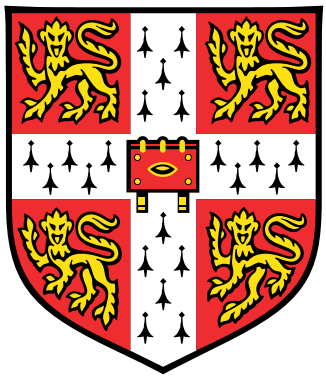
Designing a machine learning potential for molecular simulation of liquid alkanes



Max David Veit
Churchill College

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy.

14th September 2018

Supervisor: Prof. Gábor Csányi

Copyright © 2018 Max Veit.

Designing a machine learning potential for molecular simulation of liquid alkanes

Max David Veit

Summary

Molecular simulation is applied to understanding the behaviour of alkane liquids with the eventual goal of being able to predict the viscosity of an arbitrary alkane mixture from first principles. Such prediction would have numerous scientific and industrial applications, as alkanes are the largest component of fuels, lubricants, and waxes; furthermore, they form the backbones of a myriad of organic compounds. This dissertation details the creation of a potential, a model for how the atoms and molecules in the simulation interact, based on a systematic approximation of the quantum mechanical potential energy surface using machine learning. This approximation has the advantage of producing forces and energies of nearly quantum mechanical accuracy at a tiny fraction of the usual cost. It enables accurate simulation of the large systems and long timescales required for accurate prediction of properties such as the density and viscosity. The approach is developed and tested on methane, the simplest alkane, and investigations are made into potentials for longer, more complex alkanes. The results show that the approach is promising and should be pursued further to create an accurate machine learning potential for the alkanes. It could even be extended to more complex molecular liquids in the future.

Declaration

This dissertation is substantially my own work, includes nothing which is the outcome of work done in collaboration except as stated below and specified in the text, and conforms to the University of Cambridge's guidelines on plagiarism. Where reference has been made to other research this is acknowledged in the text and bibliography. It is not substantially the same as any that I have submitted, or is concurrently being submitted, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of this dissertation has already been submitted, or is currently being submitted, for any such degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. This dissertation does not exceed 65 000 words in length.

This dissertation includes the manuscript of a paper written as part of the PhD project, co-authored with my supervisor and collaborators at my funding organization, Shell. The bulk of the text was written and all the figures were made by me with input and comments provided by the co-authors; more details are provided in Chapter 4, which contains the manuscript.

Acknowledgements

First, I would like to thank my supervisor, Gábor Csányi, for advice and guidance throughout the course of this project. I would also like to thank everyone in the research group for innumerable helpful discussions.

I gratefully acknowledge Shell Global Solutions International BV for funding my studies at the University of Cambridge. I also thank the computational chemistry team at the Shell Technology Centre in Bengaluru, India, for useful discussions on hydrocarbon simulation, in addition to the collaboration already mentioned in the paper.

I would also like to thank the authors and contributors of the open software projects Python, IPython, SciPy, Matplotlib, and LaTeX, whose software allowed me to quickly and easily implement algorithms, test ideas, and explore, visualize, and communicate results. And finally, I acknowledge *The Elements of Style* by William Strunk Jr. and E. B. White, which has always offered succinct and timeless advice for improving my writing.

Finally, I thank my parents, my family, and the close friends I have made in Cambridge for their unending love and support.

Contents

Chapter 1

Introduction

Molecular simulation is a valuable tool for understanding the world we live in. We can use it to make successful predictions about the behaviour of many types of materials and apply it in concert with experimental investigations to complete our theoretical understanding of the underlying processes that govern their behaviour. In addition, predictions from simulation can be made much faster and on a much wider range of candidate materials than from experiments, in part because for experiments, each candiate material must be individually synthesized. Therefore, simulations can advance research into new materials by directing experimental investigation toward the most theoretically promising candidates.

The subject of this research is the application of molecular simulation to predict the properties of a large class of materials with enormous scientific and industrial importance: Alkanes, molecules consisting of carbon and hydrogen linked only by single bonds and having the chemical formula C_nH_{2n+2}[1], are the principal components of fuels, lubricants, and waxes. As such, they have been the subject of numerous attempts to determine their mechanical and rheological properties, as well as the mechanisms of their formation, from simulations [2–4]. The ability to accurately predict, for example, the viscosity of an arbitrary

mixture of hydrocarbons as a function of temperature would enable the computational design of new and more efficient lubricant formulations. Besides the obvious industrial applications, alkanes have also been the subject of widespread scientific interest as model systems for lipids and cell membranes [5–7], for the initial development of new potential energy surfaces [8, 9], for studying intermolecular dispersion forces [10, 11], and for the understanding of the flexibility of large molecules [12, 13].

Recent advances in molecular simulation technology are enabling the simulation of materials with a level of accuracy and capability that was never before possible. Most important to this work is the emergence of the application of machine learning to create exceptionally accurate and efficient potential energy surfaces that require few or no empirically tuned parameters [14–18]. The aim of this research is to use this technology to create a new potential for the molecular simulation of liquid-phase linear alkanes. In the short to medium term, we hope this potential will enable accurate, efficient, predictions of the viscosity of new formulations across a wide range of environmental conditions in order to better select theoretical candidate formulations for detailed experimental study. And in the long term, we hope the the new methodology established as part of this research will serve as a guide for future development of interatomic potentials, enabling researchers to easily create such accurate and efficient potentials for other, more complex molecular liquids.

