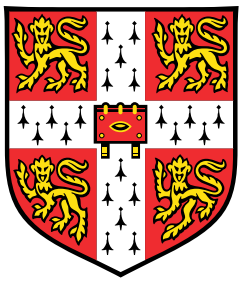
Data-driven models of water and methane



Eszter Székely

Supervisor: Prof. Gábor Csányi

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Peterhouse April 2021

Declaration

This report is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It is not the same as any work that has been submitted previously for any other degree except for some of the quantum mechanical calculations, which were performed for the author's Master thesis as indicated in the text.

Chapter 3 is based on parts of a paper in collaboration, with a table and a graph adapted from the paper. Apart from the titles of these, the text of the chapter is my own work.

Eszter Székely

April 2021

Data-driven models of water and methane

Eszter Székely

Abstract

In the field of materials modelling, traditional atomistic models seldom achieve high accuracy and speed at the same time. Recent developments using high-dimensional fits to approximate the quantum chemical potential energy surface (PES) have overcome this problem. This thesis presents such models for methane–water mixtures, in particular for methane clathrates. Since the discovery of their existence on Earth about half a century ago, methane clathrates have been subject to numerous studies motivated by industrial and environmental perspectives. This project develops atomistic models that describe methane–water interactions with high accuracy. The model development in this work focuses on the dimer and the trimer PESs, which are fitted to quantum mechanical data. The fitting methods used are the Gaussian Approximation Potentials (GAP) [1, 2] and the permutationally invariant polynomials (PIP) [3] methods, the latter applied in collaboration. The long-range electrostatic interactions are calculated using a classical force field, the modified TTM4F [4]. The resulting models are validated against quantum mechanical and experimental data. A clathrate phase diagram is calculated in the quasi-harmonic approximation using the model based on PIPs. As the fitted level, CCSD(T)-F12, is not applicable to larger systems, we compare the calculations to DMC results for the larger clusters and periodic systems. However, small systematic differences are found between the developed models and DMC; comparing different CCSD(T)-F12 versions against DMC, this inconsistency is confirmed to arise from the differences between the two quantum chemical methods. In another collaboration [5], different potential fitting methods are also compared using the same datasets and found to achieve similar accuracies when applied to only the energy differences.

Acknowledgements

First, I would like to thank my supervisor, Prof. Gábor Csányi, for his valuable advice and guidance throughout the project, and for giving me the opportunity to work on an interesting topic. Many thanks also to members of the research group for helpful discussions, particularly, Ádám Fekete, who helped to put together the TTM4Fmod+GAPs software; Albert P. Bartók, for helpful suggestions, including the inverse distance descriptor; and Alice Allen, for proofreading parts of this thesis. I would also like to thank my friends at the Department for providing company at group pizzas and both at in-person and virtual coffee breaks, during the lockdowns.

I am also grateful to collaborators, Marc Riera–Riambau and his supervisor, Prof. Francesco Paesani, for letting me use their MBX software, helping add the 3B correction fit during my visit to their group, and helpful discussions. Thanks also to collaborators working on the paper [5], namely Thuong T. Nguyen, Giulio Imbalzano, Prof. Jörg Behler, Prof. Gábor Csányi, Prof. Michele Ceriotti, Andreas W. Götz and Prof. Francesco Paesani. Also, I would like to thank Álvaro Vázquez–Mayagoitia for the quantum chemical calculations on the additional methane–water dimers. Thanks also to Chen Qu, Stephen J. Cox and Prof. Dario Alfè for providing data used in their papers.

I also appreciate the courses provided by the departmental Centre for Languages and Inter-Communication. Special thanks to Helen East, who gave suggestions on parts of this thesis as part of the writing supervisions.

I would like to acknowledge the funding from the Peterhouse Research Studentship and the funding and technical support from BP through the BP International Centre for Advanced Materials (BP-ICAM). Many thanks to my BP-ICAM-mentors, Nikos

Diamantonis, Corneliu Buda and Leslie Bolton for following this project throughout the PhD.

Finally, I would like to thank my family, who always encouraged me during my studies, and my partner, Áron, who was always by my side during the PhD.

Table of contents

Abbreviations **xiii**

1 Introduction **1**

1.1 Overview . . . 1

1.2 Data-driven models . . . 2

1.3 Data-driven methane–water models . . . 2

1.4 Methane clathrates . . . 3

1.5 Computational studies of methane clathrates . . . 6

1.6 Outline of the thesis . . . 8

2 Model components **11**

2.1 Many-body expansion (MBE) . . . 12

2.2 Quantum chemical methods . . . 13

2.2.1 Basis sets . . . 14

2.2.2 Counterpoise correction . . . 14

2.2.3 MP2 . . . 16

2.2.4 CCSD(T)-F12 . . . 16

2.3 The baseline electrostatic model: TTM4Fmod from MB-pol . . . 17

2.3.1 The TTM4Fmod-to-QUIP interface . . . 18

2.4 Gaussian Approximation Potentials (GAP) . . . 18

2.4.1 Sparsification . . . 20

2.4.2 Descriptors . . . 21

2.4.3 Datasets of fits . . . 25
2.4.4 Parameters of the TTM4Fmod+GAP model . . . 29
2.5 Permutationally Invariant Polynomials (PIP) . . . 34
2.5.1 The MB-nrg model . . . 35
2.5.2 The methane–water–water fit for the MB-nrg model . . . 36
2.5.3 The MBX-to-QUIP interface . . . 37
2.6 Differences between TTM4Fmod+GAPs and MB-nrg . . . 38
2.7 Other softwares . . . 40

3 Comparison of different potential-fitting methods 41
3.0.1 Author contribution details . . . 41
3.1 Introduction . . . 41
3.2 Datasets of the paper . . . 42
3.3 Parameters of the GAP fits . . . 42
3.4 Results . . . 42
3.5 Other comparative studies in the literature . . . 45

4 Static energy calculations for small clusters 47
4.1 The reference quantum chemical method: DMC . . . 47
4.2 Water clusters . . . 48
4.3 Methane-in-water clusters . . . 51
4.3.1 Benchmark quantum chemistry calculations . . . 55

5 Properties of periodic systems 61
5.1 Methods . . . 61
5.1.1 Generation of hydrogen-disordered structures . . . 61
5.1.2 Structures from other sources . . . 62
5.1.3 Geometry optimisation . . . 63
5.1.4 Quasi-harmonic approximation . . . 63
5.2 Ices . . . 66

5.2.1 Comparison to DMC results 66
5.2.2 Ice densities 67
5.2.3 Phase diagram prediction 67
5.3 Methane clathrates 69
5.3.1 Comparison to DMC on the sI clathrate 69
5.3.2 Clathrate densities 72
5.3.3 Phase diagram prediction 73

6 Conclusions **77**

Appendix A Supplementary graphs **81**

Appendix B Parameters of the GAP fits **83**

References **87**

Abbreviations

>3B or B3B	sum of beyond-three-body terms of many body expansion
5^x6^y	clathrate cage having x pentagonal and y hexagonal faces
AVXZ	the correlation-consistent polarized basis sets of Dunning with the X cardinality
BPNN	Behler–Parinello Neural Networks
BSCE	basis set convergence errors
BSE	basis set extrapolation
BSIE	basis set incompleteness error
BSSE	basis set superposition error
CBS	complete basis set
CCSD(T)-F12	coupled-cluster singles and doubles with perturbative triplets
CP	Counterpoise
DFT	density functional theory
DMC	diffusion Monte Carlo
fIIh	filled ice Ih
GAP	Gaussian Approximation Potentials

GP	Gaussian Process
HF	Hartree–Fock
MAE	mean absolute error
MBE	many-body expansion
MB-pol	a many-body potential for water
MB-nrg	a many-body potential for CH_4/H_2O and CO_2/H_2O systems
MD	molecular dynamics
MH-X	methane hydrate X
MP2	second order Møller–Plesset
NN	neural networks
PES	potential energy surface
PS	Partridge–Schwenke fit (for water or methane)
PIP	permutationally invariant polynomials
post-fIIh	post-filled ice Ih
PreconFIRE	Preconditioned Fast Inertial Relaxation Engine algorithm
PreconLBFGS	Preconditioned Limited-memory Broyde–Fletcher–Goldfarb–Shanno algorithm
QHA	quasi-harmonic approximation
QUIP	QUantum mechanics and Interatomic Potentials [software]
RMSE	root mean square error
sX	structure X [clathrate]

SOAP	smooth overlap of atomic positions
TTM4Fmod	a Thole-type model for water (and other small molecules in MB-nrg)
xB	x-body term of many body expansion

Chapter 1

Introduction

1.1 Overview

This project creates methane–water models by fitting quantum chemical results for the short-range interactions and using classical force fields for the long-range interactions. The motivation behind this is that whereas the quantum chemical effects are significant in the short-range interactions, the long-range interactions are dominated by electrostatics which can be described with sufficient accuracy by classical potentials. Two models are presented: one fully developed in this project, TTM4Fmod+GAPs, which uses the Gaussian Approximation Potentials (GAP) [1] fitting technique, and the second, MB-nrg, developed by collaborators [6] and improved for this project in collaboration, which uses the permutationally invariant polynomials (PIP) [3] method. These models are validated against quantum mechanical results. As the fitted level, CCSD(T)-F12 (see Section 2.2.4), would be too expensive for larger systems, we compare the results to benchmark diffusion Monte Carlo (DMC) results for larger structures and periodic systems. However, the results do not fully agree with the DMC energies, and therefore, we test CCSD(T)-F12 with different settings against DMC on smaller structures, finding some inconsistency between the results of the two methods. Physical properties of ices and clathrates are also predicted using the MB-nrg model in the quasi-harmonic approximation.

1.2 Data-driven models

Traditional models in materials modelling are seldom able to be accurate and fast enough at the same time. On the one hand, quantum chemistry can achieve very high accuracy, but this comes with a high computational cost. On the other hand, classical force fields are fast but have low accuracy. Recently, the development of machine learning force fields has managed to overcome this problem by fitting the quantum chemical results on large datasets using complex fitting tools [1, 3, 7]. The models thus developed are only fitted to quantum mechanics, so they are transferable to different physical states. In addition, their accuracy is improvable by adding more data to regions of the geometry space where higher accuracy is necessary. It is also possible to fit the quantum mechanical energies as corrections to a lower accuracy method which can be used to approximate the long-range interactions [4, 8]. This project develops new models by fitting corrections to the modified TTM4F (TTM4Fmod) [4] force field using the Gaussian Approximation Potentials (GAP) [1, 9] and the permutationally invariant polynomials (PIP) [3] methods. These methods are also compared to each other, and with the Behler–Parinello neural networks [7] method in Chapter 3.

1.3 Data-driven methane–water models

While numerous data-driven potentials have been developed for pure water systems (e.g. [10–13]), efforts to develop potentials for mixtures including methane began only in recent years [6, 14–19]. Whereas most of these models have been mostly concerned with the lower order terms of the many-body expansion [14–18], one fitted coefficients of classical potentials [19], and a recent one fitted both the lower order terms and the coefficients of force field [6].

Joel Bowman and co-workers fitted the methane–water and methane–water–water interactions using PIPs on the CCSD(T)-F12 and the MP2-F12 levels, respectively, using mixed basis sets [15, 16, 20]. Combining these terms with their earlier developed WHBB water potential, which includes fits for the monomer, dimer and trimer water

interactions [12], they calculated vibrational properties for small methane-in-water cage structures [20].

Akin–Ojo and Szalewicz fitted a potential to SAPT data for the methane–water dimer of rigid molecules [14] which they complemented by including polarisation in their 2019 paper [17]. In the same year, another model was developed by Szalewicz and co-authors [18] using the autoPES method [21] to fit CCSD(T) energies, which included dimer terms for rigid CH_4 and H_2O molecules [18]. Also in 2019, coefficients of classical potentials were fitted to less accurate quantum mechanical data (MP2) by Thakre and Jana [19].

There was no available methane–water force field with coupled cluster accuracy applicable to flexible molecules that included the long-range interactions up to the final year of this project when Riera et al. developed the MB-nrg model [6, 22] using PIP. This model also uses the TTM4Fmod potential, but here, it is also implemented for the methane molecules. For this PhD project, we complemented MB-nrg with a methane–water–water correction term in collaboration with the developers. This model is described in detail in Section 2.5.1.

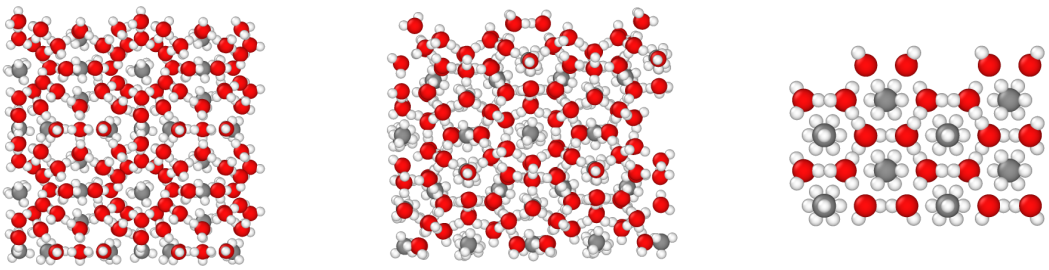
The model developed fully in this project, TTM4Fmod+GAPs, was created by fitting GAPs on the CCSD(T)-F12a/AVTZ level for the dimer and trimer interactions. As the aim was to model methane clathrates where the methane molecules are further from each other, only terms including up to one methane molecule were included. The water dimer and trimer potentials were fitted as corrections to a classical force field, the TTM4Fmod potential of MB-pol [4], and the long-range interactions for methane molecules were not included.

1.4 Methane clathrates

Methane–water systems are simple examples of apolar–polar interactions. Although these molecules do not mix under normal conditions due to their different polarities, under high pressure and low temperatures, they can mix [23] and form molecular crystals,

clathrates [24]. In the clathrates, the guest molecules, here the methanes, are in cages of water molecules. These cages are different polyhedra joining each other at their faces and having water molecules at their vertices. They can be denoted by the types and number of their faces: for example, $5^{12}6^2$ stands for a polyhedron having twelve pentagonal and two hexagonal faces [24].

Clathrates occur in large amounts within the seafloor, often near hydrocarbon resources [24, 25], so studying them has several environmental and industrial motivations. A little environmental change can cause the structures to decompose and release methane which has a 20 times stronger greenhouse effect than CO_2 [26]. Additionally, oil drills accidentally drilling into the clathrate layer might cause the clathrate to decompose, potentially resulting in environmental hazards as it happened in 2010 [27]. In oil pipelines, clathrate formation can impede the flow [28]. On the other hand, clathrates are also a possible methane resource for the energy industry and petrochemistry [28–31]. The first exploitation of clathrates started about 50 years ago at the Messoyakha field in Russia [32]. In recent years, exploitation of marine clathrates has also been attempted in Japan [29, 30] and China [31]. The possibility of storing methane in the structures has been studied [33], too. It is also promising that creating a CO_2 clathrate shell around the methane clathrate can increase its stability, a strategy which could help reduce greenhouse contributions from the two gases [34].



(a) sI, at 0.01 GPa. (b) sH_L2, at 1.5 GPa. (c) fIIh, at 2.0 GPa.

Fig. 1.1 The most common methane clathrate structures optimised using the MB-nrg model (see Section 5). The sH_L2 hexagonal structure has two CH_4 molecules in its largest cage. The carbon, oxygen and hydrogen atoms are shown with grey, red and white, respectively.

Although it is known that there are large amounts of clathrates, the estimates of this vary in the literature [25, 32], with the quantities at certain reservoirs still being explored [35,36]. The total amount of methane captured in clathrates was estimated to be around $3 \cdot 10^{15}\,\mathrm{m}^3$ ten years ago [25]. Apparently, the existence of methane clathrates on Earth was only predicted in the middle of the 20th century [32, 37] and confirmed in the 1970s and 1980s [38–40]. Given this, it is not so surprising that the full methane–water phase diagram and the kinetics of mixing and clathrate formation are not fully known yet and are still the subject of recent studies [23, 41, 42].

The most common form of methane clathrates on Earth is the so-called *structure I*: sI [24] (see Fig. 1.1). Other structures were found and described by high-pressure spectroscopic studies in the 2000s [43–49]: the *hexagonal structure*, sH; the *filled ice Ih* structure, fIIh; and the *post-filled ice Ih* structure, post-fIIh. Although some of these results were sometimes controversial – for example, some studies found structures that the other studies did not, such as sII between 0.1–0.6 GPa [43] or "structure B" between 1.6–2.1 GPa [44] – nowadays there seems to be agreement on the phase changes up to 80 GPa. Many of the spectroscopic studies in the early 2000s found two phase transitions, the sI–sH transition around 0.8–1.0 GPa [44–47, 50], and the sH–fIIh phase change around 1.9–2.0 GPa [45–50]. Later, a transition to a new structure, post-fIIh, was also described around 40 GPa [48, 49]; Schaack et al. [51] observed different steps of this transition between 30.0–50.0 GPa.

The exact filling ratio of the sH clathrate has only been determined earlier this year by Noguchi et al. [42]. In the early 2000s, even its water skeleton was not yet confirmed to be the sH [52]. The sH structure has two kinds of smaller cages: three 5^{12} cages, two $4^3 5^6 6^3$ cages, and one large cage: $5^{12}6^8$ [24]. Recently, the water structure has been confirmed to be the hexagonal clathrate structure, and there has been an agreement on the small cages being singly filled [41, 42, 47]. However, there have been different suggestions for the number of methanes in the large cage [41, 42, 47, 52]. Loveday et al. [52] suggested that their earlier spectroscopic data [53, 54] might be explained by the large cage containing five CH_4 molecules in it [52]. Kumazaki et al. [47] reported only one Raman band below

around 1.3 GPa and two above this pressure. As the two smaller cages have similar radii, the methanes in them might correspond to one peak, and the small cages might be filled singly below 1.3 GPa [47]. Kumazaki et al. explained the appearance of the new band by a change in the filling of the large cage, which they assumed to have an average of 2.5 guest molecules in it above 1.3 GPa [47]. As they approximated the filling of the large cage from the ratio of the peaks [47] and as the clathrates can have cages that are not filled, it might be different in case the small cages were not fully filled. According to the neutron diffraction experiments at 1.9 ± 0.2 GPa of Tulk et al. [41], the two smaller cavities have a bit less than one CH_4 in them while the large ($5^{12}6^8$) cavity will have a bit more than three CH_4 in it. The very recent IR and Raman spectroscopy work using D_2O of Noguchi et al. [42] showed that while all the cages are occupied by a single CH_4 molecule between 0.9–1.3 GPa, the large cage will be occupied by two CH_4 molecules between 1.3–2.0 GPa [42].

It has also been suggested that sII could be a kinetically preferred but metastable phase [55–57]. Schicks and Ripmeester [55] showed by Raman spectroscopic measurements that sI and sII can coexist between 0.003–0.009 GPa. Shin et al. [56] found that although sII was formed, it later turned into the thermodynamically stable phase at the given conditions. This finding can also explain why Chou et al. [43] saw the sII structure in their experiments. The formation of sII in the presence of other hydrocarbons was also studied by Stoporev et al. [57]. One of the models studied, MB-nrg, is applied to predict the phase diagram in the quasi-harmonic approximation and compared to the above experimental transitions (see Section 5.3.3).

1.5 Computational studies of methane clathrates

Previous computational studies have predicted some new methane clathrate phases. Vatamanu and Kusalik found a new structure, sK, while simulating clathrate formation using classical models [58]. This structure has three types of cage: the same cages as sI, 5^{12} and $5^{12}6^2$, along with a cage that has three hexagonal faces $5^{12}6^3$ [58], which is

a transition between the larger cages of the sI ($5^{12}6^2$) and sII ($5^{12}6^4$). This structure was later shown to exist as a metastable phase for Xe-clathrate by Yang et al.'s X-ray experiments; here it was termed HSI [59]. Cao and co-authors predicted some additional structures called MH-IV [60], MH-V [60] and MH-VI [61] using simulated annealing with a force field to look for possible structures and studying the stability of the different phases using DFT; however, they found only the MH-VI to appear in their computational phase diagram [61].

Clathrate phase diagram predictions have been computed previously, but to date, qualitative accuracy for the full phase diagram has not yet been achieved. Some of the studies using classical models looked at only the transitions at low pressures, and thus only included the sI clathrate, liquid water and methane gas phases [62–64]. Lenz and Ojamäe [65] also studied the low pressure region using DFT, computed properties for the sI, sII and sH clathrates and calculated the ice Ih–sI transition. The studies of Cao and co-workers [60, 61] considered a wider pressure range; they compared the stabilities of different phases using DFT. In their 2017 paper [60], they predicted two new structures having a 1:4 methane–water ratio, MH-IV and MH-V. However, looking at the chemical potentials needed to form the clathrate structures, they suggested that MH-IV would only be stable in a very narrow region, and MH-V would not be present at all. Moreover, in their later paper, Huang et al. [61] calculated positive formation enthalpies for MH-V [61]. In the same paper [61], looking at the formation enthalpies per molecules, they predicted the sK, sII, MH-VI, and fIIh phases to be stable with increasing pressure, yet only the fIIh appears in the experimental phase diagram.

The filling ratio of the large cage of sH has also been the subject of computational simulations. Alavi et al. [66] calculated free energies from molecular dynamics simulations using classical force fields. Fixing the water–water and methane–water potentials and changing only the description of the methane–methane interactions, they found that two of the methane–methane models (OPLS, Tse–Klein–McDonald) predicted five methanes and the Murad–Gubbins model predicted two methanes for filling the large cage [66]. Studying the energetics of separate cages using DFT, both Cao et al. [67] and Liu et al. [68]

suggested that the cavity would contain five CH_4 molecules. Later, however, both groups found that the cage would contain fewer methane molecules when simulating the whole clathrate structure. Looking at the formation enthalpies of the periodic structure, Cao and co-workers [61] suggested that the large cavity contained three methane molecules. Liu et al. [69], calculating cohesive and deformation energies, suggested that the structure that has four methanes in the large cage would be the most stable.

Previous computational studies mostly used methods with lower accuracy than the target level of this thesis, coupled cluster. The models used were mostly classical force fields [58, 62, 64, 66, 70–78] or density functional theory (DFT) functionals [51, 60, 61, 65, 68, 69, 79], but some studies used data-driven models [20]. Both the force fields and DFT functionals have low accuracy, and DFT functionals also have higher computational cost than the simple potentials. As explained in Section 1.3, the current data-driven models also leave room for improvement, for example, the one used in Ref. [20] does not include long-range interactions. Some of the DFT calculations used different schemes to achieve higher accuracy than simple DFT functionals: Schaack et al.'s study [51] combined DFT with PIMD to include quantum nuclear effects; and multiple papers [60,69,79,80] included the Tkatchenko–Scheffler dispersion corrections in their DFT calculations. However, Cox et al. [81] showed that even the dispersion-corrected DFT functionals can have large errors when compared to diffusion Monte Carlo (DMC) results. Thus, a model with near quantum chemical accuracy but much better scaling would serve as a valuable tool for theoretical research on clathrates and for predicting physical properties of methane–water mixtures for industrial use. This work contributes by developing such a model.

1.6 Outline of the thesis

The thesis is organised as follows: this Introduction has given an overview on potential-development efforts using data-driven techniques (Section 1.2), then has reviewed models developed for methane–water systems (Section 1.3) and described available knowledge on methane clathrate structures (Section 1.4). Chapter 2 provides descriptions of the

methods used for building the models, their inherent datasets and the reference structures. Chapter 3 compares three potential energy fitting techniques. The models described in Chapter 2 are validated against quantum mechanics on small clusters in Chapter 4, and comparisons of different quantum mechanical methods are also presented in Section 4.3.1. Calculations for periodic systems by the developed models are presented in Chapter 5. Finally, Chapter 6 summarises the study.

Chapter 2

Model components

The models are built using two different strategies for the different ranges: the short-range interactions are approximated by high-dimensional fits to the quantum mechanical potential energy surface (PES), while the long-range interactions are described by a classical model, TTM4Fmod. As the classical force field is already applied to the system, the quantum mechanical interactions are fitted as corrections to the force field, on the quantum chemistry–force field differences. The datasets include configurations of different geometries with the corresponding quantum chemical energies and sometimes the forces (see Section 2.4.3). However, as the evaluation time of quantum chemistry scales really badly with system size, a method is needed to break down the total energies to terms involving only smaller numbers of molecules and build the training datasets from these fragments. For this, we use the many-body expansion and thus avoid the need to calculate the total energy by quantum chemistry (see Section 2.1).

The quantum mechanical techniques and baseline classical force field used in the thesis are discussed in Sections 2.2 and 2.3, respectively. The underlying datasets and the main fitting tool of the project: the Gaussian Approximation Potentials (GAP) are described in Section 2.4. The other fitting method applied, the permutationally invariant polynomials (PIP), is described in Section 2.5.

2.1 Many-body expansion (MBE)

The core of the model development is the many-body expansion which is a formal expansion of the energy of a molecular system. Using this method, the short- and long-range interactions can be separated and calculated by different models [4, 8]: the long-range interactions are approximated well at the classical level, and the short-range interactions are calculated by high-dimensional fits at the quantum mechanical level. The total energy is expanded as a sum of interaction energies appearing between groups of different numbers of molecules (or atoms). Here, it is used in the molecular basis as in [82]. The total energy ($E_N(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)$) of an N-molecular system is written as a sum of X-body (XB) terms ($X = \{1, 2, ..., N\}$):

$$E_N(\boldsymbol{r}_1, ..., \boldsymbol{r}_N) = \sum_{I}^{N} E_{1\mathrm{B}}(\boldsymbol{r}_I) + \sum_{I,J}^{N} E_{2\mathrm{B}}(\boldsymbol{r}_I, \boldsymbol{r}_J) + \sum_{I,J,K}^{N} E_{3\mathrm{B}}(\boldsymbol{r}_I, \boldsymbol{r}_J, \boldsymbol{r}_K) + ... + E_{N\mathrm{B}}(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_N) \tag{2.1}$$

where E_X denotes the total energy of X molecules and $E_{X\mathrm{B}}$ stands for the X-body interaction energy of the X molecules. $E_{1\mathrm{B}}(\boldsymbol{r}_I)$ are the one-body (1B) energies of the monomers (as if they were in vacuum having the same $\boldsymbol{r}_I$ coordinates as in the original system, where $\boldsymbol{r}_I$ denotes the coordinates of all the atoms within molecule I); $E_{2\mathrm{B}}(\boldsymbol{r}_I, \boldsymbol{r}_J)$ are the interaction energies between pairs of molecules:

$$E_{2\mathrm{B}}(\boldsymbol{r}_I, \boldsymbol{r}_J) = E_2(\boldsymbol{r}_I, \boldsymbol{r}_J) - E_{1\mathrm{B}}(\boldsymbol{r}_I) - E_{1\mathrm{B}}(\boldsymbol{r}_J) \tag{2.2}$$

The higher-order $E_{X\mathrm{B}}$ terms are defined as the interaction energies that are only present between the X molecules but not between any $(X-1)$ molecular subset of the system:

$$E_{X\mathrm{B}}(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_X) = E_X(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_X) - \sum_{Y=1}^{(X-1)} \sum_{I<J<...<Y} E_{Y\mathrm{B}}(\boldsymbol{r}_I, \boldsymbol{r}_J, ..., \boldsymbol{r}_Y) \tag{2.3}$$

Using the MBE in the model development has several advantages. Firstly, the lower order terms are larger in magnitude and have lower dimensionalities, so if fitted separately, can be fitted more accurately (in a fractional sense) from less data. However, in a big structure, there are more higher-order terms, so their contribution becomes more important. Secondly, the higher-order terms usually decrease quicker with the intermolecular distances, so shorter cutoff radii can be used for them. Thirdly, calculating the databases for the higher-order terms becomes more and more computationally expensive, so to avoid high computational cost, one can use quantum mechanical methods with lower cost for the higher-order terms. Moreover, it is possible to truncate the expansion at some term, and use classical force fields or DFT functionals for the terms above it [4, 13]. When computer speed permits, the models can be systematically improved by recalculating the training sets of the fits with higher level quantum chemical methods. Similarly, corrections for the higher-order terms will be possible to fit. In this thesis, the developed models approximate quantum chemistry in the 2B and 3B terms by adding corrections to a force field fitted by either GAP or PIP. The higher-order terms are calculated by the baseline classical force field.

2.2 Quantum chemical methods

To calculate the training datasets with quantum chemistry, a quantum mechanical method is needed which has high accuracy, but a computational cost that is still tractable for the large numbers of configurations in the databases. In this work, the CCSD(T)-F12a method with the AVTZ basis set and the counterpoise correction was chosen as it has very high accuracy and large databases were already available for the systems studied. However, we later found that not all the databases were calculated at the same level, so we recalculated subsets of them (see Section 2.4.3).

2.2.1 Basis sets

The basis sets in this work are the AVXZ correlation-consistent polarized basis sets of Dunning [83] with the X cardinality, where $X = \mathrm{D}, \mathrm{T}, \mathrm{Q}$ for the double, triple and quadrupole cardinalities. These basis sets are built as sets of Gaussians [83].

