set dependence for the ring geometry suggests that CCSD(T)/aTZ should be used instead of CCSD(T)/aDZ for the 0.1 kcal/mol accuracy level. These modifications have been incorporated into Table 2.8.

We have also included in Table 2.8 a hypothetical 0.02 kcal/mol accuracy threshold. Calculations of such accuracy would be very difficult at the present time. At the MP2 level for the two-body term, one would probably have to use basis sets with the cardinal number $X = 8$ or more (which would have to be optimized) or the MP2-R12 approach in a very large basis set. Moreover, at this accuracy level, CCSD(T) would not be adequate anymore and the higher excitations would have to be included, for example using the CC method with complete triple and noniterative quadruple excitations [the CCSDT(Q) approach], as well as contributions from core electrons and the relativistic effects. On the other hand, our current calculations would be already sufficient for higher than two-body contributions. One needs to use the CCSD(T)/aQZ level for the three-body contributions, CCSD(T)/aDZ for the four-body ones, and MP2/aDZ for the five-body ones. The six-body effects could still be neglected.

The proper choice of strategy has to take into account computational requirements of various terms of the many-body expansion. We will discuss these requirements in Sec. 2.6 and then we will show in Sec. 2.7 how one can perform calculations at any

required level of accuracy using significantly less computational resources than in the canonical supermolecular approach.

2.6 Computer timings

Table 2.9 shows timing comparisons for the many-body contributions to the water hexamer interaction energy using the CCSD(T) method, the aDZ, aTZ, and aQZ basis sets, and both the HCBS and *K*CBS approaches. Let us first discuss the HCBS approach. Perhaps surprisingly, the time to calculate just the two-body contribution in the aDZ basis is significantly (2.7 times) longer than the time of the canonical supermolecular calculations of the hexamer interaction energy (given in line "hexamer + monomers"). This factor is reduced to 1.3 for the aTZ basis, but in any case the straightforward many-body approach looks like an impractical method. The timings can be well rationalized based on the scaling of the CCSD(T) method with the number of occupied and virtual orbitals (o and v , respectively). Since the hexamer-centered basis set is used in the discussed calculations, the only savings in K -body calculations vs. full hexamer calculations originate from the smaller number of occupied orbitals in the former case. Since the CCSD(T) leading term scales as $O(o^3v^4)$, this step is approximately 27 times faster for a dimer than for the hexamer. However, since there are 15 dimers, calculations for all dimers should be about two times shorter than for the hexamer. This is not the case for the timings shown in Table 2.9 due to other terms in the CCSD(T) calculations, scaling as lower powers of o . For example, the time of the integral calculation is the same in all cases and the calculation is performed repeatedly for all dimers. As K increases, the time of the calculation of a given K -body contribution initially increases (for example, for $K = 3$ the time is $1148 \cdot 327 = 821$ hours in the aTZ basis). This is due to the fact that a single calculation for a trimer is longer than for a dimer and in addition the number of trimers is larger than the number of dimers. The maximum is reached at $K = 4$ and then timings decrease due to the smaller number of K -mers. Similar relations will take place for larger clusters. For example, for $(\text{H}_2\text{O})_{24}$, the number of occupied orbitals in a dimer is 12 times smaller than for

the whole 24-mer, but there are 276 dimers. The expected savings for the leading CCSD(T) term, the speedup of $12^3/276 = 6$ times, will be partially canceled by terms scaling as a lower power of o than o^3 . Thus, for $(\text{H}_2\text{O})_{24}$ the many-body approach in the full-cluster basis set might be economical at the two-body level. However, as shown in Table 2.9 for the hexamer, the calculations of three-body contributions are a few times longer than the two-body ones, so the extension to the three-body level would not be practical.

Of course, the picture is gloomy only in the brute-force many-body approach described in the previous paragraph. As we have shown earlier, the use of the CCSD(T) level of theory and of a single basis set is a significant overkill for higher-body terms. Also, as we have shown, there is no need to use HCBS. The timings for the KCBS approach shown in Table 2.9 are about two orders of magnitude shorter than the HCBS times at the two-body level. In particular, the aQZ two-body calculations are nearly two orders of magnitude faster than the canonical supermolecular calculation in this basis set. Also the calculations of three-body contributions are less time consuming than the canonical supermolecular calculations. The obvious reason for these savings is that the dimer- and trimer-centered basis sets are three times and two times smaller, respectively, than the full hexamer basis set. However, for larger K the savings diminish and at $K = 5$ level the KCBS calculations in the aTZ basis set are actually more time consuming than the HCBS calculations. Moreover, whereas the total time needed to calculate all the many-body contributions up to a given K is equal to only the K -contribution time in the HCBS case (since all the lower K contributions are byproducts of the K -mer calculations), in the KCBS approach the total time is the sum of all times from the calculations for clusters smaller and equal to K . As a consequence, a calculation of the complete many-body expansion in the aTZ/KCBS approach would be almost nine thousand hours, 35 times longer than the canonical supermolecular calculation. The reason for the excessive high- K timings in the KCBS approach is the fact that the number of calculations is much larger now than in the HCBS approach

due to multiple versions of K -mer basis sets in the calculations for subclusters of a K -mer. For instance, for the four-body contribution, there are 20 non-equivalent trimers in the hexamer-centered basis sets but $15 \times 20 = 300$ trimers in the tetramer-centered basis set, because there are 15 non-equivalent tetramer-centered basis sets. This means that the 2-4-6 trimer calculations have to be performed not just once for all values of K as it is the case in the HCBS approach, but 15 times at the $K = 4$ level only. Clearly, the KCBS approach cannot be used for larger K (although one can devise strategies mitigating the discussed problem by restricting the number of basis sets). However, it leads to significant savings of computer time for smaller K and, in the next section, we will propose strategies for using many-body expansions at significant saving of computer time compared to the canonical supermolecular approach.

2.7 Applications of effective many-body strategies

There are several ways of applying effective many-body expansions, in particular one may want (a) to obtain results of a given accuracy with a minimum utilization of computational resources; (b) to recover results of the canonical supermolecular approach as closely as possible but with greatly reduced resources; (c) to obtain the best possible result at the maximum of available resources. In the case (a), one should basically follow the recommendation of Table 2.8 and several examples of such calculations have already been discussed. This strategy would not be working well in case (b), mainly due to the use of midbond functions. We will demonstrate case (b) strategy in recovering the results of the canonical supermolecular approach in its straightforward version from Sec. 2.3.1. In this approach, the best values could be obtained at the CCSD(T)/(TQ) level, cf. Table 2.1. We have estimated the accuracy of these results to be 0.6 kcal/mol. To recover this result as closely as possible, we have made the following changes relative to Table 2.8: we have not used the midbond functions in the two-body calculations (since these improve this term too much), increased the basis set for the three-body term to aTZ, increased the theory level for the four-body term to CCSD(T), and computed the five-body term at the MP2/aDZ level. The results are

shown in Table 2.10. The time of the calculations is still very short, 24 times shorter than using the method of Sec. 2.3.1, whereas the results of the latter calculations are reproduced to within 0.01 kcal/mol, well below the estimated uncertainty of either result. Whereas the extreme smallness of this difference is due to fortuitous cancellations of the differences in the many-body terms, if one estimates the differences of all the individual K -body contributions and sums their absolute values, the resulting discrepancy of about 0.1 kcal/mol is still very small. This is due to the fact that in both calculations the $K > 3$ contributions are converged to below 0.01 kcal/mol. The difference between the aTZ and (TQ) three-body contributions can be found in Table 2.4 and amounts to 0.04 kcal/mol (recall that the TCBS results differ negligibly from HCBS ones for three-body terms). Finally, in the two-body term, the only difference comes from the use of DCBS vs. HCBS bases. We know this difference only for the aTZ basis where it amounts to 0.26 kcal/mol. Since, as discussed earlier, such differences diminish faster than the basis set incompleteness errors, one may expect that at the (TQ) level the discrepancy will be below 0.05 kcal/mol, leading to the total estimate given above. Note that this estimate is for the discrepancy between the values obtained in the canonical supermolecular and many-body approaches. The uncertainty of the latter result with respect to the exact interaction energy is still the same as of the former, i.e., 0.6 kcal/mol.

The computer time savings given in Table 2.10 can be increased to a factor of 52 with essentially the same accuracy if one uses CCSD(T)/aDZ in the three-body calculations instead of CCSD(T)/aTZ. Alternatively, one can significantly improve the total energy by calculating the two-body term at the ((Q5),(QT)) level with bond functions. The results, shown in the last row of Table 2.10, are within 0.05 kcal/mol of our best two-body energies and the time of calculation is still 15 times shorter.

As an illustration of option (c), one can compare the timings for the most accurate calculations of the hexamer interaction energies that could currently be made, i.e., those of Sec. 2.3.3 presented in Table 2.5, which took 14887 CPU-hours (with 70% of the time spent on the three- and higher-body contributions calculated with basis

sets much larger than needed at this level of accuracy). This time is close to the 10212 CPU hours of the canonical supermolecular CCSD(T)/(TQ) calculations of Sec. 2.3.1, whereas the uncertainty of the former calculation is five times smaller. Clearly, reaching the 0.1 kcal/mol accuracy in the latter approach, i.e., the extension of the calculations to the a6Z basis set, would require vastly longer CPU time, especially since we observed a worse than theoretical scaling of the two-body term when going from the a5Z to a6Z basis set, most likely due to an increased diffuseness of the basis set.

Bates *et al.* [67] have recently applied a variant of “hybrid fragmentation” approaches to water clusters by computing total cluster interaction energies at a low level of theory and basis set and combining these results with high-level calculations of two- and three-body effects

$$E_{\text{int}}^{\text{hybrid}} = E_{\text{int}}^{\text{low}} - \sum_{K=2}^3 E_{\text{int}}[K, N]^{\text{low}} + \sum_{K=2}^3 E_{\text{int}}[K, N]^{\text{high}}. \quad (2.14)$$

This result can be viewed as a replacement of the less accurate two- and three-body effects included in $E_{\text{int}}^{\text{low}}$ by the more accurate ones computed at a higher level. The elimination of two- and three-body effects between the first and second terms in Eq. (2.14) is exact if CP method is not used, as in the work of Bates *et al.*, or if NCBS is used in calculations of these two terms. The first option is not adequate at accuracy levels we aim for and the second one is too time consuming. On the other hand, if NCBS approach is used for the first term and KCBS for the second term, the residual BSSE errors in the two-body contributions computed in small basis sets are unacceptably large, cf. Figs. 2.4 and 2.5.

One should also mention that apart from computer time savings, the many-body approach brings also substantial savings in memory and disk space requirements. In fact, due to the memory requirements, we were not able to use MOLPRO for the full hexamer aQZ calculations. We used PQS instead which is somewhat slower but utilizes a much more memory-efficient parallelization strategy.

2.8 Large water clusters

In Refs. 5, 7, Apra *et al.* obtained interaction energies for very large water clusters, up to $(\text{H}_2\text{O})_{24}$. The calculations were performed using the canonical supermolecular approach at the CCSD(T) theory level with a modified cc-pVTZ basis set (cc-pVTZ with f -functions removed) which we will denote as modTZ. The CP correction was not applied. The ability to obtain such results shows enormous improvements in both computer power and parallel algorithm efficiency achieved in recent years since the calculations were made on a peta-FLOP (floating-point operations), 223200-core machine and took 76 years of total CPU time per one $(\text{H}_2\text{O})_{24}$ calculation. These are probably the largest CCSD(T) calculations to date. Two structures with virtually identical interaction energies (labeled 308 and 316) were identified, with the structure 316 only 0.01 kcal/mol more stable than 308. Although the results have been the best available so far and are a significant step in our understanding of water clusters, the relatively small modTZ basis set which does not include any diffuse functions and the neglect of the CP correction make the conclusions of Refs. 5, 7 uncertain.

Large clusters like $(\text{H}_2\text{O})_{24}$ are the subject of recent interest (see, e.g., Ref. 68) and are an ideal case for applications of the many-body expansion. The leading two-body terms can be calculated very accurately with moderate CPU requirements using the CCSD(T) method and the CBS extrapolations. It was shown in Sec. 2.5 that the basis set incompleteness has the largest impact on the two-body energies and saturating the results with respect to the basis set brings significant improvements in the accuracy of the interaction energy of the whole cluster. We have applied the many-body expansion technique to $(\text{H}_2\text{O})_{24}$ and the values of $E_{\text{int}}[K, 24]$ contributions, $K = 2, 3, 4$, are shown in Table 2.11. All calculations were performed using the KCBS approach. We used clusters geometries from Ref. 7, 40 (listed in the Supporting Information [24]).

Similarly to Sec. 2.4, we first calculated the vertical interaction energies. As anticipated, there is indeed a significant basis set dependence of the two-body terms. The CCSD(T)/aQZ contributions are about 33 (37) kcal/mol larger in magnitude

than the TZ (modTZ) ones. The differences increase to almost 40 (44) kcal/mol for the CCSD(T)/(TQ) results (extrapolated using Eq. (2.8)). With the same estimation method as used earlier, we find the uncertainty of the two-body result to be 3.1 kcal/mol. For the three-body contribution, the basis set dependence is weaker, as it was the case for the water hexamer. In the latter case, the aDZ results were more accurate than the TZ ones and the CCSD(T)/aDZ error was about 0.5% relative to our best prediction. Therefore, the present CCSD(T)/aDZ result should have an uncertainty of about 0.3 kcal/mol. The error of the MP2/aDZ four-body contribution for the water hexamer was about 12%, which implies an uncertainty of 0.8 kcal/mol for $(\text{H}_2\text{O})_{24}$. The five- and six-body contributions constitute 0.1% of the total interaction energy of the hexamer, which gives 0.3 kcal/mol as an estimate of these effects in $(\text{H}_2\text{O})_{24}$. However, as it will be discussed below, the convergence of the many-body expansion is slower for the icosikaitetramer and the higher than four-body contributions may be more important here. Therefore, we will use 0.6 kcal/mol as the estimate. This leads to an estimate of the uncertainty of the total interaction energy of $(\text{H}_2\text{O})_{24}$ equal to 4.8 kcal/mol or about 2%. The accuracy of our result could be further improved with relatively minor extra costs by computing the two-body contributions at the MP2/a5Z level and using the hybrid approach, as well as including the midbond functions in the two-body calculations. CCSD(T)/aDZ four-body calculations instead of MP2 ones are also possible and would reduce significantly the four-body error.

The convergence of the many-body expansion is illustrated in Fig. 2.6. This convergence is very similar to that seen for the water hexamer, where the two-body terms constitute about 82%, three-body 17%, and four-body almost 1% of the total interaction energy. The corresponding values for the water icosikaitetramer in Fig. 2.6 are about 74%, 23%, and 3%. Thus, we observe a somewhat slower convergence in the latter case with higher than two-body terms constituting a larger fraction of the total interaction energy, 26% compared to 18% in the former case. Since the uncertainties of the many-body contributions estimated above are about 1%-2% of the total interaction energy, this trend is well established by our calculations.

To compare our results with those of Ref. 7, we have calculated the monomer distortion energies at the CCSD(T)/(TQ)/MCBS level. The equilibrium geometry of the water monomer was obtained by us from an optimization at the MP2/modTZ level (the geometry is given in the Supporting Information [24]). The distortion energies were added to the best vertical energies and the resulting relaxed energies are given in Table 2.12. These results show that structure 316 has a larger magnitude of the stabilization energy than structure 308, similarly to the findings of Ref. 7. In view of the huge discrepancies in the values of the interaction energies between our and Apra *et al.*'s calculations, our 0.08 kcal/mol differences between the energies of structures 308 and 316 is amazingly close to the value of 0.01 kcal/mol obtained in Ref. 7. Interestingly, the vertical interaction energy of structure 308 is 0.15 kcal/mol larger in magnitude than that of structure 316. The differences of this order are well below the absolute accuracy of our results. Although the water hexamer results show that such differences converge about an order of magnitude faster than the total interaction energies, this still gives an estimated uncertainty of about 0.5 kcal/mol, much larger than the observed energetic differences. Thus, the magnitude and the sign of the difference cannot be considered to be reliably established and we consider the two structures to be isoenergetic.

Our recommended interaction energies are about 37 kcal/mol smaller in magnitude than the values obtained by Apra *et al.* in Ref. 7. This difference is well above the estimated accuracy of our result of 4.8 kcal/mol. The major reason for the discrepancy is the smallness of the basis set and the lack of the CP correction in Ref. 7. In fact, the 37 kcal/mol difference results from a cancellation between the basis set incompleteness and superposition errors. To estimate BSSE of the results of Ref. 7, we calculated two-body CCSD(T) energies in the modTZ basis set with and without the CP corrections. These results, included in Table 2.11, show that BSSE is -84.494 and -85.192 kcal/mol for structures 316 and 308, respectively. Since the respective basis set incompleteness errors other than BSSE are 43.706 and 44.358 kcal/mol relative to our CCSD(T)/(TQ) results, the two errors partly cancel giving the observed -40.788 and -42.834 difference

between modTZ/noCP and (TQ) results. Incidentally, the CP-corrected and uncorrected modTZ results almost evenly bracket our best estimates of the two-body energy. The latter values agree very well with the -37.31 and -37.38 kcal/mol differences between the results from Ref. 7 and our best relaxed total interaction energies. The agreement is very good indeed taking into account that the estimated quantities are two-body only and the actual differences are for the total relaxed energies. Thus, our calculations provide about an order of magnitude more accurate interaction energies than the calculations of Ref. 7. Despite this improvement in accuracy, the total CPU time needed for our calculations was less than 5.5 months, or 200 times less than the calculations of Refs. 5, 7.

As an additional test, we have applied the many-body expansion approach to the $(\text{H}_2\text{O})_{16}$ cluster. For this system, significantly more accurate literature results than for $(\text{H}_2\text{O})_{24}$ are available [69]. These results were obtained using the canonical supermolecular approach at the CCSD(T)/aTZ level, albeit without any CP-correction. The energy ordering of several $(\text{H}_2\text{O})_{16}$ structures changes between MP2 and CCSD(T) methods, showing the importance of using the CCSD(T) level of theory for such clusters. The $(\text{H}_2\text{O})_{16}$ cluster has also been recently studied [67] with many-body expansions including two- and three-body energies calculated using the CCSD(T)/aTZ method and subsequently used as a correction to the MP2/aTZ interaction energies of the whole cluster. The results of our $(\text{H}_2\text{O})_{16}$ calculations, performed analogously as for the $(\text{H}_2\text{O})_{24}$ cluster, are presented in the Supplementary Material [24]. Our interaction energies should be significantly more accurate than those of Refs. 67, 69 due to our use at the two-body level of larger basis sets, CBS extrapolations, and CP corrections. Our CCSD(T) interaction energies for the four isomers differ from those of Ref. 69 by between 7.9 and 8.2 kcal/mol, whereas the estimated uncertainty of our results is about 2.8 kcal/mol. As in the case of 24-mer, this about 8 kcal/mol discrepancy results from cancellations of the larger in magnitude BSSE errors with the basis set incompleteness errors. Despite the fairly significant discrepancies in total interaction energies, we have obtained a very similar energetic ordering of the isomers as given by the canonical

supermolecular calculations of Ref. 69, later reproduced in Ref. 67. At the CCSD(T) level, the lowest isomer is 4444-a in all calculations. The relative energetic positions of the isomers boat-b and 4444-b above 4444-a computed by us agree to within 0.02 kcal/mol with the values from Ref. 69. The discrepancy is somewhat larger for the antiboat isomer, amounting to 0.28 kcal/mol, and this isomer the second lowest in our calculations rather than the third lowest in Ref. 69. Again, our more accurate results were obtained in significantly less CPU time, about 53 days, than the 47 years per isomer spent in calculations of Ref. 69. Our timings are slightly better than those of Ref. 67 (94 days per isomer). The calculation of the post-three-body effects took 6.3 CPU days in Ref. 67 whereas our calculation of four-body effects took 16 CPU days.

Since the number of subclusters in a 16-mer or 24-mer is very large, one may ask if the numerical errors in calculations of the total energies of subclusters do not significantly accumulate in calculations of many-body contributions. As discussed earlier, this should not be the case due to the very tight convergence thresholds used by us. With these thresholds, the individual total energies of subclusters appearing in Eq. (2.5) are accurate to at least 10^{-7} kcal/mol so that the possible accumulations should not show up on digits presented in the tables. We have performed a numerical experiment by increasing all the thresholds given in Sec. II by one order of magnitude. The quantity most likely to be affected by this change, the four-body HF/aDZ contribution for structure 308 in Table 2.11, which is obtained by summing over 10626 tetramers, has not changed on the digits listed in the table.

2.9 Conclusions

We have shown that one can calculate interaction energies of molecular clusters from the many-body expansion with significant computational savings compared to the canonical supermolecular approach while providing the same accuracy. Equivalently, our stratified approximation many-body approach (SAMBA) can be used for obtaining much more accurate results with similar computational costs to the canonical supermolecular approach. The savings are achieved by neglecting higher-body terms in the

many-body expansion and by calculating higher than two-body terms with a progressively lower level of theory and smaller basis sets. Significant savings also come from calculating the total energies of each K -mer and its subclusters in the basis set of this K -mer only. Although this approach substantially increases the number of calculations, a small size of each calculation leads to large computational savings, especially for methods that scale as high powers of the system size, such as CCSD(T). Therefore, such an approach is preferable over using the basis set of the whole cluster. This approach includes a small part of BSSE, but the errors committed in this way are always much smaller than the other errors resulting from the basis set incompleteness.

We presented a detailed error analysis and timings for two water hexamer structures, cage and prism, and checked it on four other hexamer structures. From the size and timings of the many-body contributions, one can derive a suitable strategy for a target accuracy goal for E_{int} . The two-body contributions are the most important. One has to calculate these terms with as good methods and basis sets as possible. It is advisable (and possible) to use the CCSD(T) method and basis sets as large as aQZ or larger including bond functions, perform CBS extrapolations, and use the hybrid approach. Three-body terms are also fairly important and should be calculated preferably with the CCSD(T) method. However, the basis sets need not be as large as in the two-body case, and the CBS extrapolations are also optional. In fact, the aDZ basis set seems to be sufficient for the three-body calculations in most cases. Four-body terms are an order of magnitude less important than the three-body ones and can be neglected if the target uncertainty is about 2% or more. However, a reasonable representation of these terms is affordable using small basis sets and the HF or MP2 levels of theory. Five-body terms can be either neglected or calculated at the HF/MP2 level only. Six-body terms can always be safely neglected. We have also analyzed the possibility of computing high- K contributions applying a hybrid canonical and many-body approach as recently implemented by Bates *et al.* [67], but it turns out that it cannot be adopted at accuracy levels that we aim at.

The detailed strategy presented above should be valid without changes for any

waters clusters as the three-body nonadditive effects contribute about 20% to the interaction energy for clusters from the trimer [25] to liquid water [39] and the four-body effects contribute only a couple percent. This strategy is robust enough to accommodate variations in the percentage contributions of three-body effects of at least a factor of two, so that it should work also for most other polar systems since for all the systems investigated to date the many-body expansion converges fast. For nonpolar systems, the many-body effects are less important and the convergence of the many-body expansion should be even faster. On the other hand, correlation effects are much more important in this case and a good description of the electron correlation by using the CCSD(T) method whenever possible is important. In some cases, the application of the MP2 method can lead to qualitative errors for the three-body terms [70]. Therefore, one can probably neglect the four- and higher-order terms but the two- and three-body ones must be calculated with the CCSD(T) approach for nonpolar systems.

We have also applied the many-body expansion strategy to $(\text{H}_2\text{O})_{16}$ and $(\text{H}_2\text{O})_{24}$. We were able to reduce the uncertainties of the previous best calculations of the interaction energies for the latter system by an order of magnitude using two orders of magnitude smaller computational resources.

Literature calculations for the water hexamer and larger clusters rarely remove BSSE. We have shown that the uncertainties of interaction energies computed without any CP corrections are dominated by BSSE. In contrast, SAMBA interaction energies are virtually free of BSSE as this error is always significantly smaller than the basis set incompleteness error provided that the two-body energies are CBS extrapolated.

Our method can be extended to still larger clusters by utilizing the asymptotic expansion of interaction energies. This expansion, with van der Waals constants computed *ab initio*, gives very accurate values of interaction energies at large intermonomer separations, as demonstrated in numerous SAPT applications [16]. Thus, two-body energies can be calculated at the CCSD(T) level only for monomers with separation smaller than about 5 Å whereas the remaining ones can be obtained from an asymptotic expansion with negligible computational resources. Similar procedures

can be applied to three-body terms. Such an approach would be somewhat similar to that proposed by Beran [71].