1.1 Molecular simulation

Molecular simulation is a family of methods to compute the macroscopic observable properties of a system through statistical averages of microscopic configurations. For example, for an equilibrium system allowed to exchange energy

and volume with its environment, a static observable such as the density can be expressed as the thermodynamic average [19, 20]

$$\langle A \rangle = \frac{\int_{\mathbb{R}^{2n}} A(\mathbf{q},\mathbf{p}) \exp(-\beta(H(\mathbf{q},\mathbf{p}) + PV))\, \mathrm{d}^n q\, \mathrm{d}^n p}{\int_{\mathbb{R}^{2n}} \exp(-\beta(H(\mathbf{q},\mathbf{p}) + PV))\, \mathrm{d}^n q\, \mathrm{d}^n p} \tag{1.1}$$

over all possible microscopic states of the system denoted by the n generalized position coordinates $\mathbf{q}$ and momenta $\mathbf{p}$, where n is the number of degrees of freedom of the microscopic system, and $\beta = \frac{1}{k_B T}$ is the inverse temperature, V the volume, and P the external pressure. The function $H(\mathbf{q},\mathbf{p})$ is the Hamiltonian, which gives the total energy of the system in some state.

For any truly macroscopic system this average would be utterly impossible to compute directly, as the integrals would be over a space of dimension on the order of 10^{24} (for example; this is roughly the number of molecules in a small glass of water). The best we can hope to do is to, when examining properties that are truly determined at the molecular scale, is to simulate a much smaller piece of the system where the dimensionality of the space to be sampled is much more manageable. We then use some computational device to make it appear as if it is embedded in an infinite quantity of the same material (since, at the molecular scale, any macroscopic system is effectively infinite). The most common such device is that of periodic boundary conditions, where one side of the simulation box is effectively glued to the other side in a three-dimensional torus topology: The system being simulated is a collection of replicas of the same small periodic box, tiled infinitely in all directions. The properties of this periodic system will be close to that of the true macroscopic bulk, converging (at various rates) as the system size increases [20]. There do exist intensive properties that exhibit so-called "finite-size effects," that is, the simulated value of the property is sensitive to the size of the simulation box (instead of stabilizing above some

relatively small box size). The most important property to show this effect is the diffusivity [21, 22], though no such effect is known for the viscosity [23, 24]. Static properties (the type expressed by Equation 1.1) are likewise expected to be unaffected.

The configuration space of the periodic system is still very large, however, and it must be sampled efficiently to obtain acceptable estimates of the desired property. Two main classes of sampling methods are available: The first, Monte Carlo (MC), uses random numbers and various biasing, rejection, or probability inversion methods to sample the Boltzmann distribution as in Equation (1.1) [20]. Molecular dynamics (or **MD**), on the other hand, evolves the system using Newton's equations of motion, which naturally samples the microcanonical ensemble, or some modification thereof, in order to sample other thermodynamic ensembles. Since MD generates a continuous trajectory within some thermodynamic ensemble, it is often more difficult than in MC methods to ensure an adequate sampling of the entire relevant configuration space, especially where "rare events" are involved (i.e., the system exhibits processes with widely separated timescales). However, MD is also in some sense the more capable of the two methods, as it allows the calculation of dynamic properties (see Section 2.3) such as the diffusivity and the viscosity. For this reason, this research only uses MD simulations, although the machine learning potentials developed in this work could equally be applied to MC simulations.

1.1.1 Components of a simulation

The most important ingredient in an MD simulation is the potential that specifies the interactions in the system. A **potential** is a function $U(\mathbf{q})$ that gives the potential energy of the system at some coordinates $\mathbf{q}$. In MD, the system

obeys Newton's equations of motion,

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = m_i \frac{\mathrm{d}^2 q_i}{\mathrm{d}t^2} = -\frac{\partial U(\mathbf{q})}{\partial q_i}, \tag{1.2}$$

which require the negative gradient of the potential (i.e. the forces) to be computed.

Molecular simulations almost always assume the Born-Oppenheimer approximation [25], which decouples the electronic and nuclear degrees of freedom. The Schrödinger equation for the electrons is solved with the nuclear coordinates fixed, generating an effective potential energy surface as a function of the nuclear coordinates. The nuclei are then evolved on this effective adiabatic **Born-Oppenheimer PES** (potential energy surface). In practice, it is never possible to obtain this surface exactly, so approximations at various levels of accuracy, theoretical rigour, and computational expense are employed (see Section 1.2). These approximations are themselves called potential energy surfaces, or simply **potentials** or **models**. Finally, the vast majority of MD simulations treat the nuclei as classical particles moving on this surface, although this is not strictly part of the Born-Oppenheimer approximation. Nevertheless, nuclei are quantum particles; the classical limit is only an approximation that becomes increasingly worse for lighter nuclei. Section 2.2 discusses techniques to go beyond this approximation and model quantum nuclear effects, most of which still involve particles moving on the Born-Oppenheimer potential energy surface.

Once the interactions are specified, the equations of motion must be discretized in time so that they can be solved numerically. Most commonly, they are solved using the velocity Verlet method, which approximately conserves the total energy of the system over a long period of time [26]. The timestep must be chosen small enough to capture the fastest motions of the system, typically the vibra-

tions of hydrogen atoms, so that the solution does not become unstable.

The standard equations of motion evolve the system under the microcanonical (NVE) ensemble, which conserves the system's total energy. Often, however, we wish to study systems that interact with their environment, exchanging energy and volume. The equations of motion can be modified to simulate constant-temperature or constant-pressure conditions; these modifications are called thermostats and barostats and are discussed in more detail in Chapter 2.

1.1.2 Challenges for prediction

The main challenge in using MD simulations to predict material properties is that current simulations either cannot reach the speed necessary to generate a good enough sample of the configuration space, or they cannot reach the accuracy in the potential necessary to reproduce the properties correctly. Current accurate, quantum mechanical methods for evaluating the potential are much too slow to sample motion at long timescales, such as passage through local minima or metastable states, or to obtain a small enough statistical uncertainty in dynamical properties. Classical analytical potentials, on the other hand, are often fast enough to sample the dynamics of very large, complex systems; however, they are inaccurate representations of the potential energy surface and these inaccuracies are reflected in the properties predicted with such simulations [3, 27]. These potentials are usually fitted to reproduce certain experimental observables at one temperature and pressure, only to have their predictions fail when they are tested at other state points. Even potentials that achieve good accuracy for some property across wide ranges of temperature and pressure often do this by simplifying or smoothing the functional form [9], which may affect other properties that were not considered at the time the potential was designed. It is still

not clear what the full impact of these approximations are, or in which regions a potential needs to be most accurate to reliably predict a given property. Some preliminary investigations into this question are done in Section 3.1, though it appears we are still far from a definitive answer.