For the datasets of the GAP fitting, AVTZ is used. When comparing different quantum mechanical settings, we also tested the correlation consistent VXZ-F12 sets [84] that were optimised for the F12-methods [84].

2.2.2 Counterpoise correction

Counterpoise correction (CP) [85] is a method for reducing the basis set superposition error (BSSE). The BSSE arises when many-body energies are calculated using different basis sets for the different fragments involved. To illustrate, the 2B energy of an AB dimer could be calculated in the following two ways [86]:

$$E_{\mathrm{2B,noCP}} = E_{\mathrm{AB}}(\phi_{\mathrm{AB}}) - E_{\mathrm{A}}(\phi_{\mathrm{A}}) - E_{\mathrm{B}}(\phi_{\mathrm{B}}) \tag{2.4}$$

$$E_{\mathrm{2B,CP}} = E_{\mathrm{AB}}(\phi_{\mathrm{AB}}) - E_{\mathrm{A}}(\phi_{\mathrm{AB}}) - E_{\mathrm{B}}(\phi_{\mathrm{AB}}) \tag{2.5}$$

where ϕ_{A} and ϕ_{B} are the basis sets of the individual molecules A and B, and ϕ_{AB} is the joint basis set of the AB dimer. The monomer energies are lower when calculated with the joint basis set, $E_{\mathrm{A}}(\phi_{\mathrm{AB}}) \leq E_{\mathrm{A}}(\phi_{\mathrm{A}})$, as the basis set extension improves on the original basis set [86, 87]. Thus in the case of Eq. (2.4), one might subtract larger monomer energies leading to an interaction energy that is lower than the real value. The CP correction uses Eq. (2.5) to calculate the interaction energy and converges to the complete basis set value from above [88]. The value of the correction can be defined as $\Delta E_{\mathrm{A,CP}} = E_{\mathrm{A}}(\phi_{\mathrm{AB}}) - E_{\mathrm{A}}(\phi_{\mathrm{A}})$.

Although it seems straightforward that the counterpoise correction eliminates BSSE which is a known source of error, this correction still causes controversies in the literature. Those arguing against using it state that not correcting BSSE works better, because

there is a possibility of BSSE and the so-called basis set convergence error (BSCE) partially cancelling each other out [88, 89]. The BSSE and BSCE constitute the basis set incompleteness error (BSIE) [88, 89]:

$$E_{\mathrm{BSIE}} = E_{\mathrm{BSSE}} + E_{\mathrm{BSCE}} \tag{2.6}$$

Using large basis sets would reduce both parts; however, it would be highly computationally expensive for large systems or large numbers of configurations. The two components (E_{BSSE} and E_{BSCE}) have opposite signs, so correcting only one of them might lead to a larger error [87–89]. A similar reasoning also led to the suggestion of Halkier et al. [90] for the so-called half–half approximation which suggests averaging the CP and noCP values:

$$E_{\mathrm{half\text{–}half}} = \frac{1}{2} \cdot (E_{\mathrm{CP}} + E_{\mathrm{noCP}}). \tag{2.7}$$

Sheng et al. [89] showed that the relative magnitudes of E_{BSSE} and E_{BSCE} change with the distance between the monomers and with the basis sets on the example of the helium dimer so the two errors do not always cancel each other out. An extensive study by Burns et al. [91] on the effects of the different CP corrections on the predictions of the MP2 and CCSD(T) methods for bimolecular complexes showed that the most accurate correction-type depends on the methods. Another extensive study in the same year by Brauer et al. focusing on the MP2-F12 and CCSD(T)-F12 methods, which also made different suggestions for the different methods and basis sets, noted that the half weight is arbitrary [88]. Even the paper of Halkier et al. [90] which first suggested the half–half method noted that, for HF (Hartree–Fock) calculations, whether the half–half or the CP method is the more accurate is system-dependent. However, using different CP versions depending on the methods and chemical systems is inconsistent. While it is not possible to correct the BSIE completely, there is no reason not to correct BSSE as it might lead to more physical results. Moreover, CP always approaches the exact result from above while noCP does not [89, 90].

Generalising the counterpoise correction for the higher-order terms of the many-body expansion can be done in different ways [92, 93]. Two possible methods are calculating it as a sum of pairwise BSSEs or the so-called "site–site function counterpoise" [92]. The site–site function CP method calculates the interaction energy using the full basis set for all the fragments involved. Thus, the CP correction can be defined as:

$$\Delta E_{\mathrm{A,CP}} = E_{\mathrm{A}}(\phi_{\mathrm{ABC...}}) - E_{\mathrm{A}}(\phi_{\mathrm{A}}). \tag{2.8}$$

In this work, this site–site function CP method is used for the CP-corrected 3B energies, so they are calculated as:

$$\begin{aligned} E_{\mathrm{3B(ABC),CP}} =& E_{\mathrm{ABC}}(\phi_{\mathrm{ABC}}) - E_{\mathrm{AB}}(\phi_{\mathrm{ABC}}) - E_{\mathrm{AC}}(\phi_{\mathrm{ABC}}) - E_{\mathrm{BC}}(\phi_{\mathrm{ABC}}) \\ &+ E_{\mathrm{A}}(\phi_{\mathrm{ABC}}) + E_{\mathrm{B}}(\phi_{\mathrm{ABC}}) + E_{\mathrm{C}}(\phi_{\mathrm{ABC}}). \end{aligned} \tag{2.9}$$

2.2.3 MP2

MP2 is second order Møller–Plesset perturbation theory [94]. Here, its density-fitted version (DF-MP2) is used to reduce computational time [95]. This method scales as N^5 and thus it is possible to calculate the forces for the fitting datasets of the thesis.

2.2.4 CCSD(T)-F12

CCSD(T)-F12 is generally regarded as the "gold-standard" of quantum chemistry [91]. The abbreviation stands for coupled cluster with singles and doubles excitations and perturbative triples, and F12 denotes a correlation factor [96, 97]. The implementation of Adler et al. [97] is included in Molpro [94]. The results of this coupled cluster method using the AVTZ basis set are more accurate than CCSD(T)/AV5Z [97]. This method scales as N^7 [98], so it has a high computational cost for larger systems. The force calculations would also have high computational cost. However, the version of Molpro used in the project (2012.1 version [94]) did not have forces available for this method;

the analytical gradients for versions of coupled cluster were only made available in the 2018.1 version [99].

The fitted quantum chemical level of the developed TTM4Fmod+GAPs model is CCSD(T)-F12a/AVTZ with the counterpoise correction [85] calculated by Molpro [94]. The F12a version is chosen because the Molpro manual suggested this version of F12 for the AVTZ basis set [99] and using larger basis sets would be more computationally expensive.

2.3 The baseline electrostatic model: TTM4Fmod from MB-pol

We chose to use the same baseline model for water as in MB-pol [4,10] that is a modified [4] TTM4F [100]. TTM4F [100] is a Thole-type model (TTM) [101,102] where the electrostatic properties of the H_2O molecule are described by three point charges placed on the hydrogen atoms and on an M site, and by inducible point dipoles placed on all the three atoms [4]. The electrostatic energy is calculated as a sum of interactions between these charges and dipoles, and a spring term corresponding to the dipoles' energies [4]. However, with the choice of parameters of the model, the energy is simplified to the sum of only two terms: interactions between the charges, and the dipoles' interactions with the electrostatic field [4]:

$$V_{\mathrm{TTM}} = \frac{1}{2}\sum_{i\neq j}\frac{q_i q_j}{(\alpha_i\alpha_j)^{1/6}}\lambda_1(u_{ij}) - \frac{1}{2}\sum_i(\boldsymbol{\mu}_i\cdot\boldsymbol{E}_i) \tag{2.10}$$

where q_i are the charges, $\boldsymbol{\mu}_i$ are the dipoles, α_i are the dipole polarizabilities, $\boldsymbol{E}_i$ is the electric field at i, $u_{ij} = r_{ij}/(\alpha_i\alpha_j)^{1/6}$, and λ_1 is a function to define the screened interaction that replaces the point charges by charge density distributions [4].

The TTM4Fmod is a modified version of TTM4F as introduced by Babin et al. [4]. They changed the damping parameter corresponding to the intra-molecular H–H dipole–dipole interaction and thus the model became more stable for distorted molecules [4].

2.3.1 The TTM4Fmod-to-QUIP interface

The TTM4Fmod model is invoked by the MB-pol plugin [4, 10, 103] for OpenMM [104] which is interfaced to ASE [105] by a code written by Ádám Fekete and the author. The PIP fits of MB-pol are turned off to get the baseline TTM4Fmod model. This interface is available at [106]. Fig. 2.1 shows the dependencies of the codes building the TTM4Fmod+GAPs model, which sums the baseline water model with the 1B CH_4 term of Schwenke and Partridge [107, 108] and the GAP corrections fitted in this project. The PIPs are not simply substituted by GAPs as the latter ones are fitted on different quantum mechanical levels and have different cutoff functions. Due to restrictions of the underlying MB-pol plugin, the force field is only applicable to orthogonal lattices. This version of TTM4Fmod does not use parallelisation, so it is slower than the version implemented in MBX (see Section 2.5.1). As published, due to license reasons, the TTM4Fmod+GAPs package can only be used for the TTM4Fmod model, and the GAPs need to be calculated by the latest version of QUIP and GAP.

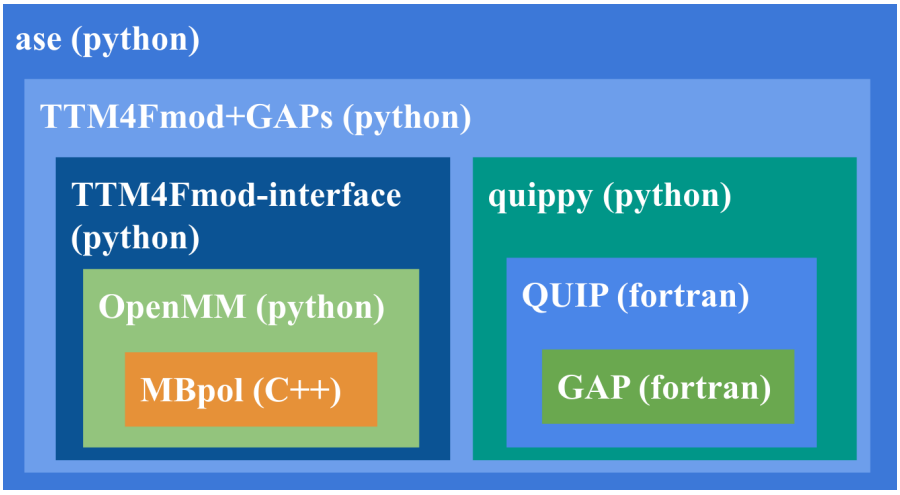


Fig. 2.1 Schematic dependency graph of the code for the TTM4Fmod+GAPs model.

2.4 Gaussian Approximation Potentials (GAP)

The Gaussian Approximation Potentials [1,2,8,9,109] is an implementation of the Gaussian Processes machine learning method for fitting potential energy surfaces (PESs), which

have been applied to many systems including materials and molecular structures [1,8,110–112]. The underlying fitting method can also be described as kernel-ridge regression [109, 113,114], so here, we will explain the technique according to this logic.

The PES is fitted in the space defined by the geometries of the training set: $\{\boldsymbol{R}\}^{(N)}$. The basis functions are defined from the geometries by the so-called descriptors: $\boldsymbol{d}(\boldsymbol{R})$ functions that describe the geometries (see Section 2.4.2). The similarity of two structures is defined by the squared exponential (ARD_SE) kernel [1, 109]; then, a kernel matrix is built having the similarity measure of $\boldsymbol{R}_I$ and $\boldsymbol{R}_J$ as its I, J element:

$$K(\boldsymbol{d}(\boldsymbol{R}_I), \boldsymbol{d}(\boldsymbol{R}_J)) = \delta^2 \exp\left(-\sum_i \frac{(d_i(\boldsymbol{R}_I) - d_i(\boldsymbol{R}_J))^2}{2\theta_i^2} \right) \tag{2.11}$$

where i runs over the elements of the descriptor $\boldsymbol{d}$, δ and θ_i are hyperparameters of the fit, θ_i being the typical decorrelation length for the descriptor element d_i [1]. The kernel will be an $N \times N$ matrix where N is the number of representative geometries. Then, the energy for a new structure, $\boldsymbol{R}_{N+1}$, can be calculated by multiplying the similarity of this new structure with the $\boldsymbol{R}_I$ representative structures with the corresponding w_I weights [8, 109, 114]:

$$E(\boldsymbol{d}(\boldsymbol{R}_{N+1})) = \sum_{I=1}^{N} w_I \cdot K(\boldsymbol{d}(\boldsymbol{R}_{N+1}), \boldsymbol{d}(\boldsymbol{R}_I)) \tag{2.12}$$

The w_I weights are determined by a least squares fit minimising the loss function:

$$l(\boldsymbol{w}) = ||K \cdot \boldsymbol{w} - \boldsymbol{E}(\boldsymbol{d})||^2 \tag{2.13}$$

where $||\boldsymbol{v}||^2 = \langle \boldsymbol{v}, \boldsymbol{v} \rangle$ for an arbitrary vector $\boldsymbol{v}$, $\langle \boldsymbol{v}, \boldsymbol{v} \rangle$ denoting the dot product. This would lead to the solution for the weights:

$$\boldsymbol{w} = K^{-1}\boldsymbol{E}(\boldsymbol{d}) \tag{2.14}$$

To regularise the linear algebra, the loss function is modified as [8, 109]:

$$l(\boldsymbol{w}) = ||K \cdot \boldsymbol{w} - \boldsymbol{E}(\boldsymbol{d})||^2 + \sigma^2 \boldsymbol{w}^T K \boldsymbol{w} \tag{2.15}$$

here σ^2 is a small, positive regularisation parameter. Then the vector of weights can be obtained by minimising Eq. (2.15), which leads to [8, 109]:

$$\boldsymbol{w} = (K + \sigma^2 \cdot I)^{-1} \boldsymbol{E}(\boldsymbol{d}) \tag{2.16}$$

where I is the $N \times N$ unit matrix.

2.4.1 Sparsification

For large datasets, the matrix inversion appearing in Eq. (2.16) can have a high computational cost and the time of evaluation (Eq. (2.12)) also increases linearly. Actually, the geometry space defined by the train set can be represented by a subset of the training geometries [1, 115], so including all the train structures for the basis functions might be unnecessary. For a fit, a smaller number of representative configurations are chosen as sparse points: $\{\boldsymbol{R}\}^{(M)}$, which define the basis functions of the fit [1,115]. In GAP [9], there are several methods implemented for choosing these structures, for instance, randomly, using the CUR matrix decomposition [116], in which an A matrix is decomposed as $A \approx CUR$ where C and R consist of columns and rows of the original matrix [116]; or defining them by the indices. For some of the fits of the project, we used the furthest point sampling (FPS) [117] technique to choose the sparse points, and specified them by the indices.

Using M sparse points, smaller covariance matrices can be defined [1]; an $M \times M$ matrix, built from the similarity measures of the sparse configurations: K_{MM}, and an $N \times M$ matrix : K_{NM}, in which the N rows correspond to the train set geometries and the M columns correspond to the geometries of the sparse subset [1]. For the fit, an

$N \times N$ matrix, Λ, is also needed, defined as [1]:

$$\Lambda = \mathrm{diag}(K_{NN} - K_{NM}K_{MM}^{-1}K_{MN}) \tag{2.17}$$

where diag denotes the diagonal elements of a matrix so that: $(\mathrm{diag}(M))_{i,j} = \delta_{i,j} \cdot M_{i,j}$. The similarity of a new structure with the M sparse structures is the vector:

$$\boldsymbol{k}_{N+1} = \{K(\boldsymbol{d}(\boldsymbol{R}_{N+1}), \boldsymbol{d}(\boldsymbol{R}_I)\}_{I=1}^{M} \tag{2.18}$$

Defining a K'_{MM} kernel including the regularisation as [1]:

$$K'_{MM} = K_{MM} + K_{MN}(\Lambda + \sigma^2 I)^{-1}K_{NM} \tag{2.19}$$

the prediction for a new structure will be [1]:

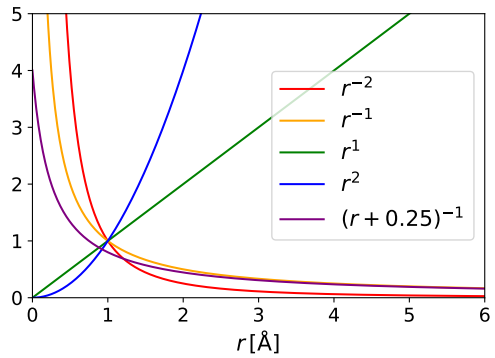
$$E(\boldsymbol{d}(\boldsymbol{R}_{N+1})) = \boldsymbol{k}_{N+1}^T K'^{-1}_{MM} K_{MN}(\Lambda + \sigma^2 I)^{-1}\boldsymbol{E}(\boldsymbol{d}) \tag{2.20}$$

2.4.2 Descriptors

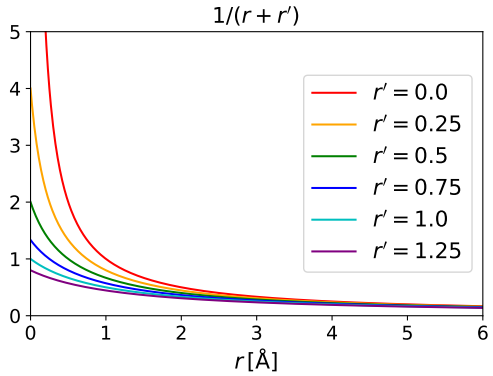
To define the kernel basis functions from the representative geometries, the structures need to be described by some functions, the so-called descriptors. As these functions are inherent to the fit of the energy functional, it is desirable that they change with the geometries as the energy would change. Thus, the descriptors should be smooth, symmetric and differentiable with respect to changes of the geometries. In this thesis two main types of descriptors are applied: a class based on interatomic distances [109] and the SOAP (*smooth overlap of atomic positions*) [118, 119] descriptor.

The interatomic distance descriptors are based on permutationally symmetrised interatomic distances or their powers. The original version is simply the set of the interatomic distances within the structure [109]:

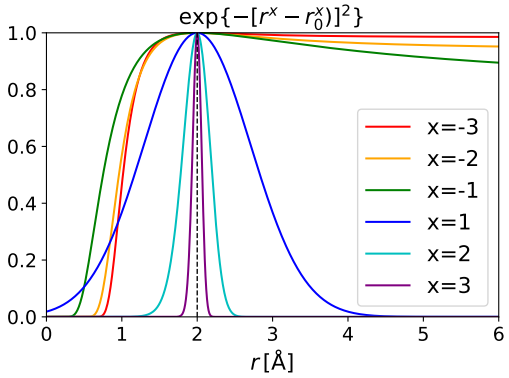
$$\boldsymbol{d}(\boldsymbol{R}) = \{r_{ij} \mid r_{ij} \in \boldsymbol{R}\} \tag{2.21}$$



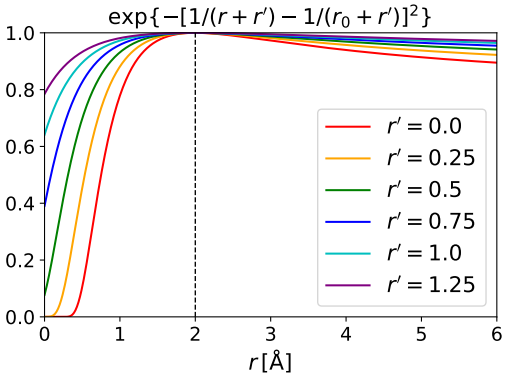
(a) Different powers of the distance and the inverse distance with a distance shift w.r.t. the distance.



(b) The inverse distance with a distance shift added, changing the distance shift w.r.t. the distance.



(c) The kernel element induced by different powers of the interatomic distance w.r.t. the distance. $r_0 = 2.0\,\text{Å}$.



(d) The kernel element induced by the inverse interatomic distance, changing the distance shift w.r.t. the distance. $r_0 = 2.0\,\text{Å}$.

Fig. 2.2

and it is made permutationally invariant for the change of the order of monomers of the same type and the order of atoms of the same species (within the monomers and if set, also between them) [9]. For the fits of this thesis, it is set to only swap the atoms within the monomers. The descriptor function is smoothly turned off at the cutoff distance using a cosine function [109] (see Eq.-s (2.24), (2.25)).

A variation of this descriptor using the inverse distances with a distance shift was suggested by A. P. Bartók during this project [120]. The inverse distance descriptor is a modification of the interatomic distance descriptor to use the inverse distances. To avoid the inverse distance being very large at short distances, we added a small shift, r', to the

distance:

$$\boldsymbol{d}(\boldsymbol{R}) = \left\{ \frac{1}{r_{ij} + r'} \mid r_{ij} \in \boldsymbol{R} \right\} \tag{2.22}$$

Different powers of the distances can be applied for the descriptor, but for the fits of this project, the inverse power seems to work best. This choice of the descriptor is convenient for fitting the PES because both the inverse distances and the interaction energy are more sensitive to changes in the molecular distances at short distances than at larger separations. Functions of the inverse distances were also used for potential energy fitting in PIP and GP by Joel Bowman and co-workers [121, 122] and in sGDML by A. Tkatchenko and co-workers [123].

The functions of the distances entering the kernel changing the power and the distance shift are shown in Figures 2.2a, 2.2b. The powers of distances are fed into the same kernel functions as the distances in the case of the interatomic distance descriptors. For the interatomic distance descriptors, the squared exponential kernel is used to calculate the similarity of two structures [1]. The equation for the kernel of the distance shifted inverse interatomic distance descriptor using the squared exponential kernel:

$$K(\boldsymbol{d}(\boldsymbol{R}_A), \boldsymbol{d}(\boldsymbol{R}_B)) = \delta^2 \exp\left(- \sum_{r_{ij} \in \boldsymbol{R}} \frac{|\frac{1}{r_{ij,A}+r'} - \frac{1}{r_{ij,B}+r'}|^2}{2\theta_{ij}^2} \cdot f_{\text{cut},\boldsymbol{R}}(\boldsymbol{R}_A) \cdot f_{\text{cut},\boldsymbol{R}}(\boldsymbol{R}_B) \right) \tag{2.23}$$

where the θ_{ij} is the typical decorrelation length for the shifted inverse interatomic distance, and the $f_{\text{cut,structure}}(\boldsymbol{R})$ cutoff function for a structure is defined as:

$$f_{\text{cut},\boldsymbol{R}}(\boldsymbol{R}) = \frac{1}{N_{\text{cut}}} \sum_{i,j \neq \text{H}} f_{\text{cut}}(r_{ij}) \tag{2.24}$$

with the sum running over the distances between the non-hydrogen atoms and N_{cut} is the number of distances included in the sum. The $f_{\text{cut}}(r_{ij})$ is a smooth cutoff function,

defined as [9, 109]:

$$f_{\text{cut}}(r_{ij}) = \begin{cases} 1 & r_{ij} < (r_{\text{cut}} - r_{\text{cut transition width}}) \\ \frac{1}{2}\Big(\cos \frac{\pi(r - r_{\text{cut}} + r_{\text{cut transition width}})}{r_{\text{cut transition width}}} + 1 \Big) & (r_{\text{cut}} - r_{\text{cut transition width}}) < r_{ij} < r_{\text{cut}} \\ 0 & r_{\text{cut}} < r_{ij} \end{cases} \tag{2.25}$$

with $r_{\text{cut transition width}}$ being a cutoff transition width, usually set to 1 Å for the interatomic distance descriptors.

The kernel element for a simple structure described by one distance (e.g. a diatomic molecule), without the cutoff function and setting $\delta = 1$ and $\theta_{i,j}^2 = 0.5$, is shown in Figures 2.2c and 2.2d, changing the power and the distance shift, respectively.

With the SOAP descriptor [118], GAP compares atomic environments, so datasets including structures of different compositions can also be fitted. The environment of atom i is described as an atomic neighbour density where the neighbour atoms are described by Gaussians centred on their relative positions [5, 118, 124]:

$$\rho_i^{\alpha}(\boldsymbol{r}) = \sum_j \exp\Big(- \frac{|\boldsymbol{r} - \boldsymbol{r}_{ij}|^2}{2\sigma_{\text{at}}^2} \Big) f_{\text{cut}}(r_{ij}). \tag{2.26}$$

where the sum is over the neighbours of species α, the neighbour densities being defined separately for the different species. Here σ_{at} is a smoothing parameter depending on the species, and the cutoff function $f_{\text{cut}}(r_{ij})$ is the same as in (2.25) [9, 109]. This atomic neighbour density can also be expanded in the space of orthogonal radial basis functions $g_n(|\boldsymbol{r}|)$ and spherical harmonics $Y_{lm}(\hat{\boldsymbol{r}})$ [5, 118, 124]:

$$\rho_i^{\alpha}(\boldsymbol{r}) = \sum_j \sum_{nlm} c_{nlm}^{(j)} g_n(|\boldsymbol{r}|) Y_{lm}(\hat{\boldsymbol{r}}). \tag{2.27}$$

The similarity of two environments is calculated as the overlap of these atomic densities and rotational invariance is enforced by integrating over three dimensional rotations,

defined by $\hat{\boldsymbol{R}}$ 3×3 matrices [119, 124]:

$$\tilde{k}(\rho_i^\alpha, \rho_j^\alpha) = \int d\hat{\boldsymbol{R}} |\rho_i^\alpha(\boldsymbol{r}) \rho_j^\alpha(\hat{\boldsymbol{R}}\boldsymbol{r}) d\boldsymbol{r}|^n \tag{2.28}$$

where the exponent $n \geq 2$. Usually, $n = 2$ is used as it would become more expensive to evaluate the kernel for larger exponents, and for $n = 1$, the angular information would be lost [118, 125].

As derived in Ref. [118], the power spectrum: $\tilde{p}^i_{nn'l} = \sum_{m=-l}^{l} c^i_{nlm} (c^i_{n'lm})^*$ can be used for calculating the kernel without evaluating the above integral [118, 124]:

$$\tilde{k}(\rho_i^\alpha, \rho_j^\alpha) = \sum_{nn'l} \tilde{p}^i_{nn'l} \tilde{p}^j_{nn'l} = \tilde{\boldsymbol{p}}^i \tilde{\boldsymbol{p}}^j. \tag{2.29}$$

where $\tilde{\boldsymbol{p}}$ is a vector created by the $\tilde{p}_{nn'l}$ elements. Then, the kernel is normalised to one, and the final atomic kernel is calculated as a positive integer power of the normalised kernel [118, 124]:

$$k(\rho_i^\alpha, \rho_j^\alpha) = \left(\frac{\tilde{k}(\rho_i^\alpha, \rho_j^\alpha)}{\sqrt{\tilde{k}(\rho_i^\alpha, \rho_i^\alpha) \cdot \tilde{k}(\rho_j^\alpha, \rho_j^\alpha)}} \right)^\zeta \tag{2.30}$$

where the exponent ζ is a small positive integer, in this work $\zeta = 2$. Using Eq. (2.29), one can also calculate the kernel using the normalised power spectrum [118, 124]: $\boldsymbol{p}^i = \tilde{\boldsymbol{p}}^i / \sqrt{\tilde{\boldsymbol{p}}^i \tilde{\boldsymbol{p}}^i}$:

$$k(\rho_i^\alpha, \rho_j^\alpha) = (\boldsymbol{p}^i \boldsymbol{p}^j)^\zeta. \tag{2.31}$$

2.4.3 Datasets of fits

For the TTM4Fmod+GAPs model, each correction term was fitted by GAP in two steps: the first on the MP2 level and the second on the CCSD(T)-F12a–MP2 differences. We used datasets from different sources; however, in the end, to make the fitted levels uniform, subsets were recalculated at the CCSD(T)-F12a/AVTZ and the DF-MP2/AVTZ levels using the counterpoise correction. In the original datasets, some [10] used only CCSD(T), some [15, 16] used mixed basis functions and some [15, 16] did not use the

counterpoise correction. The details of the fitting datasets of the GAPs are summarised in Table 2.1.

The 2B water database on the MP2/AVTZ level has 12040 configurations in it: 9040 are from Bartók et al. [8] and 3000 are sampled from a molecular dynamics (MD) simulation. The additional configurations were chosen as follows: 2000 structures were chosen by two furthest point samplings (FPSs) [117] from two snapshots of a MD simulation using a cutoff of 7.0 Å. To better sample configurations that have long OO distances, 1000 structures were added with OO distances between 6.0–7.0 Å, which were chosen as a sum of two FPSs each choosing 500 dimers from all the configurations with this distance criterion from different snapshots of a MD simulation. The CCSD(T)-F12a–MP2 2B correction was fitted on a subset of the MP2 database containing 1507 structures [11, 127], for which the quantum chemical energies were recalculated.