2.10 Acknowledgments

UG and RP acknowledge the support by the Polish Ministry of Science and Higher Education grant No. N N204 123337. This work was partly supported by the NSF Grant CHE-0848589. We thank Dr. Sotiris Xantheas for sending us the geometries of the water icosikaitetramers.

Bibliography

- [1] C. J. Burnham, J. Li, S. S. Xantheas, and M. Leslie, *J. Chem. Phys.* **110**, 4566 (1999).
- [2] K. S. Kim, P. Tarakeshwar, and J. Y. Lee, *Chem. Rev.* **100**, 4145 (2000).
- [3] K. Szalewicz, R. Bukowski, and B. Jeziorski, in *Water in Confining Geometries*, edited by V. Buch and J. P. Devlin (Springer-Verlag, Berlin, 2003) pp. 7–23.
- [4] J. K. Kazimirski and V. Buch, *J. Phys. Chem. A* **107**, 9762 (2003).
- [5] E. Apra, R. J. Harrison, W. A. deJong, A. P. Rendell, V. Tipparaju, and S. S. Xantheas, in *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis* (ACM, New York, 2010) article number 66.
- [6] D. M. Bates and G. S. Tschumper, *J. Phys. Chem. A* **113**, 3555 (2009).
- [7] E. Apra, R. J. Harrison, W. A. deJong, A. P. Rendell, V. Tipparaju, and S. S. Xantheas, Extended version of Ref. 5.
- [8] G. Chalasinski and M. M. Szczesniak, *Chem. Rev.* **94**, 1723 (1994).
- [9] K. Szalewicz, R. Bukowski, and B. Jeziorski, in *Theory and Applications of Computational Chemistry: The First 40 Years. A Volume of Technical and Historical Perspectives*, edited by C. E. Dykstra, G. Frenking, K. S. Kim, and G. E. Scuseria (Elsevier, Amsterdam, 2005) Chap. 33, pp. 919–962.
- [10] B. M. Axilrod and E. Teller, *J. Chem. Phys.* **11**, 299 (1943).
- [11] Y. Muto, *Proc. Phys. Soc. Jpn.* **17**, 629 (1943).
- [12] D. Hankins, J. W. Moskowitz, and F. H. Stillinger, *J. Chem. Phys.* **53**, 4544 (1970).
- [13] S. S. Xantheas, *J. Chem. Phys.* **100**, 7523 (1994).

- [14] S. S. Xantheas, *Chem. Phys.* **258**, 225 (2000).
- [15] B. Jeziorski, R. Moszyński, and K. Szalewicz, *Chem. Rev.* **94**, 1887 (1994).
- [16] K. Szalewicz, *Wiley Interdisciplinary Reviews—Computational Molecular Science* (2011), in press.
- [17] V. F. Lotrich and K. Szalewicz, *J. Chem. Phys.* **106**, 9668 (1997).
- [18] V. F. Lotrich and K. Szalewicz, *J. Chem. Phys.* **112**, 112 (2000).
- [19] R. Podeszwa and K. Szalewicz, *J. Chem. Phys.* **126**, 194101 (2007).
- [20] W. Klopper, M. Quack, and M. A. Suhm, *Mol. Phys.* **94**, 105 (1998).
- [21] K. Szalewicz, C. Leforestier, and A. van der Avoird, *Chem. Phys. Lett.* **482**, 1 (2009).
- [22] E. E. Dahlke and D. G. Truhlar, *J. Chem. Theory Comput.* **3**, 46 (2007).
- [23] R. A. Christie and K. D. Jordan, *Struct. Bond.* **116**, 27 (2005).
- [24] See EPAPS Document No. ?????. For more information on EPAPS, see <http://www.aip.org/pubservs/epaps.html>.
- [25] E. M. Mas, R. Bukowski, and K. Szalewicz, *J. Chem. Phys.* **118**, 4386 (2003).
- [26] B. J. Mhin, J. Kim, S. Lee, Y. Lee, and K. S. Kim, *J. Chem. Phys.* **100**, 4484 (1994).
- [27] K. Kim, K. D. Jordan, and T. S. Zwier, *J. Am. Chem. Soc.* **116**, 11568 (1994).
- [28] K. Liu, M. G. Brown, C. Carter, R. J. Saykally, J. K. Gregory, and D. C. Clary, *Nature* **381**, 501 (1996).
- [29] C. Steinbach, P. Andersson, M. Melzer, J. K. Kazimirski, U. Buck, and V. Buch, *Phys. Chem. Chem. Phys.* **6**, 3320 (2004).

- [30] M. E. Dunn, E. K. Pokon, and G. C. Shields, *J. Am. Chem. Soc.* **126**, 2647 (2004).
- [31] G. Hincapié, N. Acelas, M. Castaño, J. David, and A. Restrepo, *J. Phys. Chem. A* **114**, 7809 (2010).
- [32] Y. Wang, X. Hunag, B. C. Shepler, B. J. Braams, and J. M. Bowman, *J. Chem. Phys.* **134**, 094509 (2011).
- [33] L. Rincon, R. Almeida, and D. G. Aldea, *Int. J. Quantum Chem.* **102**, 443 (2005).
- [34] A. Hermann, R. P. Krawczyk, M. Lein, P. Schwerdtfeger, I. P. Hamilton, and J. J. P. Stewart, *Phys. Rev. A* **76**, 013202 (2007).
- [35] J. Cui, H. Liu, and K. D. Jordan, *J. Phys. Chem. B* **110**, 18872 (2006).
- [36] W. Cencek, K. Szalewicz, C. Leforestier, R. van Harrevelt, and A. van der Avoird, *Phys. Chem. Chem. Phys.* **10**, 4716 (2008).
- [37] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *J. Chem. Phys.* **128**, 094313 (2008).
- [38] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *Science* **315**, 1249 (2007).
- [39] E. M. Mas, R. Bukowski, and K. Szalewicz, *J. Chem. Phys.* **118**, 4404 (2003).
- [40] S. S. Xantheas, private communication.
- [41] R. A. Kendall, T. H. Dunning, Jr., and R. J. Harrison, *J. Chem. Phys.* **96**, 6796 (1992).
- [42] H.-J. Werner, P. J. Knowles, R. Lindh, M. Schütz, *et al.*, “Molpro, version 2009.1, a package of ab initio programs,” (2009), see <http://www.molpro.net>.

- [43] J. Baker, K. Wolinski, M. Malagoli, D. Kinghorn, P. Wolinski, G. Magyarfalvi, S. Saebo, T. Janowski, and P. Pulay, *J. Comp. Chem.* **30**, 317 (2009).
- [44] A. Halkier, W. Klopper, T. Helgaker, P. Jørgensen, and P. R. Taylor, *J. Chem. Phys.* **111**, 9157 (1999).
- [45] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, and J. Olsen, *Chem. Phys. Lett.* **302**, 437 (1999).
- [46] M. Jeziorska, R. Bukowski, W. Cencek, M. Jaszuński, B. Jeziorski, and K. Szalewicz, *Coll. Czech. Chem. Commun.* **68**, 463 (2003).
- [47] M. Jeziorska, W. Cencek, K. Patkowski, B. Jeziorski, and K. Szalewicz, *Int. J. Quantum Chem.* **108**, 2053 (2008).
- [48] S. F. Boys and F. Bernardi, *Mol. Phys.* **19**, 553 (1970).
- [49] K. Szalewicz and B. Jeziorski, *J. Chem. Phys.* **109**, 1198 (1998).
- [50] R. M. Olson, J. L. Bentz, R. A. Kendall, M. W. Schmidt, and M. S. Gordon, *J. Chem. Theory Comput.* **3**, 1312 (2007).
- [51] R. Kumar, F.-F. Wang, G. R. Jenness, and K. D. Jordan, *J. Chem. Phys.* **132**, 014309 (2010).
- [52] H.-J. Werner, F. R. Manby, and P. J. Knowles, *J. Chem. Phys.* **118**, 8149 (2003).
- [53] F. Weigend, A. Köhn, and C. Hättig, *J. Chem. Phys.* **116**, 3175 (2002).
- [54] R. Podeszwa, K. Patkowski, and K. Szalewicz, *Phys. Chem. Chem. Phys.* **12**, 5974 (2010).
- [55] F.-M. Tao and Y.-K. Pan, *J. Chem. Phys.* **97**, 4989 (1992).
- [56] K. Patkowski, R. Podeszwa, and K. Szalewicz, *J. Phys. Chem. A* **111**, 12822 (2007).

- [57] E. E. Dahlke, R. M. Olson, H. R. Leverentz, and D. G. Truhlar, *J. Phys. Chem. A* **112**, 3976 (2008).
- [58] W. Kutzelnigg and W. Klopper, *J. Chem. Phys.* **94**, 1985 (1991).
- [59] W. Klopper, F. R. Manby, S. Ten-No, and E. F. Valeev, *Int. Rev. Phys. Chem.* **25**, 427 (2006).
- [60] S. S. Xantheas, C. J. Burnham, and R. J. Harrison, *J. Chem. Phys.* **116**, 1493 (2002).
- [61] K. Patkowski and K. Szalewicz, *J. Chem. Phys.* **133**, 094304 (2010).
- [62] K. Szalewicz and B. Jeziorski, in *Molecular Interactions: From van der Waals to Strongly Bound Complexes*, edited by S. Sheiner (Wiley, 1997) pp. 3–43.
- [63] E. M. Mas and K. Szalewicz, *J. Chem. Phys.* **104**, 7606 (1996).
- [64] K. Szalewicz, G. Murdachaew, R. Bukowski, O. Akin-Ojo, and C. Leforestier, in *Lecture Series on Computer and Computational Science: ICCMSE 2006*, Vol. 6, edited by G. Maroulis and T. Simos (Brill Academic Publishers, Leiden, 2006) pp. 482–491.
- [65] X. Huang, B. J. Braams, and J. M. Bowman, *J. Phys. Chem. A* **110**, 445 (2006).
- [66] Y. Wang, S. Carter, B. J. Braams, and J. M. Bowman, *J. Chem. Phys.* **128**, 071101 (2008).
- [67] D. M. Bates, J. R. Smith, T. Janowski, and G. S. Tschumper, *J. Chem. Phys.* **135**, 044123 (2011).
- [68] S. Kazachenko and A. J. Thakkar, *Chem. Phys. Lett.* **476**, 120 (2009).
- [69] S. Yoo, E. Aprá, X. C. Zeng, and S. S. Xantheas, *J. Phys. Chem. Lett.* **1**, 3122 (2010).

[70] R. Podeszwa, J. Phys. Chem. A **112**, 8884 (2008).

[71] G. J. O. Beran, J. Chem. Phys. **130**, 164115 (2009).

Table 2.1: The total interaction energies $E_{\text{int}}^{\text{CCSD(T)}}$ (in kcal/mol) and the corresponding CBS extrapolated energies for the prism and cage water hexamers. The geometries were obtained from the CC-pol-8s+NB potential.

basis	cage		prism		$\Delta_{\text{p-c}}(X)$	$\Delta_{\text{p-c}}(\text{CBS})$
	$E_{\text{int}}^{\text{CCSD(T)}}(X)$	$E_{\text{int}}^{\text{CCSD(T)}}(\text{CBS})$	$E_{\text{int}}^{\text{CCSD(T)}}(X)$	$E_{\text{int}}^{\text{CCSD(T)}}(\text{CBS})$		
aDZ	-39.3292	-45.6578	-39.2218	-45.6733	0.1074	-0.0155
aTZ	-43.8227	-46.6085	-43.7859	-46.6478	0.0368	-0.0393
aQZ	-45.5032		-45.5089		-0.0057	

Table 2.2: The total interaction energies $E_{\text{int}}^{\text{MP2}}$ and the $\delta E_{\text{int}}^{\text{CCSD(T)}}$ contributions (in kcal/mol), as well as the corresponding CBS extrapolated energies, for the prism and cage water hexamers. The geometries were obtained from the CC-pol-8s+NB potential.

basis set	$E_{\text{int}}^{\text{MP2}}$	$E_{\text{int}}^{\text{MP2}}(\text{CBS})$	$\delta E_{\text{int}}^{\text{CCSD(T)}}$	$\delta E_{\text{int}}^{\text{CCSD(T)}}$	$E_{\text{int}}^{\text{CCSD(T)}}(X, Y)$	$\Delta_{\text{p-c}}$
	cage					
aDZ	-39.8402		0.5110			
aTZ	-43.6683	-45.2233	-0.1544	-0.4346	-44.7123 ^a	
aQZ	-45.2436	-46.2722	-0.2595	-0.3362	-46.4266 ^b	
a5Z	-45.7335	-46.2396			-45.9930 ^c	-46.5758 ^d
	prism					
aDZ	-39.5697		0.3479			
aTZ	-43.4461	-45.0440	-0.3397	-0.6292	-44.6961 ^a	0.0162
aQZ	-45.0632	-46.1247	-0.4457	-0.5231	-46.4644 ^b	-0.0378
a5Z	-45.5686	-46.0908			-46.0143 ^c	-0.0213
					-46.6139 ^d	-0.0381

^a $E_{\text{int}}^{\text{CCSD(T)}}((\text{DT}), \text{D})$.

^b $E_{\text{int}}^{\text{CCSD(T)}}((\text{TQ}), \text{T})$.

^c $E_{\text{int}}^{\text{CCSD(T)}}(5, \text{Q})$.

^d $E_{\text{int}}^{\text{CCSD(T)}}((\text{Q5}), (\text{TQ}))$.

Table 2.3: Two-body contributions, $E_{\text{int}}[2, 6]$ (in kcal/mol), to the interaction energies of the cage and prism water hexamers calculated in basis sets shown in column one with midbond functions [3s3p2d2f1g]. All calculations used dimer-centered basis sets plus midbond (DC⁺BS) unless shown otherwise. The geometries were obtained from the CC-pol-8s+NB potential.

basis set	$E_{\text{int}}^{\text{MP2}}$	$E_{\text{int}}^{\text{MP2}}(\text{CBS})$	$\delta E_{\text{int}}^{\text{CCSD(T)}}$	$\delta E_{\text{int}}^{\text{CCSD(T)}}(X, Y)$	$E_{\text{int}}^{\text{CCSD(T)}}(X, Y)$	$\Delta_{\text{p-c}}$
aTZ/HCBS	-34.8156		-0.4871		-35.3027 ^a	
aTZ/DCBS	-34.5954		-0.4501		-35.0455 ^a	
aTZ	-35.7275		-0.5717		-36.2992 ^a	
aQZ	-36.5252	-37.0985	-0.5727	-0.5734	-37.0979 ^a	
a5Z	-36.8957	-37.2693	-0.5094	-0.4429	-37.4051 ^a	
a6Z	-37.0874	-37.3505			-37.7122 ^b	
					-37.5968 ^c	
					-37.7934 ^d	
aTZ/HCBS	-34.5804		-0.7231		-35.3035 ^a	-0.0008
aTZ/DCBS	-34.3483		-0.6816		-35.0299 ^a	0.0156
aTZ	-35.5424		-0.8154		-36.3578 ^a	-0.0586
aQZ	-36.3443	-36.9215	-0.8109	-0.8075	-37.1552 ^a	-0.0573
a5Z	-36.7206	-37.1003	-0.7437	-0.6733	-37.4643 ^a	-0.0592
a6Z	-36.9153	-37.1824			-37.7736 ^b	-0.0614
					-37.6590 ^c	-0.0622
					-37.8557 ^d	-0.0623

^a Straightforward $E_{\text{int}}^{\text{CCSD(T)}}$ for the cardinal number X.

^b Extrapolated $E_{\text{int}}^{\text{CCSD(T)}}(\text{Q5})$.

^c Hybrid $E_{\text{int}}^{\text{CCSD(T)}}(6, 5)$.

^d Hybrid $E_{\text{int}}^{\text{CCSD(T)}}((56), (Q5))$.

Table 2.4: Three-body contributions, $E_{\text{int}}[3, 6]$ (in kcal/mol), to the interaction energies of the cage and prism water hexamers. All calculations used trimer-centered basis sets (no midbond) unless shown otherwise. The geometries were obtained from the CC-pol-8s+NB potential.

basis	$E_{\text{int}}^{\text{HF}}$	$E_{\text{int}}^{\text{MP2}}$	$E_{\text{int}}^{\text{MP2}}/\text{CBS}$	$\delta E_{\text{int}}^{\text{CCSD(T)}}$	$\delta E_{\text{int}}^{\text{CCSD(T)}}$	$E_{\text{int}}^{\text{CCSD(T)}}$	$\Delta_{\text{p-c}}$
	cage						
aTZ/HCBS	-8.5446	-8.4644		0.3773		-8.0871 ^a	
aTZ	-8.5458	-8.4616		0.3766		-8.0849 ^a	
aQZ	-8.5480	-8.4704	-8.4757	0.3625	0.3523	-8.1078 ^a	
						-8.1235 ^b	
a5Z	-8.5478	-8.4717	-8.4732			-8.1092 ^c	
						-8.1210 ^d	
	prism						
aTZ/HCBS	-8.4618	-8.4084		0.4428		-7.9656 ^a	0.1215
aTZ	-8.4630	-8.4036		0.4420		-7.9616 ^a	0.1233
aQZ	-8.4645	-8.4124	-8.4181	0.4276	0.4170	-7.9849 ^a	0.1229
						-8.0011 ^b	0.1224
a5Z	-8.4645	-8.4142	-8.4160			-7.9866 ^c	0.1226
						-7.9990 ^d	0.1220

^a Straightforward $E_{\text{int}}^{\text{CCSD(T)}}$ for the cardinal number X .

^b Extrapolated $E_{\text{int}}^{\text{CCSD(T)}}(\text{TQ})$.

^c Hybrid $E_{\text{int}}^{\text{CCSD(T)}}(5, \text{Q})$.

^d Hybrid $E_{\text{int}}^{\text{CCSD(T)}}(\text{Q5}, (\text{TQ}))$.

Table 2.5: Water hexamer energies from complete many-body expansion. The K -body contributions to interaction energies $E_{\text{int}}[K, 6]$ (in kcal/mol) were computed using HCBSs except as noted. The interaction energies from the canonical supermolecular approaches (the best results from Tables 2.1 and 2.2) are given for comparison. The geometries were obtained from the CC-pol-8s+NB potential.

K	level of theory	cage	prism	$\Delta_{\text{p-c}}$
2	$E_{\text{int}}^{\text{CCSD(T)}}((56),(\text{Q}5))/\text{DC}^+\text{BS}$	-37.7934	-37.8556	-0.0622
3	$E_{\text{int}}^{\text{CCSD(T)}}((\text{Q}5),(\text{TQ}))/\text{TCBS}$	-8.1210	-7.9990	0.1220
4	$E_{\text{int}}^{\text{CCSD(T)}}/\text{aTZ}$	-0.4342	-0.5737	-0.1395
5	$E_{\text{int}}^{\text{CCSD(T)}}/\text{aTZ}$	0.0027	0.0562	0.0535
6	$E_{\text{int}}^{\text{CCSD(T)}}/\text{aTZ}$	-0.0014	0.0008	0.0022
total		-46.3473	-46.3713	-0.0240
E_{int} from the canonical supermolecular approach				
	total/ (TQ)	-46.6085	-46.6478	-0.0393
	total/ $((\text{Q}5),(\text{TQ}))$	-46.5758	-46.6139	-0.0381

Table 2.6: The total vertical interaction energies $E_{\text{int}}^{\text{MP2}}$ and the $\delta E_{\text{int}}^{\text{CCSD(T)}}$ contributions (in kcal/mol), as well as the corresponding CBS extrapolated energies, for the cage and prism water hexamers. All calculations used the *KCBS* scheme. Hexamer geometries were taken from Ref. 57.

basis set	$E_{\text{int}}^{\text{MP2}}$	$E_{\text{int}}^{\text{MP2}}/\text{CBS}$	$\delta E_{\text{int}}^{\text{CCSD(T)}}$	$\delta E_{\text{int}}^{\text{CCSD(T)}/\text{CBS}}$	$E_{\text{int}}^{\text{CCSD(T)}}(X, Y)$	$\Delta_{\text{p-c}}$
cage						
haTZ/noCP	-46.6387		0.1347			
haTZ	-45.0559		0.1828			
aDZ	-41.5017		0.6861			
aTZ	-45.4982	-47.1122	0.0026	-0.2852	-46.9294 ^a	
aQZ	-47.1121	-48.1720	-0.1064	-0.1859	-48.1694 ^b	
a5Z	-47.6209	-48.1452			-47.7273 ^c	
					-48.3310 ^d	
prism						
haTZ/noCP	-46.6504		-0.0594			
haTZ	-45.1067		-0.0294			
aDZ	-41.6620		0.4795			
aTZ	-45.5665	-47.1719	-0.2131	-0.5048	-47.2013 ^a	-0.2793
aQZ	-47.1840	-48.2495	-0.3203	-0.3984	-48.4626 ^b	-0.2932
a5Z	-47.6946	-48.2211			-48.0149 ^c	-0.2876
					-48.6195 ^d	-0.2885

^a $E_{\text{int}}^{\text{CCSD(T)}}((\text{DT}), \text{D})$.

^b $E_{\text{int}}^{\text{CCSD(T)}}((\text{TQ}), \text{T})$.

^c $E_{\text{int}}^{\text{CCSD(T)}}(5, \text{Q})$.

^d $E_{\text{int}}^{\text{CCSD(T)}}((\text{Q5}), (\text{TQ}))$.

Table 2.7: MP2 and CCSD(T) distortion energies, E_{dist} of Eq. (2.12), of the monomers for the water hexamer cage and prism structures from Ref. 57 computed using MCBS basis sets. The relaxed interaction energies, $E_{\text{int}}^{\text{rel}}$ of Eq. (2.13), were calculated as the sum of the vertical ((Q5),(TQ))-extrapolated interaction energies of Table 2.6 and the (Q5)-extrapolated distortion energies. All energies are in kcal/mol.

basis set	cage		prism		$\Delta_{\text{p-c}}$
	MP2	CCSD(T)	MP2	CCSD(T)	
aQZ	2.2087	2.1620	2.2148	2.1929	0.0309
a5Z	2.2673	2.2281	2.2810	2.2649	0.0368
(Q5)	2.3055	2.2743	2.3274	2.3173	0.0430
$E_{\text{int}}^{\text{rel}}$	-45.8396	-46.0568	-45.8937	-46.3022	-0.2454
$E_{\text{int}}^{\text{rel}}$, Ref. 6	-45.80	-45.67	-45.86	-45.92	-0.25

Table 2.8: Summary of strategies for calculations employing many-body expansions for water hexamer. All calculations should be performed in K -mer-centered basis sets and bond functions should be included for two-body contributions. Calculations in bases aQZ and larger should be extrapolated.

uncertainty	two-body	three-body	four-body	five-body	six-body
10 kcal/mol or 20%	MP2/aDZ	HF/DZ	neglect	neglect	neglect
1 kcal/mol or 2%	CCSD(T)/aQZ	MP2/aDZ	neglect	neglect	neglect
0.6 kcal/mol or 1.2%	CCSD(T)/(TQ)	CCSD(T)/aDZ	MP2/aDZ	neglect	neglect
0.1 kcal/mol or 0.2%	MP2/a6Z	CCSD(T)/aTZ	MP2/aDZ	MP2/aDZ	neglect
	CCSD(T)/a5Z				
0.02 kcal/mol or 0.02%	MP2/a8Z	CCSD(T)/aQZ	CCSD(T)/aDZ	MP2/aDZ	neglect
	CCSD(T)/a7Z				
	CCSDT(Q)?				

Table 2.9: Computer timings of CCSD(T) calculations for the water hexamer (in hours of wall time, all data scaled to one core, MOLPRO package, 2.2 GHz AMD Opteron Processor). We were not able to perform calculations on the 2-6 aQZ/HCBS and 4-6-body/KCBS level in the reasonable time, thus the empty space in the table. The timings for a given K -body contribution include the complete set of calculations needed for this particular contribution. For example, the time given for $K=3$ is the time needed to calculate 6 monomers, 15 dimers, and 20 trimers in the HCBS approach. M denotes the number of K -mers one needs to calculate for a given K .

K	hexamer basis set			M	K -body basis set		
	aDZ	aTZ	aQZ		aDZ	aTZ	aQZ
2	44.89	327.52		15	0.24	5.13	128.13
3	149.69	1147.82		20	9.43	234.25	4018.99
4	281.89	2345.64		15	41.07	1051.94	
5	360.79	3006.33		6	201.41	4371.80	
6	375.62	3184.50		1	375.62	3184.50	
hexamer + monomers	16.50	252.20	10212 ^a				

^aComputed using the PQS package.