1.2 Potentials

The potential that determines the time evolution of an MD system is generally intended to be some approximation to the Born-Oppenheimer potential energy surface, or a thermal average of it. Beyond this common foundation, however, the methods for computing a potential vary from the most simplistic analytical forms to the most accurate, expensive algorithms for solving the molecular Schrödinger equation. No one potential can suit all the different problems and individual systems that molecular simulation is used to study. By looking at the way a potential is developed and how it is optimized for a specific application, we can see how to choose – or, if necessary, develop – a potential for a specific simulation.

1.2.1 Physical energy components

In a system of molecules, the Born-Oppenheimer PES is composed of several contributions that can be understood as distinct physical effects. These effects take place at different ranges of separation between the constituent atoms. Broadly, the contributions can be classified as local (or intramolecular), those between atoms connected by bonds or chains of only a few bonds, and nonlocal (or intermolecular), those taking place between different atoms in different molecules or parts of a molecule.

The most local effect is the covalent chemical bond, which is due to a lower-

ing in the total energy of a quantum system when partially filled atomic orbitals overlap. Different atoms display clear preferences in the number and orientation of the bonds they participate in; often, several configurations are possible. These configurations are local minima of the Born-Oppenheimer potential energy surface of the atom's local environment, so small deformations tend to a raise the energy of the molecule. The resulting effective forces can be expressed as functions of the bond lengths, angles, and four-body torsional profiles.

The intermolecular interactions can broadly be classified into the following effects: Electrostatic, induction, and dispersion interactions dominate at long range (large intermolecular separation), while exchange repulsion (along with the exchange-induction and exchange-dispersion terms) dominate at short range [28].

The exchange interactions dominate when molecules are close enough that their wavefunctions overlap significantly. The largest, repulsive component of the short-range exchange interaction is due to the Pauli exclusion principle, which forbids any two electrons from occupying the same quantum state. Its effect is to raise the energy of overlapping orbitals containing electrons of the same spin. It is also relevant between atoms in the same molecule, in which case it is often known as steric repulsion; this contribution is often treated together with the other (covalent) intramolecular forces. Exchange repulsion between molecules typically decays with an exponential tail as a function of intermolecular separation [28, 29] and is no longer relevant when the closest parts of a molecule become more widely separated than a few ångströms.

When molecules are far enough apart that their wavefunctions no longer have significant overlap, the only remaining interactions are electrostatics, induction, and dispersion*. The interaction of the *static* charge distributions of the two mo-

*Neither resonance nor magnetic interactions are relevant for the interaction of ordinary,

lecules is simply called the electrostatic energy. At long range, the distributions can be expanded in terms of charges, dipoles, and higher multipoles assigned to the whole molecule, leading to a classical and intuitive picture of electrostatic interactions. But this expansion is poorly convergent, even divergent, at medium to short range [28]. More importantly, it neglects a significant short-range component of the classical electrostatic energy, which is the error incurred by representing the spatially extended charge distributions by point properties. When the charge distributions of two molecules overlap, the actual electrostatic energy is smaller than that predicted by point-multipole expansions. This term, called the electrostatic penetration energy, is approximately proportional to the overlap integral of the two charge distributions. Therefore, in practice it might be grouped with the short-range components described above.

The molecular charge distributions can also deform in the electric field of another molecule or undergo quantum fluctuations. The former effect is called induction, while the latter is a purely quantum effect called dispersion, first described by van der Waals and later theoretically characterized by London [30] (hence the alternative names, “van der Waals forces” or “London dispersion forces”). Both effects are attractive. Among the true long-range interactions, dispersion is dominant (and hence most easily measurable) in systems, such as saturated hydrocarbons and noble gases, that are intrinsically nonpolar but highly polarizable [10]. Therefore, the bulk of this work will focus on characterizing the dispersion interaction in addition to short-range repulsion and covalent bonding.

In summary, the potential energy separates into several components with an

closed-shell alkane molecules in their ground states

overall energy expression:

$$E_{\text{total}} = E_{\text{1b}} + E_{\text{repulsion}} + E_{\text{dispersion}} + E_{\text{electrostatic}} + E_{\text{induction}} \tag{1.3}$$

with the intramolecular energy defined – for now – as the total energy of each molecule in the system considered in isolation, hence the "one-body" (1b) energy; the rest of the short-range terms are defined as components of the interaction *between* molecules, i.e. the "beyond-one-body" (b1b) components.

1.2.2 Classical potentials

Analytical potential energy surfaces for the interaction of atoms or molecules have been in use well before the invention of computers. They are typically based on simple functional forms derived from fundamental physical arguments, together with empirical parameters that must be optimized for each material. Such models are often called **forcefields**, with "potential" or "model" referring either to the whole model to a single term or component of the energy expression. The different physical interactions mentioned above are typically tackled separately, mirroring the real-world structure of the Born-Oppenheimer potential energy surface. For the local interactions, they contain terms that express the energy of an atom's neighbourhood in terms of bond lengths, angles between bonds, and four-body measures such as dihedral angles. The intermolecular interactions are typically modelled separately from the intramolecular ones, with terms for the electrostatic, exchange-repulsion, dispersion, and sometimes induction energies.