The 3B water MP2 dataset consists of 14554 configurations. One part of it is a subset of the dataset by Babin et al. [10], containing 11554 structures which were chosen and calculated with DF-MP2/AVTZ for the paper of Cisneros et al. [11]. (The original training set of [10] contained CCSD(T) level calculations with an AVTZ basis set extended with midbond functions [10].) 3000 additional structures were sampled from a MD simulations as a union of six FPSs, each choosing 500 configurations from different timesteps. The CCSD(T)-F12a–MP2 fit was fitted on a subset of 2654 configurations, chosen from the 11554 configurations of Cisneros et al. [11], which was calculated by CCSD(T)-F12a/AVTZ for this project. The structures are chosen as a union of two FPSs with different settings for the descriptors choosing 1500 each which lead to 2218 structures due to overlaps between the two sets. Finally, we added every 25th configurations that was not selected by the FPS samplings which resulted in 436 additional structures.

For the 2B methane–water term, the dataset of Qu et al. [15] was filtered discarding the configurations that have CO distances larger than 6.0 Å, then a subset was chosen by CUR analyses [116] which resulted in 13713 structures. This dataset was calculated at the MP2/AVTZ level. For the CCSD(T)-F12a–MP2 calculations, further subsets of this set were chosen: the first choosing all the configurations where the CO distance was

Fit	Fitted data	Config.s from sources	Notes
$E^{\mathrm{H_2O}}_{\mathrm{2B(MP2\text{-}TTM4Fmod)}}$	$\Delta E, \Delta \boldsymbol{F}$	9040 from [8] 2000 by 2 FPSs from MD with cutoff of 7.0 Å 1000 by 2 FPSs from MD, with 6Å $< r_{\mathrm{OO}} <$ 7Å	
$E^{\mathrm{H_2O}}_{\mathrm{3B(MP2\text{-}TTM4Fmod)}}$	$\Delta E, \Delta \boldsymbol{F}$	11554 subset of [10], chosen by [11] 3000 by 6 FPSs from MD step	recalculated by [11]
$E^{\mathrm{H_2O}}_{\mathrm{2B(CCSD(T)\text{-}F12\text{-}MP2)}}$	ΔE	1507 from [11]	recalculated
$E^{\mathrm{H_2O}}_{\mathrm{3B(CCSD(T)\text{-}F12\text{-}MP2)}}$	ΔE	2654 subset of [11], chosen as 2 FPSs choosing 1500 + every 25th	recalculated
$E^{\mathrm{CH_4(H_2O)}}_{\mathrm{2B(MP2)}}$	$E, \boldsymbol{F}$	12701 subset of [15] where $r_{\mathrm{CO}} <$ 6.0Å, and further subset by CUR 1000 from MD, chosen randomly by A. V.–M. [126]	recalculated calculated by A. V.–M. [126]
$E^{\mathrm{CH_4(H_2O)_2}}_{\mathrm{3B(MP2)}}$	$E, \boldsymbol{F}$	11777 subset of [16] where $r_{\mathrm{CO,min}} <$ 6.0Å and all $r_{\mathrm{CH/OH}} <$ 1.5Å then subset chosen as a union of FPS and CUR 4000 by 2 FPSs from MD	recalculated
$E^{\mathrm{CH_4(H_2O)}}_{\mathrm{2B(CCSD(T)\text{-}F12\text{-}MP2)}}$	ΔE	2328 subset of [15] chosen from MP2 set: all $r_{\mathrm{CO}} <$ 3.0Å and FPS 1000 from MD, chosen randomly by A. V.–M. [126]	recalculated calculated by A. V.–M. [126]
$E^{\mathrm{CH_4(H_2O)_2}}_{\mathrm{3B(CCSD(T)\text{-}F12\text{-}MP2)}}$	ΔE	1875 subset of [16] where $r_{\mathrm{CO,min}} <$ 6.0Å and all $r_{\mathrm{CH/OH}} <$ 1.5Å chosen by subset as a union of FPS and CUR	recalculated

Table 2.1 Details of the GAP datasets. All quantum chemical calculations used the AVTZ basis set and counterpoise correction (see Section 2.2.2); the CCSD(T)-F12 calculations used the CCSD(T)-F12a/AVTZ and the MP2 calculations used the density-fitted DF-MP2/AVTZ versions.

shorter than 3.0 Å; the second by an FPS sampling from the configurations that have CO distances between 3.0–6.0 Å. For better sampling, 1000 configurations were added to both the sets, sampled and calculated by Álvaro Vázquez–Mayagoitia [126], which were selected randomly from a molecular dynamics simulation of a methane molecule in water. Finally, the distorted configurations that have intra-molecular OH or CH distances longer than 1.5 Å were discarded. The datasets of the fits contained 13701 and and 3328 configurations for the MP2 and CCSD(T)-F12a–MP2 levels, respectively.

For the 3B methane–water–water fit, the database of Conte et al. [16] was filtered so that the structures had at least one CO distance shorter than 6.0 Å; then a subset was chosen by a union of CUR and FPS samplings; and finally, the structures that have intra-molecular OH or CH distances longer than 1.5 Å were discarded. This dataset of 11777 configurations was calculated at the MP2/AVTZ level. Then 4000 additional configurations were added by choosing twice 2000 structures by FPSs from two different snapshots of a MD simulation. The CCSD(T)-F12a–MP2 dataset was chosen similarly to the MP2 set from the database of Ref. [16]. First, the configurations that did not have CO distances shorter than 6.0 Å were discarded. Second, a union of FPS and CUR samplings resulted in 1987 configurations. Finally, the distorted structures that have intra-molecular OH or CH distances longer than 1.5 Å were discarded. Thus, 1875 configurations remained in the dataset.

When generating new structures using MD simulations, the simulations were run in Amber [128]. The periodic box contained one methane molecule and 430 water molecules. After equilibration, the configurations were sampled from an NVT simulation having 0.9 g/cm^3 density.

The datasets were split to train and test subsets, having a 95:5 ratio. The test sets were chosen randomly but uniformly along the CO/OO distances for the 2B sets and along the minimum OO and CO distances for the water 3B and methane–water–water 3B sets, respectively. The datasets and their train subsets are available at [129].

2.4.4 Parameters of the TTM4Fmod+GAP model

The model is built by fitting GAP corrections to a classical water force field, TTM4Fmod, to correct it to the CCSD(T)-F12a/AVTZ level in the 2B and 3B interactions. The 1B terms are not fitted because the water force field includes the highly accurate Partridge–Schwenke 1B potential [130] of water, and the TTM4Fmod+GAPs model includes the 1B CH_4 fit [107, 108] of Schwenke and Partridge, too.

Fitting on the forces can significantly improve the accuracy of the fits because the forces contain additional information [8]; and as the forces are the derivatives of the potential, information about the behaviour of the potential function near the data points is also included by them. Thus, fits including the forces better approximate the quantum mechanical PES, and reduce overfitting of the train data. However, the target CCSD(T)-F12a/AVTZ has a high computational cost and the forces were not available in the version of Molpro used for this project (2012.1), so the forces are only calculated at the MP2 level, and the corrections are fitted in two steps. The first steps target the MP2 level, including energy and force data, and the second steps target the CCSD(T)-F12a level, including energy data only, fitted on smaller datasets than the first steps. The datasets of the fits are described in Section 2.4.3.

The water interactions are calculated by the TTM4Fmod model, along with GAP corrections for the 2B and 3B interactions. These terms are fitted by GAP in two steps: the first targeting the MP2 level fitted on the MP2–TTM4Fmod differences including energy and force differences, and the second targeting the CCSD(T)-F12a level fitted on the CCSD(T)-F12a–MP2 differences.

The methane–water interactions are added to the model for the 1B, 2B and 3B terms, including only one methane molecule. In the clathrates, the methane molecules are further away from each other, so the terms with higher methane content are unnecessary in the first approximation. For the 1B CH_4 PES, a potential [107, 108] fitted by Schwenke and Partridge is used invoked by the code of the authors [107, 108] as downloaded from the supplementary material of [131] which is interfaced to QUIP for this project [127]. This potential is not as accurate as the Partridge–Schwenke water 1B potential [130] as

it is fitted only on the CCSD(T)/VTZ level [107]. The methane–water 2B and methane–water–water 3B terms are built similarly to the water terms with the difference that the first step is fitted directly on the MP2 level as there is no methane–water force field included. Thus, the model is only applicable to low methane concentrations currently. It would be interesting to add 2B and 3B GAPs with more methane molecules in future work.

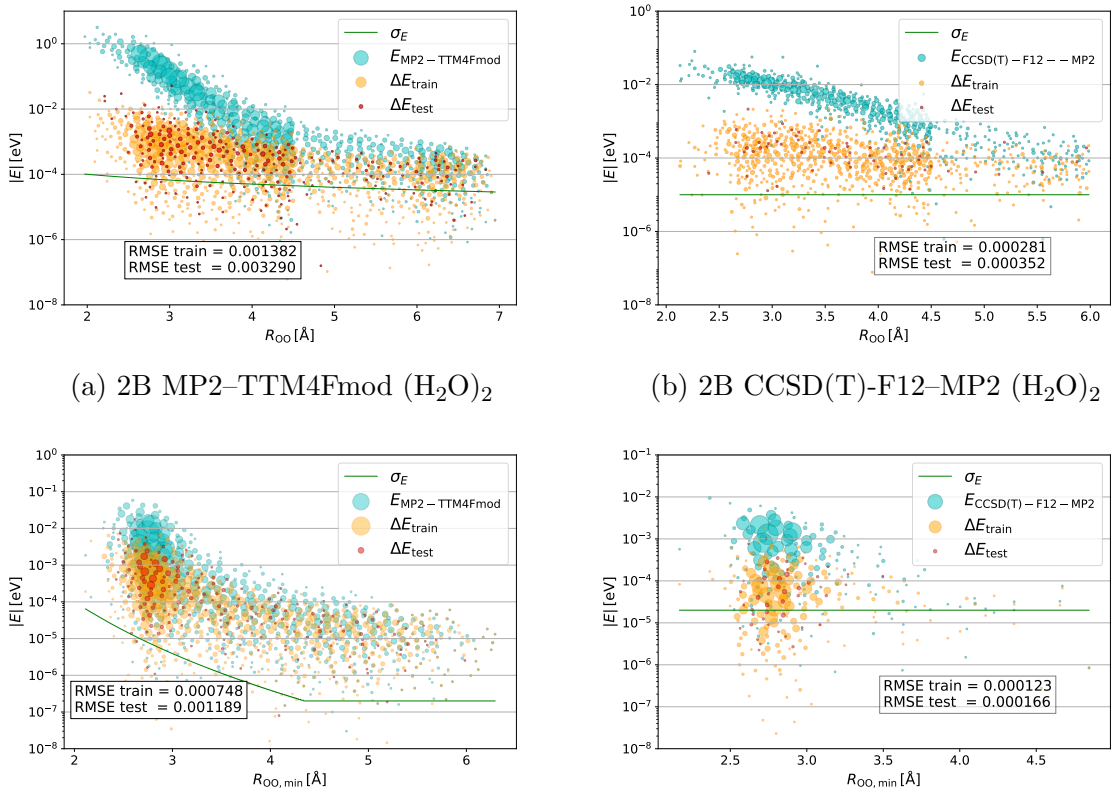
All the fits used inverse general distance descriptors with distance shifts. The hyperparameters of the fits were optimised for each fit manually. The monomer cutoffs were 1.9 Å and 1.3 Å for water and methane, respectively. The fits on the MP2 level used sparse points chosen with the FPS method [117]. As the CCSD(T)-F12a–MP2 datasets were a magnitude smaller than the MP2 ones, the fits used all the data points as sparse points, except for the methane–water–water 3B fit, which fit was already accurate enough with a smaller number of randomly chosen sparse points.

The regularising parameters of the energies (σ_E) and forces (σ_F) were scaled for all the fits on the MP2 level; and σ_E was also scaled for the 2B methane–water fit on the CCSD(T)-F12 level. These can be set for each configuration under the energy_sigma and force_sigma names in the xyz files. For the other fits, the regularising parameters were set to constant. The σ_E parameters were scaled with powers of the distances and set to each configurations as:

$$\sigma_E = \max\left(\frac{\sigma_{E,0}}{\left(\frac{r}{r_0}\right)^p}, \sigma_{E,\min}\right) \tag{2.32}$$

where r is a distance between non-hydrogen atoms (see Table 2.2), r_0 is set to 2 Å, and p is a positive integer power. In the case of the fits on the MP2 level, the regularising parameters for the forces were also scaled. For the methane–water fits, they were scaled with the distance similarly as the σ_E:

$$\sigma_F = \max\left(\frac{\sigma_{F,0}}{\left(\frac{r}{r_0}\right)^p}, \sigma_{F,\min}\right) \tag{2.33}$$



(a) 2B MP2–TTM4Fmod $(H_2O)_2$

(b) 2B CCSD(T)-F12–MP2 $(H_2O)_2$

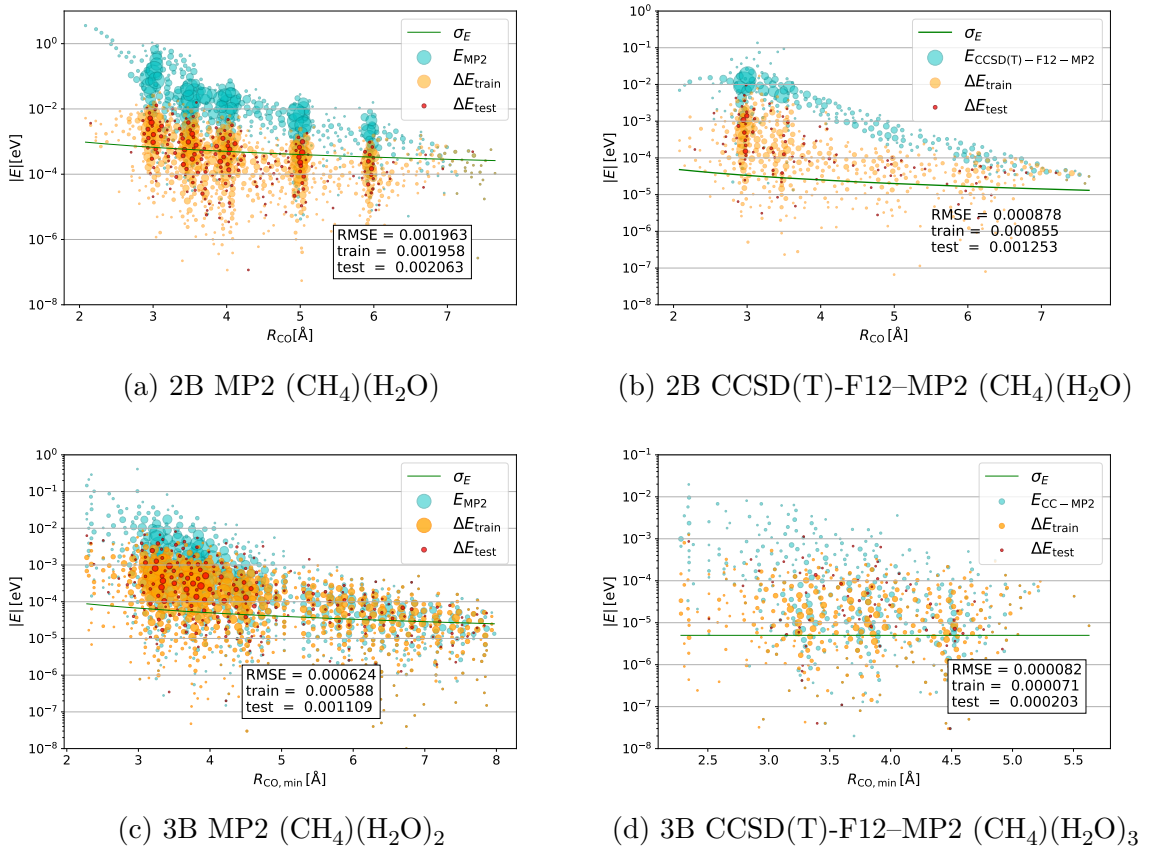
(c) 3B MP2–TTM4Fmod $(H_2O)_3$

(d) 3B CCSD(T)-F12–MP2 $(H_2O)_3$

Fig. 2.3 Energy plots of the water GAPs. In the plots, the larger circles visualise multiple points which would be at the same coordinates. The plots show the energy distribution of the train energies with light blue, and the energy error distributions of the fit for the train and test sets with orange and red, respectively. Both the energies and energy errors are shown with respect to the OO distances of the dimers and minimum OO distances of the trimers. The green σ_E is the regularising parameter of the energy.

$\sigma_{F,\mathrm{at}}$ scaled with $\lvert F_{\mathrm{at}}\rvert$	$\sigma_{E,0}$	r	p	$\sigma_{E,\mathrm{min}}$	$\sigma_{F,\mathrm{frac}}$	$\sigma_{F,\mathrm{min}}$
$E^{\mathrm{H_2O}}_{\mathrm{2B(MP2\text{-}TTM4Fmod)}}$	0.0001	r_{OO}	1	$1\cdot10^{-5}$	0.1	0.001
$E^{\mathrm{H_2O}}_{\mathrm{3B(MP2\text{-}TTM4Fmod)}}$	0.0001	$r_{\mathrm{OO,min}}$	8	$2\cdot10^{-7}$	0.01	0.0001
σ_F scaled with r	$\sigma_{E,0}$	r	p	$\sigma_{E,\mathrm{min}}$	$\sigma_{F,0}$	$\sigma_{F,\mathrm{min}}$
$E^{\mathrm{CH_4(H_2O)}}_{\mathrm{2B(MP2)}}$	0.001	r_{CO}	1	$1\cdot10^{-5}$	0.005	0.001
$E^{\mathrm{CH_4(H_2O)_2}}_{\mathrm{3B(MP2)}}$	0.0001	$r_{\mathrm{CO,min}}$	1	$1\cdot10^{-5}$	0.001	0.001
only σ_E scaled	$\sigma_{E,0}$	r	p	$\sigma_{E,\mathrm{min}}$	-	-
$E^{\mathrm{CH_4(H_2O)}}_{\mathrm{2B(CCSD(T)\text{-}F12\text{-}MP2)}}$	$5\cdot10^{-5}$	r_{CO}	1	$1\cdot10^{-6}$	-	-

Table 2.2 The parameters used to set the regularising parameters for the fits. $r_0 = 2.0\,\text{Å}$.



(a) 2B MP2 $(CH_4)(H_2O)$

(b) 2B CCSD(T)-F12–MP2 $(CH_4)(H_2O)$

(c) 3B MP2 $(CH_4)(H_2O)_2$

(d) 3B CCSD(T)-F12–MP2 $(CH_4)(H_2O)_3$

Fig. 2.4 Energy plots of the methane–water GAPs. The notation is similar to the one in Fig. 2.3, but the energies and energy errors are shown with respect to the CO distances of the dimers and minimum CO distances of the trimers.

	fitted data	r_{cutoff}	$N_{\text{sparse points}}$ ($N_{\text{configurations}}$)	$\text{RMSE}_{\text{test}}$
$E^{H_2O}_{2B(MP2\text{-}TTM4Fmod)}$	$\Delta E, \Delta \boldsymbol{F}$	7.0	2000 (11442)	0.003290
$E^{H_2O}_{3B(MP2\text{-}TTM4Fmod)}$	$\Delta E, \Delta \boldsymbol{F}$	6.0	5000 (13832)	0.001189
$E^{H_2O}_{2B(CCSD(T)\text{-}F12\text{-}MP2)}$	ΔE	7.0	1435 (1435)	0.000352
$E^{H_2O}_{3B(CCSD(T)\text{-}F12\text{-}MP2)}$	ΔE	5.0	2524 (2524)	0.000113
$E^{CH_4(H_2O)}_{2B(MP2)}$	$E, \boldsymbol{F}$	7.0	5000 (13021)	0.002063
$E^{CH_4(H_2O)_2}_{3B(MP2)}$	$E, \boldsymbol{F}$	6.0	5000 (14998)	0.001109
$E^{CH_4(H_2O)}_{2B(CCSD(T)\text{-}F12\text{-}MP2)}$	ΔE	7.0	3167 (3167)	0.001253
$E^{CH_4(H_2O)_2}_{3B(CCSD(T)\text{-}F12\text{-}MP2)}$	ΔE	5.0	700 (1787)	0.000203

Table 2.3 Details of the GAP fits. The units of the cutoff radius and RMSEs are Å and eV, respectively.

For the water interactions, they were set using the force_atom_sigma parameter, which can set them to different values for each atom; they were calculated as a fraction of the magnitude of the atomic forces:

$$\sigma_{F,\mathrm{at}} = \max\Big((|F_{\mathrm{at}}| \cdot \sigma_{F,\mathrm{frac}}), \sigma_{F,\mathrm{min}}\Big) \tag{2.34}$$

where F_{at} is the target atomic force (the 2B or 3B MP2–TTM4Fmod difference force in this case). To avoid overfitting, we used the minimum values $\sigma_{E,\mathrm{min}}$ and $\sigma_{F,\mathrm{min}}$. For the parameters used in the fits building the TTM4Fmod+GAPs model, see Table 2.2. Scaling the sigmas with the distance meant we could set a lower target accuracy for the configurations with higher energies which often occur at shorter distances. Thus, the models are more accurate for the configurations that occur more often in dynamical simulations or geometry optimisations.

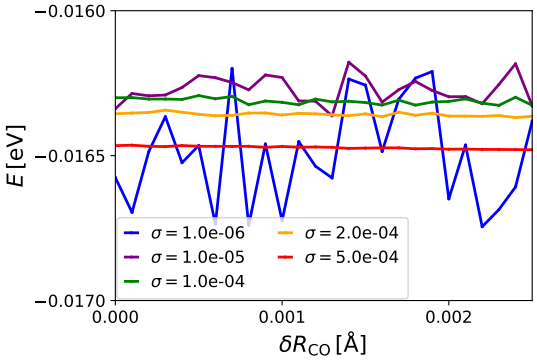


Fig. 2.5 PES scan of a methane–water dimer for different GAPs fitted on an earlier version of the MP2 database. The dimer studied is taken from an sI clathrate structure optimised by an initial version of the TTM4Fmod+GAPs model, in which the atoms are moved by very small distance steps (0.0001 Å) in the direction of the corresponding atomic forces (calculated by the initial TTM4Fmod+GAPs model), the inter-molecular CO distance of this dimer being $R_{\mathrm{CO}} = 4.049$ for the first configuration. The GAPs differ only by changing the values of the regularisation parameters, σ_E and σ_F, that are set to the same constant value in this case. All the train geometries are used as sparse points. As the distance changes are very small, a smooth potential is expected, so the GAPs that predict energies changing very quickly are presumably overfitted. The figure shows that setting the regularisation parameters to too low constant values can lead to overfitting. (Note: when using a smaller number of sparse points, it also helps reduce overfitting.)

The accuracies of the GAPs with different parameters were compared by looking at their error distributions with respect to some selected distances using the plotting tools of M. D. Veit [132], as shown shown in Figures 2.3, 2.4. We chose fits for which not only the RMSEs were low, but the errors did not have large outliers. However, while running test MD simulations with the initially optimised models, the MDs sometimes failed. We found that this happened because the potentials were not smooth due to overfitting the train sets. To reduce overfitting, we decreased the number of sparse points and increased the regularising parameters. We tested whether the fits are overfitted by the gradient tests of the QUIP package and by looking at the smoothness of PES scans (see Fig. 2.5).

The data included in the fits, the cutoffs, the number of sparse points, the number of configurations in the fitting database, and the test energy root mean square errors (RMSEs) of the GAPs building the model are shown in Table 2.3. The files of the GAPs are available at [129] along with the corresponding datasets. The exact parameters of the fits are given in Appendix B and the energy plots of the fits are shown in Figures 2.3, 2.4.

2.5 Permutationally Invariant Polynomials (PIP)

In the permutationally invariant polynomials [3], the potential energy surface is fitted in the space of polynomials built from symmetrised monomials of distances between atoms, sometimes including distances between atoms and lone electron pairs, too [4–6, 10]. The functions are permutationally symmetrised with respect to exchanging the atoms (or lone pairs) of the same type within a molecule and with respect to exchanging the molecules of the same composition [4, 6, 10, 133]. The parameters are optimised using kernel-ridge regression [5, 6]. Versions of PIP have been applied to fit the PES for numerous systems including molecules and materials [4, 6, 10, 15, 16, 134].

For the PIPs of Chapter 3, the polynomials are built from symmetrised monomials up to the fourth degree of the functions: $e^{-kr_{ij}}$, $e^{-k(r_{ij}-r_{ij}^{(0)})}$ and $e^{-kr_{ij}}/r_{ij}$ [5] where r_{ij} are the distances between atoms and in the case of the 2B fit, also the lone pairs. These

fits are the same ones that are included in MB-pol water model and are described in detail in the publications on the model [4, 10].

2.5.1 The MB-nrg model

During the course of this project, MB-pol was extended to include other molecules by the same group that developed MB-pol, and this new model is named MB-nrg [6,22,135]. The MB-nrg and the current version of MB-pol are available in the MBX software [22], and at the time of this thesis, only this version is supported [136]. This software has several advantages over the old MB-pol plugin for OpenMM: firstly, its electrostatic baseline, TTM4Fmod [4] (described in Section 2.3 for water), includes the methane molecules as well [6, 22]. Secondly, as it is implemented using OpenMP parallelisation [22], it is faster. Moreover, it was simple to interface it to QUIP on the Fortran level as it already had Fortran function calls implemented.

The water part of MB-nrg is the MB-pol model [4, 10]. The 2B water PIP was trained on a dataset of 42508 structures with energies by the CCSD(T) method with the counterpoise correction and a basis set extrapolation between the AVTZ and AVQZ basis sets supplemented by midbond functions [4]. The 3B water PIP was fitted on a dataset of 12347 trimers calculated at the CCSD(T) level using the AVTZ basis set supplemented by midbond functions [10].

The parameters of the methane–water part of the TTM4Fmod model were determined by fitting to quantum chemical data by M. Riera et al. [6]. The 1B and 2B terms including methane molecules were also published in the same paper [6]. The target level for the 1B CH_4 term was CCSD(T)-F12b using a 2-point basis set extrapolation between AVTZ and AVQZ, and this PIP was fitted on 7882 structures [6]. The 2B datasets used CCSD(T)-F12b with counterpoise correction and basis set extrapolation between AVDZ and AVTZ, and these sets consisted of 32811 and 48239 configurations for $(CH_4)(H_2O)$ and $(CH_4)_2$, respectively [6]. The 3B $(CH_4)(H_2O)_2$ term, added for this project, was fitted on 15777 configurations with counterpoise-corrected MP2/AVTZ energies and is described in the next section (Section 2.5.2). We test the model with and without this

3B correction term in Sections 4.3, 5.3.1 by simply turning off this PIP in the input json file, and we use the full complemented model for the periodic calculations using the quasi-harmonic approximation.

The polynomials of the PIPs are built from monomials of interatomic distances and for the 2B terms, distances between atoms and lone electron pairs of the water molecules [4, 6, 133]. The monomials are up to fifth degree for the 1B CH_4 term, up to fourth degree for the 2B $(H_2O)_2$, 3B $(H_2O)_3$ and 2B $(CH_4)_2$ terms, and up to third degree for the 2B $(CH_4)(H_2O)$ and 3B $(CH_4)(H_2O)_2$ terms [4, 6, 10].

In the earlier version of the model, the geometry optimisation of clathrate structures at high pressures sometimes found holes in the potential and failed. The holes are geometries where the potential predicts physically not correct, very low energies. Thus, the geometry optimisation will "fall into" these geometries and predict non-physical structures; sometimes even fail due to other errors occurring related to numerical overflow or atoms being at the same positions. To avoid this, M. Riera–Riambau added a switch according to the monomer energies [133]. When the 1B energies are higher than EMAX1B (currently set to 60 kcal/mol (2.602 eV)), the model is switching back to the TTM4Fmod electrostatics from the PIPs. This means the model will have lower accuracy for structures that include distorted molecules, which might occur at high pressures.

2.5.2 The methane–water–water fit for the MB-nrg model

A 3B methane–water–water term was added to MB-nrg during the author's visit to Francesco Paesani's group (University of California, San Diego) with the help of Marc Riera–Riambau [133]. This term is fitted on the dataset assembled for the methane–water–water MP2 GAP of this thesis, described in Section 2.4.3. We also tested fitting on the CCSD(T)-F12 dataset, but it was too small for the PIP fits.

We used the MB-nrg fitting code for optimising the PIP as M. Riera et al. [6] did for the other terms of the model describing methane interactions. The polynomials are built from monomials up to third degree of the functions: $e^{-kr_{ij}}$ where r_{ij} are interatomic distances [133].

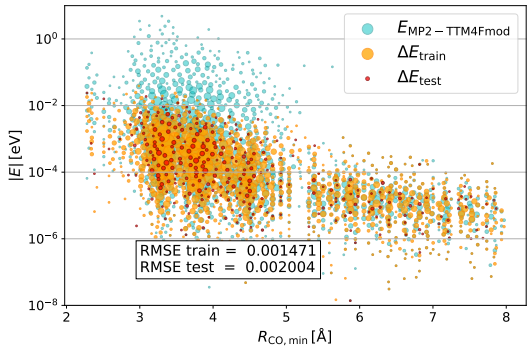


Fig. 2.6 Energy plot of the 3B methane–water–water fit for MB-nrg. The notations are similar to Fig. 2.3, but here the x axis is the minimum CO distance of the trimers.