Table 2.10: Cost-effective calculations of water hexamer interaction energies from many-body expansion. The K -body contributions to interaction energies (in kcal/mol) were calculated using the minimum-cost method providing the best possible agreement with the canonical supermolecular approach at the CCSD(T)/(TQ) level. The timings of the calculations are given in hours. All calculations used the KCBS approach (no midbond functions). The six-body contribution was neglected. The geometries were obtained from the CC-pol-8s+NB potential.

K	level of theory	cage	prism	Δ_{p-c}	time
2	CCSD(T)/(TQ)	-38.0814	-38.1587	-0.0773	133.26
3	CCSD(T)/aTZ	-8.0849	-7.9616	0.1233	234.25
4	CCSD(T)/aDZ	-0.4317	-0.5626	-0.1309	41.07
5	MP2/aDZ	0.0002	0.0460	0.0458	16.63
total		-46.5978	-46.6369	-0.0391	425.21
full hexamer	CCSD(T)/(TQ)	-46.6085	-46.6478	-0.0393	10212
two-body alternative approach (all data with midbond functions):					
2	CCSD(T)/((Q5),(TQ))mb	-37.8427	-37.9078	-0.0651	370.06

Table 2.11: K -body contributions $E_{\text{int}}[K, 24]$ (in kcal/mol) to the vertical interaction energies of the structures 316 and 308 of $(\text{H}_2\text{O})_{24}$. All calculations used the K CBS approach. Geometries of the structures were obtained from Ref. 40.

K	basis	HF	MP2	CCSD(T)
structure 316				
2	modTZ/noCP	-153.029	-243.373	-232.320
	modTZ	-110.418	-155.355	-147.826
	TZ	-109.219	-159.918	-152.576
	aTZ	-111.772	-176.976	-175.851
	aQZ	-113.076	-185.708	-185.283
	(TQ)			-191.532
3	TZ	-60.724	-60.229	
	aDZ	-59.880	-59.508	-59.002
4	TZ	-6.023		
	aDZ	-6.033	-7.001	
structure 308				
2	modTZ/noCP	-154.107	-245.078	-233.962
	modTZ	-111.101	-156.303	-148.770
	TZ	-109.870	-160.826	-153.483
	aTZ	-112.849	-178.693	-177.509
	aQZ	-114.149	-187.396	-186.904
	(TQ)			-193.128
3	TZ	-59.833	-59.151	
	aDZ	-58.862	-58.202	-57.791
4	TZ	-5.964		
	aDZ	-5.915	-6.765	

Table 2.12: Total interaction energies (in kcal/mol) for the structures 316 and 308 of $(\text{H}_2\text{O})_{24}$. The vertical interaction energies E_{int} are calculated as the sum of the best vertical many-body contributions from Table 2.11 ($E_{\text{int}}[2, 24]$ - CCSD(T)/(TQ), $E_{\text{int}}[3, 24]$ - CCSD(T)/aDZ, $E_{\text{int}}[4, 24]$ - MP2/aDZ). The relaxed interaction energies, $E_{\text{int}}^{\text{rel}}$ of Eq. (2.13), are the sum of the vertical interaction energies and of the CCSD(T)/(TQ)/MCBS distortion energies.

	structure 316	structure 308
$E_{\text{dist}}^{\text{CCSD(T)}/\text{aTZ}}$	14.942	15.196
$E_{\text{dist}}^{\text{CCSD(T)}/\text{aQZ}}$	17.594	17.815
$E_{\text{dist}}^{\text{CCSD(T)}/(\text{TQ})}$	18.815	19.046
E_{int}	-257.534	-257.684
$E_{\text{int}}^{\text{rel}}$	-238.719	-238.638
$E_{\text{int}}^{\text{rel}}$, Ref. 7	-276.03	-276.02

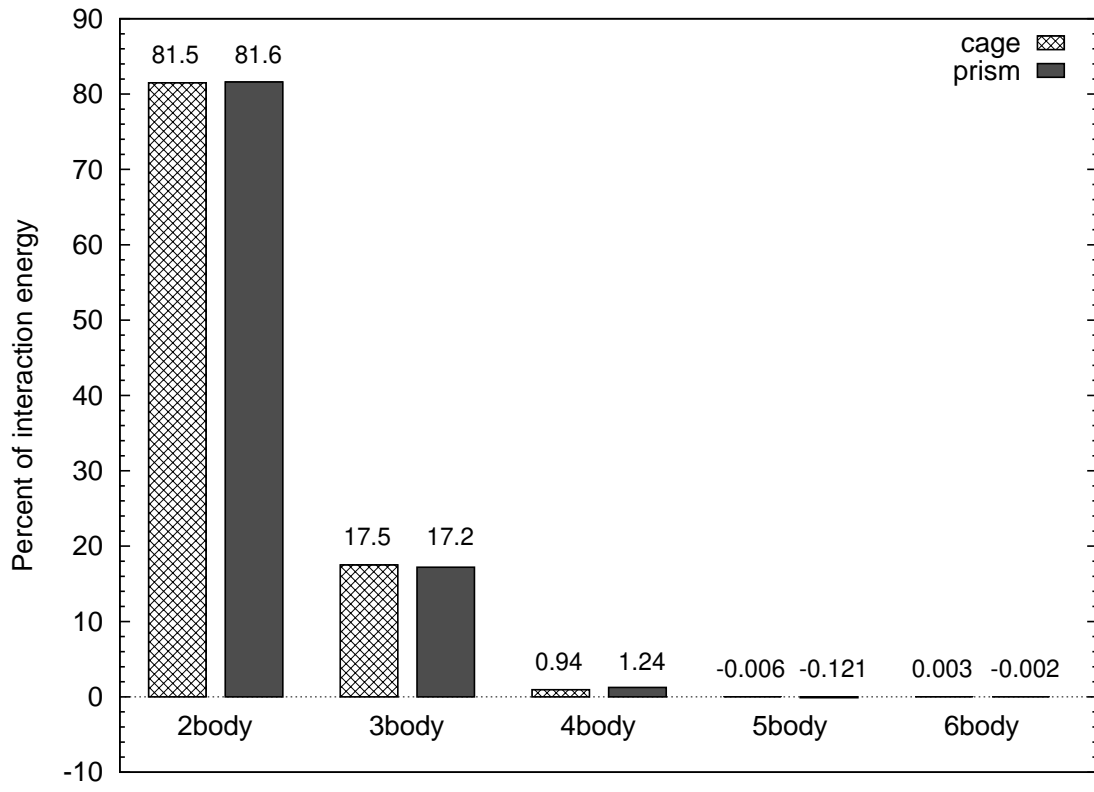


Figure 2.1: Convergence of the many-body expansion for the cage and prism structures of the water hexamer.

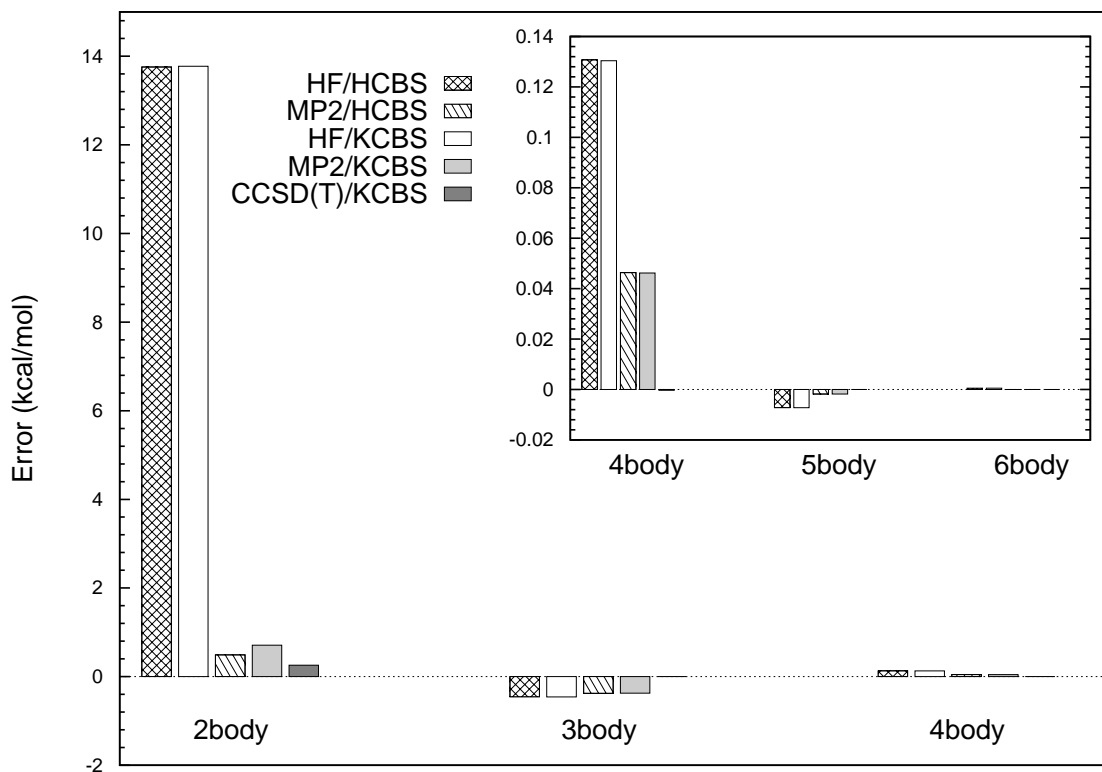


Figure 2.2: Convergence of the individual contributions in the many-body expansion with level of theory for the cage structure of water hexamer. Errors of the K -body contributions $E_{\text{int}}[K, 6]$ were computed in the aTZ basis set relative to the value of a given contribution obtained at the CCSD(T)/aTZ/HCBS level. The hatched (plain) bars represent calculations using the HCBS (KCBS) approach. Energies are in kcal/mol (including inset).

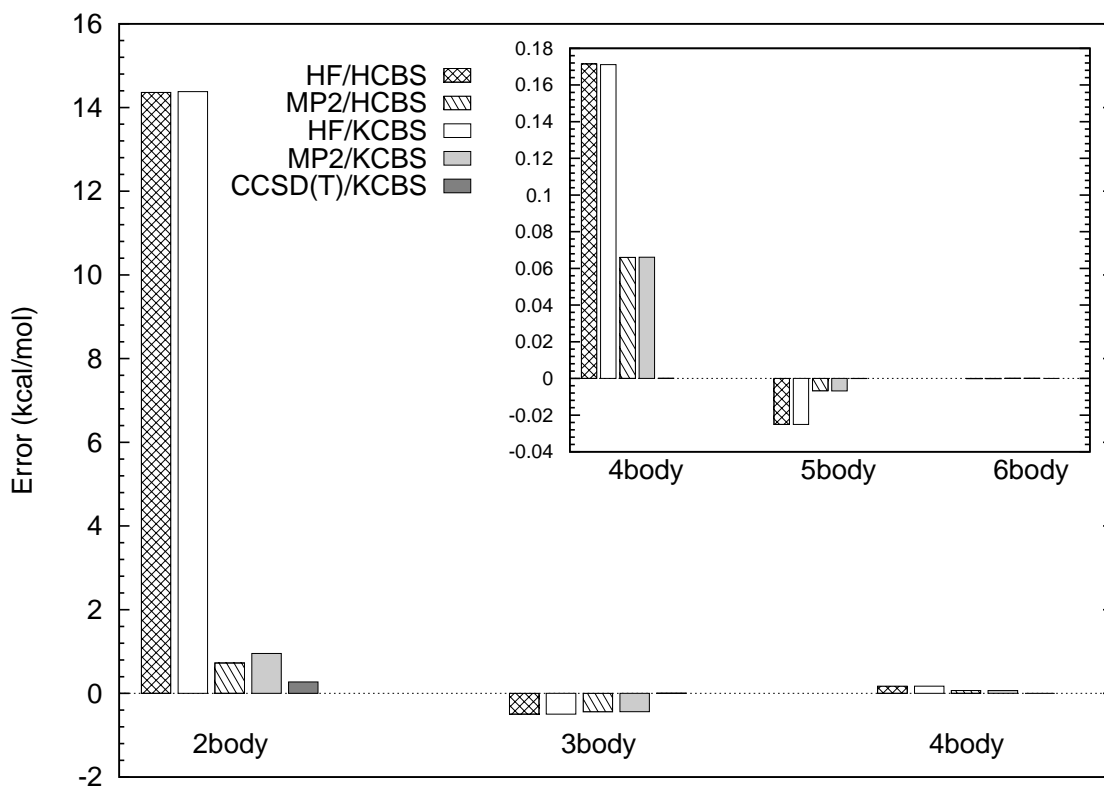


Figure 2.3: Same as Fig. 2.2 but for the prism structure.

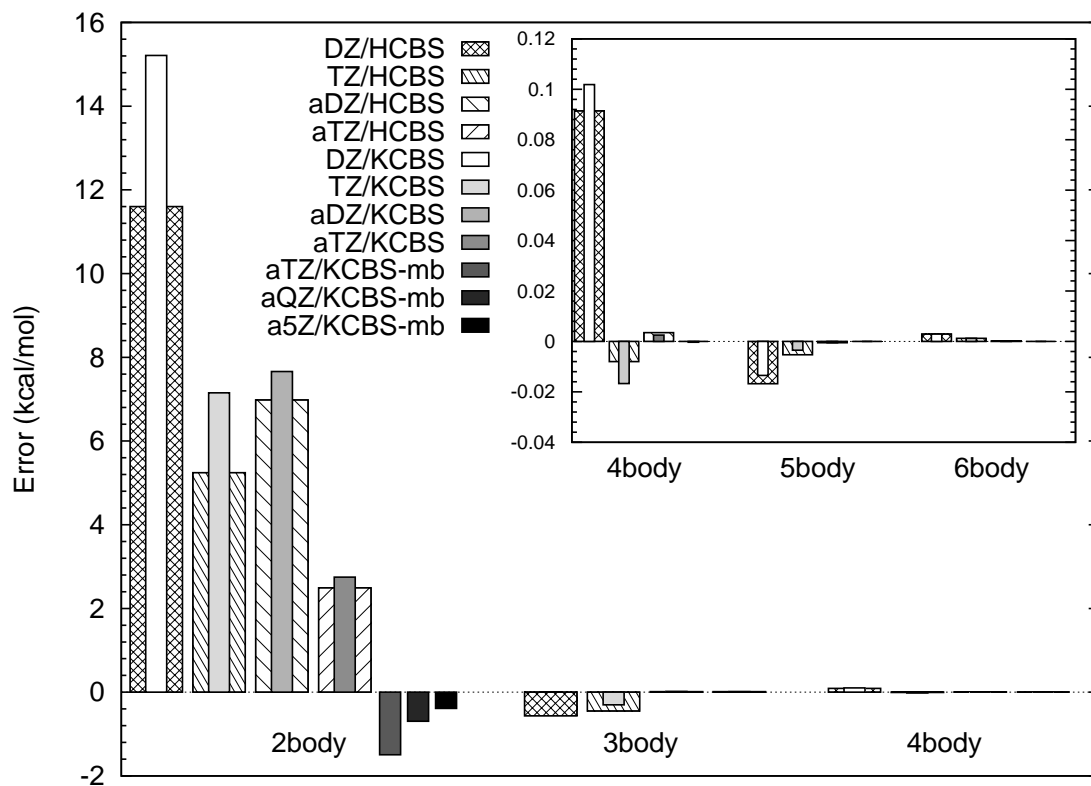


Figure 2.4: Convergence of the individual contributions in the many-body expansion with basis set size for the cage structure of water hexamer. Errors of the K -body contributions $E_{\text{int}}[K, 6]$ were computed at the CCSD(T) level of theory in the consecutive basis sets relative to the benchmark value of a given contribution taken from Table 2.5. The hatched (plain) bars represent calculations using the HCBS (KCBS) approach and “mb” denotes the use of bond functions. Energies are in kcal/mol (including inset).

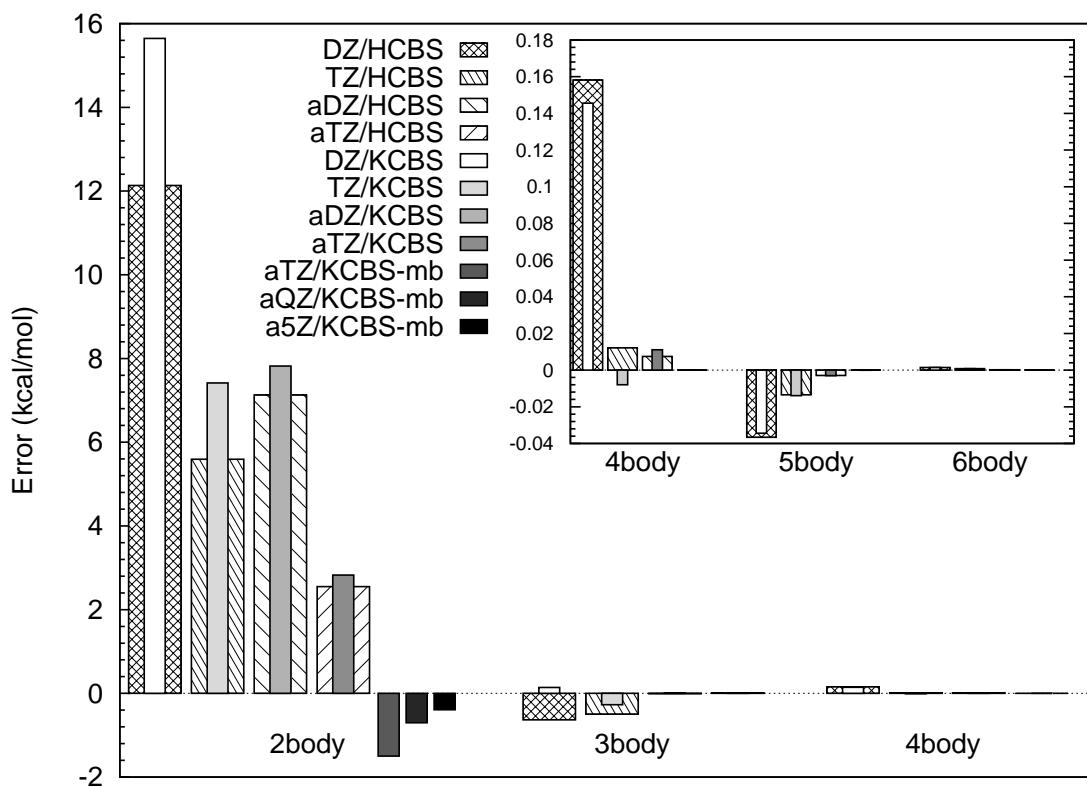


Figure 2.5: Same as Fig. 2.4 but for the prism structure.

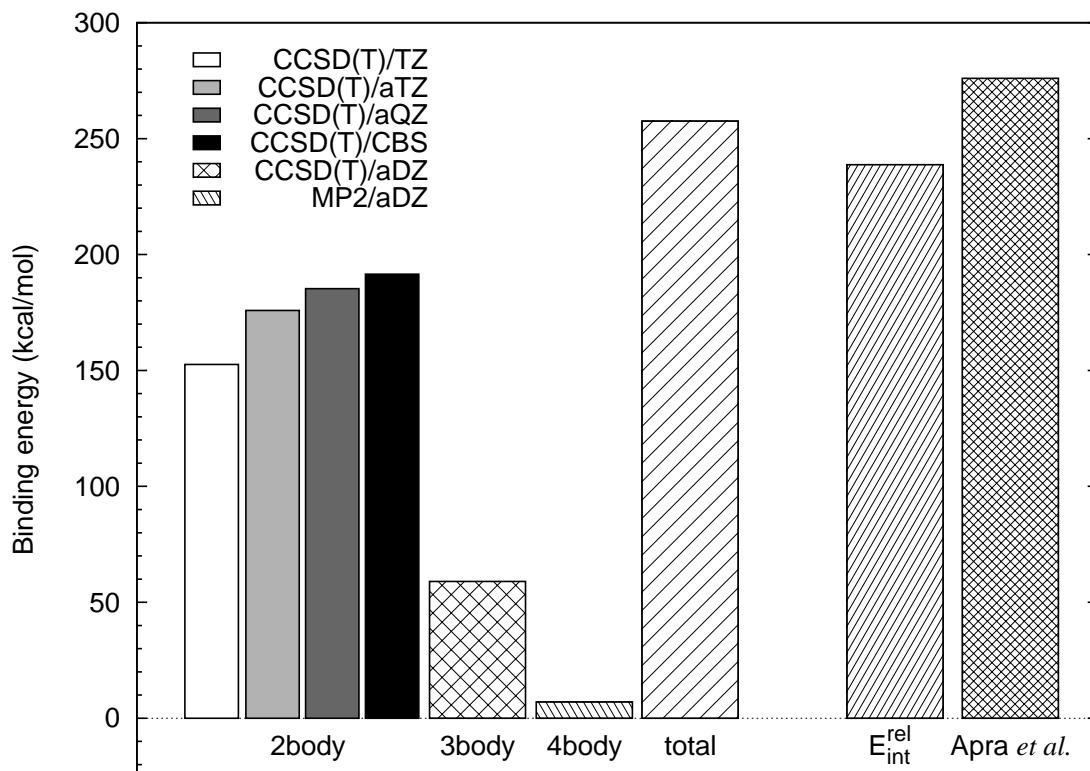


Figure 2.6: Many-body decomposition of the vertical interaction energy for the 316 structure of $(\text{H}_2\text{O})_{24}$. The first four bars represent the vertical two-body contributions in increasing-size basis sets and are followed by bars representing three- and four-body contributions. The bar “total” is the vertical interaction energy E_{int} . The subsequent bars show the relaxed interaction energies $E_{\text{int}}^{\text{rel}}$ from our work and from Apra *et al.* [7]

Date: September 1, 2024

Institute of Chemistry
University of Silesia
Szkolna 9, 40-006 Katowice, Poland

Statement on the Contribution in Publication

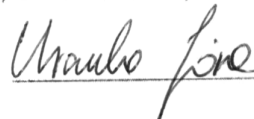
Reference: U. Góra, R. Podeszwa, W. Cencek, and K. Szalewicz, J. Chem. Phys. **135**, 224102 (2011).

Authors contributed to the publication jointly as follows:

Urszula Góra

Methodology, Software, Calculations, Data Analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing

Signature:



Rafał Podeszwa

Formal Analysis, Software, Validation, Calculations Supervision, Funding, Writing - Review & Editing

Signature:



Wojciech Cencek

Methodology, Formal Analysis, Validation, Writing - Original Draft, Writing - Review & Editing


Signature:



Krzysztof Szalewicz

Conceptualization, Methodology, Validation, Supervision, Funding Acquisition, Writing - Review & Editing

Signature:



Chapter 3

PREDICTIONS FOR WATER CLUSTERS FROM A FIRST-PRINCIPLES TWO- AND THREE-BODY FORCE FIELD¹

Abstract

A new rigid-monomer three-body potential has been developed for water by fitting it to more than 70 thousand trimer interaction energies computed *ab initio* using coupled-cluster methods and augmented triple-zeta-quality basis sets. This potential was used together with a modified form of a previously developed two-body potential and with a polarization model of four- and higher-body interactions to predict the energetics of the water trimer, hexamer, and 24-mer. Despite using the rigid-monomer approximation, these predictions agree better with flexible-monomer benchmarks than published results obtained with flexible-monomer force fields. An unexpected finding of our work is that simple polarization models predict four-body interactions to within a few percent, whereas for three-body interactions these models are known to have errors on the order of 50%.

3.1 Introduction

Water has always been attracting significant attention of theorists due to its abundance and importance for life, but also since the water monomer is a relatively small molecule so that reasonably accurate *ab initio* calculations of interaction energies could be performed. The water dimer in particular was the benchmark system for comparing the performance of various theoretical methods [1–7] at selected points on the potential energy surface. To connect to experiments, one needs a complete

¹ The text appeared in U. Góra, W. Cencek, R. Podeszwa, A. van der Avoird, and K. Szalewicz, *J. Chem. Phys.* **140**, 1941011 (2014).

potential in order to perform nuclear dynamics calculations. Such calculations for water clusters and for condensed phases of water predict observable properties such as spectra, dissociation energies, radial distribution functions, etc. The Holy Grail of theory is to develop a universal potential that can correctly predict the properties of all forms of water: from dimer to condensed phases, i.e., the predictions should be sufficiently accurate to be meaningfully confronted with experiment.