Intermolecular potential

One of the earliest examples of an intermolecular potential is the Lennard-Jones potential, which got its start when Lennard-Jones – at the time, just named

Jones – used the experimental equation of state of argon gas to fit an interatomic potential energy function formed as the sum of two inverse power terms, one for the attraction and one for the repulsion [31]. Concretely, the potential energy between two atoms was taken to be $u(r) = Ar^{-n} - Br^{-m}$, with four free parameters, A, B, m, and n, all greater than zero. The best fit to the available experimental data was found with the attractive power $m = 4$ and various larger values of the repulsive power n. Years later, however, London used the newly available quantum mechanics to study the interaction of rare-gas atoms and concluded that the first term of the attraction must go as Br^{-6} (plus eighth- and higher-power terms) [30]. Lennard-Jones then revised his model, setting $m = 6$ and proposing various values for n between 9 and 12, as an acceptable approximation to the sum of various exponential terms that more accurately model the Pauli repulsion. Eventually, the $n = 12$ form stuck (presumably for computational efficiency, as r^{-12} can obtained by simply squaring r^{-6}) and became the potential now commonly known simply as the **Lennard-Jones potential**, and more precisely as the 12-6 L-J form [26, 28]:

$$u_{\mathrm{LJ}}(r) = 4\epsilon\left(-\left(\frac{r}{\sigma}\right)^{-6} + \left(\frac{r}{\sigma}\right)^{-12}\right) \tag{1.4}$$

with just two adjustable parameters for the well depth ϵ and length scale σ. Many models take this form as the basis for their intermolecular energy function, adding an electrostatic model based on fixed charges for each atom type: The forcefields AMBER [32], OPLS [33, 34], TraPPE [9, 35], and AIREBO [36] are all based on the 12-6 repulsion-dispersion potential.

Note that this form of the potential couples the length and energy scales of the dispersion and repulsion components; most classical potentials do not treat these terms separately. But although the inverse sixth-power term remains justified

by fundamental physics, the inverse twelfth-power repulsive wall is not as good an approximation to the complex interactions that give rise to intermolecular repulsion. The twelfth-power repulsive wall is sufficient to keep molecules from getting too close to each other, but as expected, it starts giving wrong results under high-pressure conditions [37–39]. One proposal is to soften the repulsion by replacing it with a ninth-power term, the so-called 9-6 L-J potential. This form is used by forcefields such as Class II [40] and COMPASS [41], under the justification that the ninth-power repulsion is necessary to fit the interaction between closely spaced parts of the same molecule in the transition between the intramolecular and intermolecular regimes.

Other potentials choose an exponential form for the repulsion, going back to Born and Mayer's study [29] that appeared just after Lennard-Jones proposed his potential. Two more recent, fitted potential energy surfaces for the methane dimer use this form while incorporating higher-order terms in the dispersion model, up to r^{-8} in the model of Hellmann et. al. [42] and up to r^{-10} in the model of Gay et. al. [43]. The Slater-ISA model [39] was developed as a modern implementation of Born and Mayer's original reasoning, that of modelling the electron density as an exponentially decaying function and computing the repulsion as an overlap integral. Finally, the exponential repulsion form is also present in the Morse potential, on which several other potentials have been based [10, 37, 38] – although, possibly due to the potential's origin in molecular spectroscopy for describing intramolecular bonding, its long-range tail is incorrect.

The electrostatic potential is usually a Coulomb potential between fixed partial charges on atoms. It can be modified to approximately account for induction interactions, most commonly by scaling the interactions with an effective dielectric constant that models the electrostatic screening effects in the molecule and

solvent [32, 44]. More accurate models represent induction more explicitly by assigning individual polarizabilities to atoms, which requires a system of equations to be solved self-consistently to obtain the final partial charges (and possibly higher multipole moments) [26, 28, 45]. Such "polarizable" models can accurately model the many-body interactions that are crucial to determining many properties of polar liquids. The assignment of partial charges or multipoles to atoms or sites in a molecule, as well as the development of more accurate and efficient polarizable models, is an active area of research; a partial overview was given in Veit [46]. The long-range electrostatic and induction energy terms are typically very small in systems containing only saturated hydrocarbons; especially beyond about 5 Å of dimer separation, they become much smaller than the dispersion component [47]. The anisotropy of monomers such as benzene, the subject of the study just cited, can complicate matters by providing for orientations where the electrostatic energy dies out much more slowly than on average. But for the case of methane, the subject of the in-depth study of Chapter 4, this is not a problem as the molecule is highly symmetric, having an octupole as its highest permanent moment (see Section 4.1.3). It remains to check the magnitude and decay of the electrostatic terms on longer hydrocarbon chains in order to justify or, if necessary, rethink the continued neglect of these terms.

Some forcefields (e.g. AMBER) also include a special term to model hydrogen bonding; this interaction is likewise not present to a significant extent in hydrocarbons.

Intramolecular potential

Finally, a forcefield must model the intramolecular interactions, within the context of forcefields typically taken to mean interactions between atoms separated

by at most three bonds. This leaves four types of effective interaction coordinates: Bond lengths, angles, torsions (dihedral angles), and improper torsions (or out-of-plane motion), the latter two both involving groups of four atoms. Early forcefields, such as OPLS and AMBER, treated bonds and angles in the harmonic approximation, while fitting the dihedrals with a truncated Fourier series [44]. The constants of these expansions were fitted to reproduce experimental data such as heats of formation, structural parameters, and vibrational frequencies - often qualitatively, as the simple functional forms of these early potentials did not allow quantitative accuracy in all these properties [44, 48]. Other forcefields, such as the MMn series, took the approach of going beyond the harmonic, diagonal approximation: They used cubic and quartic terms to model the bond stretching and angle bending more accurately as well as introducing coupling terms between bonds, angles, and the torsional parameters. This additional complexity allowed the different degrees of freedom to interact and the forcefield to fit the experimental data more accurately [48, 49].