The energy range of the fit was set to 20 kcal/mol (0.867 eV) and the regularization parameter was set to 0.0005 kcal/mol ($2.168 \cdot 10^{-5}$ eV). The cutoff was 6.0 Å, with the cutoff transition starting at 5.0 Å. The code used the "binding energies" for weighting the configurations for the fits of the 3B energies. As the "binding energy" is only for the weights, we used the differences between the total energy of the structures and the lowest total energy of the dataset in the end. Optimising the fit, we made 20 PIP fits in MB-nrg with the same parameters and chose the one having the lowest RMSE on the low energy train set, also looking at whether the full train set RMSE is of the same magnitude. (This low energy train set was defined as the set of configurations that have binding energies which are within the energy range of the fit (0.867 eV) of the energy of the configuration having the lowest binding energy [133].) The RMSEs were 0.4786 meV for the low energy train set and 1.471 meV for the full train set. The test set's RMSE was 2.004 meV. The energy error distribution of the fit on the train and test sets is shown in Fig. 2.6.

2.5.3 The MBX-to-QUIP interface

MBX is interfaced to QUIP by calling the Fortran functions of MBX through the IPModel template of QUIP, which is also written in Fortran. For easier initialisation of the potential for different structures, python functions are written – available in

mbx_functions.py at [137], which generate the strings needed for the MBX's Fortran function. Also, the periodic MB-pol and MB-nrg models need larger minimum boxsizes than the unit cells of some of the structures of the project, so for quicker simulations, an ASE [105] Calculator is created which increases the cell only for the time of the calculation and gives back the result for the original structure (SuperCellCalcMBX in SuperCellCalcMBX.py). This function enhances the speed of dynamical simulations or geometry optimisations, for example. To initialise the nonperiodic version similarly in one line, the Calculator CalcMBX (in CalcMBXnonperiodic.py) is created.

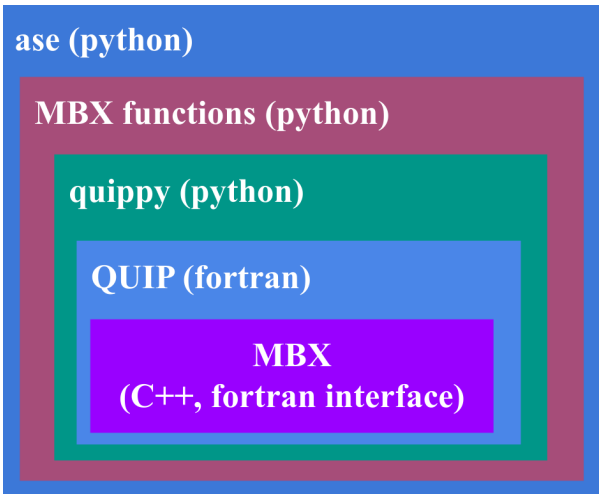


Fig. 2.7 Schematic graph of the MBX-to-QUIP interface

2.6 Differences between TTM4Fmod+GAPs and MB-nrg

The differences between the water parts of the TTM4Fmod+GAPs model and MB-nrg, called MB-pol for the water part, arise from using different fitting techniques, different databases (however, the 3B sets of GAP are subsets of MB-pol's fitting set), different cutoff functions and different levels of coupled cluster approximated by them. While the GAP model's target level is the CCSD(T)-F12a/AVTZ in each term, MB-pol's level is the CCSD(T)/CBS in the 2B [4] and the CCSD(T) level calculations with an AVTZ

basis set extended with midbond functions in the 3B [10]. Moreover, in the earlier version of MB-pol [103], from which the baseline force field is invoked for the GAP model, the cutoff of the 3B PIP had a bug and used the maximum OO distance only. This meant that the correction term was not applied to the configurations which only had two OO distances within 4.5 Å. However, this is corrected in the MB-pol version available in the MBX software [133].

The two models differ even more in the interactions involving methane molecules. While TTM4Fmod+GAPs only has the TTM4Fmod baseline for the water part, MB-nrg also has it implemented for the electrostatic interactions including methane molecules. As for the correction fits, both models have fits for the methane–water and methane–water–water interactions, and MB-nrg also has a 2B fit for the methane–methane term. The methane–water–water 3B fits use the same MP2 dataset for the two models; however, the PIP is only fitted on the MP2 level because the CCSD(T)-F12–MP2 dataset is not large enough for it. The methane–water datasets and their target levels are different. While GAP targets the CCSD(T)-F12a/AVTZ level, the PIP of MB-nrg targets the CCSD(T)-F12b level using a basis set extrapolation with the AVDZ and AVTZ basis sets. However, as the double zeta basis set has low accuracy, this extrapolation might lead to energies with lower accuracy than the pure AVTZ basis. Also, the Molpro manual suggests using the F12a version of CCSD(T)-F12 for basis sets up to triple zeta [99]. Moreover, MB-nrg uses a PIP for the 1B methane PES, whereas TTM4Fmod+GAPs includes the 1B PES of Schwenke and Partridge [107,108] which are also fitted on different levels; in this term, the level of MB-nrg has the higher accuracy. Finally, the switch function of MB-nrg for high monomer energies is not present in the GAP model.

The QUIP interface of the MBX software is faster than the one of the TTM4Fmod+GAPs model due to multiple reasons. Firstly, it is interfaced on the Fortran level, while the earlier implementation of TTM4Fmod in the MB-pol plugin for OpenMM was interfaced on the python level. Secondly, MBX uses parallelisation for the baseline force field. Finally, the PIPs can be evaluated faster than the GAPs – also because a part of the 3B GAP finding the trimers is not parallelised yet.

2.7 Other softwares

The calculations with the developed potentials and the structure manipulations are performed using the python packages: ASE [105] and quippy [9]. The Molpro software, version 2012.1 [94], is used for all the quantum mechanical calculations of the thesis. It is invoked by an interface to QUIP [9] written by A. Nichol and updated by Max D. Veit and the author of the thesis to include more options.

The data was processed in Jupyter notebooks [138], and the graphs were made using the Matplotlib [139] package. To show the error distributions of the fits on the "bubble plots" (e.g. Fig. 2.3), the python tools of Max D. Veit were used (available at [132]). The structures were visualised in the Ovito software [140].

The GenIce software [141, 142] was used to generate hydrogen-disordered structures for the ices and clathrates. The WebPlotDigitizer software [143] was applied to extract data from published graphs. The thesis is written using a template provided by the University of Cambridge [144].

Chapter 3

Comparison of different potential-fitting methods

3.0.1 Author contribution details

This chapter is based on work done with collaborators for Ref. [5]. The GAPs were fitted by the author of the thesis. Fig. 3.1 and Table 3.1 were made by the first author, Thuong T. Nguyen, and are adapted from the paper [5].

3.1 Introduction

In the recent years, several methods have been developed to approximate the quantum chemical potential energy surface by high dimensional fits. Amongst these methods, different representations of the structures and different fitting techniques were used. In this chapter, we report results of our study [5] in which we compared three popular potential fitting methods by applying them to the same datasets: the Gaussian Approximation Potentials (GAP) [1], permutationally invariant polynomials (PIP) [3] and Behler–Parinello neural networks (BPNN) [7]. This allows both a comparison of relative accuracy of each method and provide further evidence that many-body expansion (MBE) and data-driven techniques combined can create accurate potentials for water. For short

descriptions of PIP and GAP, see the corresponding sections of Chapter 2. BPNN is described by Behler and Parrinello in [7, 145], and a short description is in the paper [5].

3.2 Datasets of the paper

The fits for the comparison are performed on the two datasets that were used for the development of the MB-pol model [4, 10] (briefly described in Section 2.5.1). However, the 2B training set was modified by removing the structures having high binding energies (above 60 kcal/mol (2.6 eV)) or larger than 6.5 Å OO distances [5]. The energies fitted are coupled cluster–TTM4Fmod differences as described in Section 2.5.1. The datasets were split to train (81 %), validation (9 %) and test (10 %) sets [5], and the best fits were chosen according to their accuracy on the train and validation sets.

3.3 Parameters of the GAP fits

The 2B and 3B GAPs were both fitted according to the same strategy. The descriptors used were double SOAPs where the atomic energies are calculated as sums of two kernels with different cutoffs and smoothing parameters. (For a description of SOAP, see Section 2.4.2.) The SOAPs building up the double SOAPs had cutoffs of 4.5 Å and 6.5 Å for the 2B, and 4.5 Å and 7.0 Å for the 3B fit. The smoothing parameters were 0.4 Å and 1.0 Å for the descriptors with the shorter and longer cutoffs, respectively. The number of radial basis functions and the spherical harmonics basis band limit were both set to 10. The 2B and 3B fits used 10000 and 9000 sparse points, respectively, chosen by the CUR method [116] as implemented in QUIP. The exact parameters are given in the supplementary material of the paper [5].

3.4 Results

All three fitting methods achieved high accuracy for the two datasets. As shown in Table 3.1, the test RMSEs of the fits are of the same magnitude, and all of them are

below 4 meV. According to the test RMSEs, the best fits for both terms were achieved by PIP which used the same parameters as optimised for the MB-pol model earlier [4, 10]. The values of GAP were only about 10 % higher, while BPNN's values were about 60 % and 35 % higher for the 2B and 3B fits, respectively. However, the 3B GAP fit overfitted the train set very much: the train RMSE was only tenth of the test RMSE.

We also studied the errors of the models developed with the different techniques on small water clusters. As shown in Fig. 3.1, all the methods achieved accuracies within 0.3 kcal/mol (13.01 meV) for the summed 2B and 3B energies. Dividing by the number of molecules in the clusters, this value is within 2.2 meV / H_2O for all the structures. Note that although the 3B GAP seemed to overfit the train set, it performs well for these clusters, having lower errors than the other two methods.

As for the computational time required for the fitting, which is a one-time cost, all the three fitting codes use parallelisation, so were ready in several hours [5]. The fits were done on different computing facilities, so we only provide rough comparisons here. The PIPs required about third of the CPU hours of the time of the GAPs – GAP required 150 and 64 CPU hours for the 2B and 3B fits, respectively [5]. The BPNN code was run on GPUs, and was ready within 3 and 1 hours for the 2B and 3B fits, respectively [5].

In this paper [5], we did not study the cost of evaluation – this cost would also depend on how fast the methods collect the dimer and trimer structures of the larger systems, how many basis functions are used for a fit, and how large minimum cell sizes the models require.

Table 3.1 RMSE (in kcal/mol) per isomer on the provided training, validation, and test sets in the PIP, BPNN, and GAP short range interaction two-body (2B) and three-body (3B) energy fitting. Reproduced from Nguyen, T. T. et al., J. Chem Phys 148(241725), (2018), with the permission of AIP Publishing.

	2B			3B		
	training	validation	test	training	validation	test
PIP	0.0349	0.0449	0.0494	0.0262	0.0463	0.0465
BPNN	0.0493	0.0784	0.0792	0.0318	0.0658	0.0634
GAP	0.0176	0.0441	0.0539	0.0052	0.0514	0.0517

In summary, this study shows that the three fitting methods can achieve similar accuracies when applied to high-quality datasets which sample the configuration space well. Thus, when building data-driven models, it is probably more important to choose the underlying datasets carefully than to choose the fitting methods when choosing from state-of-the-art techniques.

For future work, it would be interesting to test these techniques on datasets including force data along with the energies. Including the forces would also reduce overfitting for the 3B GAP. Also, fitting in two steps, the first on the MP2 level including the forces and the second on the coupled cluster–MP2 differences, could be tested. Moreover, testing how well they extrapolate to structures outside the geometry space of the train set would be useful for applications. Finally, comparing the data efficiency of the techniques would also be important for future projects fitting potentials.

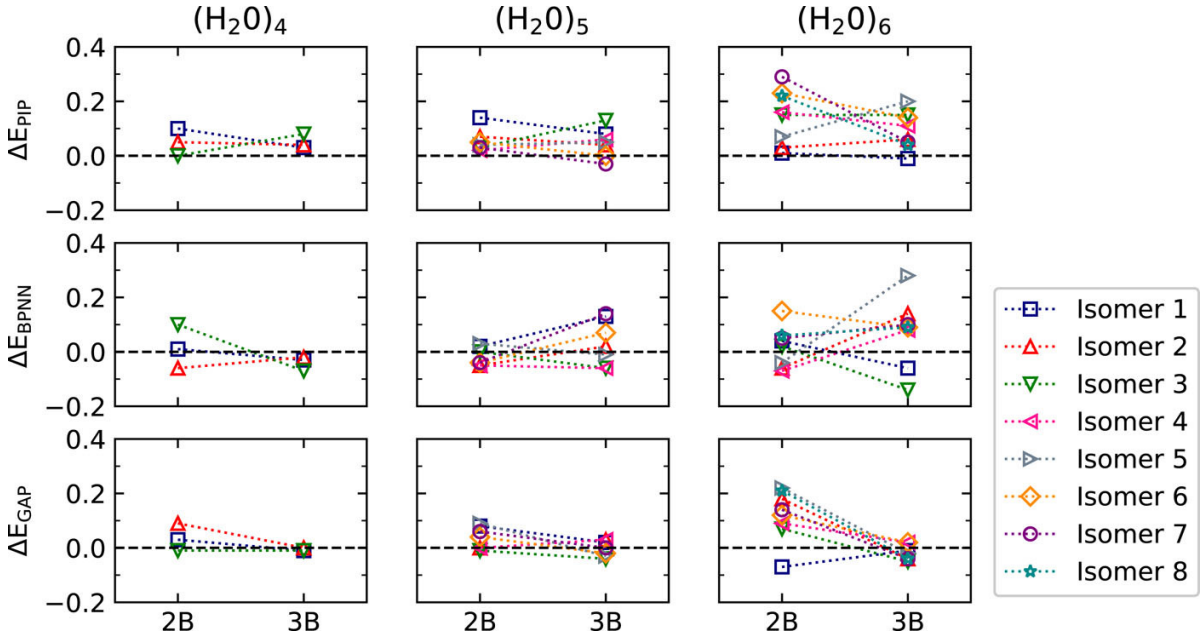


Fig. 3.1 Errors (in kcal/mol) in the 2B and 3B interaction energies calculated with PIP, BPNN, and GAP short-range potentials with respect to reference CCSD(T) values for water clusters $(H_2O)_n$, $n = 4, 5, 6$. Reproduced with permission from Nguyen, T. T. et al, J. Chem. Phys. 148(241725), (2018), with the permission of AIP Publishing.

3.5 Other comparative studies in the literature

Another study comparing neural networks and Gaussian Processes (GP) by Kamath et al. [146] fitting the PES of the formaldehyde monomer (H_2CO) was published in the same issue of J. Chem. Phys. as our study. They used the same descriptors of distances and angles for both methods and found that the GP performed better than the NN [146].

Qu et al. [122] compared PIPs, their version of GP, and PIP–GP, which is a permutationally invariant GP, on small molecules. Note that GAP also has permutationally invariant descriptors. They found the methods to perform very well, PIP-GP achieving the highest accuracy [122].

In the last year, Käser et al. [147] compared three other potential energy fitting techniques also on the example of a formaldehyde monomer, including the forces in the training data. They studied a method based on neural networks and two versions of kernel-ridge regression [147]. Their study also found that all the three techniques achieved high accuracy, the kernel-ridge methods being more accurate and better at extrapolation than the neural networks [147].

Obviously, the performance of the methods depends on the choice of descriptors, implementations of the techniques, the systems studied and the quality and size of the datasets, so it is worth testing various techniques and chemical species in the future.

Chapter 4

Static energy calculations for small clusters

4.1 The reference quantum chemical method: DMC

As the target quantum chemical method of the developed models, CCSD(T)-F12, has high computational cost and size restrictions, we compare our results to DMC (diffusion quantum Monte Carlo) energies for larger clusters. DMC is a version of QMC (quantum Monte Carlo) which is a stochastic wave function based method. It scales as N^3 [148, 149] with a large prefactor, so it is applicable to periodic systems.

DMC and coupled cluster have been thought to be comparable as in principle both the methods converge to the solution of the Schrödinger equation in the Born–Oppenheimer approximation [150, 151]; however, there are small differences between the energy results of these methods [98, 150, 152]. The selected settings affect the results of both the coupled cluster (where the options include basis sets, counterpoise correction, basis set extrapolation, using different variants of F12 or not using it at all) [91] and the DMC (where the options include fixed node approximation, finite time step, choice of pseudopotentials, modifications to Green's function) [98, 153]. Zen et al. [153] studied the size consistency of DMC and suggested a new modification to the Green's function to decrease the size consistency errors for large time steps. They showed that with a time

step of 0.005 a.u., the error can be as much as 20 meV for the methane–water dimer [153], which error was reduced when using their modification. The same time step was used by [154] which we compare the developed models and variants of the coupled cluster method to in Section 4.3; however, as tested in the paper [154], changing the time step to 0.002 a.u. only causes differences up to 8.16 meV per monomer, which is less than half of the above error. (This difference between the two papers is probably due to different other setting choices and different geometries.) The improvement of Zen et al. [153] regarding the modification of the Green's function was also used in the newer ice DMC results [155] that we compare our results to in Section 5.2.1.

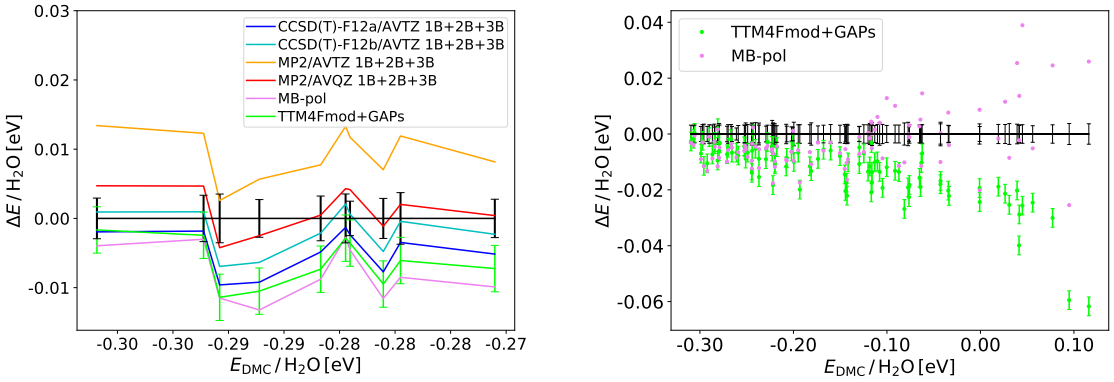
In this chapter, the water clusters and the methane–water clusters studied are from Alfè et al. [156, 157] and Gillan et al. [154], respectively; all downloaded from Dario Alfè's website [158]. The structures are indexed starting from 0 (python indexing) using the order of the downloaded files. After showing that the developed models and the fitted quantum chemical level differ systematically from DMC, we also compare other versions of CCSD(T)-F12 to DMC in Section 4.3.1.

4.2 Water clusters

The developed models are compared to the DMC results on the compressed water clusters of Alfè et al. [156, 157], which were sampled from molecular dynamics (MD) simulations of the clusters in Ref. [156]. The 1B, 2B and 3B energies of the first ten nonamers are also calculated with the fitted levels: MP2 and CCSD(T)-F12. (These quantum mechanical calculations were performed by the author for her Master's thesis [159].)

For the first ten nonamers, both MB-pol and TTM4Fmod+GAPs are more negative than DMC, as shown in Fig. 4.1a. Sometimes this difference is within the error bars for the GAP model, but mostly it is larger than the error bars. The fitted level, CCSD(T)-F12a/AVTZ is also more negative than DMC, and the GAP–CCSD(T)-F12/AVTZ differences are smaller than the predicted error bars of GAP, so the GAP–DMC differences might be consequences of the fitted level differing from DMC. Interestingly,

MP2/AVTZ seems to be systematically more positive than DMC, and the MP2/AVQZ–DMC differences are smaller and have both signs. Looking at all the 100 nonamers with the developed models in Fig. 4.1b, TTM4Fmod+GAPs is still always more negative than DMC. However, MB-pol is mostly negative for the structures having DMC energies lower than $-0.1\,\mathrm{eV}\,/\,H_2O$ but its errors have both signs for structures having DMC energies above this value. The average errors (with the mean absolute errors in parenthesis) are $-12.75\,\mathrm{meV}$ ($12.80\,\mathrm{meV}$) and $-3.264\,\mathrm{meV}$ ($7.880\,\mathrm{meV}$) per H_2O for the GAP model and MB-pol, respectively. However, as shown in Table 4.1, the errors are much larger for both models for the structures with DMC energies larger than $-0.1\,\mathrm{eV}\,/\,H_2O$.

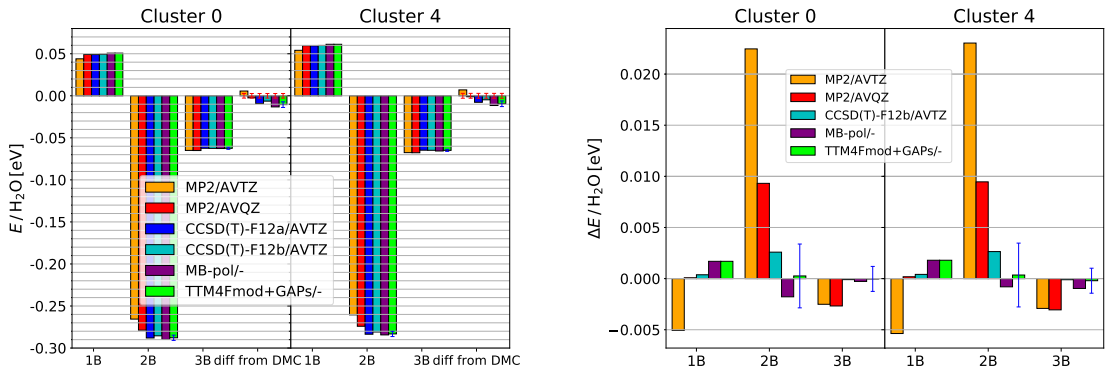


(a) (b)

Fig. 4.1 The total energy differences from DMC using different methods on the compressed water nonamers shown with respect to the DMC energies. a) shows the first 10 structures including quantum chemical calculations summing up the different body-terms up to the 3B terms (note: the lines are only to guide the eye) and b) shows all the structures using the studied data-driven models.

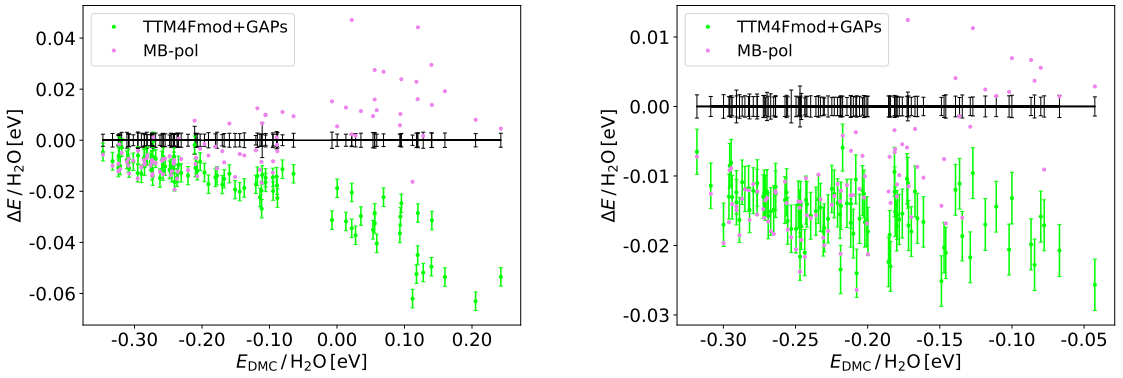
Average errors [meV]	structures 0–9	all structures	$E_{\mathrm{DMC}} < -0.1\,\mathrm{eV}$	$E_{\mathrm{DMC}} > -0.1\,\mathrm{eV}$
TTM4Fmod+GAPs	−6.256	−12.75	−8.661	−23.80
MB-pol	−7.773	3.264	−5.580	3.000
MAE [meV]	structures 0–9	all structures	$E_{\mathrm{DMC}} < -0.1\,\mathrm{eV}$	$E_{\mathrm{DMC}} > -0.1\,\mathrm{eV}$
TTM4Fmod+GAPs	6.256	12.80	8.729	23.80
MB-pol	7.773	7.880	6.635	11.25

Table 4.1 Average errors and mean absolute errors (MAE) for the 100 nonamers and for a split based on whether the DMC energy is lower or higher than $-0.1\,\mathrm{eV}\,/\,H_2O$, resulting in 73 and 27 structures in each group, respectively.



(a) Total energy (b) Differences from CCSD(T)-F12a/AVTZ

Fig. 4.2 Many-body expansion of the energy of water nonamers number 0 and 4. These structures were the ones that had the largest differences from DMC using MB-pol from the first ten structures (and largest and third largest differences using TTM4Fmod+GAPs). The DMC energies are -0.2922 and $-0.2811\,\mathrm{eV/\,H_2O}$. a) shows the summed 1B, 2B and 3B terms of the total energy, "diff from DMC" is the difference of the total energy from the DMC energy; and b) shows the differences from CCSD(T)-F12a/AVTZ in the summed 1B, 2B and 3B energies using different methods.



(a) Compressed pentadecamers (b) Compressed icosiheptamers

Fig. 4.3 The total energy differences from DMC using the TTM4Fmod+GAPs and MB-pol models on the larger compressed water clusters of Ref. [156] shown with respect to the DMC energies.

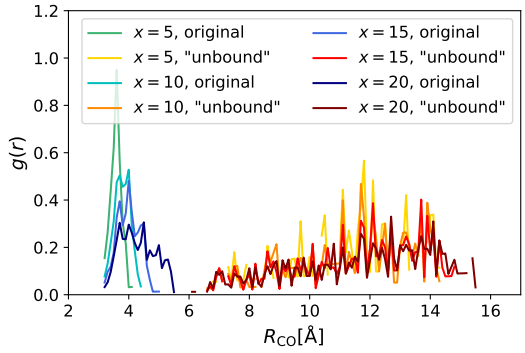
Looking at the many-body expansion of the energy of the two nonamers having the largest DMC–MB-pol differences from the first ten structures in Fig. 4.2, we see that both the summed 2B and 3B terms of the developed models agree with CCSD(T)-F12 to

within about 2 meV / H_2O accuracy, while the total differences from DMC are higher than 10 meV / H_2O. Note that the 1B energies also have differences from CCSD(T)-F12a/AVTZ of about 2 meV / H_2O but this is because the Partridge–Schwenke 1B term included in the model is fitted on a quantum chemistry method that has higher accuracy than CCSD(T)-F12a/AVTZ [130].

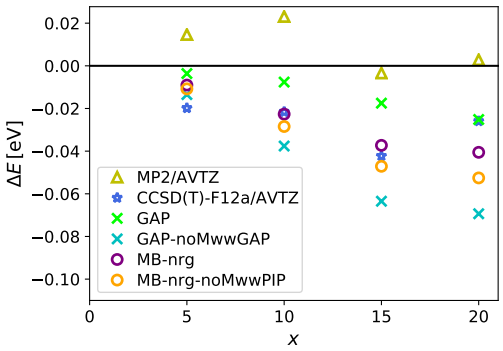
For larger structures, similar trends to Fig. 4.1b can be seen in Fig. 4.3. Both the data-driven models are generally more negative than DMC for low energies, lower than −0.10 eV / H_2O for the pentadecamers (15-mers) and lower than −0.15 eV / H_2O for the icosiheptamers (27-mers). The absolute errors are higher for both models at the more positive DMC energies. While TTM4Fmod+GAPs still has negative differences, MB-pol has differences of both signs.

4.3 Methane-in-water clusters

The methane–water interaction energies of the models are compared to DMC results on the methane-in-water structures ($CH_4(H_2O)_x$) of Gillan et al. [154] (downloaded from [158]). They sampled the methane-in-water configurations from a MD simulation using rigid molecules [154]. Firstly, a CH_4 molecule that did not have any CH_4 molecule neighbours within 7.5 Å was randomly selected, and the closest x H_2O molecules were chosen according to the CO distances [154]. Gillan et al. [154] calculated the DMC methane–water binding energies as differences between the original structure and an "unbound" structure where the methane molecule is displaced along the x coordinate by 10.58 Å [160] assuming that the interaction energy between the methane molecule and the water cage is negligible at this distance [154]. However, as shown in Fig. 4.4a, there are CO distances as short as 6.0 Å in the "unbound" structures, so the assumption of having negligible methane–water interactions might be questionable. To compare to the paper's DMC energies, we also calculated the energy differences between these two structures with the developed models.



(a) Studying the distribution of the CO distances in the original and the "unbound" $CH_4(H_2O)_x$ structures. The "unbound" structures of the paper [154] are the ones where the methane molecule is displaced along the x coordinate by 10.58 Å [160]. Ref. [154] calculates the energies of these two structures and defines the methane–water interaction energies as the differences between them.