One way of obtaining such a potential for water is to fit it in molecular dynamics simulations to reproduce as closely as possible experimental data. The potentials of this type are called empirical potentials and some well-known examples are the TIP4P [8] and SPCE [9] potentials. Since these potentials include three- and higher-body interactions in an effective way via pairwise-only terms, such empirical potentials do not work well for small clusters [10]. Modern empirical potentials are fitted simultaneously also to *ab initio* data on clusters [11, 12] and have a polarization term which partly accounts for pairwise nonadditive interactions. Thus, such potentials may perform better on small clusters, but still will probably not be able to produce sufficiently accurate results since simple polarization approximations can recover only about half of the three-body interaction energy for liquid water [13]. Another type of empirical potential can be obtained by fits to water dimer spectra [14–17]. Such potentials represent the water dimer very well, but since the fitting is done purely to the dimer properties, these potentials cannot provide any information about pairwise nonadditive interactions critical for water clusters and condensed phases [18]. Potentials of this type have been used with the polarization model of nonadditive effects, but this model had to be added *post factum* to the two-body potentials. One can also add an *ab initio* three-body nonadditive potential, but to our knowledge this option has not been tried yet.

The other way of developing a potential for water is to fit it to *ab initio* computed interaction energies. Pioneering work of this type was performed by Clementi and collaborators [1, 19, 20], but since only a few hundred grid points could be computed for the water dimer and trimer at that time, the sampling of the surface was hardly

adequate. The number of grid points was significantly increased in the next generation of first-principles potentials [21–28] to 2.5 thousand points for the dimer [25] and 7.5 thousand points for the trimer [28]. With further improvements of the dimer potential, the goal of correctly predicting properties of all forms of water, from the water dimer to liquid water, was achieved in Refs. 29–31. A good example of how theory can lead experiment in this field is the dissociation energy, D_0 , of the water dimer. In 2000, the first *ab initio* prediction [26] gave the value of 1067 cm^{-1} , while later improved calculations in 2008 [31] and 2009 [32] gave 1111 and 1104 cm^{-1} , respectively. This quantity was accurately measured only in 2011 [33] and the result of $1105 \pm 10 \text{ cm}^{-1}$ agreed very well with prior theoretical predictions. Despite this striking agreement on D_0 , there is still need for improvements of water potentials, in particular in the pairwise nonadditive part. An especially challenging subject are the anomalous properties of liquid water such as the high boiling temperature, anomalous density-temperature dependence, high dielectric constant, and many others. Another possible peculiarity to investigate is the existence of a liquid-liquid critical point in supercooled water [34]. Molecular simulations aimed at predicting these properties are very sensitive to the quality of the applied intermolecular potentials.

First-principles potentials are commonly based on the many-body expansion of the N -body interaction energy

$$E_{\text{int}} = E_{\text{int}}[2, N] + E_{\text{int}}[3, N] + \dots + E_{\text{int}}[N, N], \quad (3.1)$$

where $E_{\text{int}}[2, N]$ is the sum of pair interactions in the N -body cluster and $E_{\text{int}}[K, N]$ with $K > 2$ are the pairwise-nonadditive K -body contributions. This expansion utilizes the so-called vertical interaction energies, i.e., the energies relative to the monomer energies at the same geometries as in the N -body cluster. One can add to this expansion the one-body term which is the difference between these monomer energies and the sum of the energies of equilibrium-geometry (r_e) monomers. The extended expansion defines then the so-called relaxed interaction energy [35]. One should note that the

latter quantity is essentially the total electronic energy of the complex, only shifted by a constant. All the interaction energies in the present paper will be vertical unless noted otherwise. The expansion of Eq. (3.1) converges sufficiently fast to eliminate, except in very rare cases, any need to deal with more than a few initial terms for any given system size N . A fundamental advantage of *ab initio* potentials is the straightforward separation of the interaction energies into the K -body contributions. Empirical models, on the other hand, are fitted to bulk physical properties and if the higher-body terms are somehow effectively approximated through a two-body potential, properties such as second virial coefficients that depend only on the physical pair interactions, are described poorly. In water, three-body effects are absolutely essential and contribute as much as 16% to the liquid energy [36]. In some clusters, this percentage is even larger, for instance 23% in the 24-mer [37]. Also the four-body effects cannot be neglected in high-accuracy calculations, as their contribution is about 1% in the hexamer and 3% in the 24-mer [37]. Higher than four-body effects typically account for only a few tenths of a percent, but still can become relevant in problems such as establishing the relative energetic ordering of close-lying cluster structures. Most of the *ab initio* calculations leading to water potentials have been devoted to the first term in the expansion (3.1), recent work are Refs. 22–26, 29–32, 38–52. The number of papers devoted to the second term is much smaller, only Refs. 20, 28, 36, 46, 47, 53. No four-body potentials exist for any system, even an atomic one.

The reason that it is very difficult to develop higher terms in the many-body expansion from first principles is the so-called “dimensionality curse”. The successive terms $E_{\text{int}}[K, N]$ of the expansion of Eq. (3.1) are functions of $3KL - 6$ relative coordinates, where L is the number of atoms in the monomer. If these functions are to be obtained by an analytic fit to calculated *ab initio* energies, the problem quickly becomes intractable because the number of dimensions precludes any reasonable coverage of the total space with calculated data points. A solution is to calculate the interaction energies on-the-fly for any geometry generated in nuclear dynamics simulations, instead of producing global analytic functions. Obviously, only low-level *ab*

initio methods are sufficiently fast for this purpose which limits the predictive power of this approach. Therefore, the current state-of-the-art in accurate predictions for water is to employ two- and three-body potentials fitted to *ab initio* interaction energies and approximate the higher-body terms by polarization models which account for the asymptotic induction component (by computing the Coulomb interactions between the permanent multipole moments and the induced multipole moments of the monomers, see Sec. 3.5). To our knowledge, there were no published investigations of how well this approximation works for four and higher-body effects (the present work will provide such information). As already mentioned, simple polarization models recover only about 50% of the total three-body contribution in liquid water [13]. This shows that one cannot avoid construction of first-principles three-body potentials for accurate water simulations.

A two-body potential with rigid monomers is 6-dimensional, whereas the inclusion of the intramonomer degrees of freedom results for triatomic monomers in a 12-dimensional potential. With current computational power, it is possible to represent reasonably well the 12-dimensional surface by a set of grid points [41, 44, 48, 52], although as many as 250 thousand such points may be required [41]. On the other hand, three-body flexible-monomer potentials are 21-dimensional. Generation of such fits would require calculations of 2.1 million data points using a mere 2 points per dimension and as many as 10 billion data points with 3 points per dimension. The flexible-monomer three-body potentials of Wang *et al.*, fitted to 30 thousand points [46] or 40 thousand points [47], use only about 1.6 points per dimension. For the 6-dimensional rigid-monomer water dimer, this number of points per dimension would result in a total of only 20 grid points, clearly an inadequate number. Since it is unclear whether 30–40 thousand points are adequate for the water trimer with flexible monomers, we have decided to use the rigid-monomer approximation, as assumed in Ref. 28. With the 12 resulting degrees of freedom, the about 70 thousand grid points that we have used in our computations correspond to about 2.5 points per dimension. Whereas such sampling still seems barely adequate, this number of points is one order of magnitude larger

than used in the development of the rigid-monomer SAPT-3B nonadditive three-body water potential in Ref. 28. We will analyze our results to shed light on the question whether it is more beneficial, with a given number of grid points, to obtain a more accurate rigid-monomer potential or a possibly less accurate flexible-monomer one. It should be noted that the choice of an optimal rigid-monomer geometry is crucial: it has been established that the use of the average geometry in the lowest rovibrational state leads to much more accurate results than the use of the equilibrium geometry of the monomers [4, 54]. The rigid-monomer potentials have some obvious limitations, for example one cannot use such potentials to predict the shifts of intramonomer rovibrational frequencies upon complexation. However, the class of problems where such potentials work well is broad and we will demonstrate this in particular for the structure and energetics of water clusters.

The rigid-monomer SAPT-3B three-body potential of Mas *et al.* [28] was based on 7533 data points computed using symmetry-adapted perturbation theory (SAPT) [55–59] at the level equivalent to the Hartree-Fock (HF) method and a moderate-size [5s3p2d1f/3s2p] basis set. The computed nonadditive interaction energies were fitted to a physically motivated analytic formula containing representations of the short-range exchange contributions and damped induction terms of the same form as in polarization models. This three-body potential was initially combined with the two-body SAPT-5s potential from Ref. 25 (this combination was denoted as SAPT-5s+3B) and then with the CC-pol-5s potential of Refs. 29, 30 (CC-pol-5s+3B). The latter two-body potential was fitted to dimer interaction energies computed using the coupled-cluster method with single, double and non-iterative triple excitations [CCSD(T)] extrapolated to the complete basis set (CBS) limit. The CC-pol-5s+3B potential was used to predict trimer spectra in Ref. 60 and achieved very good agreement with experiment, much better than in the case of other potentials. An early version of the three-body potential, restricted to the trimer tunneling path, was used to predict trimer spectra in Ref. 26. These strictly two- plus three-body potentials were extended by adding a polarization model describing four- and higher-body nonadditive effects (such potentials were denoted by

SAPT-5s+NB and CC-pol-5s+NB) and used in simulations of liquid water in Refs. 29, 31, 36. As already mentioned, the latter work provided a uniformly accurate description of all forms of water.

The project described in the present paper started from an application of the CC-pol-8s potential of Ref. 43 combined with the three-body potential of Ref. 28 and higher-body polarization effects (CC-pol-8s+NB) in water-cluster calculations of Ref. 37. CC-pol-8s was fitted to the same set of interaction energies as CC-pol-5s, but uses a more elaborate functional form with 8 rather than 5 symmetry-unique sites per monomer. CC-pol-8s is still the most accurate rigid-monomer two-body water potential available. The abbreviation “pol” reflects the fact that a self-consistent two-body polarization term is explicitly included in the potential. When predictions of the CC-pol-8s+NB potential were compared to *ab initio* decompositions of cluster energies, it was found that the three-body contribution clearly dominates the overall error (with respect to benchmark *ab initio* results). This was not surprising in view of the fact that a rather limited number of data points and a moderate level of theory were used in Ref. 28. Therefore, the primary aim of the present project was to develop a significantly more accurate rigid-monomer three-body potential by calculating an order of magnitude more points at a much higher level of theory. This development is described in Secs. 3.2, 3.3, and 3.6. The functional form of the fit has also been significantly changed compared to that of Ref. 28. In particular, a more sophisticated polarization model was developed and optimized to partly reproduce four-body effects, see Sec. 3.5. To use consistently the same polarization model for all K -body terms, we have refitted the CC-pol-8s [43] two-body potential. Furthermore, as reported in Sec. 3.4, we used additional *ab initio* data points computed in Ref. 13 to improve the accuracy in the repulsive wall region and introduced a very short-distance damping of site-site functions. Section 3.7 describes applications of the complete new N -body model to the water trimer, hexamer, and 24-mer. The new potential was also used to calculate the trimer spectrum, this work will be described in a separate paper.

The combination of the two-body and three-body potentials described above

could be performed in several ways. In fact, since the CC-pol- ls potentials are polarizable, one could iterate the polarization model over all monomers in an N -body cluster, thus approximating the pairwise nonadditive interaction energies by polarization terms only. Another option is to use a straight sum of the two-body and three-body potentials, CC-pol- $ls+3B$, equivalent to a truncation of the expansion of Eq. (3.1) after the second term. Finally, one can add to the CC-pol- $ls+3B$ potential higher than three-body effects by iterating the polarization model over all N monomers and subtracting from the result the two- and three-body polarization components, which leads to the CC-pol- $ls+NB$ potentials. To avoid confusion with previous work and to simplify the notation, we introduce here a new nomenclature for the potentials developed in this work. The names are composed of the stem “CCpol” followed by the digits “2” and/or “3” (depending on which K -body terms are present), optionally followed by a plus sign if higher-body effects are treated by the polarization model. Thus, “CCpol2” stands for the two-body potential only (including the two-body polarization) and is the only possible choice in the case of the water dimer or it can be used to determine purely two-body effects in larger clusters or in the bulk. Similarly, “CCpol3” stands for the three-body pairwise nonadditive potential. “CCpol2+” adds higher-than-two-body polarization effects (but not the complete three-body potential). Similarly, “CCpol23” will denote a pure two- plus three-body potential, whereas “CCpol23+” includes polarization effects beyond the three-body level and is our most complete force field for systems larger than the trimer.

3.2 Choice of trimer configurations

The number and choice of the trimer configurations (grid points) that are used in the fitting process significantly impact the quality of the resulting fit. Since the number of points is always limited by the costs of *ab initio* calculations, the optimal selection of such points is critical. The configurations used in Ref. 28 served as an initial guide, and in particular we calculated nonadditive interaction energies for all 7533 grid points from Ref. 28. We have used the same $\langle r \rangle_0$ (averaged over the ground

rovibrational state) rigid-monomer geometry as Ref. 28, which originates from Ref. 4.

Since the set of Ref. 28 contained mostly trimers with intermonomer separations close to those in trimer’s equilibrium structure, we first augmented it with 7315 geometries with larger separations, selected from trimer configurations present in water clusters from the tetramer to the 21-mer taken from the Cambridge Cluster Database [61]. These geometries were optimized using the TIP5P[62] potential. For each cluster of size N , all possible $\binom{N}{3}$ trimers were generated. The rigid-monomer geometry in TIP5P is different from ours, therefore we have placed our monomers in these trimers in such a way that the centers of mass (COM), the bisectors of the HOH angle, and obviously the molecular planes coincide with the TIP5P monomers.

To ensure that we do not completely miss some regions of configuration space, we generated another set of grid points randomly, with the sampling of COM-COM separations R restricted to the range 2.6–7 Å. To avoid placing grid points close to the existing ones, we created a sorted list of distances between all the atoms in the trimers and compared them one by one. We defined a “distance” between trimers using an extension of formulas (21) and (22) from Ref. 63. We selected 30,000 points that were farthest from the previously chosen ones. Although the algorithm is not fully permutationally invariant, so that the maximum distance criterion is not satisfied, it removes all duplicates and we checked by inspection that it removed most of the trimers similar to the existing ones.

To improve the description of the region relevant for liquid water, 8520 additional points were selected from snapshots of a converged molecular dynamics (MD) simulation at ambient conditions performed in Ref. 13. This was done in the same way as described in Sec. IV of that reference except that 71 snapshots spaced by 5 ps were used.

Since one of the intended applications of our potential are calculations for the water hexamer, a subject of significant recent interest [53, 64, 65], a number of trimer configurations were taken from hexamer structures. First, a set of 14,500 trimer configurations was selected from snapshots of quantum diffusion Monte Carlo simulations

performed by us for the hexamer with the CC-pol-8s two-body potential, an early version of the three-body potential developed in the present work, and the CC-pol-8s polarization model for higher-body effects. This set should improve the description of regions relevant for the rovibrational motions in the hexamer and other water clusters. To improve the description of the regions near the hexamer local minima, another set of configurations was generated as follows. First we optimized the geometries of the cage, prism, book, boat, bag, and ring isomers using the CC-pol-8s+NB potential, starting from the configurations taken from Ref. 66. Since the latter configurations included flexible monomers, we “projected” our monomers similarly as in the case of the TIP5P potential, except that our rigid-monomer bisector now coincides with the line connecting the position of the oxygen atom with the midpoint of the segment connecting the two hydrogens. We will refer to cluster geometries produced in this way as geometries with “rigidized” monomers. The geometries were optimized using a simple Powell [67] algorithm, changing all six coordinates (three center-of-mass coordinates and three Euler angles) of a single water molecule at a time and going through all six molecules in cycles until the energy was converged to at least 10^{-6} kcal/mol. The procedure usually converges to the minimum structure closest to the starting point. All the 120 trimers present in the hexamers thus obtained were extracted (these trimers were not included in our data set used for fitting the potential). Then 2400 trimers were created from this set by adding small random increments of either sign to the coordinates: between 0.03 and 1 bohr for the COM-COM distances and between 1 and 10 degrees for the Euler angles.

During the initial fitting of the potential to the set of three-body nonadditive interaction energies computed at the 70,268 grid points described above, we found that the fit was not sufficiently accurate for very small intermonomer separations. Therefore, 1188 points were added in this region as described in Sec. 3.6.2.

Altogether, a total of 71,456 trimer interaction energies were used in the fitting process, almost ten times the number of points used in Ref. 28. As already stated, this corresponds to 2.54 points per dimension. The fit of Wang *et al.* [47] used a comparable

number of grid points, about 40,000, but these points had to cover a 21-dimensional space (which amounts to 1.66 points per dimension).

3.3 *Ab initio* calculations

For all calculations in the present paper, the MOLPRO suite of programs [68] was used. We used the MOLPRO’s 1 hartree = 627.5096 kcal/mol conversion factor. For each grid point described in the previous section, we performed *ab initio* supermolecular calculations of the vertical three-body nonadditive interaction energy. In the counterpoise (CP) corrected approach which removes the basis set superposition error [35, 69], this quantity is defined for a trimer consisting of monomers A, B, and C as

$$E_{\text{int}}[3, 3] = E_{\text{ABC}} - E_{\text{AB}} - E_{\text{AC}} - E_{\text{BC}} + E_{\text{A}} + E_{\text{B}} + E_{\text{C}}, \quad (3.2)$$

where the energies on the right-hand-side are the total energies of the indicated systems, all energies are computed in the full trimer basis set, and the positions of the monomers and of the ghost sites are in all calculations the same as in the trimer. The frozen-core approximation was used in all calculations of the correlation energies unless noted otherwise.

Although the individual calculations of the energies in Eq. (3.2) are nowadays not very demanding computationally even for fairly large basis sets and high levels of theory, the large number of grid points puts severe restrictions on the basis set size and computational methods. Therefore, we spent some time testing various combinations of these two elements. We performed this testing on 40 water trimers extracted from the cage and prism structures of the hexamer as described in Sec. 3.2. We have used the standard augmented correlation-consistent basis sets [70], aug-cc-pVXZ (further abbreviated as aXZ) with $X = 2, \dots, 5$, as well as the ‘half-augmented’ triple-zeta basis set (haTZ) that includes augmentation only on the oxygen atom. This basis set was recommended for the water hexamer by Bates and Tschumper [71] as a good compromise between size and performance.

The authors of Ref. 28 compared the performance of the HF, MP2 (second-order perturbation theory based on the Møller-Plesset partition of the Hamiltonian), and CCSD(T) methods for the water trimer. Although at that time the conclusion was that the HF level is sufficiently accurate, with the currently desired accuracy we have decided to use the CCSD(T) level as this method is known to provide very reliable interaction energies in virtually all applications. However, due to its n^7 scaling, CCSD(T) is computationally much more expensive than MP2 which scales as n^5 . Therefore, we also investigated a hybrid approach defined as

$$E_{\text{int}}^{\text{CCSD(T)}}/(X-X') = E_{\text{int}}^{\text{MP2}}(X) + \delta E_{\text{int}}^{\text{CCSD(T)}}(X'), \quad (3.3)$$

with $X > X'$, where

$$\delta E_{\text{int}}^{\text{CCSD(T)}} = E_{\text{int}}^{\text{CCSD(T)}} - E_{\text{int}}^{\text{MP2}}, \quad (3.4)$$

with both quantities computed in the same basis set.

To construct benchmarks for comparisons, we calculated interaction energies at the CBS level for all trimers. We used an extension of the hybrid method that additionally separates the HF interaction energy resulting in the quantity

$$\delta E_{\text{int}}^{\text{MP2}} = E_{\text{int}}^{\text{MP2}} - E_{\text{int}}^{\text{HF}}. \quad (3.5)$$

We used the extrapolations to CBS limits tested extensively in Refs. 72, 73. For the HF interaction energy, the extrapolation formula was:

$$E_{\text{int}}^{\text{HF}}(X) = E_{\text{int}}^{\text{HF}}(\text{CBS}) + Ae^{-\alpha X}, \quad (3.6)$$

where $\alpha = 1.63$, as recommended in Ref. 73, and A is an adjustable parameter. The correlation energies were extrapolated as

$$E_{\text{int}}^{\text{corr}}(X) = E_{\text{int}}^{\text{corr}}(\text{CBS}) + BX^{-3}, \quad (3.7)$$

where $E_{\text{int}}^{\text{corr}}(X)$ is given by Eq. (3.4) or Eq. (3.5) calculated with the aug-cc-pVXZ basis set. The parameters in the extrapolation formulas can be obtained by solving sets of linear equations resulting from writing these formulas for X and $X - 1$. In cases where it will be relevant to indicate the basis set used to obtain the CBS limit, we will replace “(CBS)” by the cardinal numbers involved, for example, “(T4) \equiv (34)” will indicate extrapolations with $X - 1 = 3$ and $X = 4$. Note that the symbols Method($X'X$) with $X' = X - 1$ and Method/($X-X'$) denote two different computational approaches. At the HF and MP2 levels, we used $X = 5$, whereas $X = 4$ was used at CCSD(T) level. The total benchmark energy was then calculated as the sum of the three CBS values

$$E_{\text{int}}(\text{CBS}) = E_{\text{int}}^{\text{HF}}(\text{CBS}) + \delta E_{\text{int}}^{\text{MP2}}(\text{CBS}) + \delta E_{\text{int}}^{\text{CCSD(T)}}(\text{CBS}). \quad (3.8)$$

The results computed in smaller basis sets are compared to the benchmarks in Table 3.1. The maximum absolute error (MAE) of the straightforward CCSD(T) calculations in the aQZ basis relative to the CBS benchmarks is 0.003 kcal/mol. Thus, our CBS results are probably at least that accurate relative to the exact values at the CCSD(T) frozen-core level. Such level of accuracy is actually difficult to reach at the two-body level, see Table III in Ref. 37. The total errors in recovering the values of $E_{\text{int}}[3, 6]$ are below 0.2% in the aQZ basis. Somewhat surprisingly, the analogous errors in the aDZ basis are below 0.5% only. However, MAE’s are as much as 11 (16) times larger for the cage (prism) trimers in the aDZ compared aQZ bases. The aTZ basis would be a good choice in terms of accuracy, but CCSD(T)/aTZ are too expensive for the calculations on the complete set of grid points. Therefore, the choice is really only between aDZ and haTZ at the CCSD(T) level. For the total three-body contribution, the two bases produce errors very similar in magnitude, however, the MAE’s are a factor 3–4 smaller in the latter case. Thus, the choice between these two basis sets would be difficult when the costs of calculations are taken into account (haTZ with 74 functions per monomer is still much larger than aDZ with 41 functions per monomer). However, it turned out that the hybrid approach, listed in the last column

for each hexamer, offers actually the best performance among these three cases, with the smallest magnitude of the total error and with the MAE of 0.009 (0.013) kcal/mol for the cage (prism) trimers. In fact, this approach performs comparably to CCSD(T) in the aTZ basis set. Thus, we have chosen the hybrid approach in aTZ/aDZ bases, i.e., $X = T$ and $X' = D$ in Eq. (3.3). Since MP2 calculations are much faster than CCSD(T) ones, the hybrid approach saves significant amounts of computer time. Overall, a calculation of $E_{\text{int}}^{\text{CCSD(T)}/(\text{T-D})}$ is 4 times faster than a calculation of $E_{\text{int}}^{\text{CCSD(T)}/\text{aTZ}}$ and takes about 1 hour on a single core of the 2.4 GHz Opteron processor. The root-mean-square error (RMSE) of the $E_{\text{int}}^{\text{CCSD(T)}/(\text{T-D})}$ values in Table 3.1 relative to the CBS results is 0.004 and 0.006 kcal/mol for the cage and prism isomers, respectively. Clearly, our hybrid approach is more than adequate at the current accuracy level of water potentials. For comparison, the calculations of Wang *et al.* [46, 47] used the aTZ basis set at the MP2 level. Results at this level of theory are also shown in Table 3.1. As one can see, the RMSE's of the MP2/aTZ approach are 5-6 times larger than in the $E_{\text{int}}^{\text{CCSD(T)}/(\text{T-D})}$ approach, whereas the corresponding errors in the total three-body contribution are 11-14 times larger in magnitude. Clearly, the addition of the $\delta E_{\text{int}}^{\text{CCSD(T)}/\text{aDZ}}$ contributions, despite the small size of the basis set, dramatically improves the agreement with the benchmarks.

A further comparison of the performance of various methods is provided in Table 3.2. The magnitudes of the errors in the total three-body nonadditive contribution to the hexamer interaction energies at the HF and MP2 CBS levels are about 0.4-0.5 kcal/mol, or only 1% of the hexamer interaction energy, but are clearly too large for investigations of subtle effects as the cage-prism energy difference that amounts to 0.25 kcal/mol [37, 71]. Interestingly enough, the use of MP2 gives only a negligible improvement over the HF level of theory. In contrast, going up to the CCSD(T) level gives a substantial contribution to the hexamer energy. There are no estimates of beyond CCSD(T) contributions to three-body nonadditive energies, but most likely such contributions to the hexamer energies will be on the order of 0.01 kcal/mol. The excellent agreement of the CCSD(T)(34) results with the CBS benchmarks shows that

the hybrid approach used in the calculations of benchmarks, with HF and MP2 extrapolations at the (45) level, was not needed (i.e., the CCSD(T)(34) benchmarks would have been sufficiently accurate). The remaining columns of Table 3.2 will be discussed in Sec. 3.7.1.