More recently, it has become practical to fit forcefields directly to a quantum mechanical potential energy surface – that is, a potential energy surface computed by approximately solving the equations of quantum mechanics (see Section 1.2.3). The creators of perhaps the first potential to use this approach, the Class II forcefield [40], argued that fitting the quantum mechanically derived Born-Oppenheimer PES is a more systematic approach capable of higher accuracy. The Class II energy equation follows a fairly standard intermolecular form (9-6 L-J plus Coulomb), but the intramolecular terms have been made more general and flexible to allow a close fit to quantum mechanical data. Like the MMn series, they include anharmonic bond and angle terms via a polynomial expansion (up to the fourth power of the bond and angle distorsions), as well as

off-diagonal coupling terms that model the coupling between two adjacent bond stretches, between bond and angle distorsions, between two adjacent angles, and finally between a torsional angle and all of the involved bonds and angles, mostly as an energy term proportional to the product of each of the involved distorsions. Note that this forcefield still only models the molecule close to equilibrium; very large distortions (dissociations or chemical changes) are not included.

Other forcefields have since followed in the same vein, fitting some or all of their intramolecular parameters using quantum mechanical calculations: COMPASS [41], which uses the Class II energy expression, and a refit of OPLS for long hydrocarbons [7] are more recent developments, as is the method of Hessian matrix projection [50] for systematic derivation of parameters for forcefields limited to the harmonic, diagonal form.

This approach is finding use in determining the intermolecular component of forcefields as well: The methane dimer potential of Li and Chao [51, 52] uses a standard 12-6 L-J form between each of the carbon and hydrogen atoms of the rigid methane molecule, with the parameters adjusted to fit accurate quantum mechanical data. Here again the constraint of simple functional forms prevents a more accurate fit. A more complex, physically motivated expression was used in the methane dimer potentials of Hellmann et. al. [42] and Gay et. al. [43] (as mentioned above). The earlier work of fitting intermolecular potentials to spectroscopic measurements [53] can even be seen as a forerunner to this approach, since the goal was to fit the correct quantum mechanical potential energy surface (only through experiments rather than explicit calculation).

This gradual change in forcefield fitting paradigm brings with it an important distinction that is not often mentioned: The quantum mechanical potential energy surface does *not* include quantum nuclear effects by itself, while the ex-

perimental data naturally does. This means that forcefields fit to experimental data incude these effects in an average way, while forcefields fit to the quantum PES must include it on top of the fit - either by explicitly including quantum nuclear effects in the simulations, by adding an approximate correction (as in the Hellmann methane dimer potential [42]), or even with something as simple as a constant scaling factor [41].

Summary

The set of classical potentials is large, diverse, and still expanding; it contains models as simple as the early, diagonal-harmonic forcefields and as complex as the explicit many-body local energy potentials (REBO [54] and AIREBO [36]). Their use of parameterized, analytical functional forms generally conveys them the advantage of computational efficiency, since these functions are typically very fast and easy to evaluate – with the efficiency decreasing as the complexity and need for special computational procedures increases. The use of simple, physics-based functional forms also usually makes a potential more transferable, that is, more easily applicable to new systems. This may be because the simpler forms are better at approximately capturing the generally applicable, underlying physical principles rather than tailoring their application to a specific material or class of materials.

In return for this efficiency and transferability, they have to constrain themselves to a small subspace of energy landscapes, a space that can only partially approximate the true Born-Oppenheimer PES. The fitting process is typically long and difficult, requiring considerable human input, as the goal is not uniquely defined – far from optimizing a simple objective function (e.g. an energy

error), care must instead be taken to balance accuracy in the desired properties with transferability across chemical compounds [41]. Even simple potentials still must deal with a moderate degree of coupling between the various energy components [40], so a small change in one parameter may affect properties across a wide range of systems.

Analytical potentials have been the tool of choice for decades in biomolecular and liquid simulation, perhaps because until recently no other approach could reach the length and time scales required. But the potentials described here must always make a compromise between accuracy and generality – both in terms of the variety of systems they can treat and the range of environmental conditions (temperature and pressure) for which they remain accurate.

1.2.3 Quantum-mechanical methods

In parallel with the development of analytical forcefields to understand liquids and large, biological molecules, there have been great advances in solving the equations of quantum mechanics on smaller systems to derive the Born-Oppenheimer potential energy surface from first principles. These methods are much more granular than analytical potentials – they treat electrons explicitly – and are much more expensive as well, meaning the development of these methods has generally advanced in the study of small molecules and crystalline materials. These methods all essentially attempt to solve the Schrödinger equation, which governs the behaviour of all common materials*. They range from the most accurate level of quantum chemistry, which is so computationally expensive that it is only feasible for small clusters of molecules, to the many variants of

*The Schrödinger equation does neglect relativistic effects; several methods exist to incorporate relativistic contributions in materials where these are important, but these corrections are almost negligibly small in atoms as light as carbon [55].

density functional theory, where the energy is obtained through various approximate functionals of the electron density, an approximation that allows much larger systems to be treated.

Quantum chemistry is a hierarchy of methods, also known as the wavefunction-based methods, starting with the Hartree-Fock self-consistent solution of the many-electron interaction problem. This method uses a fully antisymmetrical Slater determinant as its solution wavefunction, so it treats exchange exactly. However, the true ground state of a many-electron system includes contributions from many more Slater determinants that together lower the total energy. This difference from these contributions is termed the correlation energy. Quantum chemistry treats correlation in various ways, from perturbation theory (MP2) to explicit inclusion of higher Slater determinants (CI and coupled-cluster), or both (such as CCSD(T)) [56]. While quantum chemistry can only treat very small systems at an acceptable accuracy (The cost of CCSD(T), for example, scales with the seventh power of the number of atoms [56]; a single CCSD(T) calculation on the methane dimer may already require, as a representative figure, 10 minutes on 16 cores with the aug-cc-pVTZ basis set and the F12 correction), the hierarchy of methods converges systematically to the exact solution of the molecular Schrödinger equation and thus provides a standard class of reference methods for molecular energies.