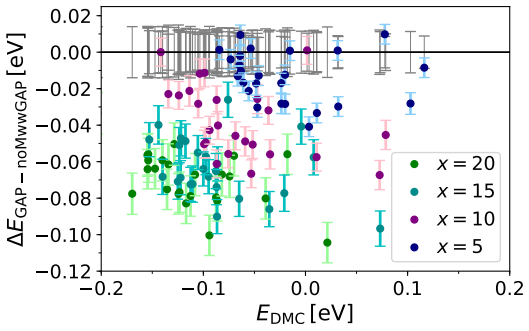
(b) The average differences from the DMC methane–water interaction energy using different methods w.r.t. the number of water molecules in the $CH_4(H_2O)_x$ clusters. While the developed models were calculated for all the 25 clusters for each x, the quantum mechanical methods are calculated only for the structures shown in Fig. 4.6 (2 structures for $x = 5, 10, 15$ and 1 structure for $x = 20$).

Fig. 4.4

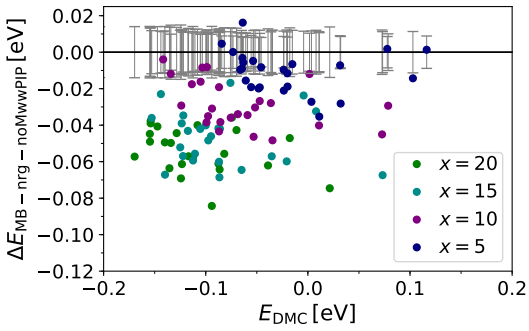
Another possible source of error, in this case mentioned by the authors [154], is that using a smaller time step for the DMC calculations would cause differences in the results for the $CH_4(H_2O)_{10}$ clusters of between 5.44–8.16 meV, which would lie within the error bars of DMC [154].

In Figures 4.4b, 4.5, 4.6, different methods are compared to the DMC methane–water interaction energy. For the quantum mechanical methods and the GAP model, the energy of the structures is calculated as the sum of the methane–water 2B and methane–water–water 3B energies. For the two MB-nrg versions, the electrostatic baseline includes the methane–water interactions as well, so the energy is the total energy calculated by the models. In both cases, we calculate the interaction energies as the differences between the original and the "unbound" structures. The methane–water 2B and methane–water–water 3B energies are also shown in Fig. 4.6 to compare the different methods. As the goal was to see where the differences come from, the quantum chemical calculations were performed on the clusters for which the differences between the initial GAP-predictions

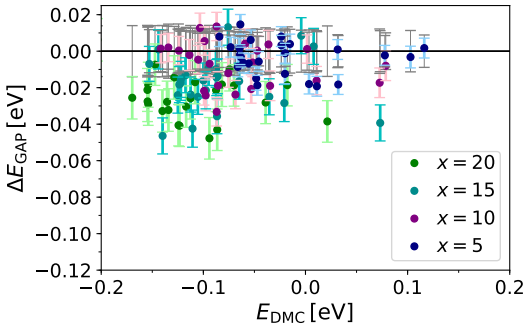
and DMC were largest. (Note: some of the CCSD(T)-F12 2B and 3B calculations for the smaller clusters were performed for the Master thesis of the author [159].)



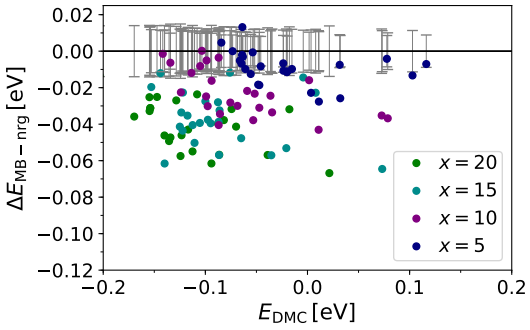
(a) TTM4Fmod+GAPs without the $CH_4(H_2O)_2$ fits



(b) MB-nrg without the $CH_4(H_2O)_2$ fit



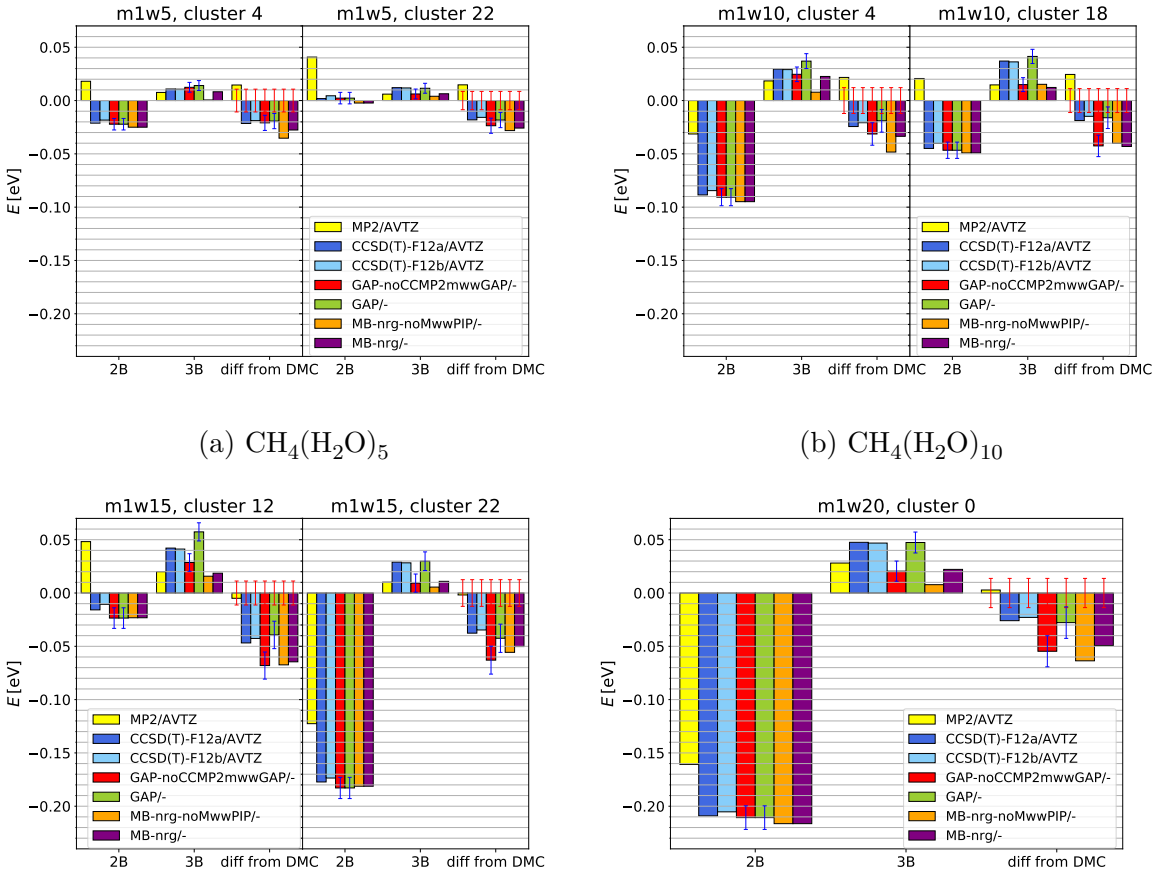
(c) TTM4Fmod+GAPs



(d) MB-nrg

Fig. 4.5 The difference in the methane–water binding energy between the developed models and DMC w.r.t. the DMC values in the $CH_4(H_2O)_x$ clusters.

In Fig. 4.5, we can see that the energies of the developed models are mostly more negative than the DMC energies. Figures 4.5a, 4.5b show the results using the models without the 3B methane–water–water fit, which models show poorer agreement with DMC than the models including this fit. Showing the average energy differences in Fig. 4.4b, the differences from DMC increase with the size of the structures and this is in agreement with the behaviour of the CCSD(T)-F12a/AVTZ energies (where calculated, see caption of figure), calculated as a sum of the many-body terms up to the 3B level. Interestingly, MP2/AVTZ has different errors: for $x = 5$ and $x = 10$, it is more positive



(a) $CH_4(H_2O)_5$

(b) $CH_4(H_2O)_{10}$

(c) $CH_4(H_2O)_{15}$

(d) $CH_4(H_2O)_{20}$

Fig. 4.6 Many-body expansion of the methane–water binding energy for seven of the $CH_4(H_2O)_{x=5,10,15,20}$ clusters (indices shown in the graph titles). The sum of the 2B and 3B energies are shown with MP2, CCSD(T)-F12a and CCSD(T)-F12b using the AVTZ basis set and with the MB-nrg and GAP models using different amounts of the fitted corrections. The "diff from DMC" is the difference of the total binding energy from the DMC value (this is approximated using the sum of the 2B and 3B terms for CCSD(T)-F12 and MP2). The blue error bars are the error bars of the GAP models (calculated from the test set RMSEs) and the red error bars are the error bars of the DMC results.

than DMC but for the larger structures it is within the error bars of DMC. Comparing the summed 2B and 3B energies of MP2 and CCSD(T)-F12 in Fig. 4.6, it seems that the differences between their energies are mostly due to the large differences in the 2B energies: these are always more positive for MP2 than for CCSD(T)-F12. The 3B energies of MP2 and CCSD(T)-F12 also differ (in the other direction) but those differences are much smaller – maybe because the absolute 3B energies are smaller, too.

Looking at the clusters studied in Fig. 4.6, the sum of the 2B and 3B GAP energies by the TTM4Fmod+GAPs (shown with green) are close to the CCSD(T)-F12a/AVTZ values. Also, the summed MB-nrg 2B energies are close to the CCSD(T)-F12 values, while the summed 3B terms of MB-nrg are close to its fitted MP2 level. For these seven structures, the MP2 3B terms seem to be always more negative than CCSD(T)-F12, which explains why the MB-nrg energies have larger negative energy differences from DMC than the GAP model. However, it is still a smaller error for most of the structures than the error of the original MB-nrg, which did not have the 3B methane–water–water correction fit (shown with orange). It would be interesting for future work to add a 3B term to MB-nrg fitted on the coupled cluster level, too. For comparison, a GAP version without the 3B CCSD(T)-F12–MP2 correction is also shown (with red) which has only a 3B term fitted on the same MP2 dataset as the 3B PIP of MB-nrg. As expected, this GAP version has similar summed 3B energies as the original MB-nrg version.

4.3.1 Benchmark quantum chemistry calculations

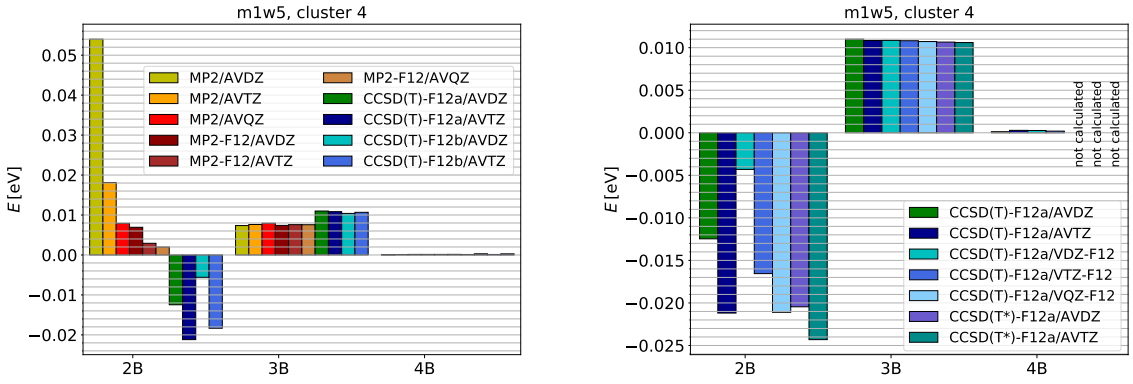
A $CH_4(H_2O)_5$ cluster

As there seemed to be a systematic difference between the fitted quantum chemistry and DMC, we also performed benchmark calculations with other versions of the MP2 and coupled cluster methods.

Fig. 4.7a shows the many-body expansion of the methane–water interaction energy in the $CH_4(H_2O)_5$ cluster (number four) up to the 4B level calculated by the MP2, MP2-F12, CCSD(T)-F12a/b methods with AVDZ and AVTZ basis sets. Fig. 4.7b also

shows the many-body expansion for CCSD(T)-F12 using the VXZ-F12 basis sets and the CCSD(T*)-F12a which scales the triples energy contribution using the MP2-F12 calculation (automatically by Molpro [94]). While the summed 2B energies differ when changing AVXZ to VXZ-F12, the summed 3B energies are very similar. Also, the 4B energies are very small with all the methods where calculated, so the differences from DMC are not due to truncating the many-body expansion at the 3B level. As shown in Fig. 4.6, the CCSD(T)-F12a–DMC difference is about −20 meV for the $CH_4(H_2O)_5$ cluster, so a method that is comparable to DMC would be 20 meV more positive than CCSD(T)-F12a/AVTZ. The CCSD(T)-F12a calculations using the less accurate double zeta basis sets have the largest differences from CCSD(T)-F12a/AVTZ, with the CCSD(T)-F12a/VDZ-F12 result having almost that much difference. All the other studied coupled cluster results are within 5 meV of the CCSD(T)-F12a/AVTZ result, so they are still at least 15 meV more negative than DMC. According to these tests, it seems that the CCSD(T)-F12 and DMC significantly differ when using at least triple zeta basis sets.

Looking at the less accurate MP2 results in Fig. 4.7a, we see a large variety in the summed 2B energies. The MP2-F12 2B values are closer to the CCSD(T)-F12 values than MP2, though they are still positive while the CCSD(T)-F12 values are negative.



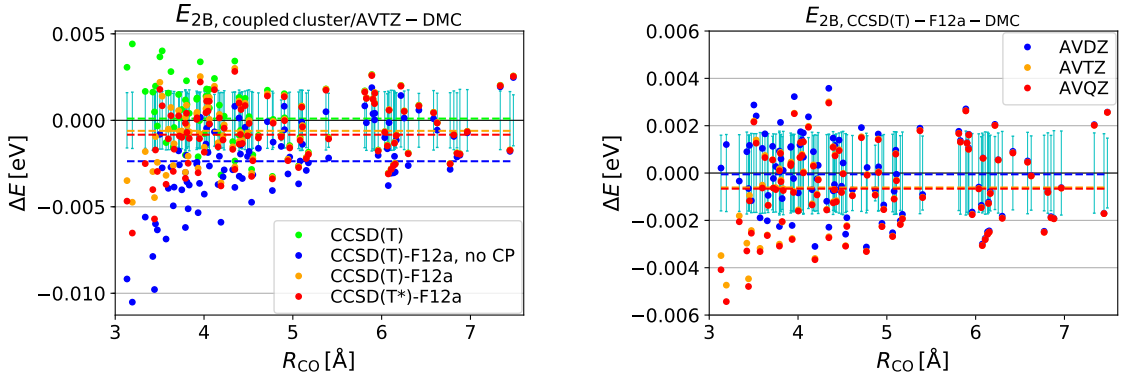
(a) Different quantum mechanical methods using the AVXZ basis sets.

(b) The CCSD(T)-F12a energies using different settings and basis sets.

Fig. 4.7 Many-body expansion of the methane–water binding energy for one $CH_4(H_2O)_5$ cluster including the 4B terms that are found to be very small using different quantum chemical methods with the AVXZ and VXZ-F12 basis sets.

Methane–water dimers

To examine the coupled cluster–DMC differences in more detail, we compare the 2B energy of the different quantum chemical methods on the $(CH_4)(H_2O)$ dimers of Gillan et al. [154]. These dimers were sampled from a high-temperature MD using DFT, and the monomers were modified to their equilibrium gas phase geometries [154]. Different versions of coupled cluster are tested against DMC along with the one used in the datasets: CCSD(T)-F12a/AVTZ with the counterpoise (CP) correction, and we find that there seems to be a systematic negative difference for most of the coupled cluster methods except for the ones with the lowest accuracy.



(a) (b)

Fig. 4.8 Coupled cluster–DMC 2B energy differences w.r.t. the carbon–oxygen distance of the methane–water dimers (100 configurations from Ref. [154]) using different coupled cluster settings. The dashed lines show the average differences.

The DMC 2B energies were calculated [154] by subtracting the energies of the isolated monomers from the total energy. Here, unless noted otherwise, the 2B energies are calculated with the CP correction to correct the basis set superposition error (BSSE) (see Section 2.2.2). Note that this is what the original paper [154] did when comparing different quantum chemistry methods to DMC, too. Using the CP correction causes large differences in the monomer energies which are subtracted to achieve the 2B energies: with the CCSD(T)-F12a/AVTZ method, the maximum differences between the monomer energies using extended basis sets (of the CP approximation) and the monomer basis sets are 4.6 meV and 1.3 meV for H_2O and CH_4, respectively. When not using the CP

correction, the CCSD(T)-F12a/AVTZ results will have even larger differences from DMC than the version using it has (see Fig. 4.8a).

Testing different coupled cluster methods using the AVTZ basis set in Fig. 4.8a, the energies are almost always more negative than DMC except for the not so accurate CCSD(T). Fig. 4.8b shows that CCSD(T)-F12a/AVXZ is systematically more negative in the methane–water 2B energy when using larger than double zeta basis sets. Using the VXZ-F12 basis sets would only change little on these differences (as shown in Fig. 4.9c); and indeed, the average differences from DMC would still be negative for larger than double zeta basis sets: -0.21 meV and -0.52 meV for the VTZ-F12 and VQZ-F12 basis sets, respectively.

Improving the basis set from AVTZ to AVQZ would cause smaller than 1 meV differences per dimer as shown in Fig. 4.9b. With the AVTZ basis set, the Molpro manual [94] suggests using the CCSD(T)-F12a version, but we also show the differences between using the -F12a or -F12b varieties of the method in Fig. 4.9a, and the differences for AVTZ are mostly smaller than 1.5 meV for this change, too.

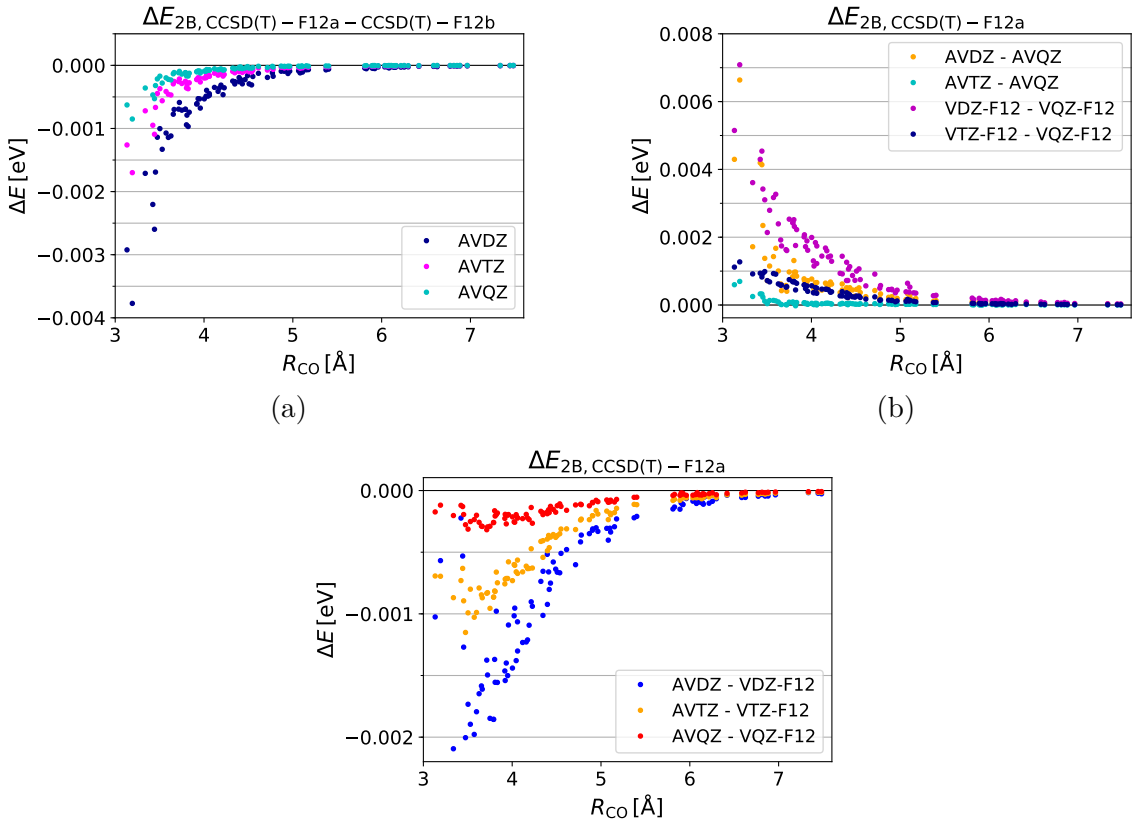


Fig. 4.9 Energy differences between using different CCSD(T)-F12 settings w.r.t. the carbon–oxygen distance of the methane–water dimers (100 configurations from Ref. [154]). a) shows CCSD(T)-F12a–CCSD(T)-F12b differences using AVXZ basis sets, b) shows CCSD(T)-F12a differences between using XZ or QZ basis sets and c) shows differences between using AVXZ and VXZ-F12 basis sets.

Chapter 5

Properties of periodic systems

5.1 Methods

5.1.1 Generation of hydrogen-disordered structures

The GenIce software [141, 142] is used to generate hydrogen-disordered structures for the ices: Ih, VI, VII, VIII and XI; and the clathrates: structure I (sI), structure II (sII), structure III (sIII), structure IV (sIV), structure VII (sVII), hexagonal structures (sH), filled ice C0 (c0te) [161], hydrogen-ordered hydrate C1 / filled ice II (c1te) [161], filled ice Ic / C2 (c2te) [161]. We chose the structures which are known to exist for methane clathrate. Also, we included some other structures to test if the model finds any of them more stable than the known structures.

The sH structure has three different cages: 5^{12}, $4^3 5^6 6^3$, $5^{12}6^8$, and we studied five different fillings. (As explained in Section 1.4, $4^3 5^6 6^3$ stands for a cage having three tetragonal, six pentagonal and three hexagonal faces.) As the largest cage, $5^{12}6^8$, is thought to be too large for one CH_4 molecule [43] and different numbers of methanes have been suggested for this cage [41, 42, 47, 52, 61, 66–69], we tested having 0, 1, 2 and 3 methane molecules in it, named as "sH_v2", "sH_full", "sH_L2" and "sH_L3", respectively. Additionally, we tested the structure where only the 5^{12} cage was filled ("sH"). The structures having only empty or singly filled cages were generated by GenIce:

sH, sH_v2 and sH_full. To create initial geometries for the structures with multiply occupied large cages, we modified the fully occupied structure. For the doubly-occupied large cage (sH_L2), we placed a methane-dimer (having 3.0 Å CC distance) on the position of the CH_4 molecule's C atom. For the triply-occupied large cage (sH_L3), we placed a methane-trimer on the same position having 2.5 Å CC distances along the x axis.

We used the random seed 0 for the hydrogen orientations, but we also tested how large errors this can lead to. Structures generated using different random seeds were geometry optimised for the Ih, VI and sII structures, and we looked at how large the differences are between the lowest enthalpy structure and the one generated with the 0 random seed. For the ices Ih and VI, the most stable structures amongst the 1–20 configurations have enthalpies 5.9 meV / H_2O and 2.3 meV / H_2O lower than the structures using the 0 random seed. Amongst the 1–9 random orientations for sII, the lowest enthalpy is only 0.43 meV / H_2O lower than the enthalpy of the structure generated with the 0 random seed. Using the hydrogen orientations having the lowest enthalpies would probably improve the accuracy of the results; however, it would need much more computational time to minimise all the structures starting from different random seeds. (One might also consider looking at the different random seed configurations at various pressures to see if the same orientation is appropriate there.) Using the lowest Gibbs free energy structures would need even more computer time as the quasi-harmonic calculations have high computational cost.

5.1.2 Structures from other sources

The filled ice Ih (fIIh) structure is from the Supporting Information of Ref. [161]. We also studied the MH-IV structure which is from the data of Cao et al. [60]'s Supporting Information. Note: for their MH-V structure, the angle information was missing in the Supporting Information; however, the formation enthalpies for this phase are positive (as calculated by the same group in [61]), so it is unlikely to appear in the phase diagram. Their MH-VI structure [61] was not orthogonal, so was left out from this study.

The structural data for the sK (also called: HSI) clathrate of [58] missed two oxygen atoms in the paper's Supporting Information and the original structure was not available at the time of this study. However, the sIV / HS1 clathrate of GenIce [142] seems to be equivalent with this structure.

The structures for the DMC comparisons are from the studies of Santra et al. [162] and Raza et al. [163] for the ices, and from Cox et al.'s paper [81] for the sI clathrate (the structures of the DMC calculations of [81] were not available, so we use the ones provided by Stephen J. Cox [164], see details in Section 5.3.1).

5.1.3 Geometry optimisation

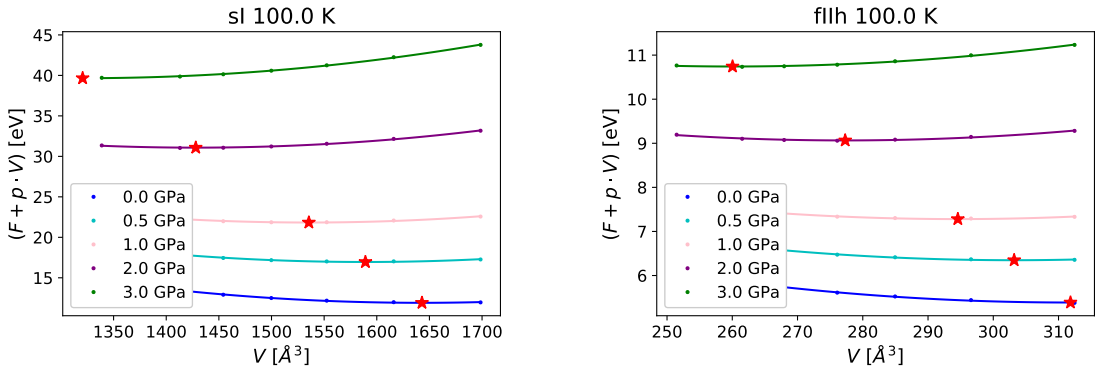
The structure optimisation was done in ase [105] using preconditioned optimiser methods [165]: PreconFIRE and PreconLBFGS. (The un-preconditioned FIRE and LBFGS methods are described in [166] and [167].) When the structures were far from the equilibrium, the PreconFIRE method was run, then the PreconLBFGS was run firstly with the Armijo and then with the Wolfe line search methods. The optimisations were run using variable cells with the cells kept orthogonal because the earlier version of the MBX software did not have non-orthogonal cells. The optimisations were performed changing the pressures gradually, each starting from the structure of the previous step.

5.1.4 Quasi-harmonic approximation

Phonon calculations are performed using the phonopy software of Togo and Tanaka [168, 169] on the geometry-optimised structures. In phonopy, the phonon free energies are calculated in the harmonic approximation using the finite displacement method.

In the harmonic approximation, the free energy of the system is approximated as [168]:

$$F(T,V) = U(V) + F_{\text{phonon}}(T,V) \tag{5.1}$$



(a) (b)

Fig. 5.1 Illustrating the fit according to Eq. 5.4. The red stars show the minimum Gibbs free energy (G) at the corresponding volume (V) for the given pressure and temperature. When the minimum of the curve is outside the calculated range, the T, p condition is considered to be outside the validity range, such as the 3 GPa fit for the sI phase. Each $G(T, p)$ point is determined by a separate fit.

Phonopy does not account for the entropy of mixing, so it is added as:

$$\Delta S_{\text{mix}} = -k_{\text{B}} N_{\text{A}} \sum_i n_i \ln(x_i) \tag{5.2}$$

where k_B is the Boltzmann constant, N_{A} is the Avogadro constant, i runs over the different species (CH_4 and H_2O), n_i is the amount of the given species and x_i is the corresponding mole fraction.

Then the free energy is:

$$F(T, V) = U(V) + F_{\text{phonon}}(T, V) - T\Delta S_{\text{mix}} \tag{5.3}$$

In the quasi-harmonic approximation (QHA), the Gibbs free energy is calculated by transforming the (T, V) variables to the (T, p) variables [168]:

$$G(T, p) = \min_V (F(T, V) + p \cdot V) \tag{5.4}$$

To determine the Gibbs free energy and the corresponding volume according to this transformation, we fitted the $a \cdot (V - V_0)^2 + c$ equation to the $(F(T, V) + p \cdot V), V$ curves

at different temperatures and pressures (see Fig. 5.1). To include the curves' minima within the fitted data, we needed to add calculations for structures geometry-optimised at negative pressures (−0.5 GPa, sometimes −1.0 GPa) and sometimes at pressures higher than the target pressure of the calculations (4.0 GPa). The pressures for the geometry-optimised structures for the quasi-harmonic calculations are spaced by 0.5 GPa.