To check the importance of correlating core electrons, we have computed the three-body interaction energies for all trimers extracted from the hexamer with the aug-cc-pCVTZ basis set [75]. The magnitude of the largest correction for a trimer amounted to 0.002 (0.002) kcal/mol and the sum for all trimers was -0.011 (-0.013) kcal/mol for the cage (prism) hexamer. These errors are a few times smaller than the errors of the hybrid approach selected for our work, so the neglect of correlation effects involving the core electrons is justified. However, when a still more accurate future potential will be developed, these effects will have to be included.

3.4 Two-body fit

The form of the two-body CCpol2 fit is similar to that of the CC-pol-8s fit of Ref. 43, namely

$$V_2 = \sum_{a \in A} \sum_{b \in B} \tilde{u}_{ab}(r_{ab}) + V_2^{\text{ind}}(AB), \quad (3.9)$$

where \tilde{u}_{ab} are site-site functions depending only on the distances r_{ab} between sites associated with monomers A and B and $V_2^{\text{ind}}(AB)$ represents the induction interaction. The $\tilde{u}_{ab}(r_{ab})$ functions can be written as

$$\tilde{u}_{ab}(r_{ab}) = u_{ab}(r_{ab})d_{ab}(r_{ab}), \quad (3.10)$$

where $u_{ab}(r_{ab})$ has the form of Eq. (2) in Ref. 43 and $d_{ab}(r_{ab})$ are (very) short-range damping functions equal to 1 if $u_{ab}(r_{ab}) > 0$, and otherwise

$$d_{ab}(r_{ab}) = \{1 + \exp[-\gamma(r_{ab} - \tilde{r})]\}^{-1}, \quad (3.11)$$

with γ and \tilde{r} being adjustable parameters. Note that this damping is different from the damping already contained in the asymptotic terms of the functions u_{ab} and sets in at much smaller R than the latter damping. The induction interaction $V_2^{\text{ind}}(AB)$ is calculated with a new polarization model, described in Sec. 3.5, which is more elaborate than that used in CC-pol-8s. After adopting this polarization model, the adjustable parameters of the site-site part in the fit of Eq. (3.9) were fitted in the same way as described in Sect. II.E of Ref. 43, with the damping turned off (*i.e.*, all the functions $d_{ab}(r_{ab})$ set to one). However, the original set of 2510 data points was enlarged by adding 706 short-separation dimer geometries computed in Ref. 13 in order to improve the description of the repulsive wall. The *ab initio* approach used in these calculations was the same as in Refs. 29, 30. The RMSE of the new V_2 fit relative to the training set of interaction energies is 0.081 kcal/mol on the whole set of 3216 points and 0.011 kcal/mol for negative interaction energies, *i.e.*, very similar to the error of the fit developed in Ref. 43.

In the next step, we switched on the d_{ab} damping, keeping the other, previously optimized parameters fixed. We selected this approach instead of performing a simultaneous optimization of all parameters since we did not want to change the known very good behavior of the two-body fit in the physical region. The reason for introducing the additional damping factor d_{ab} was only to improve the very small R behavior of the total CCpol23 fit (*i.e.*, with the inclusion of the $V_3[3,3]$ part described in Sec. 3.6). The total interaction energy should be repulsive at very small R , but the $V_3[3,3]$ fits have a tendency to collapse there to unphysical, strongly negative values. We were unable to fully control this behavior of $V_3[3,3]$ alone (as described in Sec. 3.6), so we fixed this problem by accelerating the increase of V_2 for R going to zero. This acceleration takes place only in the region not relevant for the intended physical applications. In this way, the sum of these terms, *i.e.*, the CCpol23 potential, behaves reasonably. It may seem senseless to work on the behavior of the potential in an unphysical region, but in molecular simulations this region is occasionally sampled. If the potential found in such sampling is strongly repulsive, as it should be, this sampling has virtually no effect

on the simulations. However, if the potential is strongly attractive due to artifacts of the potential functions, the simulation may collapse. At such very small intermonomer separations, the V_2^{ind} term is small compared to the sum of the functions u_{ab} (this means that we do not observe the so-called polarization catastrophe). Therefore, we have not introduced any additional damping in V_2^{ind} . The functions u_{ab} are both positive and negative and despite strong cancellations between them (as it commonly happens for fits with many terms), their sum may become unphysical for very small R for some intermonomer orientations. We found that damping of the negative contributions in this region has the desired effect on the total CCpol23 fit. With the damping parameters chosen as described below, the functions $u_{ab}(r_{ab})$ are non-negligibly affected only for r_{ab} smaller than a fraction of 1 Å, i.e., their behavior is unchanged in the physical region. The damping strength is controlled by the parameters γ and \tilde{r} in Eq. (3.11): $\tilde{r} = 0$ and a very large γ implies no damping, increasing \tilde{r} turns the damping on at the pertinent site-site distances, while decreasing γ makes the effect more diffuse so that the damping effectively starts at larger distances. The values of these parameters were selected in the following way. First, we set γ to 100 bohr⁻¹ and increased \tilde{r} until the RMSE of the V_2 fit for all the points with interaction energies between +10 and +30 kcal/mol started deteriorating. This occurred at $\tilde{r} = 1.4$ bohr. Then, we gradually lowered γ with the same RMSE criterion, arriving in this way at $\gamma = 20$ bohr⁻¹. The damped V_2 fit has an RMSE equal to 0.25 kcal/mol between +10 and +30 kcal/mol, compared to 0.16 kcal/mol without damping, while the accuracy below +10 kcal/mol is not affected: both the damped and undamped fits have an RMSE of 0.011 kcal/mol for interaction energies below zero and an RMSE of 0.034 kcal/mol for those in the range from 0 to +10 kcal/mol. The RMSE of the damped fit on all 3216 points is very large, 222 kcal/mol, whereas for the undamped fit it is only 0.081 kcal/mol. Since virtually the whole RMSE in the former case originates from errors coming from the interaction energies above +30 kcal/mol (corresponding to kT at 15,000 K), which are practically not sampled in simulations near (and below) room temperature, this error is inconsequential for the intended applications of our potential.

We have not tested CCpol2 on the dimer characteristic points and spectra since it is numerically so close to CC-pol-8s in all physically relevant regions that the results computed with CCpol2 should be nearly identical to those computed with CC-pol-8s [43].

3.5 Many-body induction energy model

We will now define the polarization model used in our potential. This model was applied in the two-body component already described in Sec. 3.4, the three-body component that will be described in Sec. 3.6, and alone to represent four- and higher-body effects. The polarization model will be defined below in the general N -body context, special cases of two and three bodies follow immediately.

The polarization model represents the asymptotic induction energy of N molecules. It is often called classical polarization model but in fact the formalism is the same in quantum mechanics. This model can also be damped to account for the charge-overlap effects in induction interactions. In polarization models, the electric field due to the multipole moments of the charge distribution on isolated monomers (called permanent multipole moments) induces multipole moments on each monomer (of course, the permanent moments of a given monomer do not contribute to the field that induces the moments on this monomer). These induced moments, in turn, create an electric field that is added to the original field and induces additional moments. This procedure is iterated until convergence. The converged fields and induced multipole moments (and thus, the polarization energy) can also be found by solving a system of equations. In the simplest case, where only the induced dipole moments are considered and the permanent multipole moments are approximated by a set of distributed (partial) charges, the polarization energy of a system of N molecules can be written as

$$V_N^{\text{ind}} = -\frac{1}{2} \sum_{i=1}^N \sum_{k=1}^{N_{\text{pol}}^i} \mathbf{E}_{ik}^0 \cdot \boldsymbol{\mu}_{ik}^{\text{ind}}, \quad (3.12)$$

where \mathbf{E}_{ik}^0 is the electric field generated on the k th polarizable center of molecule i by

permanent distributed charges on all the other molecules, and

$$\boldsymbol{\mu}_{ik}^{\text{ind}} = \alpha_{ik} \mathbf{E}_{ik} \quad (3.13)$$

is the dipole moment induced on this polarizable center by the *total* electric field \mathbf{E}_{ik} generated by other molecules. The quantity α_{ik} is the polarizability (considered here to be isotropic) of the k th center of molecule i such that

$$\alpha_i = \sum_{k=1}^{N_{\text{pol}}^i} \alpha_{ik} \quad (3.14)$$

is the total molecular polarizability. The total electric field on each center is the sum of permanent and induced components

$$\mathbf{E}_{ik} = \mathbf{E}_{ik}^0 + \sum_{j \neq i}^N \sum_{l=1}^{N_{\text{pol}}^j} f_3(\delta_3, r_{ik,jl}) \mathbf{T}_{ik,jl} \boldsymbol{\mu}_{jl}^{\text{ind}}, \quad (3.15)$$

where

$$\mathbf{T}_{ik,jl} = -\frac{1}{r_{ik,jl}^3} \left(1 - 3 \frac{\mathbf{r}_{ik,jl} \otimes \mathbf{r}_{ik,jl}}{r_{ik,jl}^2} \right) \quad (3.16)$$

is the dipole-dipole interaction matrix, vector $\mathbf{r}_{ik,jl}$ points from the k th polarization center of molecule i to the l th polarization center of molecule j , and \otimes denotes the vector direct product. The factor $f_n(\delta, r)$ in Eq. (3.15) is the Tang-Toennies damping function [76]

$$f_n(\delta, r) = 1 - e^{-\delta r} \sum_{m=0}^n \frac{(\delta r)^m}{m!}, \quad (3.17)$$

which continuously decays to zero at small r . The fields from the permanent charges q_l are also damped:

$$\mathbf{E}_{ik}^0 = - \sum_{j \neq i}^N \sum_{l=1}^{N_{\text{q}}^j} f_2(\delta_2, r_{ik,jl}) q_l \frac{\mathbf{r}_{ik,jl}}{r_{ik,jl}^3}, \quad (3.18)$$

where N_{q}^j is the number of partial charges on molecule j .

The polarization model defined above is a generalization of that used in the nonadditive three-body [28, 36], CC-pol [29–31], and CC-pol-8s [43] water potentials, where it was restricted to just $N_{\text{pol}}^i = 1$. Note that these past polarization models were damped at the three-body level but not at the two-body level. In the present work, we used three polarization centers located on the atoms of each monomer, with the polarizability values $\alpha_1 = 6.5186$ a.u. and $\alpha_2 = \alpha_3 = 1.5507$ a.u. The values are chosen in such a way that the total polarizability α is equal to 9.62 a.u. (the benchmark CCSD(T) result of Ref. 13), while the ratio of the oxygen to hydrogen values is 4.2036 (the value calculated [77] using the CamCASP code [78], with an asymptotically corrected PBE0 density functional [79, 80] and in the doubly augmented daug-cc-pVTZ basis set [81]). The conversion factor of 1 hartree = 627.51 kcal/mol was used in this case. The damping constants δ_2 and δ_3 were optimized on a training set containing both three-body and four-body nonadditive energies. In the former case, these were pure induction and exchange-induction energies including the overlap effects. In the latter case, we used the complete four-body nonadditive contributions, hoping that in this way our polarization model will effectively improve the description of the four-body interactions. Specifically, let us denote by σ_i , $i = 1, \dots, 6$, the RMSE’s of the nonadditive four-body energies for all 15 tetramers contained in each of the hexamer structures: prism, cage, book, bag, boat, and ring, respectively, calculated with the polarization model and relative to the $E_{\text{int}}^{\text{CCSD(T)}}/(\text{T-D})$ values. Additionally, let σ_7 stand for the RMSE of the total nonadditive four-body contributions, $E_{\text{int}}[4, 6]$, in the six structures. Finally, let σ_8 be the RMSE of the polarization model with respect to the following sum of three-body SAPT corrections

$$E_{\text{ind}}^{(20)}[3, 3] + E_{\text{exch-ind}}^{(20)}[3, 3] + \delta E^{\text{HF}}[3, 3] = E_{\text{int}}^{\text{HF}}[3, 3] - E_{\text{exch}}^{(10)}[3, 3] \quad (3.19)$$

for the 5704 water trimer geometries computed in Ref. 28. [After the fit was completed, we found that we had erroneously included also the 1829 trimers for which $E_{\text{exch}}^{(10)}$ was not computed and was set to zero in the data set. However, the 1829-point subset

consists mostly of large trimers for which $E_{\text{exch}}^{(10)}$ is very small in magnitude, so we have not corrected this error.] One may question the inclusion of the exchange-induction energies in Eq. (3.19) since these energies decay exponentially. However, at the two-body level it has been shown that the exchange-induction energy to a large extent cancels the purely exponential overlap component of the induction energy (see Ref. 82) which is also not a part of the polarization model. The values $\delta_2 = 1.65 \text{ bohr}^{-1}$ and $\delta_3 = 1.55 \text{ bohr}^{-1}$ were found by minimizing the sum $\sum_{i=1}^8 \sigma_i$. For the final model, the resulting values of σ_i are 0.020, 0.034, 0.018, 0.023, 0.009, 0.045, 0.071, and 0.203 kcal/mol.

Table 3.3 compares the predictions of the final model with those of the model from Ref. 28 and with CCSD(T) results for the four-, five-, and six-body interaction energies in the six hexamers. The three-center polarization model leads to a modest improvement over the one-center model in the four-body energies: the RMSE on all hexamers relative to the CCSD(T) benchmarks is 0.059 vs. 0.076 kcal/mol. The relative errors for individual hexamers range between -0.6% and 12.2% with the average magnitude of the relative errors amounting to 5.8% for the three-center model. This is a surprisingly small relative error, much smaller than in the case of three-body energies [13] where the accuracy of simple polarization models is only about 50%. To our knowledge, this fact has not yet been noted in literature and is of significant importance in developing many-body force fields.

For five- and six-body nonadditive interactions, the RMSE's of the three-center (one-center) models are 0.025 (0.019) and 0.005 (0.004) kcal/mol, respectively. Thus, the one-center model actually performs slightly better, but these differences are negligible. The overall accuracy is not as good as in the four-body case, but it is reasonable for the contributions that are of more significant size: the magnitudes of relative errors of the three-center model for the contributions larger in magnitude than 0.03 kcal/mol are in the range of 23-37%.

Interestingly, the sum of the four- to six-body contributions is recovered significantly better than any individual component. The RMSE for the three-center model

is only 0.046 kcal/mol and the average magnitude of the relative errors is 3.5%. The many-body effects beyond the polarization interactions are not that small for individual K -mers within each hexamer, but there are significant cancellations.

We then tested the performance of the polarization model on the set of 600 non-additive three-body interaction energies for trimer configurations extracted from the ambient-conditions liquid water MD simulations of Ref. 13 (see Sec. 3.2). These energies, ranging from -0.9 to 1.0 kcal/mol, were computed at the hybrid CCSD(T)/(T-D) level and are quite well reproduced by our new polarization model. The value of the RMSE relative to *ab initio* benchmarks is 0.073 kcal/mol, much improved compared to the single-center polarization models included in the CC-pol-8s' and CC-dpol-8s' potentials of Ref. 13 which both give an RMSE of 0.107 kcal/mol [13]. This improvement was achieved despite the fact that the new polarization model was partly optimized for four-body effects. This performance indicates that the three-center model is a more physically sound representation of the induction effects in water than the one-center model.

3.6 Nonadditive three-body fit

3.6.1 Functional form of fit

In analogy with Eq. (18) of Ref. 28, our present three-body fit is the sum of three components,

$$V_3[3, 3] = F_{S^3} + F_{S^2} + V_3^{\text{ind}}[3, 3]. \quad (3.20)$$

The term $V_3^{\text{ind}}[3, 3] = V_3^{\text{ind}} - V_2^{\text{ind}}(AB) - V_2^{\text{ind}}(BC) - V_2^{\text{ind}}(AC)$ is the nonadditive induction energy from the new polarization model described in Sec. 3.5. The F_{S^2} term is the interaction energy contribution of the same form as in Ref. 28 that is due to single two-electron permutations between monomer pairs, while F_{S^3} represents mostly the interaction energy due to cyclic permutations involving three electrons from three different monomers, but also other residual effects not accounted for by the other terms. Also, nonadditive dispersion energy which is nonnegligible for this system only

for short and medium-range intermolecular distances (for the near-equilibrium trimers, this contribution is on the order of only 0.01 kcal/mol [18]) is modeled by this term. Instead of the Legendre polynomial expansion employed in Ref. 28, we used a trimer generalization of the exponential site-site expansion of CC-pol-8s (see Eqs. (1) and (2) of Ref. 43). Specifically,

$$F_{S^3} = \sum_{a \in A} \sum_{b \in B} \sum_{c \in C} f_{ab}(r_{ab}) f_{ac}(r_{ac}) f_{bc}(r_{bc}) \sum_{k=0}^1 \sum_{l=0}^1 \sum_{m=0}^1 c_{klm}^{(abc)} r_{ab}^k r_{ac}^l r_{bc}^m \quad (3.21)$$

where

$$f_{ab}(r_{ab}) = \{1 + \exp[-\gamma_{ab}(r_{ab} - \tilde{r}_{ab})]\}^{-1} \exp(-\beta_{ab} r_{ab}) \quad (3.22)$$

and similarly for $f_{ac}(r_{ac})$ and $f_{bc}(r_{bc})$. As before, r_{ab} denotes the distance between site a in molecule A and site b in molecule B . The nonlinear adjustable parameters γ_{ab} , β_{ab} , and \tilde{r}_{ab} , as well as the linear ones, $c_{klm}^{(abc)}$, can be obtained from the optimization of an appropriate least-squares functional, with linear parameters obtained in each step of the nonlinear optimization by solving the set of equations of the linear least-square problem. The first factor in the function $f_{ab}(r_{ab})$ is a damping function of the same form as $d_{ab}(r_{ab})$ in Eq. (3.11) but, in contrast to the two-body fit, we used parameters γ_{ab} and \tilde{r}_{ab} that were independently optimized for each site-site pair. The accuracy of the expansion (3.21) depends critically on the number and location of monomer sites. In Ref. 43, the new sites in the dimer expansion were successively added and their positions optimized until reaching 25 (8 symmetry-unique) sites per molecule. Since a similar optimization procedure would be extremely expensive in the trimer case, we used the first 17 (6 symmetry-unique, including the O and H atoms) site positions obtained in Ref. 43. After grouping the symmetry-equivalent terms as described below, Eq. (3.21) includes 364 independent linear parameters. For 6 unique sites per molecule, the number of nonlinear fit parameters amounts to 63 (21 for β_{ab} , γ_{ab} , and \tilde{r}_{ab} each).

Since the symmetry operations of the water monomer transform a given site into one of its equivalents, the nonadditive three-body potential for the water trimer

with rigid monomers should be invariant to permutations of symmetry-equivalent sites within each monomer. In addition, the potential should be invariant to the six (including the identity) permutations of complete monomers A, B, and C. Thus, the terms in Eq. (3.21) can be separated into groups such that each term in a given group has the same linear coefficient. Also, to realize the symmetry conditions, the functions of Eq. (3.22) should be identical for all a 's and all b 's that are symmetry equivalent. Thus, there are only $6 \times 7/2 = 21$ different such functions. We used a simple algorithm (executed only during initialization of the potential subroutine) to impose this symmetry. Initially, an integer-valued array $I(a, b, c, k, l, m)$ is filled with zeros and a counter N_{lin} is set to one. In a loop going through all possible $17^3 \times 2^3 = 39,304$ terms of the sum in Eq. (3.21), if $I(a, b, c, k, l, m)$ is still equal to zero for a given combination of a, b, c, k, l, m , it is set to N_{lin} . If $I(a, b, c, k, l, m)$ is not equal to zero, this combination is skipped because it has been handled before. Then each of the parameters a, b, c is identified as belonging to a group of symmetry equivalent sites and $I(a', b', c', k, l, m)$ are set to N_{lin} for all a', b', c' of the same type as a, b, c , respectively. At the same time, the ABC permutational symmetry is realized by setting to N_{lin} all the six cases resulting from the permutations of the three monomers. Finally, the value of N_{lin} is increased by one. In actual calculations using Eq. (3.21), all terms with the same value of $I(a, b, c, k, l, m) = n$ are added together and associated with the linear parameter c_n .

3.6.2 Fitting of three-body potential

The first stage of the fitting procedure used the initial set of 70,268 trimer geometries obtained as described in Sec. 3.2. The data points were unweighted (all weights set to one) and no damping was applied, *i.e.*, the factors d_{ab} in Eq. (3.22) were all equal to one. Several hundred different fits were generated by using different (randomly generated) starting values of the nonlinear parameters and then optimizing them using Powell's algorithm [67]. The induction component was held fixed. Several of the most accurate fits, those with RMSE's (relative to all *ab initio* three-body

nonadditive interaction energies) equal to about 0.02 kcal/mol, were tested at short-distance configurations in the following way. For each point on a three-dimensional grid of intermonomer distances from 1.8 to 3.0 Å with steps of 0.2 Å, we generated one million different trimer geometries by randomly choosing the orientations of the monomers. At each geometry, the two- and three-body energies were evaluated from the fits V_2 (also with $d_{ab} = 1$) and $V_3[3, 3]$. The 5940 geometries corresponding to trimer interaction energies that were deemed “most unphysical” were selected. These were not just very negative energies, but we looked in particular at negative energies that were strongly dominated by $V_3[3, 3]$. In the next step, we checked if switching on the damping in both fits can eliminate the unphysical behavior at short distances. To this end, a new series of $V_3[3, 3]$ fits were generated with all parameters optimized (including the damping parameters) and tested on the same 5940 points in the presence of the damped two-body fit. The number and the magnitude of negative interaction energies at short distances were significantly reduced, but not sufficiently enough. Therefore, we decided to enlarge the training data set to encompass a number of very small trimers. To this end we calculated three-body energies at the 5940 problematic points using the CCSD(T) method and the aDZ basis set. A subset of randomly chosen 1188 points was added to the main data set, while the remaining 4752 points were used for testing purposes. The fitting process was then repeated. However, to prevent deterioration of the fit accuracy in the main, physically relevant region, the following weighting factors were used:

$$w = \begin{cases} 1, & E_{\min} < E < E_{\max} \\ (E_{\min} - E + 1)^{-3}, & E < E_{\min} \\ (E - E_{\max} + 1)^{-3}, & E > E_{\max}, \end{cases} \quad (3.23)$$

where $E_{\min} = -1.51$ kcal/mol and $E_{\max} = 0.69$ kcal/mol are the values of the lowest and highest three-body energies occurring in the 60 trimers present in the three lowest-energy hexamer structures (cage, prism, and book). Such a choice assigns progressively lower weights to trimers with energies far from the physically relevant region (which were included in the training set only to enforce a qualitatively correct behavior in

high-energy regions). The addition of these 1188 points to the training set made the fit to behave well enough for all 5940 points. To select a small subset of “finalists” out of a large set of generated fits (differing, again, by the starting values of the nonlinear parameters), we used several criteria: the overall RMSE, the RMSE on the set of 60 hexamer trimers mentioned above (not included in the training set), and the magnitude of the difference between the total three-body energies in the cage and prism structures. The final fit was selected based on the smallest errors on the testing set of 4752 short-distance geometries. Its RMSE on the initial set of 70,268 geometries amounts to 0.0184 kcal/mol and on the 60 trimers extracted from hexamers to 0.0145 kcal/mol. The former RMSE may be compared to the typical values of nonadditive three-body energies in our 70,268 set which range from -3.52 to 1.94 kcal/mol and to the 0.47 kcal/mol value of their root-mean square. An RMSE of about 0.02 kcal/mol is consistent with that of the two-body potential which is 0.01 kcal/mol (for negative interaction energies) for a single dimer, so it amounts to 0.02 kcal/mol per trimer if the errors are added in squares. Such uncertainties are also consistent with the uncertainties of the *ab initio* calculations estimated in Ref. 43 to be about 0.05 kcal/mol for the water dimer near its van der Waals minimum. Of course, the accuracy of the fit could have been increased easily by using a more elaborate fit function, but such a fit would also take more time in MD simulations. The RMSE of CCpol3 can be compared with that of the fit of Mas *et al.* [28] which was 0.07 kcal/mol and Wang *et al.* [47] which was 0.15 kcal/mol for the 5th-order fit and 0.042 kcal/mol for the 6th-order fit, in all cases relative to the training data set used in a given reference.