Density functional theory (DFT), on the other hand, attempts a more efficient solution of the molecular Hamiltonian by recasting the Schrödinger equation in terms of the electron density. In principle, the electron density contains all the information necessary to reconstruct the Hamiltoninan and the total energy, so these methods only need to minimize the total energy with respect to the electron density. The main problem is that the functional that maps the electron

density to the total energy is not known exactly. An immense array of different approximations exists; the most successful ones include a fraction of exact exchange from the Hartree-Fock method and are fitted either to databases of molecules or to simple physical models [57]. While DFT is generally much more efficient than quantum chemistry models of comparable accuracy, and can treat much larger systems than quantum chemistry can, it has serious shortcomings in the treatment of dispersion, since the local formulation of the most common types of functionals neglects the long-range electron correlations that give rise to this interaction between well-separated molecules [28, 58].

Several methods are available to compute a dispersion correction on top of the DFT energy. One of the simplest is a pairwise additive model with fixed dispersion coefficients (called C^6 coefficients, where the dispersion correction between two atoms i and j goes as $u_{ij}(r_{ij}) = -C^6_{ij} r^{-6}_{ij}$), combined with a damping function to avoid the singularity at short range; the Grimme D2 model [59] is of this form. The dispersion model of Tkatchenko and Scheffler [60] (also known as T-S) computes its pairwise coefficients from the DFT-derived electron density, giving it the ability to correctly model changes in the dispersion coefficients due to distorsions in the molecular geometry and changes in the molecular environment. It has been shown to give accurate pairwise coefficients when compared against reference values computed using both high-level theory and experimental values. The newer D3 model of Grimme and co-workers [61] also adds geometry dependence, albeit with a relatively simple analytical function, though it also adds many-body effects that are known to be important in dispersion-bound systems. Finally, the MBD model of Tkatchenko et. al. [62] extends the original T-S model to also include these many-body effects. It is also possible to compute dispersion interactions by modifying the DFT functional itself [63–65]. A recent review [58]

covers the diverse array of methods that have been developed to augment DFT with dispersion interactions.

1.2.4 Machine learning potentials

Many applications require the accuracy of a quantum-mechanical method but cannot actually afford to use one, either because many evaluations of the potential are required, or because the system size is so large, or both. For example, the long-term goal of this research is the calculation of the viscosity of a liquid composed of large hydrocarbon molecules. Such a calculation requires very large system sizes and multiple independent, long simulations to achieve acceptable statistical and sampling errors: an example simulation from [3] used 100 n-hexadecane molecules and several independent simulations on nanosecond timescales (millions of timesteps). Such size and time scales are typical of simulation of liquids in general, in order to control the statistical fluctuations inherent to their relatively unstructured nature. Furthermore, the viscosity, as with other transport properties, is extremely sensitive to the accuracy of the inter- and intramolecular force models used in the simulation especially at higher pressures [3, 27, 66]. Ideally we would like a potential whose accuracy on each type of interaction can be systematically measured and controlled as in the quantum-mechanical methods described above. However, the limited dimensionality of the space in which traditional forcefields are fit makes this kind of fine-grained control difficult, as parameters must be carefully tuned to balance the error across various target properties while maintaining transferability to different compounds.

Clearly there is a wide gap between the two worlds of quantum methods, with their systematic convergence and accuracy, and of analytical potentials,

with their unrivalled computational efficiency. But in the past decade, there has been significant progress towards bridging this gap with the help of machine learning. The idea of using quantum data to fit an analytical potential is not new, as mentioned above, and neither is machine learning, which has been used to automatically classify and fit data – including chemical data [67] – almost as long as there have been computers to use for automation [68]. But the same forces that have brought machine learning into widespread use in recent times – the availability of massive amounts of data, along with the computing power necessary to fit larger, more complex models – have likewise made it possible to fit accurate potential energy surfaces directly to a sample of quantum mechanical calculations without needing to assume an underlying functional form. This works because the Born-Oppenheimer potential energy surface is smooth, i.e. similar molecular configurations have similar energies*. Therefore, a quantum-mechanical calculation at one geometry gives us information about similar geometries. Machine learning methods exploit this similarity by interpolating between the precomputed samples to construct a direct approximation of the PES. Unlike the traditional sense of an "interpolation", this approximation need not go exactly through the (possibly noisy) data points. The process of constructing the best possible interpolation for given data is known as **fitting** (or sometimes "training", more formally **regression**). Evaluating the interpolated model for some new point on the PES is orders of magnitude more efficient than using the quantum method directly; a speedup of thousands or millions is common. When these models are used in long, demanding simulations, the computer time saved in this way quickly justifies the initial computational cost of computing the training data and constructing the interpolant.

*Certain "exotic" quantum phenomena, such as level crossings, can occasionally disrupt this smoothness. But closed-shell molecules in their ground states typically do not encounter such phenomena in the regions of the potential energy surface where they are stable.

Many machine learning methods have been applied to fit the potential energy surfaces of atoms or molecules. The oldest is artificial neural networks (ANNs), which were used in an early application to fit the energies of various hydrocarbon and carbon-nitrogen compounds [67], and in a more recent incarnation to model liquid water [16, 18], ionic solids [69, 70], and properties of molecules [71]. Another method, the one that will be used for this work, is based on Gaussian processes [72, 73]. These can be interpreted either as a Bayesian estimate of the PES from available data or as a least-squares linear fit in the transformed space of similarity functions, or **kernels**, that relate two molecular geometries. This latter view relates this method to kernel ridge regression (KRR) with a radial basis, a third type of machine learning method that gives equivalent predictions to Gaussian processes under certain conditions [74]. KRR has likewise been applied to predicting the properties of small molecules [75, 76], materials, or both [77, 78]. Both methods are also related to neural networks: A neural network with one hidden layer and certain forms of the switching function converges to a Gaussian process as the number of hidden nodes is increased towards infinity [79].