Due to the calculated data, our predictions' upper pressure limit is 2.0 GPa for most of the structures. As shown in Fig. 5.1, the validity for the pressure range is decided by whether a point is within the range of the data included in the curve fitting. This means there are predictions for these pressures but as they are outside the calculated data, their accuracy is lower. For the temperatures up to 135 K, the limit of validity for the QHA is 2 GPa for the ices Ih and XI. It is also 2 GPa for all the clathrates, except for sH_L2, fIIh, MH-IV and c1te, the limit for these structures is 3.0 GPa. At 0 K, the limit is slightly lower, 1.9 GPa for sII, sH, sH_v2, sH_full, sIII, sIV and c2te. Additionally, for sVII, it is even lower, 1.8 GPa, even when including calculations optimised at 4.0 GPa. However, these phases are not predicted to exist in the QHA (see Fig. 5.5a), so the lower accuracy does not cause any issues. As all the phases that appear in the phase diagrams have validity limits of at least 2.0 GPa, we will use this limit for the predicted phase diagrams. This validity range could be widened in future work by including more calculations; however, the current version of the studied potential has also lower accuracy at high pressures.

All the calculations were run using a phonopy supercell 1x1x1 because the phonopy software would reorder the atoms of the structure within itself when building larger supercells, which would cause problems in running the MB-nrg calculations through the MBX software. (MBX needs to have the atoms grouped according to the molecules in a specified order.) Thus, the unit cells contained at least 12 molecules for the clathrates (fIIh having the smallest cell) and 16 molecules for the ices (Ih and XI having the smallest cells). For the finite displacement, the atoms were displaced by 0.02 Å.

Note that the QHA assumes that the molecules only vibrate about their equilibrium positions. As the methane molecule in the hexagonal clathrate's largest cage behaves as

a quasi-rotor when singly filled even below 40 K [42], this approximation might not work well for the sH_full structure. Moreover, this approximation is supposed to be valid only up to 1/2–2/3 of the melting temperature [170–172]. The decomposition temperatures are between 300–323 K depending on the pressure for sI [173], around 323 K for the hexagonal phase [173], and between 340–400 K [174] at the studied pressures for fIIh. Thus, the validity limit of the QHA might be between 150–160 K for sI, 160 K for the hexagonal clathrate, and between 170–200 K for fIIh. For the reference ices which are used to adjust the methane-to-water ratios, this temperature limit will be even lower: in the case of ice Ih, around 135 K at ambient pressure.

5.2 Ices

5.2.1 Comparison to DMC results

The interaction energies of the developed models are compared to DMC energies on the ice structures Ih [162], VIII [162], XIc [163] and XIh [163], downloaded from D. Alfè's website [158]. The DMC technology has been improved in recent years by Zen et al. [153, 155], so we include their DMC results too for the structures calculated (they used the same geometries as Santra et al. [162]). Zen et al. [155] presented two DMC results, and we chose to use the more accurate ones, which used large supercells. The differences between the old and new DMC versions for the ices Ih and VIII are higher than the error bars of the DMC calculations (the difference being greater than 25 meV / H_2O for ice VIII), so the earlier results of [163] for the ices XIh and XIc could possibly be improved, too.

The TTM4Fmod+GAPs model is almost always more positive than DMC, being within the sum of the error bars of GAP and DMC for the earlier results of the ices Ih, VIII and XIh, but having a larger error for ice XIc. TTM4Fmod+GAPs and MB-pol are close to each other for the first three structures, but not for ice XIc, where the GAP model has a larger difference from the DMC value. MB-pol is mostly closer to DMC than GAP, except for the earlier DMC result for ice VIII, where GAP is very close to

the DMC result. Interestingly, both the data-driven models have higher errors (about 20–25 meV/H_2O) compared to the two newer DMC values.

method	Ih	VIII	XIh	XIc
DMC [162, 163]	-0.5960 ± 0.0050	-0.5670 ± 0.0040	-0.5890 ± 0.0040	-0.5950 ± 0.0040
DMC [155]	-0.6146 ± 0.0052	-0.5939 ± 0.0062	-	-
experiment [162] *	-0.6100	-0.5770	-	-
experiment [155] *	-0.6094	-0.5949	-	-
TTMF4mod+GAPs	-0.5894 ± 0.0059	-0.5668 ± 0.0158	-0.5711 ± 0.0103	-0.5517 ± 0.0105
MB-pol	-0.5938	-0.5730	-0.5863	-0.5840

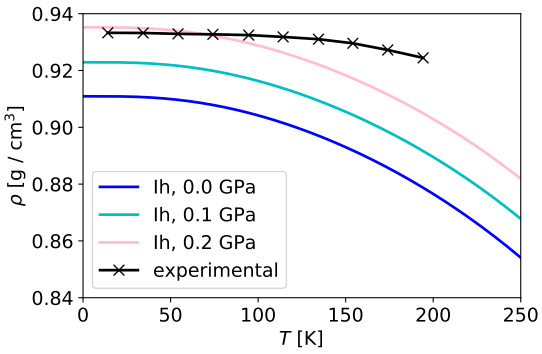
Table 5.1 Ice interaction energies (in eV). The DMC energies are from [155, 162, 163]; for the earlier version of DMC, the ices Ih and VIII are from [162] and XIc and XIh are from [163]. The earlier DMC values are taken from the xyz files published on the website [158]. The updated values are from [155]. *For the experimental values of Ih and VIII, we show the recalculated values with zero point effects removed from the experimental values of [175] (by [162] and [155]). As explained in the Supporting Information of [155], there is an uncertainty of ≥ 0.01 eV for their estimated value of VIII.

5.2.2 Ice densities

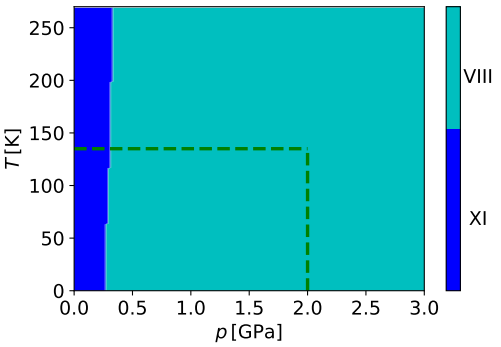
The densities (see Fig. 5.2a) are calculated from the volumes corresponding to the minimum G in the QHA using MB-pol. Comparing to experimental data of Brill and Tippe [176] (extracted from Fukusako [177]'s Fig. 4.), the predicted densities of the model are about 2.5–5 % lower. The larger differences are at the higher temperatures, and this is as expected because the QHA is only a good approximation up to about two thirds or half of the melting temperature [170–172]. Up to 135 K, the densities are within 3.6 % of the experimental data.

5.2.3 Phase diagram prediction

Studying the Ih, VI, VII, VIII and XI ice structures in the quasi-harmonic approximation using MB-pol, the model predicts the ice XI to ice VIII phase transition at around 0.25 GPa (see Fig. 5.2b), which is between the experimental values for the XI–II and XV–VIII transitions (0.08 GPa and 1.40 GPa for 50 K, and 0.09 and 1.46 GPa for 100 K, as extracted from [179]). Ice XI is the proton-ordered pair of ice Ih [180], which is stable



(a) The ice Ih density at different pressures w.r.t. temperature. The experimental values are calculated [178] from the experimental lattice constants of Brill and Tippe [176], extracted from [177]'s Fig. 4.



(b) The calculated ice phase diagram. The green dashed lines show the estimated upper limits of validity of the approximation, as explained in Section 5.1.4.

Fig. 5.2 Predictions for ice properties using MB-pol in the quasi-harmonic approximation.

at low temperatures (below 70 K). According to experiments, ice Ih is stable above about 70 K [179]; however, predicting the proton-ordered pair of this ice is an acceptable result. The predicted Gibbs free energy difference between ice Ih and ice XI is less than 5 meV / H_2O above 150 K at 0.0 GPa with an ice XI to ice Ih phase transition predicted around 273 K, and the maximum of the Ih–XI Gibbs free energy difference being around 6 meV below 150 K (see Fig. A.2). Note that this difference is of the same magnitude as the enthalpy difference caused by using different random seeds for the hydrogen orientations of Ih, 5.9 meV / H_2O (see Section 5.1.1). We do not show temperatures above 270 K in the phase diagrams because the liquid water phase is not included in this study.

The experimental ice XI/Ih to ice VIII transition would also have two intermediate phases: the ices II and VI or XV or β-XV at low temperatures [179]. We did not include the ices II, XV and β-XV in these calculations which are only stable in small areas of the experimental phase diagram, as the aim of these calculations was to create reference phases for the clathrate phase diagram. (Ice XV is only the proton-ordered pair of ice VI [180].) As noted above, the QHA only works well up to about half of the melting temperature, so the results are not expected to be accurate above about 135 K at atmospheric pressure.

To improve the accuracy of this phase diagram, different hydrogen orientations should be calculated for each structure. Additionally, to get predictions for the ices stable in smaller areas of the phase diagram, the corresponding ices would need to be included in the calculations. Running path integral molecular dynamics (PIMD) would improve the predictions at high temperatures, and the liquid phase could be calculated, too.

5.3 Methane clathrates

5.3.1 Comparison to DMC on the sI clathrate

The developed models are compared to DMC results for the sI clathrate by Cox et al. [81]. The DMC values of [81] were calculated on the optPBE-vdW-optimised structures after re-optimising them in Quantum Espresso [81, 164]; however, these original structures were not available at the time of our study [181]. Thus, we performed the calculations on the optPBE-vdW-optimised structures received from Stephen J. Cox [164]. The energies are calculated as in Ref. [81], so the total cohesive energy is the total energy minus the equilibrium monomer energies; the water cohesive energy is the total energy of the water skeleton subtracting the equilibrium water monomer energies; finally, the methane–water interaction energy is the difference between the energies of the full structure and the water skeleton, with the energies of the equilibrium methane monomers subtracted [81].

For the water interactions in the clathrate (see Fig. 5.3a), both TTM4Fmod+GAPs and MB-pol are about 10 meV / H_2O more positive than DMC. This positive difference is similar to the trend for the ices, where the developed models were also either more positive than the DMC values or close to them (see Table 5.1).

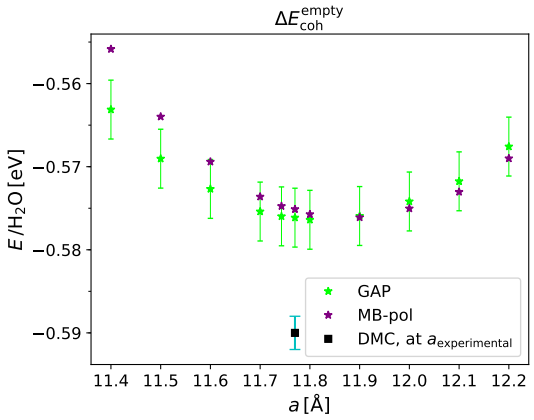
As for the total interaction energy (shown in Fig. 5.3c), MB-nrg performs well, while the TTM4Fmod+GAPs model has higher errors. Adding the 3B methane–water–water term improves the MB-nrg result; however, the GAP model becomes more negative when adding this 3B term. Moreover, the full TTM4Fmod+GAPs' lattice constant is lower than the one of DMC. The main difference between the two models is that MB-nrg has the electrostatic baseline implemented for the methanes too. Moreover, MB-nrg also

has a correction fit for the methane–methane interactions that the GAP model does not include; however, the sum of this term only accounts for about $-1.64\,\mathrm{meV}\,/\,H_2O$ molecules for MB-nrg (and $-1.45\,\mathrm{meV}$ for TTM-nrg) at $a = 11.4\,\text{Å}$, where the CH_4 molecules are the closest to each other. The cutoff for the methane–water 2B correction term is longer for the MB-nrg (9.0 Å) than for GAP (7.0 Å). Additionally, the target levels of the methane–water 2B and 3B correction fits differ. The 2B corrections approximate different versions of CCSD(T)-F12 (see Section 2.6). The difference between the 3B terms of the two models is that TTM4Fmod+GAPs has the CCSD(T)-F12–MP2 correction as well, while MB-nrg has only a correction targeting the MP2 level. Including the CCSD(T)-F12–MP2 methane–water–water GAP decreases the total cohesive energy by about 3–4 $\mathrm{meV}\,/\,H_2O$, which is only a small fraction of the GAP model–MB-nrg difference. Interestingly, the force field (TTM-nrg) included in the MBX software also achieves values very close to the ones of DMC.

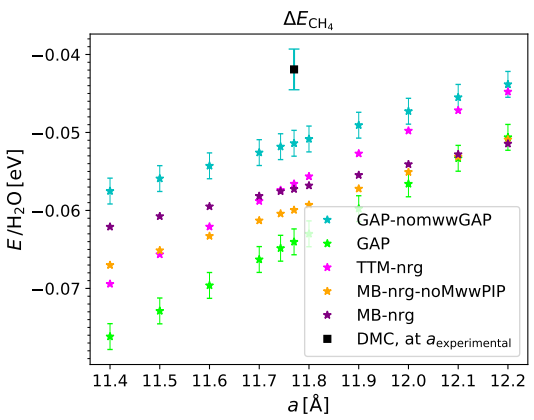
Note that the experimental lattice constant, $11.77 \pm 0.01\,\text{Å}$ (measured at 5.2 K in D_2O) [182], is also smaller than DMC's value, but the GAP model's lattice constant is even lower than this value. However, the TTM4Fmod+GAPs model being more negative than DMC is in agreement with the results for small clusters in Figures 4.1, 4.4b.

Note also that this DMC result might also be improved by the method of A. Zen et al. [153], who suggested a new modification to the Green's function and thus, decreased the time-step errors. The results of Ref. [155] using this modification for ices were sometimes even $25\,\mathrm{meV}\,/\,H_2O$ lower than the ones of the earlier DMC calculations of Ref. [162].

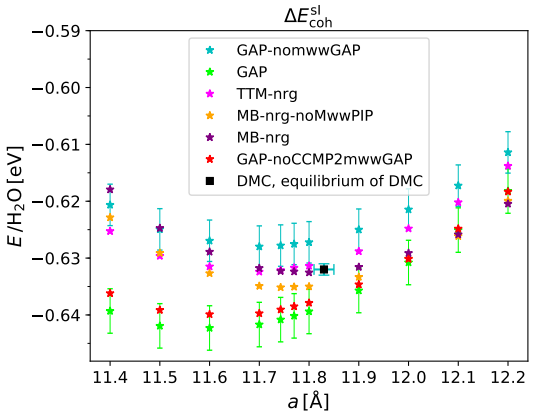
Nevertheless, the reported results of the developed models are more accurate (when compared to DMC) than those of the best performing DFT calculations of Ref. [81]'s Fig. 2, where most of the DFT functionals had errors higher than $125\,\mathrm{meV}\,/\,H_2O$ for the total interaction energy, except PBE and revPBE-vdW. Additionally, some of the functionals predict lattice constants below the experimental value, with errors between 2–7 % in the case of LDA, PBE-D2, PBE-vdW$^{\mathrm{TS}}$, optB88-vdW and optB86b-vdW [81]. Moreover, as explained in Ref. [81], the DFT functionals producing the seemingly best results are not



(a) The cohesive energy of the empty clathrate divided by the number of H_2O molecules



(b) The methane–water interaction energy divided by the number of H_2O molecules



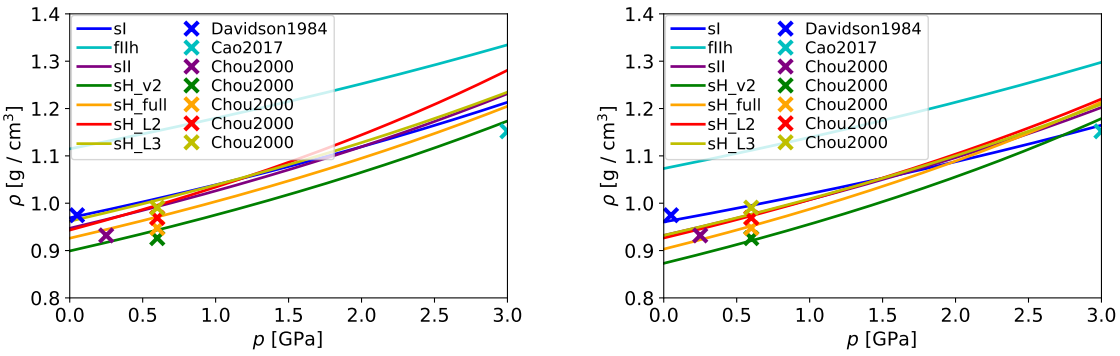
(c) The total cohesive energy divided by the number of H_2O molecules

Fig. 5.3 Comparing the developed models on different parts of the sI clathrate's cohesive energy to DMC benchmarks with respect to the lattice constant. Different parts of the TTM4Fmod+GAPs model are shown: the model without the methane–water–water 3B term is with cyan, and the model with only the MP2 fit for the methane–water–water 3B term is with red. The MB-nrg is also shown without the 3B methane–water–water PIP of this thesis, as published in [6] (orange). The TTM-nrg model [6] is also shown (with magenta) which is the MB-pol model with the electrostatic interactions of TTM4Fmod for the interactions involving methane molecules.

accurate everywhere: although the cohesive energy of revPBE-vdW is close to DMC, it predicts a higher lattice constant (12.077 Å), and has a too negative methane insertion energy [81]. Also, PBE's good result, being only about 10 meV / H_2O more negative than DMC, is due to cancellation of errors between the water–water and methane–water interactions [81]. In the developed models, each term is fitted separately, so there is no cancellation of errors between the different parts of the energy by construction.

5.3.2 Clathrate densities

The clathrate densities are studied in the QHA using the MB-nrg model (see Fig. 5.4). The experimental densities are the values calculated from the experimental results of Davidson et al. [182] for sI, from the results of Chou et al. [43] for sII and the hexagonal structures, and recalculated by Cao et al. [60] from the results of Loveday et al. [45] for the fIIh structure.



(a) $T = 5\,\mathrm{K}$ (b) $T = 293\,\mathrm{K}$

Fig. 5.4 Predicted clathrate densities at 5 K and 293 K w.r.t. pressure. The experimental densities (shown by x-s) are the values calculated from the lattice constant for sI by Davidson et al. [182] that was result of a neutron diffraction study at 5.2 K and the pressure was not specified; calculated from the volumes of Chou et al. [43] that used x-ray spectroscopy at 298.15 K for sII and the hexagonal structure (where the methane:water ratio was unspecified so we used the volume for all the versions of sH); and recalculated by Cao et al. [60] from the neutron diffraction results of Loveday et al. [45], at room temperature.

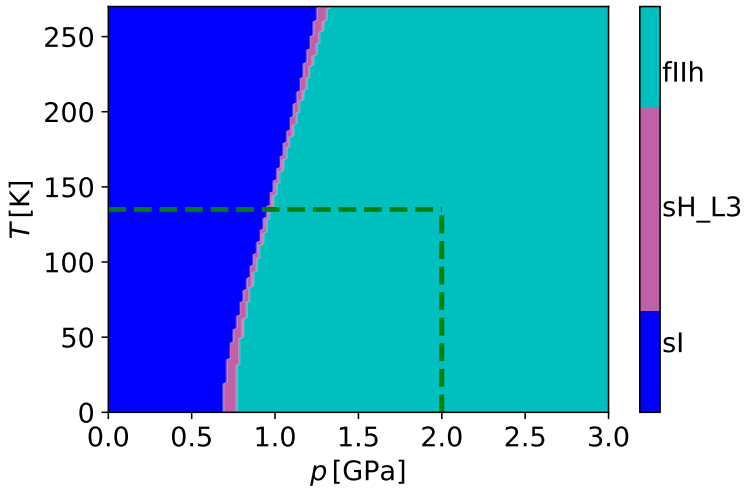
We show calculations at 5 K and 293 K, but the temperature had mostly little effects on the densities at the lower pressures in our calculations. The predictions are generally close to the experimental values (within 2 %), except for the high-pressure structure, fIIh, for which the predicted density is about 12 % higher than the experimental value (comparing to results calculated at 5 K for sI, and at 293 K for the other structures). Note that the large error for fIIh might be coming from being above the QHA validity temperature (170 K) for the structure. These results suggest that while the MB-nrg model combined with the QHA is very accurate at low pressures, it is less accurate at higher pressures. This is also what we saw when optimising the fIIh structure, and the model found holes before the addition of the switch function; a function which switches back to the baseline model for structures with distorted monomers in them (see Section 2.5.1).

5.3.3 Phase diagram prediction

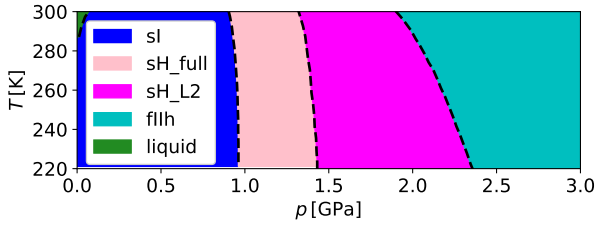
The phase diagram (shown in Fig. 5.5a) is predicted in the QHA using the full MB-nrg model. The structures studied are: sI, sII, sH, sH_v2, sH_full, sH_L2, sH_L3, fIIh, sIII, sIV, sVII, MH-IV, c0te, c1te and c2te. The most stable phase at the given conditions is decided by which clathrate phase has the lowest Gibbs free energy per $CH_4(H_2O)_{5.75}$ formula. This water-to-methane ratio is chosen to be the same as the one of the sI clathrate because in the experimental phase diagram, this phase has the highest water content. To achieve this constant ratio, we virtually complement the systems with the ice stable at the given conditions (see Fig. 5.2b) by adding its Gibbs free energy in the appropriate numbers. (From the studied structures, only the sH, sH_v2, sVII and c1te have higher water-to-methane ratios, for which the Gibbs free energy of the ice is subtracted.) We choose to look at the stability this way because then we only need to have the reference ice phases and not the methane phases. For methane, the reference phases would include the liquid, for which the free energy calculation would need a different calculation method, *path integral molecular dynamics* (PIMD), which requires longer computation time. Also, the accuracy of the water part of MB-nrg (MB-pol) has

been tested more extensively than the accuracy of MB-nrg for methane as this model is very recent. Moreover, while the model has the 3B correction term for water, it has only correction terms up to 2B for pure methane, so its accuracy might not be sufficient when going to phases where the methanes can get closer to each other than in the clathrates.

Comparing to the experimental phase diagram of Noguchi et al. [42] (see Fig. 5.5b), which combined IR and Raman spectroscopy experiments, the predicted phase diagram captures some aspects of the experimental results well: that sI is stable below 0.9 GPa, and fIIh is stable above 2.0 GPa; however, it only predicts the hexagonal phase to be stable in a small pressure range between them and with a higher methane content than the experimental result [42], although this methane content agrees with the experimental result of Tulk et al. [41]. According to the most recent experiments [42], the hexagonal phase would be stable between 0.9–2.0 GPa with an increase in the methane content from one to two in the large case at around 1.3 GPa. The MB-nrg model predicts the sI–sH_L3 transition at 0.9 GPa at 100 K (and around 1.3–1.4 GPa between 250–270 K), but already predicts the fIIh phase to be more stable than the sH_L3 slightly above the transition pressure. Missing the hexagonal phase almost completely might be partly explained by the guest molecule being able to rotate in the large cage of sH_full [42], and rotations are not accounted for in the QHA. Predicting the large cage of sH to contain three methanes is an error probably due to the approximation, but the previous computational papers also predicted three or more methane molecules for this cage [61, 66–69], except for a classical force field, the Murad–Gubbins model [66]. The current phase diagram prediction agrees better with experiments than the previous computational prediction by Huang et al. [61], which only agreed with the experimental results in finding the fIIh structure at high pressures. Huang et al. [61] looked at enthalpies; interestingly, when looking at the enthalpies per $CH_4(H_2O)_{5.75}$ formula, the predictions using MB-nrg also differ for the low pressures, predicting sII to be slightly more stable than sI (see Fig. A.1). Moreover, the transition pressures between the phases are lower than when looking at the Gibbs free energies. Predicting the sII phase is similar to Huang et al.'s results who predicted the sK, sH_L3, MH-VI and fIIh phases to be stable with increasing the



(a) The predicted clathrate phase diagram using the MB-nrg model in the quasi-harmonic approximation. The green dashed lines show the estimated upper limits of validity of the approximation, as explained in Section 5.1.4.



(b) Experimental clathrate phase diagram, reproduced from the corresponding area of Noguchi, N. et al. (2021). *J. Phys. Chem. C*, 125(1), 189–200, Fig. 1., which used results of Shimizu et al. [46], Loveday et al. [53] and Kurnosov et al. [173]. The black dashed lines are extracted from Ref. [42]'s Figure 1. using the WebPlotDigitizer tool [143]. Note that this graph only covers a smaller temperature range than Fig. 5.5a.

Fig. 5.5

pressure [61]. Note: here, we did not include MH-VI and Huang et al. [61] did not include sII; but sK is a transition phase between sI and sII. Thus, probably the results of Huang et al. would also be different if calculating the Gibbs free energies.

To improve these results, one might need to include different hydrogen orientations in the geometry optimisation and run the quasi-harmonic calculations on the structures having the lowest energies. Of course, this would increase the computational cost. Additionally, studying non-integer filling ratios might change these predictions but would be even more computationally expensive. Including the ices which are stable only in small areas of the water phase diagram might change the corresponding areas of this phase diagram, too. Moreover, path integral molecular dynamics simulations could be run, this technique includes quantum nuclear effects and is also reliable at higher temperatures than the QHA. The accuracy at high pressures could be improved using more stable correction terms, which are less likely to have holes. The stability of the PIPs could be enhanced for example by using more robust monomials for all the PIPs, such as the ones included in the methane–water–water 3B PIP [133]; using datasets which better represent the distorted configurations; or including the forces in the fits might also help. Thus, the model would not need to switch back to the baseline force field for structures having high monomer energies, which occurred during the geometry optimisation process of fIIh. Additionally, MB-nrg could be made more accurate and reliable for higher methane concentrations by adding 3B correction terms which include more than one methane molecule. Using a uniformly accurate target level for the different correction terms of MB-nrg might also improve the results.

Chapter 6

Conclusions

This work has presented the first two many-body methane–water models which achieve quantum mechanical accuracy, one developed fully in this project and the other improved in collaboration. The models were built by adding data-driven correction terms to classical force fields, correcting them to quantum chemical (mostly CCSD(T)-F12) accuracy. The model developed fully in the project, TTM4Fmod+GAPs, used the Gaussian Approximation Potentials (GAP) method for the fitting. The second model, MB-nrg, was developed by Marc Riera et al. [6] using the the permutationally invariant polynomials (PIP) method, and we added the 3B methane–water–water correction term in collaboration with Marc Riera–Riambau and Francesco Paesani [133]. Note that the water part of this model is the MB-pol [10].

While developing the TTM4Fmod+GAPs model, the geometrical representations of the structures were changed to use an inverse distance descriptor with a distance shift which made the fits more accurate. In addition, the datasets were complemented to better represent the geometry space, and also recalculated at a uniform quantum chemical level to make the target level of the fits consistent. One of these datasets was also used for adding the 3B methane–water–water term to the MB-nrg model.

The two models were tested against DMC results, and they achieved similarly good accuracies for small molecular clusters, but the PIP model (MB-pol/MB-nrg) was found to be more accurate for periodic systems. While benchmarking quantum mechanics on

small clusters, we demonstrated that there is an inconsistency between CCSD(T)-F12 and DMC. The difference between CCSD(T)-F12 and DMC was found to increase with system size (see Fig. 4.4b), and thus might be even larger for periodic systems. DMC is currently being improved (e.g. [153]), so it might provide more useful benchmarks in the future. However, in the light of our findings, comparing data-driven models to earlier DMC results might not be the best way to verify their accuracy. In the future, it would be interesting to check which CCSD(T)-F12 settings would be comparable with the improved DMC technique.

The complemented MB-nrg model has several advantages over the GAP model. Firstly, it is faster. Secondly, its electrostatic baseline also includes methane molecules. Moreover, it has a methane–methane 2B correction term, so it is applicable to systems with higher methane concentrations. However, for distorted configurations, it was found to be necessary to switch to the baseline force field in order to avoid holes in the correction terms, which might cause lower accuracy at high pressures.

Results of the faster model, MB-pol/MB-nrg, were also compared to experimental results. In the quasi-harmonic approximation (QHA), the predicted densities of the ice Ih and the clathrates sI, sII and hexagonal structure were within 3.6 % of the experimental data, while the high-pressure value for fIIh was about 12 % higher than the experimental value; however, the temperature for this experiment was above the estimated validity limit of QHA for fIIh. Thus, the MB-pol/MB-nrg model combined with the QHA is expected to be very accurate at low pressures and less accurate at higher pressures.

Phase diagram calculations were also performed in the QHA using the MB-pol/MB-nrg model. As the aim was to determine the clathrate phase diagram, only the more important ice phases were included. The model predicted the proton-ordered pair of ice Ih: ice XI instead of the Ih phase, but it captured the XI-to-VIII transition at a pressure between the experimental values for the XI–II and XV–VIII transitions. For the clathrate phase diagram, the phases sI and fIIh were predicted to be stable with a phase transition around 1 GPa, while the hexagonal phase was predicted to be stable in a very small region. This is only in qualitative agreement with the experiments. These

results could possibly be improved by running path integral molecular dynamics (PIMD) simulations, calculating structures with different hydrogen orientations and with various cage-filling ratios, and adding 3B correction terms including more than one methane molecules.