3.7 Application to clusters

3.7.1 Water trimer

The main result of this work is a new three-body pairwise nonadditive potential (CCpol3). It was first tested on the 40 trimers selected from the cage and prism hexamers (not included in the training data set) and the results are shown in Table 3.2. As one can see, CCpol3 performs very well on these trimers, with the errors of the fit

with respect to the CBS benchmarks about 3–5 times larger than the errors of the *ab initio* results at the CCSD(T)/(T-D) level of theory used to produce the training set. The performance of CCpol3 is still better on the total three-body contribution to each hexamer energy, with the error even slightly smaller for the prism and two times larger for the cage than the CCSD(T)/(T-D) errors. CCpol3 recovers the benchmark energies in Table 3.2 much better than any other potential. The SAPT-3B potential of Mas *et al.* [28] works reasonably well for the cage, but gives a large error for the prism (the origin of this error will be discussed below). These errors are consistent with the HF theory level and modest size basis sets used in Ref. 28. The WHBB5 potential [47] is of similarly accuracy as SAPT-3B, but the more flexible fit used in WHBB6 significantly improves the accuracy. This potential performs overall better than the HBB2-pol potential [53]. WHBB6, the best performing potential from the literature, gives an RMSE relative to the CBS energies for the cage (prism) isomers of 0.036 (0.031), whereas CCpol3 gives 0.019 (0.019) kcal/mol. We can also compare our results to the DPP2 water model of Kumar *et al.* [83] The nonadditive three-body part of this model was fitted to CCSD(T)-level values for the trimers extracted from the cage, prism, book, and ring hexamers. For the two former hexamers, the DPP2 values from Table V of Ref. 83 have errors only of -0.22 and -0.19 kcal/mol, respectively, relative to the sums of three-body nonadditive energies for each hexamer from their training set (however, the latter values are different by 1.5 kcal/mol from our benchmarks). These deviations compare favorably to the performance of most methods in Table 3.2.

Table 3.4 examines the performance of the CCpol23 potential at the stationary points of the water trimer. The geometries of the stationary points were optimized using CCpol23 and are listed in the Supplementary Information [74]. At each point, CBS interaction energies were computed as a sum of two-body and three-body contributions. The latter contributions were computed as for the trimers investigated in Table 3.1. The former contributions were computed also in the hybrid approach defined by Eqs. (3.6), (3.7), and (3.8), but with $X = 6$ at the HF and MP2 levels and $X = 4$ at the CCSD(T)

frozen-core level. The midbond $3s3p2d2f1g$ basis set, the same as used in Ref. 37, was applied in the dimer calculations. The use of such larger basis sets was necessary since the two-body energies converge slower than the three-body energies. In contrast to all other calculations presented here, only a dimer-centered basis set was applied in the two-body calculations, which is justified at the CBS level with such large values of X [37]. Based on the results in Table III of Ref. 37, one can estimate that such a CBS limit for the total two-body contribution to the trimer interaction energy should be accurate to about 0.01 kcal/mol relative to the exact CCSD(T) frozen-core value. We have then computed the all-electron interaction energies using Eq. (3.6) at the HF level and Eq. (3.7) at the MP2 level with bases aug-cc-pCVTZ and aug-cc-pCVQZ [75]. The $\delta E_{\text{int}}^{\text{CCSD(T)}}$ term was calculated in the aug-cc-pCVTZ basis without any extrapolation. The resulting correction to the frozen-core approximation ranged from -0.118 kcal/mol for the global minimum to -0.092 kcal/mol for the bifurcated transition state. Thus, the inclusion of this effect is absolutely necessary for predicting the total energies at the 0.01 kcal/mol accuracy. Since the core correction is converged to better than 0.01 kcal/mol in the basis set used, the overall accuracy of our all-electron CCSD(T) results is the same as that of the frozen-core values, i.e., 0.01 kcal/mol. Since the three-body nonadditive energies at the CBS level were estimated in Sec. 3.3 to be accurate to at least 0.003 kcal/mol, the overall error of the CBS value is determined by the two-body component.

The *ab initio* interaction energies used as the data set to fit the CCpol2 potential were computed using all electrons but in aug-cc-pVXZ bases which were optimized in frozen core calculations. To check the effects of using such basis sets, we computed the correction to the frozen-core results using the aug-cc-pVTZ and aug-cc-pVQZ bases (i.e., the ones used in Ref. 30 instead of the aug-cc-pCVTZ and aug-cc-pCVQZ ones and found that the results for the six characteristic points were different only by from 0.004 to 0.005 kcal/mol, i.e., negligibly. Thus, the core correction is well reproduced in valence-optimized bases provided that the CBS extrapolations are used.

Table 3.4 shows that the CCpol23 fit recovers the CBS trimer energies with an

RMSE of 0.108 kcal/mol and a maximum absolute error of 0.134 kcal/mol. All the CCpol23 interaction energies lie above the CBS ones. The smallest error is for the highest stationary point, which must be fortuitous. One should recall here that the two-body part of this potential was fitted to *ab initio* interaction energies computed at the following levels: HF-aQZ, MP2-(TQ), CCSD(T)-aTZ, i.e., significantly lower than the level of our current benchmarks. For the global minimum, the total two-body contribution at this level is -12.613 kcal/mol, whereas the CBS result with all electrons correlated is -12.692 kcal/mol. The difference of 0.089 kcal/mol constitutes 66% of the difference between CCpol23 and CBS values. The remaining part of the difference, 0.046 kcal/mol, is consistent with uncertainties resulting from the fitting process: 3×0.011 kcal/mol from the two-body part and 0.018 kcal/mol from the three-body part.

One can also evaluate the performance of CCpol23 by comparing the barriers on the surface, i.e., the differences between the energies of stationary points and the minimum energy. Such barriers are very important for the spectra of the trimer. As Table 3.4 shows, the RMSE of CCpol23 relative to the CBS barriers is 21 cm^{-1} or 0.06 kcal/mol. The error is largest, 0.126 kcal/mol, for the barrier to the highest stationary point which is related to the fact that the energy of this point is so well reproduced that there is virtually no cancellations of errors.

Whereas the CCpol23 predictions agree very well with CBS benchmarks, one may ask how close are the stationary-point CCpol23 geometries to those on flexible-monomer potential energy surface. The 21-dimensional optimizations were performed by Anderson *et al.* [84] at the MP2 level. The simplest test for assessing the closeness of geometries was to compute the CCpol23 interaction energies at "rigidized" Anderson's *et al.* geometries, i.e., for each of the six stationary-point trimers from Ref. 84, we have constructed a trimer with monomers at the geometry used in CCpol23, in way described in Sec. 3.2. The geometries optimized in Ref. 84 were provided to us by the authors of that work (see Ref. 85: note that the original supplementary material of Ref. 84 contains misprints). These rigidized trimer interaction energies turned out to

be very close to the CCpol23 values in the first column of Table 3.4: the differences range from 0.009 to 0.027 kcal/mol, and are below 0.2% of the interaction energies. For the barriers, the differences are between 1.1 and 6.2 cm^{-1} . These comparisons indicate that the two sets of geometries are indeed very close.

Compared to the small differences in barrier heights at CCpol23-optimized and rigidized Anderson *et al.* geometries, both the CCpol23 and CBS barrier heights in Table 3.4 are relatively far from the results of Anderson *et al.*, with RMSE's of 52 and 36 cm^{-1} , respectively. Thus, most of these differences must be due to different monomer geometries in the rigid-monomer and flexible-monomer structures. To check this hypothesis, we have computed a "monomer flexibility correction" $\Delta E_{\text{R}\rightarrow\text{F}}$ defined in the same way as in Ref. 43, i.e., as difference between the total electronic energies of the original trimer and the trimer with rigidized monomers. These corrections were computed using Eqs. (3.6), (3.7), and (3.8) with (45) extrapolations at the HF and MP2 levels, and (34) extrapolations at the CCSD(T) level. The frozen-core approximation was used at the correlated levels. Note that one may view the $\Delta E_{\text{R}\rightarrow\text{F}}$ correction as describing the energetic effect of trimer geometry optimization starting from a rigid-monomer stationary point and relaxing intramonomer coordinates under the conditions of keeping constant the molecular plane, COM, and the position of the line through oxygen and the midpoint of the segment connecting two hydrogen.

The comparisons utilizing the $\Delta E_{\text{R}\rightarrow\text{F}}$ corrections are presented in Table 3.4 for both the CBS and CCpol23 barriers. The barriers from Ref. 84 listed in the table are not the "best estimates" from that work (which include some post-CCSD(T) contributions), but their results at the extrapolated CCSD(T) level. These energies were computed using bases up to a6Z at the MP2 level and up to aQZ at CCSD(T) level and performing several different types of extrapolations which led to estimated uncertainties of the barriers between 6 and 16 cm^{-1} . As the results in Table 3.4 show, the CBS and CCpol23 barriers corrected for monomer-flexibility effects are in excellent agreement with the results from Ref. 84, with RMSE's relative to the values from Ref. 84 almost to within the uncertainties of the latter quantities. These RMSE's

(0.02 and 0.05 kcal/mol, respectively) are also of the size expected from estimates of the uncertainties of our calculations. This agreement is consistent with the finding discussed above that the rigid-monomer CCpol23 potential predicts the intermolecular geometries of the trimer stationary points very well. The largest discrepancies with the Anderson *et al.* results shed some light on accuracy of the CCpol23 potential. For the bifurcated transition state where the difference between CCpol23 and CBS barriers is the largest in magnitude, the CBS barrier corrected for monomer-flexibility effects agrees well with Anderson’s *et al.* value, showing that despite the discrepancy, the geometry of this transition point is accurate. On the other hand, the CBS barrier corrected for monomer-flexibility effects shows the largest discrepancy with Anderson’s *et al.* result for the C_{3h} structure which is a third-order stationary point. It is possible that for this point our geometry optimization was not completely converged. The other possibility is that this point is sensitive to the $\delta E^{\text{CCSD(T)}}$ contribution which was not included in the optimizations of Ref. 84.

One can also compare interaction energies in a similar way, although this comparison is less straightforward than for barriers which are just differences of total electronic energies. In the case of interaction energies, the reference points are different: equilibrium monomers in the case of the interaction energies of Anderson *et al.* [84] and $\langle r \rangle_0$ geometries in the case of CCpol23 and our CBS limit benchmarks. The total trimer interaction energies at the global minima are -15.89 kcal/mol [84] (relaxed CP-corrected energy at CCSD(T) level), -16.06 kcal/mol (CCpol23, vertical energy), and -16.20 kcal/mol (CBS, vertical energy). The total deformation energy of Anderson’s *et al.* structure is 0.40 kcal/mol (computed by us at the CCSD(T)/a5Z level), so that the corresponding vertical interaction energy is -16.29 kcal/mol. These differences may seem small taking into account that the energy of three $\langle r \rangle_0$ monomers lies 0.67 kcal/mol (Ref. 86) above the energy of three r_e monomers. One reason is a partial cancellation of contributions. Let’s measure the energies from the energy of the global trimer minimum with flexible monomers. The energy of the corresponding trimer with

monomers rigidized to the $\langle r \rangle_0$ geometry is at 0.39 kcal/mol (i.e., equal to the magnitude of the monomer-flexibility correction $E_{R \rightarrow F}$). The CBS vertical interaction energy for this geometry is -16.17 kcal/mol (estimated from the value in Table 3.4 and the CCpol23 difference of energies between CCpol23-optimized and Anderson’s *et al.* structures), so that three $\langle r \rangle_0$ monomers are at 16.56 kcal/mol. Subtracting 0.67 kcal/mol from this value gives 15.89 kcal/mol, the magnitude of the relaxed interaction energy of Anderson *et al.* Thus, the partial cancellations in this energetic “cycle” explain to some extent the good agreement of the rigid- and flexible-monomer approaches observed here and later on. The other reason for the closeness of rigid- and flexible-monomer predictions for the equilibrium trimer is that the hydrogen-bonded OH bond length in the latter approach, close to 0.972 Å for all monomers, happens to be the same as the $\langle r \rangle_0$ value. Thus, the “binding ring” is almost identical in the flexible-monomer and rigid $\langle r \rangle_0$ trimer minima. The free OH bonds in flexible-monomer approach are virtually unchanged from the equilibrium value of 0.959 Å and therefore are quite different than in the $\langle r \rangle_0$ monomers, but obviously contribute much less to interaction energy. One may finally note that the increase of the length of the OH bond participating in the hydrogen bond amounting to 0.013 Å is much larger than for the dimer where it is only 0.005 Å (Ref. 7).

A comparison of the performance of the CCpol23 potential with literature potentials on the trimer stationary points optimized using the CCpol23 potential is shown in Fig. 3.1. The quantities compared are the total trimer interaction energies and the CBS benchmarks are those described above. As we already know from Table 3.4, the CCpol23 predictions are remarkably close to the benchmarks. By contrast, the potentials from the literature give predictions with RMSE’s relative to the benchmarks (at CCpol23-optimized geometries) between 0.31 and 0.53 kcal/mol, several times larger than the 0.11 kcal/mol RMSE of CCpol23. Surprisingly, CC-pol-8s+NB performs better than other published potentials, in fact as well as CCpol23 as its RMSE is also 0.11 kcal/mol, probably because trimer tunneling paths were well represented in its training data base. Also surprisingly, WHBB5 predicts the trimer interaction energies slightly

better than WHBB6 despite the simpler form of the fitting function and despite the opposite performance on the trimers extracted from the hexamer. Note that the shape of the diagram is very similar for each of the six stationary points, which stems from the fact that all the structures have a similar, near-equilateral triangle oxygen skeleton and differ mainly (except for one case) by flipping of the non-bonded hydrogen atoms.

We next tested the performance of CCpol3 on a set of 600 random geometries from MD simulations of Ref. 13 which were used already in Sec. 3.5. These geometries were part of our training set, but constituted only a small fraction of the total number of points. The comparisons are made here again on the nonadditive energies only, rather than on the trimer interaction energies. The results for CCpol3 and literature fits are summarized in Table 3.5. This table shows that CCpol3 performs very well in this test, with an RMSE relative to $E_{\text{int}}^{\text{CCSD(T)}}/(\text{T-D})$ values of 0.0154 kcal/mol, very close to the RMSE on the training set and almost 3 times improved relative to the SAPT-3B potential of Ref. 28. Surprisingly, the HHB2-pol potential [53] produces an RMSE only slightly better than that of SAPT-3B, whereas the RMSE of the WHBB potentials [47] is about 1.5 times larger than that of SAPT-3B and very close to that of our pure polarization model described in Sec. 3.5.

To appreciate the significance of these RMSE values, we should compare them to the root mean square value of the three-body energy for the 600 trimers which amounts to 0.187 kcal/mol [13]. Thus, the simple polarization models considered in Ref. 13 result in roughly 50% errors. The use of the current three-center polarization model reduces this error to 39% whereas the use of CCpol23 reduces it to 8%. Most likely the 39% value is close to how well one can reproduce the three-body nonadditive energies with a polarization model based only on asymptotic information. One should mention here, however, that an effective polarization model can be constructed to better reproduce the total three-body nonadditive energies if it is fitted to these values. Such a model is a part of the DPP2 water potential of Kumar *et al.* [83]. Its functional form is fairly close to that in our three-center model, but the parameters were fitted simultaneously to the water monomer polarizability and to a set of *ab initio* computed

total three-body nonadditive energies for trimers extracted from low-energy isomers of the hexamer. As discussed before, this model reproduces very well the sums of the three-body energies for the isomers from their training set. Obviously, this is achieved in a partly unphysical manner as the nonadditive first-order exchange terms, which are significant for all nonasymptotic separations, have a different functional dependence than the polarization terms for which the form of the model is valid.

Scatter plots of WHBB6, HBB2-pol, and our new three-body fit energies as functions of the 600 benchmark energies are presented in Fig. 3.2. The horizontal band at the WHBB6 fit energies equal to zero results from the fact that the WHBB potentials neglect the nonadditive three-body energy if any of the distances between oxygen atoms in the trimer is larger than 6 Å (this occurs for 211 points out of the 600). It is worth noting that, among the 600 geometries, the largest absolute error of the CCpol3 fit amounts to 0.099 kcal/mol and there are only 8 points with errors larger than 0.05 kcal/mol. For WHBB and HBB2-pol, the largest absolute errors (the number of points with errors larger than 0.05 kcal/mol) are 0.411 kcal/mol (125) and 0.218 kcal/mol (63), respectively.

3.7.2 Water hexamer

As mentioned before, the hexamer is the smallest water cluster with stable “three-dimensional” forms (in the sense that the oxygen atoms are highly non-planar). It exists in several low-energy local-minima structures, which has led to a long controversy regarding the most stable isomer [37, 53, 64–66, 71, 87, 88]. Therefore, accurate predictions of the relative energies of various hexamer structures have been recognized as one of the most important tests of water potentials. We investigated six of the lowest structures (prism, cage, book, bag, ring, boat) often considered in the literature, as high-quality comparative benchmarks are available for these configurations. Dahlke *et al.* [66] optimized the geometries at the MP2 level in the haTZ basis set and evaluated the hexamer energies using CCSD(T) in the same basis. Bates and Tschumper [71] used geometries from Ref. 66 and performed analogous calculations for the $\delta E_{\text{int}}^{\text{CCSD(T)}}$

contributions, but their MP2 energies were calculated using the MP2-R12 method [89] which should give results close to the MP2 CBS limit. Note that there are some nomenclature differences in the literature regarding the water hexamer. We follow the convention adopted in Ref. 66, while in Ref. 71 the names “book-1”, “cyclic-boat-1”, and “cyclic-chair” are used for book, boat, and ring, respectively, and two more structures are considered (“book-2” and “cyclic-boat-2”) differing just by the orientation of the free hydrogen atoms at some monomers.

A comparison of the performance of various potentials on the hexamer isomers is presented in Fig. 3.3. The total hexamer interaction energies are given relative to the energy of the prism configuration obtained with a given method. The results for the WHBB5, WHBB6, and HBB2-pol potentials were taken from supplementary material of Refs. 47 and 53. The energies of hexamer isomers were optimized in these references varying all coordinates (i.e., with flexible monomers) using the appropriate potentials. The interaction energies plotted for these potentials as well as the benchmark energies taken from Ref. 66, 71 are the relaxed ones [35]. Geometry optimizations with the CC-pol-8s+NB and CCpol23+ potentials were performed by us and the interaction energies corresponding to these potentials are the vertical ones. Hence, each interaction energy in Fig. 3.3 is self-contained, i.e., was obtained completely within the given approach (except for the benchmark results which were computed at geometries optimized in smaller bases and not including the $\delta E_{\text{int}}^{\text{CCSD(T)}}$ contribution).

Figure 3.3 shows that the interaction energies from CCpol23+ and both benchmarks are very close in all cases, although CCpol23+ is noticeably closer to the CBS-level benchmarks of Bates and Tschumper [71] (0.01 to 0.03 kcal/mol) than to the haTZ results of Dahlke *et al.* [66] (0.03 to 0.06 kcal/mol). This is not accidental, as both the two-body and the three-body parts of CCpol23+ were fitted to energies obtained with basis sets much larger than haTZ. This excellent performance partly reflects the fact that several trimers similar to those present in hexamer structures were used in the development of CCpol23+, as described in Sec. 3.2 (the trimers from the actual hexamer structures used in Fig. 3.3 were not part of our fit data set: it contained

trimers from TIP4P hexamers and distorted trimers from CC-pol-8s+NB optimizations of the hexamers). This level of agreement is remarkable since we compare here relaxed interaction energies of the benchmarks at flexible-monomer geometries with vertical interaction energies of CCpol23+ at rigid-monomer geometries. As discussed in detail for the trimer case, one reason is that the lowering of the total energy resulting from accounting for the monomer-flexibility effects is partly canceled by the subtracting the equilibrium isolated monomer energies.

The HBB2-pol potential [53] works best among the published potentials. It predicts consistently somewhat too large gaps between the prism and the other structures, with errors (relative to the CBS benchmark) in the range of 0.14–0.35 kcal/mol, about an order of magnitude larger than the errors of CCpol23+. The errors of the WHBB6 potential [47] are in the range of 0.25–1.40 kcal/mol, i.e., a few times larger than in the case of HBB2-pol. The errors of the WHBB5 potential [47] are still larger, although for the cage it performs slightly better than WHBB6.

The CC-pol-8s+NB potential gives predictions generally of similar quality to WHBB5, except for the cage structure where it performs worse. Thus, the CC-pol-8s+NB potential fares much worse here than in the trimer tests. Since the two-body components of CC-pol-8s+NB and of CCpol23+ are almost identical and four- and higher-body effects are too small to account for this effect, the bulk of the difference must stem from the lower quality of the three-body fit of Ref. 28. We were able to understand this behavior by analyzing the results for the trimers extracted from the cage and prism hexamers (see the extended version of Table 3.2 in the Supplementary Information [74]). CC-pol-8s+NB performs well on stationary states of the trimer since all such states are similar to the minimum structure that was well represented in the training set of Ref. 28 and most of them lie on the tunneling paths also extensively explored in Refs. 24, 28. By contrast, some of the trimers present in hexamer structures were virtually absent from the training set of Ref. 28. In particular, the two trimers forming the top and the bottom of the prism hexamer contain one water molecule that

is a double-donor of hydrogen bonds. These trimers are poorly predicted by SAPT-3B, with errors with respect to the CBS benchmarks of -0.40 and -0.32 kcal/mol, respectively. These two trimers alone make the prism energy computed from the CC-pol-8s+NB potential significantly too negative and—with the cage energy not suffering of this problem—lead to an excessive cage-prism gap.

In contrast to the comparisons for the trimers, the performance of CCpol23+ relative to flexible-monomer literature potentials, demonstrated in Fig. 3.3, could not have been anticipated. For the trimers, such comparisons were made for rigid monomers (except for some comparisons to the *ab initio* benchmarks at the trimer characteristic points). Therefore the good performance of CCpol23 was expected since it is easier to fit a 12-dimensional than a 21-dimensional potential. For the hexamer, the rigid-monomer CCpol23+ potential achieves a better agreement with the flexible-monomer benchmarks than any flexible-monomer potential. This fact leads to the following answer to the question posed in the Introduction: monomer flexibility effects are less important than an accurate description of each K -body contribution at the rigid-monomer level with the current state-of-the-art of the *ab initio* methods, at least for cluster equilibrium structures.

3.7.3 24-mer

The $(\text{H}_2\text{O})_{24}$ cluster was the subject of an extensive CCSD(T) calculation within the conventional supermolecular framework, *i.e.*, it included the calculation of the whole cluster energy, which took 76 years of CPU time [90, 91]. The authors of that work identified two energetically low structures, labeled 308 and 316, and found their total energies differing by only 0.01 kcal/mol, with structure 316 being more stable. As impressive as this calculation is from the computational point of view, due to the small basis set used (cc-pVTZ with the f functions removed) and the lack of a CP correction, its predictive value is rather limited. An alternative way of calculating the energies of large clusters, the “stratified approximation” many-body approach (SAMBA), was proposed and applied to several water clusters by Góra *et al.* [37] The idea of this

method consists of calculating K -body contributions to the interaction energy (with $K \geq 1$) separately, in basis sets limited to K monomers in each case, thus avoiding the use of the exceedingly large basis sets of the whole cluster. Only low- K contributions need to be considered since the many-body expansion converges fast. In Ref. 37, the consecutive K -body contributions to the energy difference $E_{316} - E_{308}$ were found from *ab initio* calculations to be -0.23 , 1.60 , -1.21 , and -0.24 kcal/mol for $K = 1, 2, 3$, and 4 , respectively, giving the sum equal to -0.08 kcal/mol. Although the uncertainty of the total energy predictions for each structure was estimated in Ref. 37 to be 4.8 kcal/mol, the analysis of the respective differences shows that the value of -0.08 kcal/mol is probably accurate to within 0.1 – 0.2 kcal/mol. The remaining (higher than four-body) contributions to $E_{316} - E_{308}$ were estimated to be probably around 0.3 kcal/mol and not higher than 0.6 kcal/mol. Thus, the SAMBA method provides a rather accurate estimate of the value of $E_{316} - E_{308}$ for such a large cluster although, due to the almost equal energies, the question about the most stable structure remains open. Although the total CPU time of the SAMBA calculations for $(\text{H}_2\text{O})_{24}$ was 200 times shorter than the CPU time needed for the calculations in Ref. 91, it was still a large computational effort. It is therefore an important question how accurate can be the predictions of first-principles water potentials applied to such clusters, as such results can be obtained in mere seconds.