The method of Gaussian processes applied to PESs is called Gaussian approximation potentials (GAP) [15]. It uses a Gaussian process formulated in such a way that total energies, energy gradients (forces and stresses), and most recently second derivatives (Hessians; see Section 5.2.1) can all be used in the fitting, with various weighting factors that control the relative importance of each type of input data. It has been used to fit potential energy surfaces for crystalline solids such as silicon and carbon [15], tungsten [80], iron [81], and boron [82]. It has also found success in less structured systems, such as water [17, 83] and amorphous systems [84–87], and even in fitting coarse-grained

models [88] and potentials of mean force [89] for higher-level representations of biomolecular systems. There is also considerable interest from other groups in general, transferable molecular potentials [75] and highly accurate modeling of liquid water [18]. Finally, recent progress has also been made in modeling multicomponent systems [87, 90], across different chemical compounds [71, 77] and even across different classes of materials [78], thus approaching the level of flexibility currently offered by full quantum methods.

The GAP method excels in describing condensed systems where complex, many-body interactions are dominant. Its flexibility allows it to absorb errors that would otherwise be made by insufficient description of the physics or truncation of analytical expansions. This flexibility can also be a weakness, however; interpolation with many-body descriptors tends to fail more spectacularly than smooth analytical forms in regions of configurational space where insufficient data is present (see e.g. Section 5.3). Furthermore, the computational expense of a many-body GAP model is still much higher than that of simple, analytical forcefields or even *ab initio* polarizable models, especially when the cost of generating the training database is considered (see Section 4.1.4). This cost generally limits it to simulating systems of at most a few thousands of atoms [87].

One of the most challenging tasks in fitting a potential energy surface for a new material is to find good descriptors or representations of the molecular geometry. For kernel-based learning methods (KRR and GAP), this task is more specifically one of finding a good kernel function that adequately measures the similarity of any two geometries. In order to be useful for fitting, this function must satisfy certain properties: In addition to having the same symmetries as the energy function itself (invariance to translation, rotation, and permutation of like atoms), it must also be able to adequately distinguish dissimilar geomet-

ries [91]. A more detailed discussion of available descriptors and kernel functions is given in Section 3.2.1.

Another significant challenge is to choose the sample of the potential energy surface so that it adequately covers the space of geometries that could be encountered during a simulation. A typical approach is to perform an MD simulation with a simpler potential and extract snapshots at regular intervals. However, care must still be taken to cover the relevant range of simulation conditions, as evidenced by the sampling procedure and evaluation covered in Section 4.1.3.

In summary, the goal of this work is to create a model, based on the GAP framework, that describes the potential energy surface of interacting alkane models with a high, adjustable degree of accuracy. Such a model would bridge the gap between expensive quantum-mechanical simulations and insufficiently accurate classical potentials, enabling simulations that can accurately predict, from first principles, the viscosity of an arbitrary mixture of alkanes under a wide variety of ambient conditions. This work presents important steps towards such a model, with the theoretical background to systematic potential development discussed in Chapter 3 and developed into a general, extensible framework, demonstrated with an application to build an accurate potential for liquid methane, presented in Chapter 4. Finally, Chapter 5 shows preliminary work toward an intramolecular potential for arbitrary-length alkanes, thus completing the potential and making it useful for realistic simulations of the viscosity of longer alkanes or even complex mixtures.

Chapter 2

Molecular simulation methods

Once the potential is specified, it remains to choose a method that optimally samples the thermodynamic ensemble (to compute static properties defined by means of Equation (1.1)) or allows computation of dynamic observables, such as the diffusivity or viscosity. This chapter explores these methods with the view of validating a new potential, that is, computing its predictions of the properties of a known material and comparing these predictions to experimental values. In contrast with the traditional paradigm of fitting an empirical potential directly to experimental data, the aim here is to create a potential from fundamental physical principles, then *measure* its success by how well it describes the real material.

The optimal strategy generally tends to be different for computing static versus dynamic observables, so these are often treated in separate simulations. Since static properties are arguably simpler to compute, they will be the primary focus of this research with the aim of later attempting the more complex and expensive dynamical properties.

2.1 Static properties

One of the most basic properties that a potential for liquids should reproduce accurately is the density as a function of pressure and temperature (the liquid's equation of state). This relationship can be measured directly, reliably, and accurately even for compressed liquids [92].

In principle, the density should also be simple to predict from simulation. If the system can be made to sample a constant-pressure, constant-temperature ensemble (called NpT after the thermodynamic variables that are held constant; N refers to the number of particles) for some given pressure and temperature, then the density at that state point can be straightforwardly determined from the simulation cell. It is the sampling of the proper thermodynamic ensemble that presents difficulties: Ordinary molecular dynamics, or the straightforward evolution of Newton's equations of motion, does not allow the exchange of energy and volume with the environment and therefore samples the NVE ensemble. In order to sample the NpT ensemble instead, the equations of motion must be modified to allow the system to exchange energy and volume with a simulated external bath so as to achieve the desired temperature and pressure.

2.1.1 Thermostats

Modifications that regulate a system's temperature and hence make it sample the NVT ensemble are known as thermostats. Many different thermostats have been developed, but they all commonly have parameters to control independently the temperature of the enviroment and the time the system will take to relax to this target temperature. These parameters affect the speed and reliability of equlibration, but not (within a reasonable range) the final equilibrium

properties.

Perhaps the most popular thermostat is the Nosé-Hoover thermostat, which adds an extra degree of freedom to the Hamiltonian that couples all the particles to a virtual thermal reservoir [93, 94]. This thermostat is effective at regulating the system's temperature without significantly perturbing its dynamics and the resulting transport properties. It is also deterministic, which can be a downside – the thermostat is non-ergodic for many systems [19, 95]. This means Nosé-Hoover simulations may get stuck in a small subset of state points and fail to explore the majority of relevant configurations, leading to wrong predictions of equilibrium properties. One possible solution to this problem, at the cost of additional complexity, is to add thermostats to the thermostat, creating a chain of dynamical variables [96].