In collaboration [5], we also compared three popular potential energy fitting methods: GAP, PIP and the Behler–Parinello neural networks (BPNN), for which the author of the thesis optimised the GAP fits. We found that the three methods are able to fit the same water 2B and 3B datasets with the corresponding energies with similarly high accuracy; this suggests that when having high-quality datasets, any of these methods can be applied. It would also be interesting to compare the accuracies of the different fitting methods on datasets that also include the forces along with the energies. The results of this collaboration [5] also suggest that changing the GAP to PIP in this project probably does not affect the results much. Of course, the other differences, for example having different datasets with different target levels, having the baseline force field implemented for both molecules and having the 2B methane–methane term fitted, change the results.

In future work, the phase diagrams could be improved by running PIMD simulations, which could, in particular, increase the accuracy at high temperatures. Additionally, running calculations for different hydrogen orientations and different cage-filling ratios could make the predictions more realistic. To make the developed models applicable to higher methane concentrations, 3B correction terms with higher methane content could be added. For the TTM4Fmod+GAPs model to be applicable to the higher concentrations, one would also need to add a baseline force field that includes the methane molecules as well, such as the TTM4Fmod in the MB-nrg model. Additionally, the speed of the GAP model could be increased by using parallelisation for the 3B descriptor's trimer finding routine, and by using a parallel implementation of the baseline force field. Turning to MB-nrg, its accuracy might be increased by using a uniform target quantum chemical level for its correction fits, but this might not cause changes significant enough to be worth the computational cost of the re-calculation of the datasets. However, for its 2B terms including methane, it would be interesting to test the model using the AVTZ basis

set instead of the basis set extrapolation using the low accuracy AVDZ set. Moreover, improving its correction fits to avoid having holes and thus the need to switch back to the baseline force field would make the MB-nrg model more accurate for high-pressure structures. Finally, the MB-nrg model could also be used to predict physical properties of clathrates and other methane–water mixtures, and to study the kinetics of clathrate formation and decomposition.

Appendix A

Supplementary graphs

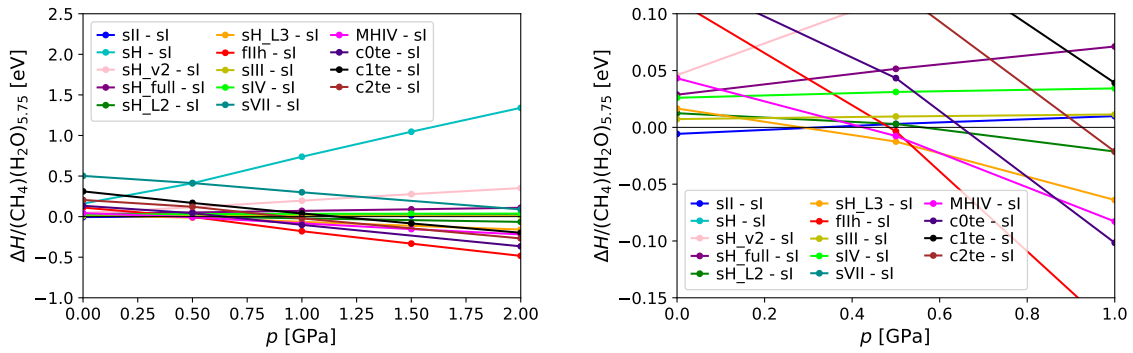
Thermodynamics calculations

Enthalpy differences

The enthalpy is calculated for the geometry-optimised structures at pressure p as:

$$H = U + p \cdot V \tag{A.1}$$

where p is calculated for the structures (slightly differing from the target pressures).

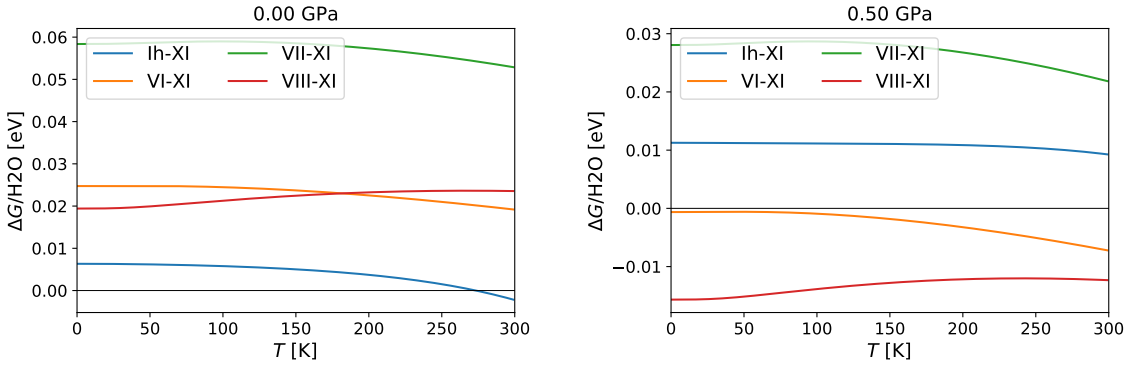


(a) (b)

Fig. A.1 Clathrate enthalpy differences between the different clathrate structures and sI. b) is the enlarged version to show the details at low pressures.

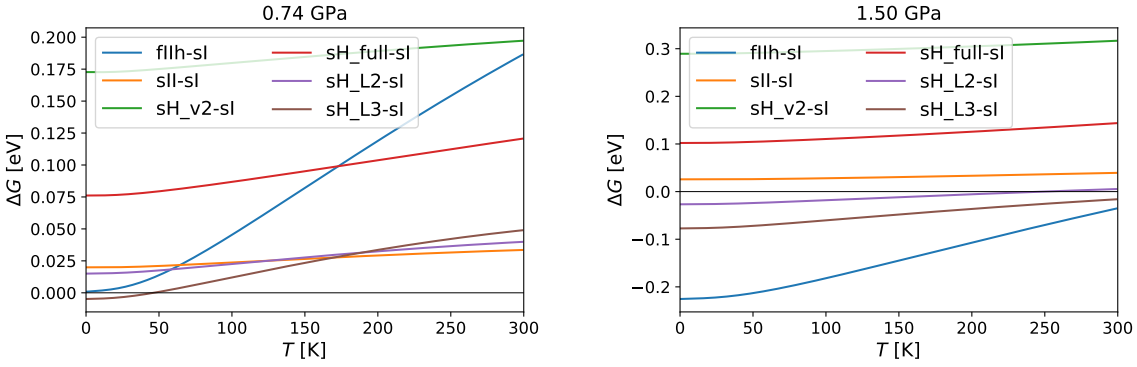
The enthalpies are compared per $CH_4(H_2O)_{5.75}$ formula, complemented by the enthalpies of the ices having the lowest enthalpies at the given pressure, similarly to the Gibbs free energies (see Sec. 5.3.3).

Gibbs free energy differences



(a) 0.0 GPa (b) 0.5 GPa

Fig. A.2 Ice Gibbs free energy differences between different ices and ice XI



(a) 0.74 GPa (b) 1.5 GPa

Fig. A.3 Clathrate Gibbs free energy differences (per $CH_4(H_2O)_{5.75}$ formula, as explained in Section 5.3.3) between the experimental clathrate structures and sI.

Appendix B

Parameters of the GAP fits

These parameters are written after the gap_fit command. The input files and the fitted GAP files can be found at the repository of the University of Cambridge [129].

Water fits

2B MP2–TTM4Fmod

```
at_file=ww_mp2_ttm_12040Subset0.95sigma_E0.0001d1.0F_at0.1min1e-05_0.001.
    ↪ xyz
gap={water_dimer covariance_type=ARD_SE sparse_method=INDEX_FILE
sparse_file=fps_2000_tht1.0pow-1_cut7.0_from_ww_mp2_ttm_12040Subset0.95.
    ↪ xyz.txt
cutoff=7.0 n_sparse=2000 theta_file=theta2.0_0.5.dat delta=1.0
cutoff_transition_width=1.0 monomer_cutoff=1.9 OHH_ordercheck=F power=-1
dist_shift=0.25 } default_sigma={1 1 1 1} sparse_jitter=1.0e-10 e0=0.0
energy_parameter_name=E_MP2_TTM_2b force_parameter_name=F_MP2_TTM_2b
sparse_separate_file=T do_copy_at_file=F gp_file=${xml_file_name}
```

2B CCSD(T)-F12–MP2

```
at_file=train_cc_mp2_2bSubset0.95.xyz gap={water_dimer
covariance_type=ARD_SE sparse_method=RANDOM cutoff=7.0 n_sparse=1435
theta_file=theta2.0_0.5.dat delta=0.05 cutoff_transition_width=1.0
monomer_cutoff=1.9 OHH_ordercheck=F power=-1 dist_shift=0.25 }
default_sigma={0.00001 1 1 1} sparse_jitter=1.0e-10
e0=0.0 energy_parameter_name=E_CC_MP2_2B sparse_separate_file=T
do_copy_at_file=F gp_file=${xml_file_name}
```

3B MP2–TTM4Fmod

```
at_file=www_mp2_ttm_14554Subset0.95sigma_E0.0001d8.0F0.01_min2e-07_0.0001.
    ↪ xyz
gap={general_trimer covariance_type=ARD_SE sparse_method=INDEX_FILE
sparse_file=fps_5000_tht1.0_pow-1_cut6.0_from_www_mp2_ttm_14554Subset0.95.
    ↪ xyz.txt
cutoff=6.0 cutoff_transition_width=1.0 monomer_one_cutoff=1.9
monomer_two_cutoff=1.9 monomer_three_cutoff=1.9
signature_one={{8 1 1}} signature_two={{8 1 1}} signature_three={{8 1 1}}
n_sparse=5000  atom_ordercheck=F strict=F
theta_file=theta3b_1.0_0.5.dat delta=0.05 power=-1 dist_shift=0.1 }
default_sigma={1 1 1 1} sparse_jitter=1.0e-10 e0=0.0
energy_parameter_name=E_MP2_TTM_3b force_parameter_name=F_MP2_TTM_3b
sparse_separate_file=T do_copy_at_file=F gp_file=${xml_file_name}
```

3B CCSD(T)-F12–MP2

```
at_file=www_ccmp2_2654Subset0.95.xyz gap={general_trimer
covariance_type=ARD_SE sparse_method=RANDOM cutoff=5.0
cutoff_transition_width=1.0
monomer_one_cutoff=1.9 monomer_two_cutoff=1.9 monomer_three_cutoff=1.9
signature_one={{8 1 1}} signature_two={{8 1 1}} signature_three={{8 1 1}}
n_sparse=2524 theta_file=theta3b_1.0_0.5.dat delta=0.01 power=-1
dist_shift=0.25 atom_ordercheck=F strict=F }
```

```
default_sigma={0.00002 1 1 1} sparse_jitter=1.0e-10
e0=0.0 energy_parameter_name=E_CC_MP2_3B sparse_separate_file=T
do_copy_at_file=F gp_file=${xml_file_name}
```

Methane–water interaction fits

2B $(CH_4)(H_2O)$ MP2

```
at_file=mw_mp2Filt1.5Subset0.95sigma_E0.001F0.005d1.0_min1e-05_0.001.xyz
gap={general_dimer covariance_type=ARD_SE sparse_method=INDEX_FILE
sparse_file=fps_5000_tht1.0_cut7.0_from_mw_mp2atz_conc_filt1.5
    ↪ _minus_subset_5.0percent.xyz.txt
cutoff=7.0 n_sparse=5000 theta_file=theta1.0_0.5.dat delta=5.0
cutoff_transition_width=1.0 monomer_one_cutoff=1.3 monomer_two_cutoff=1.9
internal_swaps_only=T signature_one={{6 1 1 1 1}} signature_two={{8 1 1}}
strict=F mpifind=F atom_ordercheck=F double_count=F power=-1
dist_shift=0.5 } default_sigma={1 1 1 1} sparse_jitter=1.0e-10 e0=0.0
energy_parameter_name=E_MP2_2b force_parameter_name=F_MP2_2b
sparse_separate_file=T do_copy_at_file=F gp_file=${xml_file_name}
```

2B $(CH_4)(H_2O)$ CCSD(T)-F12–MP2

```
at_file=mw_cca_mp2_3328Subset0.95sigma_E5e-05d1.0min1e-06.xyz
gap={general_dimer covariance_type=ARD_SE sparse_method=RANDOM cutoff=7.0
n_sparse=3167 theta_file=theta1.0_0.5.dat delta=0.01
cutoff_transition_width=1.0 monomer_one_cutoff=1.3 monomer_two_cutoff=1.9
internal_swaps_only=T power=-1 dist_shift=0.5 signature_one={{6 1 1 1 1}}
signature_two={{8 1 1}} strict=F mpifind=F atom_ordercheck=F
double_count=F } default_sigma={1 1 1 1} sparse_jitter=1.0e-10
e0=0.0 energy_parameter_name=E_CC_MP2_2b
force_parameter_name= sparse_separate_file=T do_copy_at_file=F
gp_file=${xml_file_name}
```

3B $(CH_4)(H_2O)_2$ MP2

```
at_file=mww_mp2_15777Subset0.95sigma_E0.0001F0.001d1.0_min1e-05_0.001.xyz
gap={general_trimer covariance_type=ARD_SE sparse_method=INDEX_FILE
sparse_file=fps_5000_tht1.0pow-1_cut5.0_from_mww_mp2_15777Subset0.95.xyz.
    ↪ txt
cutoff=6.0 n_sparse=5000 theta_file=theta3b_3.0_0.5.dat delta=0.1
cutoff_transition_width=1.0 monomer_one_cutoff=1.3 monomer_two_cutoff=1.9
monomer_three_cutoff=1.9 signature_one={{6 1 1 1 1}}
signature_two={{8 1 1}} signature_three={{8 1 1}} strict=F
atom_ordercheck=F power=-1 dist_shift=0.25
mpifind=F } default_sigma={1 1 1 1} sparse_jitter=1.0e-10 e0=0.0
energy_parameter_name=E_MP2_3B force_parameter_name=F_MP2_3B
sparse_separate_file=T do_copy_at_file=F gp_file=${xml_file_name}
```

3B $(CH_4)(H_2O)_2$ CCSD(T)-F12–MP2

```
at_file=mww1875ccmp2Subset0.95.xyz gap={general_trimer
covariance_type=ARD_SE sparse_method=RANDOM
cutoff=5.0 n_sparse=700 theta_file=theta3b_3.0_0.5.dat
delta=0.02 cutoff_transition_width=1.0 monomer_one_cutoff=1.3
monomer_two_cutoff=1.9 monomer_three_cutoff=1.9
signature_one={{6 1 1 1 1}} signature_two={{8 1 1}}
signature_three={{8 1 1}} strict=F atom_ordercheck=F
power=-1 dist_shift=0.6 mpifind=F } default_sigma={0.000005 1 1 1}
sparse_jitter=1.0e-10 e0=0.0 energy_parameter_name=E_CC_MP2_3B
force_parameter_name= sparse_separate_file=T do_copy_at_file=F gp_file=${
    ↪ xml_file_name}
```

References

[1] A. P. Bartók, Gaussian Approximation Potential: an interatomic potential derived from first principles Quantum Mechanics, Ph.D. thesis, University of Cambridge, Engineering Dept, Trumpington St, Cambridge CB2 1PZ (2010).

[2] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, Phys. Rev. Lett. 104 (13) (2010) 136403. doi:10.1103/PhysRevLett.104.136403.

[3] B. J. Braams, J. M. Bowman, Permutationally invariant potential energy surfaces in high dimensionality, Int. Rev. Phys. Chem. 28 (4) (2009) 577–606. doi:10.1080/01442350903234923.

[4] V. Babin, C. Leforestier, F. Paesani, Development of a "first principles" water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient, J. Chem. Theory Comput. 9 (12) (2013) 5395–5403. doi:10.1021/ct400863t.

[5] T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, F. Paesani, Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions, J. Chem. Phys. 148 (24) (2018) 241725. doi:10.1063/1.5024577.

[6] M. Riera, A. Hirales, R. Ghosh, F. Paesani, Data-Driven Many-Body Models with Chemical Accuracy for CH4/H2O Mixtures, J. Phys. Chem. B 124 (49) (2020) 11207–11221. doi:10.1021/acs.jpcb.0c08728.

[7] J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett. 98 (2007) 146401. doi:10.1103/PhysRevLett.98.146401.

[8] A. P. Bartók, M. J. Gillan, F. R. Manby, G. Csányi, Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water, Phys. Rev. B Condens. Matter Mater. Phys. 88 (5) (2013) 054104. doi:10.1103/PhysRevB.88.054104.

[9] GAP Software, www.libatoms.org (2016).

[10] V. Babin, G. R. Medders, F. Paesani, Development of a "first principles" water potential with flexible monomers. II: Trimer potential energy surface, third virial

coefficient, and small clusters, J. Chem. Theory Comput. 10 (4) (2014) 1599–1607. doi:10.1021/ct500079y.

[11] G. A. Cisneros, K. T. Wikfeldt, L. Ojamäe, J. Lu, Y. Xu, H. Torabifard, A. P. Bartók, G. Csányi, V. Molinero, F. Paesani, Modeling Molecular Interactions in Water: From Pairwise to Many-Body Potential Energy Functions, Chem. Rev. 116 (13) (2016) 7501–7528. doi:10.1021/acs.chemrev.5b00644.

[12] Y. Wang, B. C. Shepler, B. J. Braams, J. M. Bowman, Full-dimensional, ab initio potential energy and dipole moment surfaces for water, J. Chem. Phys. 131 (5) (2009). doi:10.1063/1.3196178.

[13] M. J. Gillan, D. Alfè, A. P. Bartók, G. Csányi, First-principles energetics of water clusters and ice: A many-body analysis, J. Chem. Phys. 139 (24) (2013). doi:10.1063/1.4852182.

[14] O. Akin-Ojo, K. Szalewicz, Potential energy surface and second virial coefficient of methane-water from ab initio calculations, J. Chem. Phys. 123 (13) (2005) 134311–14314. doi:10.1063/1.2033667.

[15] C. Qu, R. Conte, P. L. Houston, J. M. Bowman, "Plug and play" full-dimensional ab initio potential energy and dipole moment surfaces and anharmonic vibrational analysis for CH_4–H_2O, Phys. Chem. Chem. Phys. 17 (12) (2015) 8172. doi:10.1039/c4cp05913a.

[16] R. Conte, C. Qu, J. M. Bowman, Permutationally invariant fitting of many-body, non-covalent interactions with application to three-body methane-water-water, J. Chem. Theory Comput. 11 (4) (2015) 1631–1638. doi:10.1021/acs.jctc.5b00091.

[17] O. Akin-Ojo, K. Szalewicz, Does a pair of methane molecules aggregate in water?, J. Chem. Phys. 150 (8) (2019). doi:10.1063/1.5083826.

[18] M. P. Metz, K. Szalewicz, J. Sarka, R. Tóbiás, A. G. Császár, E. Mátyus, Molecular dimers of methane clathrates: Ab initio potential energy surfaces and variational vibrational states, Phys. Chem. Chem. Phys. 21 (25) (2019) 13504–13525. doi:10.1039/c9cp00993k.

[19] N. Thakre, A. K. Jana, Computing Anisotropic Cavity Potential for Clathrate Hydrates, J. Phys. Chem. A 123 (13) (2019) 2762–2770. doi:10.1021/acs.jpca.8b12335.

[20] C. Qu, J. M. Bowman, Ab Initio, Embedded Local-Monomer Calculations of Methane Vibrational Energies in Clathrate Hydrates, J. Phys. Chem. C 120 (6) (2016) 3167–3175. doi:10.1021/acs.jpcc.5b11117.

[21] M. P. Metz, K. Piszczatowski, K. Szalewicz, Automatic Generation of Intermolecular Potential Energy Surfaces, J. Chem. Theory Comput. 12 (12) (2016) 5895–5919. doi:10.1021/acs.jctc.6b00913.

[22] M. Riera-Riambau, MBX, https://github.com/chemphys/MBX (2019).

[23] C. G. Pruteanu, G. J. Ackland, W. C. K. Poon, J. S. Loveday, When immiscible becomes miscible — methane in water at high pressures, Sci. Adv. 3 (8) (2017) 1–6. doi:10.1126/sciadv.1700240.

[24] E. D. Sloan, C. A. Koh, Clathrate Hydrates of Natural Gases, CRC Press, Taylor & Francis Group, 2008.

[25] R. Boswell, T. S. Collett, Current perspectives on gas hydrate resources, Energy Environ. Sci. 4 (4) (2011) 1206–1215. doi:10.1039/C0EE00203H.

[26] F. W. Taylor, The greenhouse effect and climate change, Rep. Rog. Phys. 54 (1991) 881–918. doi:10.1029/RG027i001p00115.

[27] R. Pallardy, Deepwater Horizon oil spill of 2010, https://www.britannica.com/event/Deepwater-Horizon-oil-spill-of-2010, accessed: 2017-08-28 (2010).

[28] A. K. Sum, C. A. Koh, E. D. Sloan, Clathrate Hydrates: From Laboratory Science to Engineering Practice, Ind. Eng. Chem. Res. 48 (16) (2009) 7457–7465. doi:10.1021/ie900679m.

[29] K. Yamamoto, Y. Terao, T. Fujii, T. Ikawa, M. Seki, M. Matsuzawa, T. Kanno, et al., Operational overview of the first offshore production test of methane hydrates in the Eastern Nankai Trough (2014). doi:10.4043/25243-MS.

[30] K. Yamamoto, X. X. Wang, M. Tamaki, K. Suzuki, The second offshore production of methane hydrate in the Nankai Trough and gas production behavior from a heterogeneous methane hydrate reservoir, RSC Adv. 9 (45) (2019) 25987–26013. doi:10.1039/c9ra00755e.

[31] L. Chen, Y. Feng, J. Okajima, A. Komiya, S. Maruyama, Production behavior and numerical analysis for 2017 methane hydrate extraction test of Shenhu, South China Sea, J. Nat. Gas Sci. Eng. 53 (January) (2018) 55–66. doi:10.1016/j.jngse.2018.02.029.

[32] Y. F. Makogon, R. Y. Omelchenko, Commercial gas production from Messoyakha deposit in hydrate conditions, J. Nat. Gas Sci. Eng. 11 (2013) 1–6. doi:10.1016/j.jngse.2012.08.002.

[33] A. Kumar, H. P. Veluswamy, R. Kumar, P. Linga, Direct use of seawater for rapid methane storage via clathrate (sII) hydrates, Appl. Energy 235 (August 2018) (2019) 21–30. doi:10.1016/j.apenergy.2018.10.085.

[34] A. M. Gambelli, An experimental description of the double positive effect of CO2 injection in methane hydrate deposits in terms of climate change mitigation, Chem. Eng. Sci. 233 (2021) 116430. doi:10.1016/j.ces.2020.116430.

[35] J. Wang, S. Wu, Y. Yao, Quantifying gas hydrate from microbial methane in the South China Sea, J. Asian Earth Sci. 168 (January) (2018) 48–56. doi:10.1016/j.jseaes.2018.01.020.

[36] Z. Tayber, A. Meilijson, Z. Ben-Avraham, Y. Makovsky, Methane hydrate stability and potential resource in the levant basin, southeastern mediterranean sea, Geosci. J. 9 (7) (2019) 306. doi:10.3390/geosciences9070306.

[37] Y. Makogon, Features of Natural Gas Fields Exploitation in Permafrost Zone, Gazov. Promyshlennost 9 (1966) 1–17.

[38] C. Bily, J. W. L. Dick, Naturally Occurring Gas Hydrates in the Mackenzie Delta, N.W.T., Bulletin of Canadian Petroleum Geology 22 (3) (1974) 340–352.

[39] K. A. Kvenvolden, Methane hydrate - A major reservoir of carbon in the shallow geosphere?, Chem. Geol. 71 (1-3) (1988) 41–51. doi:10.1016/0009-2541(88)90104-0.

[40] K. A. Knenvolden, M. A. McMenamin, Hydrates of Natural Gas: A Review of Their Geologic Occurence, U.S. Geol. Surv. Circ. 825 (1980) 1–11.

[41] C. A. Tulk, D. D. Klug, A. M. Dos Santos, G. Karotis, M. Guthrie, J. J. Molaison, N. Pradhan, Cage occupancies in the high pressure structure H methane hydrate: A neutron diffraction study, J. Chem. Phys. 136 (5) (2012). doi:10.1063/1.3679875.

[42] N. Noguchi, T. Yonezawa, Y. Yokoi, T. Tokunaga, T. Moriwaki, Y. Ikemoto, H. Okamura, Infrared and Raman Spectroscopic Study of Methane Clathrate Hydrates at Low Temperatures and High Pressures: Dynamics and Cage Occupancy of Methane, J. Phys. Chem. C 125 (1) (2021) 189–200. doi:10.1021/acs.jpcc.0c09315.

[43] I.-M. Chou, A. Sharma, R. C. Burruss, J. Shu, H.-k. Mao, R. J. Hemley, A. F. Goncharov, L. A. Stern, S. H. Kirby, Transformations in methane hydrates, Proc. Natl. Acad. Sci. U. S. A. 97 (25) (2000) 13484–13487. doi:10.1073/pnas.250466497.

[44] H. Hirai, Y. Uchihara, H. Fujihisa, M. Sakashita, E. Katoh, K. Aoki, K. Nagashima, Y. Yamamoto, T. Yagi, High-pressure structures of methane hydrate observed up to 8 GPa at room temperature, J. Chem. Phys. 115 (15) (2001) 7066–7070. doi:10.1063/1.1403690.

[45] J. S. Loveday, R. J. Nelmes, M. Guthrie, D. D. Klug, J. S. Tse, Transition from Cage Clathrate to Filled Ice: The Structure of Methane Hydrate III, Phys. Rev. Lett. 87 (21) (2001) 215501. doi:10.1103/PhysRevLett.87.215501.

[46] H. Shimizu, T. Kumazaki, T. Kume, S. Sasaki, In situ observations of high-pressure phase transformations in a synthetic methane hydrate, J. Phys. Chem. B 106 (1) (2002) 30–33. doi:10.1021/jp013010a.

[47] T. Kumazaki, Y. Kito, S. Sasaki, T. Kume, H. Shimizu, Single-crystal growth of the high-pressure phase II of methane hydrate and its Raman scattering study, Chem. Phys. Lett. 388 (1) (2004) 18–22. doi:10.1016/j.cplett.2004.02.064.

[48] H. Hirai, S. I. Machida, T. Kawamura, Y. Yamamoto, T. Yagi, Stabilizing of methane hydrate and transition to a new high-pressure structure at 40 GPa, Am. Mineral. 91 (5-6) (2006) 826–830. doi:10.2138/am.2006.1991.

[49] S. I. Machida, H. Hirai, T. Kawamura, Y. Yamamoto, T. Yagi, A new high-pressure structure of methane hydrate surviving to 86 GPa and its implications for the interiors of giant icy planets, Phys. Earth Planet. Inter. 155 (1-2) (2006) 170–176. doi:10.1016/j.pepi.2005.12.008.

[50] S. Sasaki, Y. Kito, T. Kume, H. Shimizu, High-pressure Raman study on the guest vibration in the host cage of methane hydrate structure I, Chem. Phys. Lett. 444 (1) (2007) 91–95. doi:10.1016/j.cplett.2007.07.018.

[51] S. Schaack, U. Ranieri, P. Depondt, R. Gaal, W. F. Kuhs, P. Gillet, F. Finocchi, L. E. Bove, Observation of methane filled hexagonal ice stable up to 150 GPa, Proc. Natl. Acad. Sci. U. S. A. 116 (33) (2019) 16204–16209. doi:10.1073/pnas.1904911116.

[52] J. S. Loveday, R. J. Nelmes, D. D. Klug, J. S. Tse, S. Desgreniers, Structural systematics in the clathrate hydrates under pressure, Can. J. Phys. 81 (1-2) (2003) 539–544. doi:10.1139/p03-040.

[53] J. S. Loveday, R. J. Nelmes, M. Guthrie, High-pressure transitions in methane hydrate, Chem. Phys. Lett. 350 (2001) 459–465.

[54] J. S. Loveday, R. J. Nelmes, M. Guthrie, S. a. Belmonte, D. R. Allan, D. D. Klug, J. S. Tse, Y. P. Handa, Stable methane hydrate above 2 GPa and the source of Titan's atmospheric methane, Nature 410 (6829) (2001) 661–663. doi:10.1038/35070513.

[55] J. M. Schicks, J. A. Ripmeester, The Coexistence of Two Different Methane Hydrate Phases under Moderate Pressure and Temperature Conditions: Kinetic versus Thermodynamic Products, Angew. Chemie 116 (25) (2004) 3372–3375. doi:10.1002/ange.200453898.

[56] W. Shin, S. Park, H. Ro, D. Y. Koh, J. Seol, H. Lee, Spectroscopic confirmation of metastable structure formation occurring in natural gas hydrates, Chem. - An Asian J. 7 (10) (2012) 2235–2238. doi:10.1002/asia.201200040.

[57] A. S. Stoporev, A. G. Ogienko, A. A. Sizikov, A. P. Semenov, D. S. Kopitsyn, V. A. Vinokurov, L. I. Svarovskaya, L. K. Altunina, A. Y. Manakov, Unexpected formation of sII methane hydrate in some water-in-oil emulsions: Different reasons for the same phenomenon, J. Nat. Gas Sci. Eng. 60 (2018) 284–293. doi:10.1016/j.jngse.2018.10.020.