Figure 3.4 shows the differences of the K -body contributions and total interaction energies between the two 24-mers. All K -body contributions shown are vertical whereas the total energy differences are relaxed for flexible-monomer approaches and vertical for rigid-monomers approaches. The total values of $E_{316} - E_{308}$ are 13.95 , 15.62 , 1.20 , 0.68 , and 0.47 kcal/mol from the WHBB5, WHBB6, HBB2-pol, CC-pol-8s+NB, and CCpol23+ potentials, respectively. When we take into account the maximum uncertainty estimate of 0.8 kcal/mol for the benchmark value discussed above, it is seen that both the CCpol-8s+NB and CCpol23+ potentials are within the benchmark error bars, HBB2-pol is reasonably close, while the WHBB potentials are far from it (again, as for the trimer stationary points, the 6th-degree fit performs worse than the

simpler one). Since the $(\text{H}_2\text{O})_{24}$ cluster contains a large number of trimers (2024), the inadequacies of the three-body WHBB fits add up to the values of +12.2 kcal/mol in the case of WHBB5 and +13.8 kcal/mol in the case of WHBB6.

Table 3.6 analyzes the many-body expansion for the 24-mer in more detail, in particular we give the four- and higher-body terms separately for each isomer. We have not included the flexible-monomer potentials from the literature since their polarization models are of similar complexity as the 1-center model included in CC-pol-8s. The many-body expansion was truncated in Ref. 37 at $K = 4$ (the SAMBA contribution in the set of columns denoted in Fig. 3.4 as “>3-body” is the pure four-body contribution), so only terms up this K can be compared. As one can see in Table 3.6, the two-body energies of both isomers are nearly identical for the two potentials and differ from SAMBA results by about 8 kcal/mol or 4%. The 8 kcal/mol differences cancel to a fraction of kcal/mol in the 316-308 difference quantities. In the three-body case, the predictions of the CCpol23+ agree with SAMBA significantly better than those of CC-pol-8s+NB, the discrepancies are about 7–8 and 13 kcal/mol, respectively, and again a significant cancellation of errors takes place in the calculations of difference quantities. In these comparisons one should take into account that all the energies in Table 3.6 (except for the last column) are vertical, whereas the geometries are different in the SAMBA and potential calculations.

The last column of Table 3.6 lists the relaxed interaction energies. The average deformation correction per monomer of 0.8 kcal/mol is about six times larger than in the global minimum of the water trimer. One reason is that in the investigated 24-mers a large fraction of monomers have both hydrogens participating in hydrogen bonds, which leads to elongation of both intramonomer OH bonds. In contrast, in the trimer only one OH bond is elongated. Second, this elongation increases with cluster size in compact clusters (the OH bond length in ice is 1.01 Å). Finally, the optimizations of 24-mers were performed [92] at the MP/aDZ level and the length of the isolated monomer OH bond is overestimated at this level by 0.006 Å.

We have already seen the very good performance of both polarization models on

the tetramers extracted from hexamers, as discussed in Sec. 3.5. This performance is equally good for the 24-mer, with errors of the three-center model amounting to 7% for isomer 316 and 0.6% for 308. Although the 316-308 difference predicted by CCpol23+ is of different sign than predicted by SAMBA, the magnitude of the discrepancy amounts only to 0.4 kcal/mol, i.e., is rather small.

If the terms with $K = 2-4$ are added together, the differences between the CCpol23+ and SAMBA predictions, i.e., the sums of differences discussed above, are about 16 kcal/mol. However, if the monomer distortion corrections from Ref. 37 are added to the SAMBA results, the discrepancies decrease to only about 3 kcal/mol.

For the $K > 4$ -body effect, we do not have *ab initio* results to compare with. However, in view of the good performance at the four-body level and the previous observations for the hexamer, the N -body polarization models included in the CC-pol-8s and CCpol23+ potentials should provide reasonable estimates of these nonadditive many-body effects. Table 3.6 shows that the five- and higher-body effects give a very small contribution to the total interaction energies of both isomers, on the order of 0.1%. Thus, the magnitude of these estimates is in agreement with the estimates made in Ref. 37. In contrast to the lower- K contributions, the high- K terms do not cancel out between the 316 and 308 isomers and the difference is of the same order of magnitude as the contributions for individual monomers. One may be tempted to add these differences to the SAMBA prediction (which was truncated at four-body interactions and with the one-body term included gives $E_{316} - E_{308} = -0.08$ kcal/mol). Such an addition gives the amended SAMBA value of $E_{316} - E_{308} = +0.11$ kcal/mol, in even better agreement with the CCpol23+ value of $E_{316} - E_{308} = +0.47$ kcal/mol.

The results for the water 24-mer show that the performance of the CCpol23+ potential in predictions for clusters such as $(\text{H}_2\text{O})_{24}$ is competitive with the most advanced applicable *ab initio* methods, while the latter require (in this case) an about seven orders of magnitude larger computational effort. Analogously to the case of the hexamer, it should be pointed out that the inclusion of the monomer-flexibility effects makes little difference in the quality of predictions, despite the large deformations of

monomers. The CC-pol-8s+NB and CCpol23+ potentials give such good predictions despite being evaluated at the rigidized geometries of Ref. 91.

3.8 Summary and Conclusions

A new first-principles three-body pairwise-nonadditive interaction energy potential has been developed for water using the rigid-monomer approximation with monomers in their average rovibrational ground-state geometry. This 12-dimensional surface was fitted to 71,456 *ab initio* three-body nonadditive interaction energies obtained using a hybrid approach that combines results computed at the CCSD(T) level using the aug-cc-pVDZ basis with those computed at the MP2 level using the aug-cc-pVTZ basis. This level of *ab initio* theory gives more than sufficient accuracy as the RMSE's relative to CBS benchmarks on trimers extracted from equilibrium hexamers are only 0.004 and 0.006 kcal/mol for the cage and prism hexamers, respectively.

The functional form of the fit was motivated by the SAPT decomposition of the nonadditive energy into physical components. It included a new, three-center damped polarization model. This model alone recovered relatively well the three-body nonadditive effects as its RMSE on 600 trimers selected randomly from snapshots of MD simulations in ambient conditions was 0.073 kcal/mol. This accuracy is similar to that of some recent potentials fitted to *ab initio* calculations such as WHBB5 [47] and significantly better than that of single-center polarization potentials such as those used in Ref. 13 which gave an RMSE of 0.107 kcal/mol. This polarization model was combined with terms describing exchanges of electrons between monomers. Significant care was taken to ensure that the fit behaves physically at very short separations between nuclei, a feature important for molecular simulations. The final fit, denoted as CCpol3, uses 63 nonlinear and 364 linear adjustable parameters and its RMSE on the subset of 70,268 grid points that excludes geometries with very small intermonomer separations is 0.0145 kcal/mol. To make our three-body potential consistent with the two-body one, the CC-pol-8s potential from Ref. 43 was refitted using the new polarization model and additional damping functions for very short separations. The

accuracy of the new two-body fit, denoted as CCpol2, is unchanged compared to CCpol-8s in the physical region.

As a byproduct of this work, we have found that the polarization model recovers very well the four-body nonadditive interaction energies, with errors on the order of 10% rather than the 50% errors observed in Ref. 13 in the three-body case. Apparently, this very positive fact has not been noted earlier: it is of significant importance in developing many-body force fields. In the case of water hexamers, we have also compared the performance of the polarization model for five- and six-body contributions. Whereas this performance was significantly worse than for the four-body terms, the worst performance was for contributions of negligible magnitude. The sum of the four- through six-body effects was recovered by the polarization model of CCpol23+ with the average magnitude of the relative error over the six isomers amounting to only 3.5%.

CCpol3 was first tested on the nonadditive interaction energies of water trimers. On the set of trimers contained in the cage and prism hexamers, the RMSE of the fit is 0.019 kcal/mol for each isomer. On the set of 600 hexamers from MD simulations, the RMSE is 0.0154 kcal/mol, 2.4 times smaller than given by the HBB2-pol potential [53] which performed best of the published potentials.

We then tested the sum of CCpol3 and of CCpol2, denoted as CCpol23, on the total trimer interaction energies at the six trimer stationary geometries. The geometries optimized using CCpol23 were found to be close to those of the *ab initio* optimizations of Ref. 84 with flexible monomers after the latter geometries were rigidized. The RMSE of the CCpol23 predictions relative to the CBS benchmarks computed by us is only 0.11 kcal/mol and it is a few times smaller than given by the best potentials from the literature.

The remaining tests were performed on the water hexamer and 24-mer using the CCpol23+ potential, i.e., CCpol23 plus the four- and higher-body interactions represented by the polarization model. In contrast to the tests described so far in this section, we compared to cluster energies computed using flexible-monomer approaches

for those methods that do not use rigid-monomer approximation. Somewhat surprisingly, despite the rigid-monomer approximation, the energetic predictions of CCpol23+ are closer to the best *ab initio* benchmarks (which include monomer-flexibility effects) than the predictions of the best published flexible-monomer models. We believe that the main reason for this behavior is the high accuracy of our three-body potential. For hexamers, the energies of the five considered isomers relative to the energy of the lowest-energy prism isomer are within 0.01–0.03 kcal/mol of the Bates-Tschumper [71] benchmarks, i.e., are within the uncertainties of the *ab initio* calculations. The second-best prediction is given by the HBB2-pol potential, but the errors of 0.14–0.35 kcal/mol are larger by an order of magnitude. For the 24-mer, only CCpol23+ predicts the difference between the two lowest-energy “308” and “316” isomers, estimated from *ab initio* calculations to be -0.1 ± 0.8 kcal/mol, within the error bars. It gives the value of 0.47 kcal/mol for this difference, whereas the best performing flexible-monomer potential, HBB2-pol, predicts 1.20 kcal/mol.

The comparison for hexamers and 24-mers suggests that at the currently possible level of accuracy, the residual errors in intermolecular part of flexible-monomer potentials are larger than the monomer-flexibility effects on cluster energetics. Hence, at this point it is more advantageous from a physical point of view to improve the accuracy of trimer rigid-monomer potentials rather than to develop flexible-monomer potentials which cannot be sufficiently accurate in their dependence on the intermolecular coordinates due to a too low density of grid points. This is in contrast to the two-body case where one can now develop very accurate 12-dimensional potentials [41, 44, 48, 52, 86]

Our work also indicates an issue concerning the accuracy of the potentials used in recent investigations of hexamer structures aimed at determining the energetic order of isomers [53, 64, 65]. Since the difference in the interaction energies of the prism and cage isomers is as small as 0.25 kcal/mol [37, 71], only CCpol23+ is sufficiently accurate among existing potentials to correctly recover this quantity.

Since the CCpol23+ accuracy for the hexamer is close to that of the state-of-the-art *ab initio* benchmarks, whereas for the 24-mer it is likely higher, this potential can be used to generate high-accuracy benchmarks for large water clusters with costs completely negligible compared to any *ab initio* calculations. We also expect this potential to find broad applications in predicting the properties of the condensed phases of water, including the anomalous properties. The results for the clusters suggest that the rigid-monomer character of the potential does not impair the accuracy of the CCpol23+ predictions significantly, although for systems with large distortion of the monomer geometries the potential may not work that well.

3.9 Acknowledgments

The authors would like to thank Omololu Akin-Ojo for several discussions and for providing the set of trimers from MD simulations. This research was supported by the NSF grant CHE-1152899. RP would like to acknowledge support from Polish National Science Centre grant No. DEC-2012/05/B/ST4/00086.

Bibliography

- [1] O. Matsuoka, E. Clementi, and M. Yoshimine, *J. Chem. Phys.* **64**, 1351 (1976).
- [2] B. Jeziorski and M. van Hemert, *Mol. Phys.* **31**, 713 (1976).
- [3] K. Szalewicz, S. J. Cole, W. Kolos, and R. J. Bartlett, *J. Chem. Phys.* **89**, 3662 (1988).
- [4] E. M. Mas and K. Szalewicz, *J. Chem. Phys.* **104**, 7606 (1996).
- [5] W. Klopper, J. G. C. M. van Duijneveldt-van de Rijdt, and F. B. van Duijneveldt, *Phys. Chem. Chem. Phys.* **2**, 2227 (2000).
- [6] G. S. Tschumper, M. L. Leininger, B. C. Hoffman, E. F. Valeev, H. F. Schaefer III, and M. Quack, *J. Chem. Phys.* **116**, 690 (2002).
- [7] J. R. Lane, *J. Chem. Theory Comput.* **9**, 316 (2012).
- [8] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).
- [9] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, *J. Phys. Chem.* **91**, 6269 (1987).
- [10] K. Szalewicz, C. Leforestier, and A. van der Avoird, *Chem. Phys. Lett.* **482**, 1 (2009).
- [11] L.-P. Wang, T. Head-Gordon, J. W. Ponder, P. Ren, J. D. Chodera, P. K. Eastman, T. J. Martinez, and V. S. Pande, *J. Phys. Chem. B* **117**, 9956 (2013).
- [12] K. T. Wikfeldt, E. R. Batista, F. D. Vilazc, and H. Jonsson, *Phys. Chem. Chem. Phys.* **15**, 16542 (2013).
- [13] O. Akin-Ojo and K. Szalewicz, *J. Chem. Phys.* **138**, 024316 (2013).

- [14] R. S. Fellers, C. Leforestier, L. B. Braly, M. G. Brown, and R. J. Saykally, *Science* **284**, 945 (1999).
- [15] R. S. Fellers, L. B. Braly, R. J. Saykally, and C. Leforestier, *J. Chem. Phys.* **110**, 6306 (1999).
- [16] N. Goldman, R. S. Fellers, M. G. Brown, L. B. Braly, C. J. Keoshian, C. Leforestier, and R. J. Saykally, *J. Chem. Phys.* **116**, 10148 (2002).
- [17] N. Goldman, C. Leforestier, and R. J. Saykally, *Phil. Trans. Royal Soc. (London), Ser. A* **363**, 493 (2005).
- [18] K. Szalewicz, R. Bukowski, and B. Jeziorski, in *Theory and Applications of Computational Chemistry: The First 40 Years. A Volume of Technical and Historical Perspectives*, edited by C. E. Dykstra, G. Frenking, K. S. Kim, and G. E. Scuseria (Elsevier, Amsterdam, 2005) Chap. 33, pp. 919–962.
- [19] E. Clementi and P. Habitz, *J. Phys. Chem.* **87**, 2815 (1983).
- [20] U. Niesar, G. Corongiu, E. Clementi, G. R. Kneller, and D. K. Bhattacharya, *J. Phys. Chem.* **94**, 7949 (1990).
- [21] C. Millot and A. J. Stone, *Mol. Phys.* **77**, 439 (1992).
- [22] E. M. Mas, K. Szalewicz, R. Bukowski, and B. Jeziorski, *J. Chem. Phys.* **107**, 4207 (1997).
- [23] C. Millot, J. C. Soetens, M. T. C. M. Costa, M. P. Hodges, and A. J. Stone, *J. Phys. Chem.* **102**, 754 (1998).
- [24] G. C. Groenenboom, E. M. Mas, R. Bukowski, K. Szalewicz, P. E. S. Wormer, and A. van der Avoird, *Phys. Rev. Lett.* **84**, 4072 (2000).
- [25] E. M. Mas, R. Bukowski, K. Szalewicz, G. C. Groenenboom, P. E. S. Wormer, and A. van der Avoird, *J. Chem. Phys.* **113**, 6687 (2000).

- [26] G. C. Groenenboom, P. E. S. Wormer, A. van der Avoird, E. M. Mas, R. Bukowski, and K. Szalewicz, *J. Chem. Phys.* **113**, 6702 (2000).
- [27] M. J. Smit, G. C. Groenenboom, P. E. S. Wormer, A. van der Avoird, R. Bukowski, and K. Szalewicz, *J. Phys. Chem. A* **105**, 6212 (2001).
- [28] E. M. Mas, R. Bukowski, and K. Szalewicz, *J. Chem. Phys.* **118**, 4386 (2003).
- [29] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *Science* **315**, 1249 (2007).
- [30] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *J. Chem. Phys.* **128**, 094313 (2008).
- [31] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *J. Chem. Phys.* **128**, 094314 (2008).
- [32] A. Shank, Y. Wang, A. Kaledin, B. J. Braams, and J. M. Bowman, *J. Chem. Phys.* **130**, 144314 (2009).
- [33] B. E. Rocher-Casterline, L. C. Ch'ng, A. K. Mollner, and H. Reisler, *J. Chem. Phys.* **134**, 211101 (2011).
- [34] O. Mishima and H. E. Stanley, *Nature* **392**, 164 (1998).
- [35] K. Szalewicz and B. Jeziorski, *J. Chem. Phys.* **109**, 1198 (1998).
- [36] E. M. Mas, R. Bukowski, and K. Szalewicz, *J. Chem. Phys.* **118**, 4404 (2003).
- [37] U. Góra, R. Podeszwa, W. Cencek, and K. Szalewicz, *J. Chem. Phys.* **135**, 224102 (2011).
- [38] C. J. Burnham and S. S. Xantheas, *J. Chem. Phys.* **116**, 1500 (2002).
- [39] C. J. Burnham and S. S. Xantheas, *J. Chem. Phys.* **116**, 5115 (2002).
- [40] G. S. Fanourgakis and S. S. Xantheas, *J. Phys. Chem. A* **110**, 4100 (2006).

- [41] K. Szalewicz, G. Murdachaew, R. Bukowski, O. Akin-Ojo, and C. Leforestier, in *Lecture Series on Computer and Computational Science: ICCMSE 2006*, Vol. 6, edited by G. Maroulis and T. Simos (Brill Academic Publishers, Leiden, 2006) pp. 482–491.
- [42] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *J. Chem. Phys.* **125**, 044301 (2006).
- [43] W. Cencek, K. Szalewicz, C. Leforestier, R. van Harrevelt, and A. van der Avoird, *Phys. Chem. Chem. Phys.* **10**, 4716 (2008).
- [44] X. Huang, B. J. Braams, and J. M. Bowman, *J. Phys. Chem. A* **110**, 445 (2006).
- [45] X. Huang, B. J. Braams, J. M. Bowman, R. E. A. Kelly, J. Tennyson, G. C. Groenenboom, and A. van der Avoird, *J. Chem. Phys.* **128**, 034312 (2008).
- [46] Y. Wang, B. C. Shepler, B. J. Braams, and J. M. Bowman, *J. Chem. Phys.* **131**, 054511 (2009).
- [47] Y. Wang, X. Hunag, B. C. Shepler, B. J. Braams, and J. M. Bowman, *J. Chem. Phys.* **134**, 094509 (2011).
- [48] C. Leforestier, K. Szalewicz, and A. van der Avoird, *J. Chem. Phys.* **137**, 014305 (2012).
- [49] V. Babin, G. R. Medders, and F. Paesani, *J. Phys. Chem. Lett.* **3**, 3765 (2012).
- [50] G. R. Medders, V. Babin, and F. Paesani, *J. Chem. Theory Comput.* **9**, 1103 (2013).
- [51] G. R. Medders and F. Paesani, *J. Chem. Theory Comput.* **9**, 4844 (2013).
- [52] V. Babin, C. Leforestier, and F. Paesani, *J. Chem. Theory Comput.* **9**, 5395 (2013).
- [53] V. Babin and F. Paesani, *Chem. Phys. Lett.* **580**, 1 (2013).

- [54] M. Jeziorska, P. Jankowski, K. Szalewicz, and B. Jeziorski, *J. Chem. Phys.* **113**, 2957 (2000).
- [55] B. Jeziorski, R. Moszyński, and K. Szalewicz, *Chem. Rev.* **94**, 1887 (1994).
- [56] K. Szalewicz, *Wiley Interdisc. Rev.-Comp. Mol. Sci.* **2**, 254 (2012).
- [57] V. F. Lotrich and K. Szalewicz, *J. Chem. Phys.* **106**, 9668 (1997).
- [58] V. F. Lotrich and K. Szalewicz, *J. Phys. Chem.* **106**, 9688 (1997).
- [59] V. F. Lotrich and K. Szalewicz, *Phys. Rev. Lett.* **79**, 1301 (1997).
- [60] A. van der Avoird and K. Szalewicz, *J. Chem. Phys.* **128**, 014302 (2008).
- [61] T. James, D. Wales, and J. Hernández-Rojas, *Chem. Phys. Lett.* **415**, 302 (2005).
- [62] M. Mahoney and W. Jorgensen, *J. Chem. Phys.* **112**, 8910 (2000).
- [63] W. Cencek, M. Jeziorska, O. Akin-Ojo, and K. Szalewicz, *J. Phys. Chem. A* **111**, 11311 (2007).
- [64] C. Perez, M. T. Muckle, D. P. Zaleski, N. A. Seifert, B. Temelso, G. C. Shields, Z. Kisiel, and B. H. Pate, *Science* **336**, 897 (2012).
- [65] Y. M. Wang, V. Babin, J. M. Bowman, and F. Paesani, *J. Am. Chem. Soc.* **134**, 11116 (2012).
- [66] E. E. Dahlke, R. M. Olson, H. R. Leverentz, and D. G. Truhlar, *J. Phys. Chem. A* **112**, 3976 (2008).
- [67] M. J. D. Powell, *Comp. J.* **7**, 155 (1964).
- [68] H.-J. Werner, P. J. Knowles, R. Lindh, M. Schütz, *et al.*, “Molpro, version 2009.1, a package of ab initio programs,” (2009), see <http://www.molpro.net>.
- [69] S. F. Boys and F. Bernardi, *Mol. Phys.* **19**, 553 (1970).

- [70] R. A. Kendall, T. H. Dunning, Jr., and R. J. Harrison, *J. Chem. Phys.* **96**, 6796 (1992).
- [71] D. M. Bates and G. S. Tschumper, *J. Phys. Chem. A* **113**, 3555 (2009).
- [72] A. Halkier, W. Klopper, T. Helgaker, P. Jørgensen, and P. R. Taylor, *J. Chem. Phys.* **111**, 9157 (1999).
- [73] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, and J. Olsen, *Chem. Phys. Lett.* **302**, 437 (1999).
- [74] See EPAPS Document No. ??????. For more information on EPAPS, see <http://www.aip.org/pubservs/epaps.html>.
- [75] D. E. Woon and T. H. Dunning Jr., *J. Chem. Phys.* **103**, 4572 (1995).
- [76] K. T. Tang and J. P. Toennies, *J. Chem. Phys.* **80**, 3726 (1984).
- [77] A. J. Misquitta, private communication (2009).
- [78] A. J. Misquitta and A. J. Stone, “CamCASP: a program for studying intermolecular interactions and for calculations of molecular properties in distributed form,” University of Cambridge, UK (2010).
- [79] J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [80] C. Adamo and V. Barone, *J. Chem. Phys.* **110**, 6158 (1999).
- [81] D. E. Woon and T. H. Dunning Jr., *J. Chem. Phys.* **100**, 2975 (1994).
- [82] V. F. Lotrich, H. L. Williams, K. Szalewicz, B. Jeziorski, R. Moszynski, P. E. S. Wormer, and A. van der Avoird, *J. Chem. Phys.* **103**, 6076 (1995).
- [83] R. Kumar, F. F. Wang, G. R. Jenness, and K. D. Jordan, *J. Chem. Phys.* **132**, 014309 (2010).

- [84] J. A. Anderson, K. Crager, L. Fedoroff, and G. S. Tschumper, *J. Chem. Phys.* **121**, 11023 (2004).
- [85] G. S. Tschumper, http://quantum.chem.olemiss.edu/026_DATA.
- [86] P. Jankowski, G. Murdachaew, R. Bukowski, O. Akin-Ojo, C. Leforestier, and K. Szalewicz, “*Ab initio* water pair potential with flexible monomers,” (2014), submitted.
- [87] K. Liu, M. G. Brown, C. Carter, R. J. Saykally, J. K. Gregory, and D. C. Clary, *Nature* **381**, 501 (1996).
- [88] S. S. Xantheas, C. J. Burnham, and R. J. Harrison, *J. Chem. Phys.* **116**, 1493 (2002).
- [89] W. Kutzelnigg and W. Klopper, *J. Chem. Phys.* **94**, 1985 (1991).
- [90] E. Apra, R. J. Harrison, W. A. deJong, A. P. Rendell, V. Tipparaju, and S. S. Xantheas, in *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis* (ACM, New York, 2010) article number 66.
- [91] E. Apra, R. J. Harrison, W. A. deJong, A. P. Rendell, V. Tipparaju, and S. S. Xantheas, in *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis; extended version* (ACM, New York, 2010) article number 66 (extended version).
- [92] S. Yoo, M. V. Kirov, and S. S. Xantheas, *J. Am. Chem. Soc.* **131**, 7564 (2009).