[58] J. Vatamanu, P. G. Kusalik, Unusual Crystalline and Polycrystalline Structures in Methane Hydrates, J. Am. Chem. Soc. 128 (49) (2006) 15588–15589. doi:10.1021/ja066515t.

[59] L. Yang, C. A. Tulk, D. D. Klug, I. L. Moudrakovski, C. I. Ratcliffe, J. A. Ripmeester, B. C. Chakoumakos, L. Ehmd, C. D. Martin, J. B. Parise, Synthesis and characterization of a new structure of gas hydrate, Proc. Natl. Acad. Sci. U. S. A. 106 (15) (2009) 6060–6064. doi:10.1073/pnas.0809342106.

[60] X. Cao, Y. Huang, X. Jiang, Y. Su, J. Zhao, Phase diagram of water-methane by first-principles thermodynamics: discovery of MH-IV and MH-V hydrates, Phys. Chem. Chem. Phys. 19 (24) (2017) 15996–16002. doi:10.1039/C7CP01147D.

[61] Y. Huang, K. Li, X. Jiang, Y. Su, X. Cao, J. Zhao, Phase Diagram of Methane Hydrates and Discovery of MH-VI Hydrate, J. Phys. Chem. A 122 (28) (2018) 6007–6013. doi:10.1021/acs.jpca.8b02590.

[62] L. Jensen, K. Thomsen, N. Von Solms, S. Wierzchowski, M. R. Walsh, C. A. Koh, E. D. Sloan, D. T. Wu, A. K. Sum, Calculation of liquid water-hydrate-methane vapor phase equilibria from molecular simulations, J. Phys. Chem. B 114 (17) (2010) 5775–5782. doi:10.1021/jp911032q.

[63] M. M. Conde, C. Vega, Determining the three-phase coexistence line in methane hydrates using computer simulations, J. Chem. Phys. 133 (6) (2010). doi:10.1063/1.3466751.

[64] D. Jin, B. Coasne, Molecular Simulation of the Phase Diagram of Methane Hydrate: Free Energy Calculations, Direct Coexistence Method, and Hyperparallel Tempering, Langmuir 33 (42) (2017) 11217–11230. doi:10.1021/acs.langmuir.7b02238.

[65] A. Lenz, L. Ojamäe, Structures of the I-, II- and H-Methane clathrates and the ice-methane clathrate phase transition from quantum-chemical modeling with force-field thermal corrections, J. Phys. Chem. A 115 (23) (2011) 6169–6176. doi:10.1021/jp111328v.

[66] S. Alavi, J. A. Ripmeester, D. D. Klug, Molecular dynamics study of the stability of methane structure H clathrate hydrates, J. Chem. Phys. 126 (12) (2007). doi:10.1063/1.2710261.

[67] X. Cao, Y. Su, Y. Liu, J. Zhao, C. Liu, Storage capacity and vibration frequencies of guest molecules in CH 4 and CO2 hydrates by first-principles calculations, J. Phys. Chem. A 118 (1) (2014) 215–222. doi:10.1021/jp408763z.

[68] J. Liu, H. Liu, J. Xu, G. Chen, J. Zhang, S. Wang, Structure and stability of multiply occupied methane clathrate hydrates, Chem. Phys. Lett. 637 (2015) 110–114. doi:10.1016/j.cplett.2015.08.010.

[69] J. Liu, Y. Yan, J. Zhang, J. Xu, G. Chen, J. Hou, Theoretical investigation of storage capacity of hydrocarbon gas in sH hydrate, Chem. Phys. 525 (2019) 110393. doi:10.1016/j.chemphys.2019.110393.

[70] Y. Okano, K. Yasuoka, Free-energy calculation of structure-H hydrates, J. Chem. Phys. 124 (2) (2006) 1–9. doi:10.1063/1.2150430.

[71] J. Vatamanu, P. G. Kusalik, Molecular Insights into the Heterogeneous Crystal Growth of sI Methane Hydrate, J. Phys. Chem. B 110 (32) (2006) 15896–15904. doi:10.1021/jp0616841.

[72] H. Jiang, K. D. Jordan, C. E. Taylor, Molecular dynamics simulations of methane hydrate using polarizable force fields, J. Phys. Chem. B 111 (23) (2007) 6486–6492. doi:10.1021/jp068505k.

[73] S. J. Wierzchowski, P. A. Monson, Calculation of free energies and chemical potentials for gas hydrates using Monte Carlo simulations, J. Phys. Chem. B 111 (25) (2007) 7274–7282. doi:10.1021/jp068325a.

[74] Jiang, Hao, E. M. Myshakin, K. D. Jordan, R. P. Warzinski, Molecular Dynamics Simulations of the Thermal Conductivity of Methane Hydrate, J. Phys. Chem. B 113 (112) (2008) 10207–10216. doi:10.1021/jp807208z.

[75] S. A. Bagherzadeh, P. Englezos, S. Alavi, J. A. Ripmeester, Molecular simulation of non-equilibrium methane hydrate decomposition process, J. Chem. Thermodyn. 44 (1) (2012) 13–19. doi:10.1016/j.jct.2011.08.021.

[76] P. Pirzadeh, P. G. Kusalik, Molecular insights into clathrate hydrate nucleation at an ice-solution interface, J. Am. Chem. Soc. 135 (19) (2013) 7278–7287. doi:10.1021/ja400521e.

[77] A. H. Nguyen, M. A. Koc, T. D. Shepherd, V. Molinero, Structure of the ice-clathrate interface, J. Phys. Chem. C 119 (8) (2015) 4104–4117. doi:10.1021/jp511749q.

[78] S. J. Cox, D. J. Taylor, T. G. Youngs, A. K. Soper, T. S. Totton, R. G. Chapman, M. Arjmandi, M. G. Hodges, N. T. Skipper, A. Michaelides, Formation of Methane Hydrate in the Presence of Natural and Synthetic Nanoparticles, J. Am. Chem. Soc. 140 (9) (2018) 3277–3284. doi:10.1021/jacs.7b12050.

[79] J. Liu, J. Hou, H. Liu, M. Liu, J. Xu, G. Chen, J. Zhang, Molecular mechanism of formation of the face-sharing double cages in structure-I methane hydrate, Chem. Phys. Lett. 691 (2018) 155–162. doi:10.1016/j.cplett.2017.11.013.

[80] J. Hou, J. Liu, J. Xu, J. Zhong, Y. Yan, J. Zhang, Two-dimensional methane hydrate: Plum-pudding structure and sandwich structure, Chem. Phys. Lett. 725 (2019) 38–44. doi:10.1016/j.cplett.2019.04.006.

[81] S. J. Cox, M. D. Towler, D. Alfè, A. Michaelides, Benchmarking the performance of density functional theory and point charge force fields in their description of sI methane hydrate against diffusion Monte Carlo, J. Chem. Phys. 140 (17) (2014) 1–5. doi:10.1063/1.4871873.

[82] D. Hankins, J. W. Moskowitz, F. H. Stillinger, Water Molecule Interactions, J. Chem. Phys. 53 (12) (1970) 4544–4554. doi:10.1063/1.1673986.

[83] T. H. Dunning, Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen, J. Chem. Phys. 90 (2) (1989) 1007–1023. doi:10.1063/1.456153.

[84] K. A. Peterson, T. B. Adler, H. J. Werner, Systematically convergent basis sets for explicitly correlated wavefunctions: The atoms H, He, B-Ne, and Al-Ar, J. Chem. Phys. 128 (8) (2008) 084102. doi:10.1063/1.2831537.

[85] S. Boys, F. Bernardi, The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors, Mol. Phys. 19 (4) (1970) 553–566. doi:10.1080/00268977000101561.

[86] A. J. Stone, The Theory of Intermolecular Forces, Clarendon Press, Oxford, 1996.

[87] F. B. Van Duijneveldt, J. G. van Duijneveldt-van de Rijdt, J. H. van Lenthe, State of the Art in Counterpoise Theory, Chem. Rev. 94 (1994) 1873–1885. doi:10.1021/cr00031a007.

[88] B. Brauer, M. K. Kesharwani, J. M. Martin, Some observations on counterpoise corrections for explicitly correlated calculations on noncovalent interactions, J. Chem. Theory Comput. 10 (9) (2014) 3791–3799. doi:10.1021/ct500513b.

[89] X. W. Sheng, L. Mentel, O. V. Gritsenko, E. J. Baerends, Counterpoise Correction is Not Useful for Short and Van der Waals Distances but May Be Useful at Long Range, J. Comput. Chem. 32 (2011) 2896–2901. doi:10.1002/jcc.21872.

[90] A. Halkier, W. Klopper, T. Helgaker, P. Jørgensen, P. R. Taylor, Basis set convergence of the interaction energy of hydrogen-bonded complexes, J. Chem. Phys. 111 (20) (1999) 9157–9167. doi:10.1063/1.479830.

[91] L. A. Burns, M. S. Marshall, C. D. Sherrill, Comparing counterpoise-corrected, uncorrected, and averaged binding energies for benchmarking noncovalent interactions, J. Chem. Theory Comput. 10 (1) (2014) 49–57. doi:10.1021/ct400149j.

[92] B. H. Wells, S. Wilson, Van der Waals Interaction Potentials: Many-Body Basis Set Superposition Effects, Chem. Phys. Lett. 101 (4,5) (1983) 429–434. doi:10.1016/0009-2614(83)87508-3.

[93] J. M. Martin, J. P. François, R. Gijbels, Combined bond-polarization basis sets for accurate determination of dissociation energies - Part 3: Basis set superposition error in polyatomic systems, Theor. Chim. Acta 76 (3) (1989) 195–209. doi:10.1007/BF00527473.

[94] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut, K. R. Shamasundar, T. B. Adler, R. D. Amos, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, E. Goll, C. Hampel, A. Hesselmann, G. Hetzer, T. Hrenar, G. Jansen, C. Köppl, Y. Liu, A. W. Lloyd, R. A. Mata, A. J. May, S. J. McNicholas, W. Meyer, M. E. Mura, A. Nicklass, D. P. O'Neill, P. Palmieri, D. Peng, K. Pflüger, R. Pitzer, M. Reiher, T. Shiozaki, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, M. Wang, Molpro, version 2012.1, a package of ab initio programs, see http://www.molpro.net (2012).

[95] H.-J. Werner, F. R. Manby, P. J. Knowles, Fast linear scaling second-order Møller-Plesset perturbation theory (MP2) using local and density fitting approximations, J. Chem. Phys. 118 (18) (2003) 8149–8160. doi:10.1063/1.1564816.

[96] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, Molpro: A general-purpose quantum chemistry program package, Wiley Interdiscip. Rev. Comput. Mol. Sci. 2 (2) (2012) 242–253. doi:10.1002/wcms.82.

[97] T. B. Adler, G. Knizia, H.-J. Werner, A simple and efficient CCSD(T)-F12 approximation, J. Chem. Phys. 127 (22) (2007) 221106. doi:10.1063/1.2817618.

[98] Y. S. Al-Hamdani, A. Tkatchenko, Understanding non-covalent interactions in larger molecular complexes from first principles, J. Chem. Phys. 150 (1) (2019) 010901. doi:10.1063/1.5075487.

[99] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, W. Györffy, D. Kats, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut, K. R. Shamasundar, T. B. Adler, R. D. Amos, S. J. Bennie, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, E. Goll, C. Hampel, A. Hesselmann, G. Hetzer, T. Hrenar, G. Jansen, C. Köppl, S. J. R. Lee, Y. Liu, A. W. Lloyd, Q. Ma, R. A. Mata, A. J. May, S. J. McNicholas, W. Meyer, T. F. Miller III, M. E. Mura, A. Nicklass, D. P. O'Neill, P. Palmieri, D. Peng, K. Pflüger, R. Pitzer, M. Reiher, T. Shiozaki, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, M. Wang, M. Welborn, MOLPRO, version 2019.2, a package of ab initio programs, see https://www.molpro.net (2019).

[100] C. J. Burnham, D. J. Anick, P. K. Mankoo, G. F. Reiter, The Vibrational Proton Potential in Bulk Liquid Water and Ice, J. Chem. Phys. 128 (15) (2008) 154519. doi:10.1063/1.2895750.

[101] B. T. Thole, Molecular polarizabilities calculated with a modified dipole interaction, Chem. Phys. 59 (3) (1981) 341–350. doi:10.1016/0301-0104(81)85176-2.

[102] C. J. Burnham, J. Li, S. S. Xantheas, M. Leslie, The parametrization of a Thole-type all-atom polarizable water model from first principles and its application to the study of water clusters (n=2-21) and the phonon spectrum of ice Ih, J. Chem. Phys. 110 (9) (1999) 4566–4581. doi:10.1063/1.478797.

[103] MB-pol plugin for OpenMM, https://github.com/paesanilab/mbpol_openmm_plugin (2018).

[104] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L. P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, V. S. Pande, OpenMM 7: Rapid development of high performance algorithms for molecular dynamics, PLoS Comput. Biol. 13 (7) (2017) 1–17. doi:10.1371/journal.pcbi.1005659.

[105] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, K. W. Jacobsen, The atomic simulation environment—a python library for working with atoms, J. Phys. Condens. Matter 29 (27) (2017) 273002. doi:10.1088/1361-648X/aa680e.

[106] E. Székely, A. Fekete, Code for TTM4Fmod+GAPs, https://github.com/eszter137/TTM4FmodGAPs (2021).

[107] D. W. Schwenke, Towards accurate ab initio predictions of the vibrational spectrum of methane, Spectrochim. Acta A Mol. Biomol. Spectrosc. 58 (4) (2002) 849–861. doi:10.1016/S1386-1425(01)00673-4.

[108] D. W. Schwenke, H. Partridge, Vibrational energy levels for CH_4 from an ab initio potential, Spectrochim. Acta A Mol. Biomol. Spectrosc. 57 (4) (2001) 887–895. doi:10.1016/S1386-1425(00)00451-0.

[109] A. P. Bartók, G. Csányi, Gaussian approximation potentials: A brief tutorial introduction, Int. J. Quantum Chem. 115 (16) (2015) 1051–1057. doi:10.1002/qua.24927.

[110] M. Veit, S. K. Jain, S. Bonakala, I. Rudra, D. Hohl, G. Csányi, Equation of State of Fluid Methane from First Principles with Machine Learning Potentials, J. Chem. Theory Comput. 15 (4) (2019) 2574–2586. doi:10.1021/acs.jctc.8b01242.

[111] S. Fujikake, V. L. Deringer, T. H. Lee, M. Krynski, S. R. Elliott, G. Csányi, Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures, J. Chem. Phys. 148 (24) (2018) 241714. doi:10.1063/1.5016317.

[112] V. L. Deringer, N. Bernstein, G. Csányi, C. B. Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, S. R. Elliott, Origins of structural and electronic transitions in disordered silicon, Nature 589 (2021) 59–64. doi:10.1038/s41586-020-03072-z.

[113] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

[114] A. Grisafi, D. M. Wilkins, G. Csányi, M. Ceriotti, Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems, Phys. Rev. Lett. 120 (3) (2018) 36002. doi:10.1103/PhysRevLett.120.036002.

[115] E. Snelson, Z. Ghahramani, Sparse Gaussian Processes Using Pseudo-Inputs, Adv. Neural. Inf. Process. Syst. 18 (2006) 1259–1266.

[116] M. W. Mahoney, P. Drineas, CUR matrix decompositions for improved data analysis, Proc. Natl. Acad. Sci. U. S. A. 106 (3) (2009) 697–702. doi:10.1073/pnas.0803205106.

[117] M. Ceriotti, G. A. Tribello, M. Parrinello, Demonstrating the transferability and the descriptive power of sketch-map, J. Chem. Theory Comput. 9 (3) (2013) 1521–1532. doi:10.1021/ct3010563.

[118] A. P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Phys. Rev. B 87 (18) (2013) 184115. doi:10.1103/PhysRevB.87.184115.

[119] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. Kermode, G. Csányi, M. Ceriotti, Machine Learning Unifies the Modelling of Materials and Molecules, Sci. Adv. 3 (12) (2017) 1–9. doi:10.1126/sciadv.1701816.

[120] A. P. Bartók, personal communication (2018).

[121] A. Brown, B. J. Braams, K. Christoffel, Z. Jin, J. M. Bowman, Classical and quasiclassical spectral analysis of CH_5^+ using an ab initio potential energy surface, J. Chem. Phys. 119 (17) (2003) 8790. doi:10.1063/1.1622379.

[122] C. Qu, Q. Yu, B. L. Van Hoozen, J. M. Bowman, R. A. Vargas-Hernandez, R. A. Vargas-Hernández, Assessing Gaussian Process Regression and Permutationally Invariant Polynomial Approaches to Represent High-Dimensional Potential Energy Surfaces, J. Chem. Theory Comput. 14 (7) (2018) 3381–3396. doi:10.1021/acs.jctc.8b00298.

[123] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, K. R. Müller, Machine learning of accurate energy-conserving molecular force fields, Sci. Adv. 3 (5) (2017). doi:10.1126/sciadv.1603015.

[124] A. P. Bartók, J. Kermode, N. Bernstein, G. Csányi, Machine learning a general purpose interatomic potential for silicon, Phys. Rev. X 8 (4) (2018) 41048. doi:10.1103/PhysRevX.8.041048.

[125] M. Ceriotti, M. J. Willatt, G. Csányi, Machine Learning of Atomic-Scale Properties Based on Physical Principles, Handb. Mater. Model. (2018) 1–27doi:10.1007/978-3-319-42913-7_68-1.

[126] A. Vázquez-Mayagoitia, personal communication (2018).

[127] A. P. Bartók, personal communication (2017).

[128] D. Case, R. Betz, D. Cerutti, T. Cheatham, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. Lee, S. LeGrand, P. Li, C.Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. Merz, G. Monard, H. Nguyen, I. H.T. Nguyen, A. Onufriev, D. Roe, A. Roitberg, C. Sagui, C. Simmerling, W. Botello-Smith, J. Swails, R. Walker, J. Wang, R. Wolf, X. Wu, L. Xiao, P. Kollman, AMBER 2016, University of California, San Francisco, 2016.

[129] E. Székely, Methane–water GAP-models, their datasets and parameters (2021). doi:10.17863/CAM.65236.

[130] H. Partridge, D. W. Schwenke, The determination of an accurate isotope dependent potential energy surface for water from extensive ab initio calculations and experimental data, J. Chem. Phys. 106 (11) (1997) 4618–4639. doi:10.1063/1.473987.

[131] X. G. Wang, T. Carrington, Using experimental data and a contracted basis Lanczos method to determine an accurate methane potential energy surface from a least squares optimization, J. Chem. Phys. 141 (15) (2014) 154106. doi:10.1063/1.4896569.

[132] M. D. Veit, Python plotting tools, https://github.com/max-veit/pyutils (2016).

[133] M. Riera-Riambau, personal communication (2020).

[134] C. van der Oord, G. Csányi, G. Dusson, C. Ortner, Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials, Mach. Learn. Sci. Technol. 1 (1) (2020) 015004. doi:10.1088/2632-2153/ab527c.

[135] M. Riera, E. P. Yeh, F. Paesani, Data-Driven Many-Body Models for Molecular Fluids: CO2/H2O Mixtures as a Case Study, J. Chem. Theory Comput. 16 (4) (2020) 2246–2257. doi:10.1021/acs.jctc.9b01175.

[136] F. Paesani, personal communication (2020).

[137] M. Riera-Riambau, E. Székely, Git clone for MBX, with QUIP-interface and quippy functions added, https://github.com/eszter137/MBX/tree/master-dev (2019).

[138] F. Pérez, B. E. Granger, IPython: a system for interactive scientific computing, Comput. Sci. Eng. 9 (3) (2007) 21–29. doi:10.1109/MCSE.2007.53.

[139] J. D. Hunter, Matplotlib: A 2d graphics environment, Comput. Sci. Eng. 9 (3) (2007) 90–95. doi:10.1109/MCSE.2007.55.

[140] A. Stukowski, Visualization and analysis of atomistic simulation data with OVITO – the Open Visualization Tool, Model. Simul. Mater. Sci. Eng 18 (1) (2010). doi:10.1088/0965-0393/18/1/015012.

[141] M. Matsumoto, T. Yagasaki, H. Tanaka, GenIce: Hydrogen-Disordered Ice Generator, J. Comput. Chem. 39 (2017) 61–64. doi:10.1002/jcc.25077.

[142] M. Matsumoto, GenIce, https://github.com/vitroid/GenIce (2017).

[143] A. Rohatgi, Webplotdigitizer: Version 4.4, https://automeris.io/WebPlotDigitizer (2020).

[144] K. Kumar, PhD/MPhil Thesis - a LaTeX Template, http://www-h.eng.cam.ac.uk/help/tpl/textprocessing/ThesisStyle/ (2021).

[145] J. Behler, Constructing high-dimensional neural network potentials: A tutorial review, Int. J. Quantum Chem. 115 (16) (2015) 1032–1050. doi:10.1002/qua.24890.

[146] A. Kamath, R. A. Vargas-Hernández, R. V. Krems, T. Carrington, S. Manzhos, Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy, J. Chem. Phys. 148 (24) (2018) 241702. doi:10.1063/1.5003074.

[147] S. Käser, D. Koner, A. S. Christensen, O. A. Von Lilienfeld, M. Meuwly, Machine Learning Models of Vibrating H2CO: Comparing Reproducing Kernels, FCHL, and PhysNet, J. Phys. Chem. A 124 (42) (2020) 8853–8865. doi:10.1021/acs.jpca.0c05979.

[148] W. M. C. Foulkes, L. Mitas, R. J. Needs, G. Rajagopal, Quantum Monte Carlo simulations of solids, Rev. Mod. Phys. 73 (1) (2001) 33–83. doi:10.1103/RevModPhys.73.33.

[149] A. Zen, J. G. Brandenburg, A. Michaelides, D. Alfè, A new scheme for fixed node diffusion quantum Monte Carlo with pseudopotentials: Improving reproducibility and reducing the trial-wave-function bias, J. Chem. Phys. 151 (13) (2019) 134105. doi:10.1063/1.5119729.

[150] M. A. Morales, J. McMinis, B. K. Clark, J. Kim, G. E. Scuseria, Multideterminant wave functions in quantum Monte Carlo, J. Chem. Theory Comput. 8 (7) (2012) 2181–2188. doi:10.1021/ct3003404.

[151] M. Dubecký, P. Jurečka, R. Derian, P. Hobza, M. Otyepka, L. Mitas, Quantum Monte Carlo methods describe noncovalent interactions with subchemical accuracy, J. Chem. Theory Comput. 9 (10) (2013) 4287–4292. doi:10.1021/ct4006739.

[152] F.-F. Wang, M. J. Deible, K. D. Jordan, Benchmark Study of the Interaction Energy for an $(H_2O)_{16}$ Cluster: Quantum Monte Carlo and Complete Basis Set Limit MP2 Results, J. Phys. Chem. A 117 (32) (2013) 7606–7611. doi:10.1021/jp404541c.

[153] A. Zen, S. Sorella, M. J. Gillan, A. Michaelides, D. Alfè, Boosting the accuracy and speed of quantum Monte Carlo: Size consistency and time step, Phys. Rev. B 93 (24) (2016) 241118. doi:10.1103/PhysRevB.93.241118.

[154] M. J. Gillan, D. Alfè, F. R. Manby, Energy benchmarks for methane-water systems from quantum Monte Carlo and second-order Møller-Plesset calculations, J. Chem. Phys. 143 (10) (2015) 102812. doi:10.1063/1.4926444.

[155] A. Zen, J. G. Brandenburg, J. Klimeš, A. Tkatchenko, D. Alfè, A. Michaelides, Fast and accurate quantum Monte Carlo for molecular crystals, Proc. Natl. Acad. Sci. U. S. A. 115 (8) (2018) 1724–1729. doi:10.1073/pnas.1715434115.

[156] D. Alfè, A. P. Bartók, G. Csányi, M. J. Gillan, Communication: Energy benchmarking with quantum Monte Carlo for water nano-droplets and bulk liquid water, J. Chem. Phys. 138 (22) (2013) 221102. doi:10.1063/1.4810882.

[157] D. Alfè, A. P. Bartók, G. Csányi, M. J. Gillan, Analyzing the errors of DFT approximations for compressed water systems, J. Chem. Phys. 141 (1) (2014) 014104. doi:10.1063/1.4885440.

[158] D. Alfè, Quantum monte carlo benchmark energies of water systems, http://www.homepages.ucl.ac.uk/~ucfbdxa/qmcwater.htm (2017).

[159] E. Székely, Machine learning models of water and methane, Master's thesis, University of Cambridge, Engineering Dept, Trumpington St, Cambridge CB2 1PZ (2017).

[160] D. Alfè, personal communication (2018).

[161] P. Teeratchanan, A. Hermann, Computational phase diagrams of noble gas hydrates under pressure, J. Chem. Phys. 143 (15) (2015) 154507. doi:10.1063/1.4933371.

[162] B. Santra, J. Klimeš, D. Alfè, A. Tkatchenko, B. Slater, A. Michaelides, R. Car, M. Scheffler, Hydrogen bonds and van der Waals forces in ice at ambient and high pressures, Phys. Rev. Lett. 107 (18) (2011) 1–5. doi:10.1103/PhysRevLett.107.185701.

[163] Z. Raza, D. Alfè, C. G. Salzmann, J. Klimeš, A. Michaelides, B. Slater, Proton ordering in cubic ice and hexagonal ice; A potential new ice phase - XIc, Phys. Chem. Chem. Phys. 13 (44) (2011) 19788–19795. doi:10.1039/c1cp22506e.

[164] S. Cox, personal communication (2017).

[165] D. Packwood, J. Kermode, L. Mones, N. Bernstein, J. Woolley, N. Gould, C. Ortner, G. Csányi, A universal preconditioner for simulating condensed phase materials, J. Chem. Phys. 144 (16) (2016). doi:10.1063/1.4947024.

[166] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, P. Gumbsch, Structural relaxation made simple, Phys. Rev. Lett. 97 (17) (2006) 1–4. doi:10.1103/PhysRevLett.97.170201.

[167] J. Nocedal, S. J. Wright, Numerical optimization, 2nd Edition, Springer series in operations research, Springer, New York ; London, 2006.

[168] A. Togo, I. Tanaka, First principles phonon calculations in materials science, Scr. Mater. 108 (2015) 1–5. doi:10.1016/j.scriptamat.2015.07.021.

[169] A. Togo, Phonopy, https://phonopy.github.io/phonopy/ (2009).

[170] J. M. Skelton, Phonons & Phonopy: Pro Tips, https://www.slideshare.net/jmskelton/phonons-phonopy-pro-tips-2014 (2014).

[171] M. Youssef, Re: [Phonopy-users] Abnormal Behavior of Bulk Modulus At High Temperature [Electronic mailing list message], Retrieved from https://sourceforge.net/p/phonopy/mailman/message/35269276/ (2016).

[172] L. T. Kong, Phonon dispersion measured directly from molecular dynamics simulations, Comput. Phys. Commun. 182 (10) (2011) 2201–2207. doi:10.1016/j.cpc.2011.04.019.

[173] A. Kurnosov, L. Dubrovinsky, A. Kuznetsov, V. Dmitriev, High-pressure/high-temperature behavior of the methane-ammonia-water system up to 3 GPa, Zeitschrift fur Naturforsch. - Sect. B J. Chem. Sci. 61 (12) (2006) 1573–1576. doi:10.1515/znb-2006-1215.

[174] H. Kadobayashi, H. Ohfuji, H. Hirai, M. Ohtake, Y. Yamamoto, Stability of methane hydrate at high-pressure and high-temperature of up to 40 GPa and 573 K, J. Phys. Conf. Ser. 1609 (1) (2020) 0–5. doi:10.1088/1742-6596/1609/1/012007.

[175] E. Whalley, Energies of the phases of ice at zero temperature and pressure, J. Chem. Phys. 81 (9) (1984) 4087–4092. doi:10.1063/1.448153.

[176] R. Brill, A. Tippe, Gitterparameter von Eis I bei tiefen Temperaturen, Acta Cryst. 23 (3) (1967) 343–345. doi:https://doi.org/10.1107/S0365110X67002774.

[177] S. Fukusako, Thermophysical properties of ice, snow, and sea ice, Int. J. Thermophys. 11 (2) (1990) 353–372. doi:10.1007/BF01133567.

[178] P. V. Hobbs, Ice physics, Clarendon Press, Oxford, 1974.

[179] M. Chaplin, Water Phase Diagram plot, http://webhome.phy.duke.edu/~hsg/763/table-images/water-phase-diagram.html, accessed: 2021-04-08 (2021).

[180] A. Reinhardt, B. Cheng, Quantum-mechanical exploration of the phase diagram of water, Nat. Commun. 12 (1) (2021). doi:10.1038/s41467-020-20821-w.

[181] M. Towler, personal communication (2017).

[182] D. W. Davidson, Y. P. Handa, C. I. Ratcliffe, J. S. Tse, B. M. Powell, The ability of small molecules to form clathrate hydrates of structure II, Nature 311 (1984) 142–143. doi:10.1038/311142a0.