Table 3.1: Comparisons of errors of three-body nonadditive interaction energies (in kcal/mol) computed relative to the CBS extrapolated values. The latter energies were obtained in a hybrid way as defined by Eqs. (3.6), (3.7), and (3.8) with $X = 5$ at the HF and MP2 levels and $X = 4$ at the CCSD(T) level. All the calculations were performed in trimer-centered basis sets. The trimer geometries were extracted from cage and prism hexamers optimized using the CC-pol-8s+NB potential. There are 20 trimers in each hexamer. The geometries of all trimers are given in the Supplementary Information [74]. ‘Sum’ corresponds to the sum of the signed errors for individual trimers and is equal to the error in $E_{\text{int}}[3,6]$ for the corresponding hexamers. The root mean square errors (RMSE), mean absolute errors (MAE), and maximum absolute errors (MAXE) are also given. The $E_{\text{int}}[3,6](\text{CBS})$ values are equal to -8.121 and -7.999 kcal/mol for cage and prism hexamers, respectively. An extended version of this table containing the contributions for individual trimers is included in the Supplementary Information [74].

	MP2	CCSD(T)			hybrid	
	aTZ	aDZ	haTZ	aTZ	aQZ	aT-aD
			cage			
Sum	-0.341	0.039	-0.036	0.036	0.013	0.031
RMSE	0.024	0.012	0.004	0.003	0.001	0.004
MAE	0.018	0.010	0.003	0.003	0.001	0.004
MAXE	0.057	0.033	0.009	0.007	0.003	0.009
			prism			
Sum	-0.405	0.032	-0.040	0.037	0.014	0.028
RMSE	0.028	0.019	0.007	0.004	0.002	0.006
MAE	0.020	0.014	0.005	0.003	0.001	0.004
MAXE	0.073	0.048	0.019	0.008	0.003	0.013

Table 3.3: Comparison of the four-, five-, and six-body nonadditive energies (in kcal/mol) of various structures of the water hexamer (optimized using the CC-pol-8s+NB potential of Ref. 43), calculated with CCSD(T), with one-center polarization model from Ref. 28, and with the current three-center polarization model. The CCSD(T) calculations were performed in the aTZ basis set using full hexamer-centered bases for each cluster. The values in parentheses correspond to $E_{\text{int}}^{\text{CCSD(T)}/(\text{T-D})}$ calculations in tetramer-centered basis sets (these values were the data used in the optimizations of the polarization potential).

	prism			cage			book		
	4-B	5-B	6-B	4-B	5-B	6-B	4-B	5-B	6-B
CCSD(T)	-0.5737 (-0.5708)	0.0562	0.00076	-0.4342 (-0.4332)	0.0027	-0.0014	-0.8692 (-0.8667)	-0.0311	-0.0028
1-center	-0.7037	0.0397	0.00193	-0.4402	0.0081	-0.0018	-0.8167	-0.0349	-0.0037
3-center	-0.6435	0.0354	0.00096	-0.4516	0.0129	-0.0025	-0.8638	-0.0423	-0.0049
	bag			ring			boat		
	4-B	5-B	6-B	4-B	5-B	6-B	4-B	5-B	6-B
CCSD(T)	-0.9204 (-0.9172)	-0.0141	0.00567	-1.4529 (-1.4487 ^a)	-0.1494	-0.0079	-1.3392 (-1.3347)	-0.1287	-0.0078
1-center	-0.8427	-0.0314	0.00395	-1.4357	-0.1848	-0.0164	-1.2486	-0.1468	-0.0132
3-center	-0.8386	-0.0402	0.00485	-1.3969	-0.1887	-0.0176	-1.2648	-0.1584	-0.0143

^a An incorrect value of this energy equal to -1.2811 kcal/mol was used during optimization.

Table 3.4: Trimer stationary-point interaction energies (in kcal/mol) and barriers relative to the energy at the minimum (in cm^{-1}) at geometries optimized with the CCpol23 potential. The levels of CBS calculations are described in the text and include all-electron correlation in the two-body part. The monomer-flexibility correction $\Delta E_{\text{R}\rightarrow\text{F}}$ [in cm^{-1}] relative to its value at the minimum is defined as in Ref. 43. The notation for minima, stationary points (SP), and transition states (TS) follows Ref. 84. The conversion factor from kcal/mol to cm^{-1} was 349.7550.

stationary point	E_{int}		ΔE_{int}		E_{int}		ΔE_{int}		$\Delta E_{\text{R}\rightarrow\text{F}}$		ΔE_{int}	
	CCpol23	CCpol23	CCpol23	CCpol23	CBS	CBS	CBS	CBS	CBS	CBS	CCpol23	Ref. 84
minimum [uud]	-16.061		104.4		-16.196							
$\text{C}_1 \rightarrow \text{C}_1$ TS [udp]	-15.763		104.4		-15.887		108.1		-21.0		83.4	82
C_3 minimum [uuu]	-15.268		277.3		-15.389		282.0		-14.3		263.0	269
$\text{C}_1 \rightarrow \text{C}_3$ TS [uup]	-15.216		295.6		-15.336		300.7		-23.5		272.1	275
$\text{C}_{3\text{h}}$ SP [ppp]	-14.598		511.8		-14.687		527.8		-70.0		441.8	440
Bifurcated TS [upbi]	-13.829		780.8		-13.838		824.6		-61.2		719.7	760
RMSE	0.108 ^a		21 ^a								8 ^b	

^a With respect to CBS results.

^b With respect to the barriers of Ref. 84.

Table 3.5: Comparison of RMSE's (in kcal/mol) of nonadditive three-body energies on 600 trimers selected from snapshots of MD simulations in Ref. 13.

polarization model (Ref. 13)	0.107
polarization model (present work)	0.0734
SAPT-3B (Ref. 28)	0.0418
WHBB5 (Ref. 47)	0.0735
WHBB6 (Ref. 47)	0.0642
HBB2-pol (Ref. 53)	0.0374
CCpol3	0.0154

Table 3.6: Many-body contributions (in kcal/mol) to the interaction energies of the 316 and 308 water 24-mers. SAMBA results are from Ref. 37.

	isomer 316								
	2-B	3-B	4-B	5-B	>5-B	>4-B	\sum_2^4	\sum_1^4	
CC-pol-8s+NB	-183.856	-45.675	-6.664	-0.349	0.086	-0.263	-236.195		
CCpol23+	-183.888	-51.251	-6.498	-0.320	0.075	-0.245	-241.637		
SAMBA	-191.532	-59.002	-7.001				-257.534	-238.719	
	isomer 308								
CC-pol-8s+NB	-184.800	-44.365	-7.381	-0.611	0.019	-0.592	-236.546		
CCpol23+	-184.889	-50.310	-6.722	-0.461	0.030	-0.431	-241.921		
SAMBA	-193.128	-57.791	-6.765				-257.864	-238.638	
	316–308								
CC-pol-8s+NB	0.945	-1.310	0.716	0.262	0.067	0.329	0.351		
CCpol23+	1.001	-0.942	0.224	0.141	0.045	0.186	0.284		
SAMBA	1.596	-1.211	-0.236				0.150	-0.081	

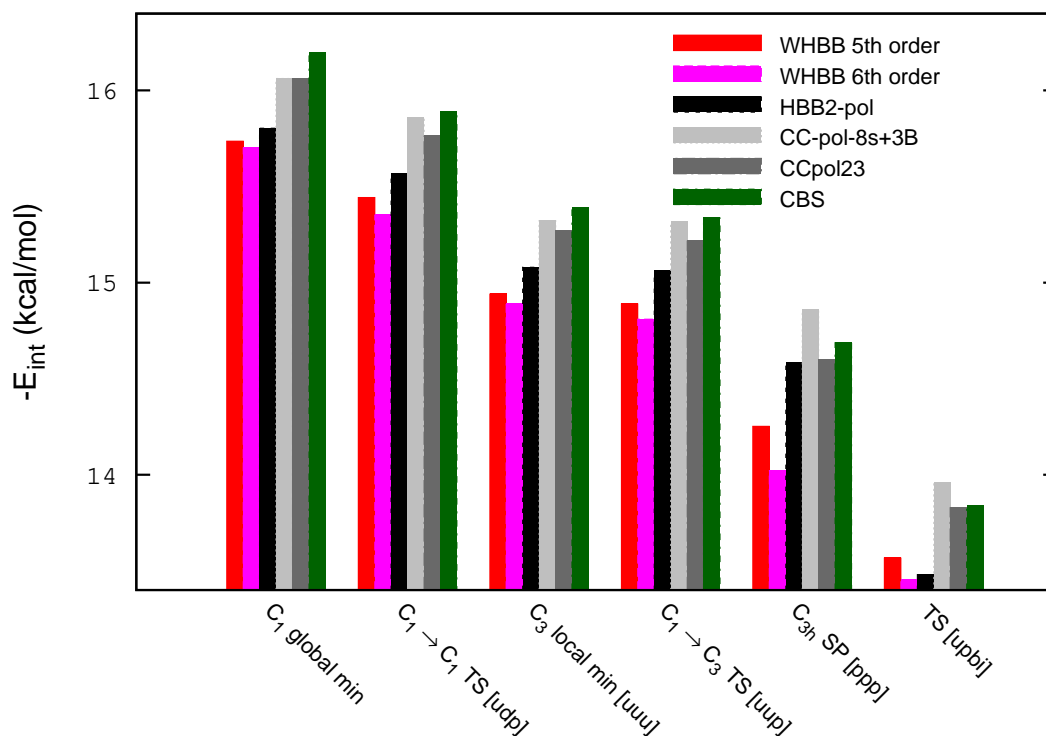


Figure 3.1: Comparison of the total interaction energies of the water trimer computed using CCpol23 and potentials from the literature. The geometries are those of stationary points optimized using CCpol23 and are the same for all methods. The CBS results were obtained as described in the text. The WHBB5 and WHBB6 potentials are from Ref. 47, whereas HBB2-pol is from Ref. 53. The vertical interaction energies produced by these potentials were computed by us.

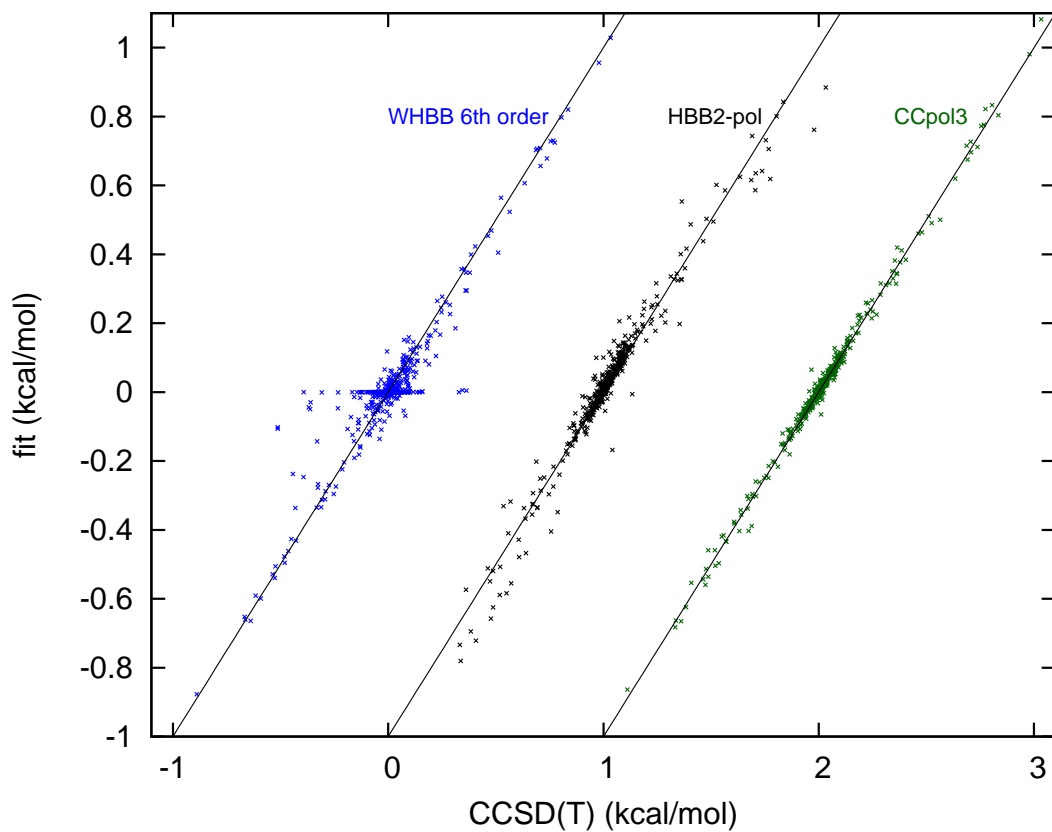


Figure 3.2: Nonadditive three-body energies of 600 trimer configurations selected from snapshots of MD simulations in Ref. 13 calculated from WHBB6 [47], HBB2-pol [53], and CCpol3 fits and compared with CCSD(T)/(T-D) values. The straight lines represent the ideal case (*i.e.*, fit energies equal to CCSD(T) energies). Note that the CCSD(T) energies are augmented by 1 and 2 kcal/mol in the HBB2-pol and CCpol3 plots, respectively.

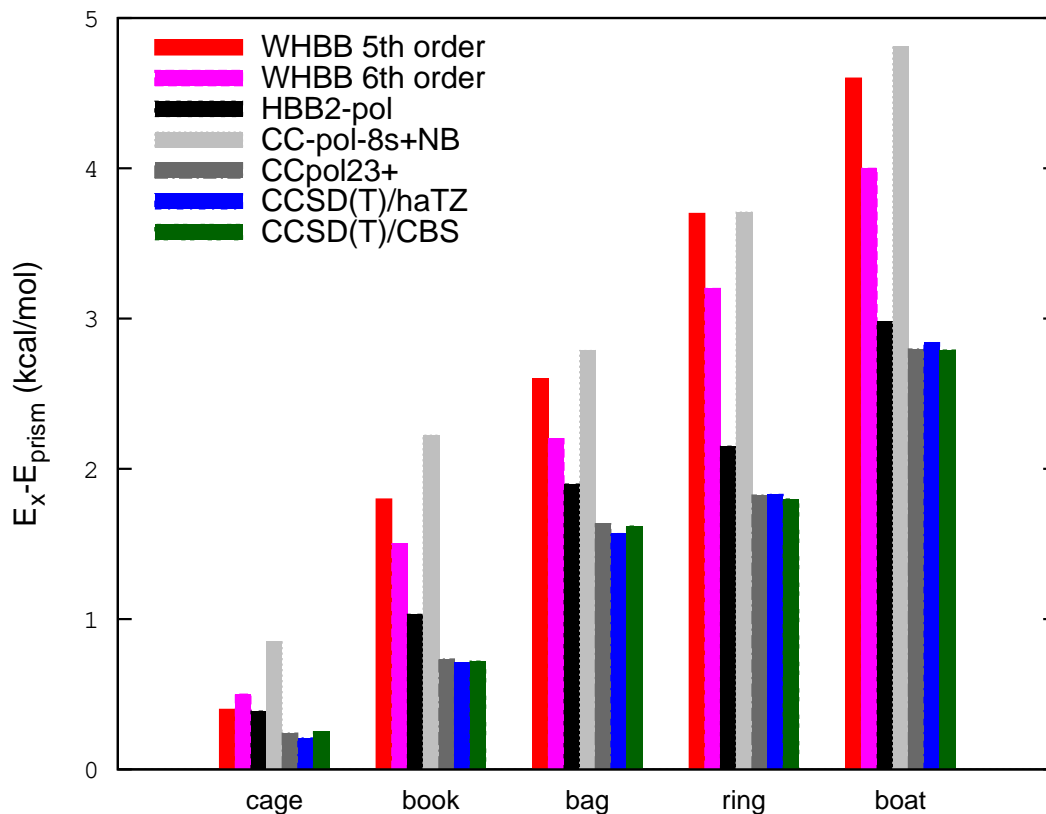


Figure 3.3: Energies of various structures of the water hexamer relative to the lowest structure (prism). In each case, the geometry optimization was performed with the same method as the subsequent energy calculation, except for the CCSD(T) energies which were obtained at MP2-optimized geometries. The WHBB and HBB2-pol results are taken from supplementary material of Refs. 47 and 53, respectively. The CCSD(T) results in the haTZ basis are from Ref. 66. The results denoted as CCSD(T)/CBS are from Ref. 71 and were computed using the MP2-R12 method plus $\delta E_{\text{int}}^{\text{CCSD(T)}/\text{haTZ}}$. The flexible-monomer results include monomer-relaxation corrections.

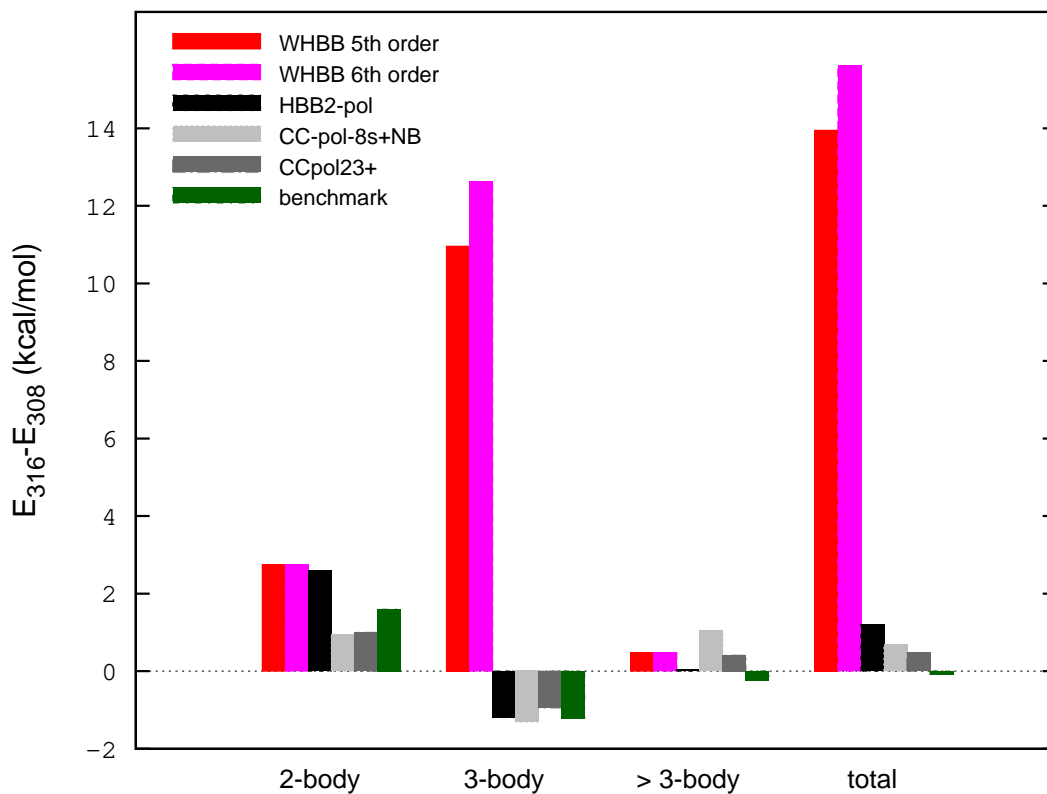


Figure 3.4: Differences of the K -body and total interaction energies between the “316” and “308” structures of the water 24-mer. The benchmark *ab initio* results are taken from Ref. 37. The K -body contributions shown are in all cases the vertical interaction energies. The $K > 3$ -body contribution is the pure four-body effect in the case of the benchmark. The total interaction energy differences from the flexible-monomer potentials as well as the *ab initio* benchmarks for this quantity include the one-body term. The WHBB and HBB2-pol values were obtained by us at the geometries of Ref. 91. The CC-pol-8s+NB and CCpol23+ results were obtained for the “rigidized” structures, *i.e.*, with all monomers within the clusters transformed to their $\langle r \rangle_0$ -monomer geometries and the total interaction energy differences for these potentials do not include any one-body terms.

Date: September 1, 2024

Statement on the Contribution in Publication

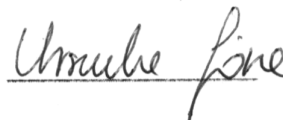
Reference: U. Góra, W. Cencek, R. Podeszwa, A. van der Avoird, and K. Szalewicz, *J. Chem. Phys.* **140**, 1941011 (2014).

Authors contributed to the publication jointly as follows:

Urszula Góra

Methodology, Software, Calculations, Data Analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing

Signature:



Wojciech Cencek

Methodology, Software, Calculations, Validation, Writing - Original Draft, Writing - Review & Editing

Signature:



Rafał Podeszwa

Formal Analysis, Validation, Calculations Supervision, Funding, Writing - Review & Editing

Signature:



Ad van der Avoird

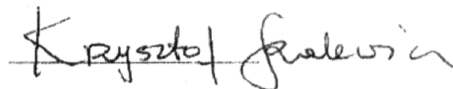
Methodology, Validation, Writing - Review & Editing

Signature:

Krzysztof Szalewicz

Conceptualization, Methodology, Validation, Supervision, Funding Acquisition, Writing - Review & Editing

Signature:



Date: September 1, 2024

Institute of Chemistry
University of Silesia
Szkolna 9, 40-006 Katowice, Poland

Statement on the Contribution in Publication

Reference: U. Góra, W. Cencek, R. Podeszwa, A. van der Avoird, and K. Szalewicz, J. Chem. Phys. **140**, 1941011 (2014).

Authors contributed to the publication jointly as follows:

Urszula Góra

Methodology, Software, Calculations, Data Analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing

Signature: _____

Wojciech Cencek

Methodology, Software, Calculations, Validation, Writing - Original Draft, Writing - Review & Editing

Signature: _____

Rafał Podeszwa

Formal Analysis, Validation, Calculations Supervision, Funding, Writing - Review & Editing

Signature: _____

Ad van der Avoird

Methodology, Validation, Writing - Review & Editing

Signature: _____



Krzysztof Szalewicz

Conceptualization, Methodology, Validation, Supervision, Funding Acquisition, Writing - Review & Editing

Signature: _____

Chapter 4

APPENDIX

SCHOLARSHIPS AND FELLOWSHIPS

- 2011 - 2012 4-month scholarship in Department of Physics and Astronomy,
University of Delaware, Newark, DE, USA
- 2011 Conference award for young researchers at ISTCP,
Tokyo, Japan
- 2010 4-month scholarship in Department of Physics and Astronomy,
University of Delaware, Newark, DE, USA
- 2009 4-month scholarship in Department of Physics and Astronomy,
University of Delaware, Newark, DE, USA
- 2009 - 2012 participation in Polish Ministry of Science and Higher Education grant
No. N N204 123337
- 2008 Conference award for young researchers at Theory and Applications Computa-
tional Chemistry Conference, Shanghai, China
- 2005 1-month scholarship in Slovak Academy of Sciences, Bratislava, Slovakia
- 2002–2003 6-month Socrates-Erasmus scholarship in Aarhus University, Denmark

PUBLICATIONS

- Urszula Góra, Wojciech Cencek, Rafał Podeszwa,
Ad van der Avoird, and Krzysztof Szalewicz,
*Predictions for water clusters from a first-principles two- and three-body force
field*,
Journal of Chemical Physics **140**, 194101 (2014). Citations: 59.
- Urszula Góra, Rafał Podeszwa, Wojciech Cencek, and Krzysztof Szalewicz, *In-
teraction energies of large clusters from many-body expansion*,
Journal of Chemical Physics **135**, 224102 (2011). Citations: 132.

PRESENTATIONS

- 09/2011 7th Congress of the International Society for Theoretical Chemical Physics, Tokyo, Japan; lecture: *Revealing mysteries of the water hexamer*
- 10/2008 XIV European Seminar on Computational Methods in Quantum Chemistry, Elba, Italy; poster: *Towards the faster convergence in equation of motion coupled cluster method*
- 09/2008 International Conference on the Theory and Applications of Computational Chemistry, Shanghai, China; poster: *Excitation energy calculations with $R12$ corrections*
- 04/2008 EuroQUAM Inauguration Conference, Barcelona, Spain; poster: *How does EOM-CCSD- $R12$ work?*
- 09/2007 Central European Symposium on Theoretical Chemistry, Litschau, Austria; poster: *Explicitly correlated coupled cluster calculations of excitation energy, Pilot results*
- 09/2007 Symposium on Advanced Methods of Quantum Chemistry and Physics, Toruń; poster: *Explicitly correlated coupled cluster calculations of Excitation Energy, The scheme*
- 09/2006 Central European Symposium on Theoretical Chemistry, Zakopane; poster: *Explicitly correlated coupled cluster calculations of excitation energy, Programming scheme*
- 06-07/2006 The 9th Sostrup Summer School, Aarhus, Denmark; poster: *Explicitly correlated coupled cluster implementation for excitation energies*
- 09/2003 Central European Symposium on Theoretical Chemistry, Nove Hradky, Czech Republic; poster: *Multipole moments and polarizabilities of weak molecular complexes*