The Langevin thermostat, on the other hand, does not have this problem. It simulates the effect of a heat bath by adding random perturbative as well as drag forces to each of the particles. For static observables such as the density, the Langevin thermostat delivers reliable results because it also converges a system to the canonical distribution [26]. But the random perturbations, which are key to this thermostat's ergodicity, can also interfere with the dynamical properties of the system. For this reason, a hybrid of the two thermostats has been proposed: the Nosé-Hoover-Langevin (NHL) thermostat uses the Nosé-Hoover extra dynamical variable, itself regulated by a Langevin thermostat [19], in order to preserve the system's dynamics while maintaining ergodic sampling. A more recent modification has emerged as an appealing alternative to Nosé-Hoover chains [97]; this is the thermostat used by default in the group's QUIP software [98].

The Langevin thermostat is, in fact, only one of a larger class of stochastic

thermostats. The standard Langevin thermostat essentially adds white noise to each of the system's momentum variables. It is possible, however, to use noise with a different frequency profile in order to better target the equilibration of the system's vibrational modes [99]; such a thermostat is called a coloured-noise or a generalized Langevin equation (GLE) thermostat [100]. While coloured noise formally makes the equations of motion non-Markovian (dependent on the system's history), the equations can be recast in a Markovian form with the help of additional, auxiliary dynamical variables.

2.1.2 Barostats

A system's pressure may be regulated in a similar manner as its temperature. The Nosé-Hoover approach may be adapted to create a barostat, where the external degree of freedom is now a pressure reservoir coupled to the box's volume. Some barostats can vary each of the box's cell parameters independently in order to allow for anisotropic stress tensors but for homogeneous liquids and gases the simple isotropic version suffices [20, 101]. As with thermostats, the barostat's target pressure and relaxation time can be independently adjusted.

Thermostats and barostats can be combined to generate the isothermal-isobaric (NpT) statistical ensemble. In particular, this is shown to work with a Langevin thermostat and Nosé-Hoover–style barostat [102], although the barostat could just as well be controlled with the Langevin or generalized Langevin equations. With a suitable thermostat and barostat in place, the density of a system can be predicted by running a simulation for a long enough time. Neither the thermostat nor the barostat can bring the system exactly to its target temperature or pressure, though, since temperature and pressure are macroscopic averages and a small system will always randomly oscillate around those val-

ues. The prediction is best obtained by taking some length of simulation time after the initial equilibration and taking the average of the density over that time. The necessary averaging time can be rigorously determined based on the desired precision [103]; in practice, however, it is just as often determined by experience and intuition.

2.2 Quantum nuclear effects

As previously mentioned, a complete treatment of molecular materials must also treat the nuclei as quantum particles. The Born-Oppenheimer approximation itself only separates the electronic and nuclear wavefunctions; the reduction of the nuclear wavefunction to its classical limit is an additional approximation, made in the vast majority of molecular simulation studies, but justified only in the limit of large (heavy) nuclei and high temperatures (large ratio of thermal to zero-point energy) [104]. Compressed liquid methane is generally regarded to be neither; especially the presence of hydrogen in the materials of interest generally indicates that the extent of quantum nuclear effects must be evaluated [105].

One way to estimate the influence of quantum nuclear effects is to use approximate or semiempirical corrections, especially those accounting for zero-point vibrational energy (ZPVE). This effect is analogous to the non-zero ground-state energy of a quantum harmonic oscillator, which is used to approximate a chemical bond near equilibrium. An example of this approach was used in Hellmann et al. [42]: The increase in C-H bond length, as well as the increase in the polarizability of the molecule as a whole [106, 107], were both explicitly accounted for in the forcefield parameterization; an additional dynamical quantum correction was applied to accurately compute the second virial coefficient. Such semiempirical corrections, while founded in rigorous physical arguments, are much less

suitable for condensed-phase systems and high pressures where the simplifying physical assumptions (rare-gas approximation) begin to break down and additional quantum effects beyond the ZPVE become important [108].

A more explicit and flexible approach to including quantum nuclear effects in MD simulations makes use of an isomorphism between a path-integral approximation to the quantum partition function and an extended classical system [104, 109]; this approach is known as path-integral molecular dynamics or **PIMD**. It can account for ZPVE and tunnelling effects, though the nuclear exchange is still neglected [104, 108]. The extended classical system takes the form of several replicas of the original system, corresponding atoms joined cyclically between the replicas in a ring-polymer structure, hence the more specific name of ring-polymer molecular dynamics (RPMD) [110]. Coupled with an appropriate stochastic thermostat [111], the dynamics converges to the quantum Boltzmann distribution as the number of replicas is increased, as well as enabling the accurate calculation of other properties such as vibrational spectra [112].

More recent techniques have drastically reduced the additional cost associated with PIMD, including ring polymer contraction [113] to deal with the slower-varying force components in a less expensive way, coloured-noise thermostats [105, 114] that mimic the effect of quantum fluctuations and reduce the number of replicas needed to converge to the quantum distribution, and techniques that make use of a higher-order expansion of the partition function [115] to improve the convergence with respect to the number of replicas. This work makes use of the coloured-noise thermostats in particular, as they offer the largest improvement in efficiency with minimal additional complexity.

2.3 Dynamical properties

The static properties of a system are straightforward to compute in comparision to dynamical properties such as diffusivity and viscosity. This type of property measures the macroscopically averaged rate of microscopically stochastic transport through the bulk: In the case of diffusivity, it is the rate of individual particle movement; in the case of viscosity, it is the rate of momentum transport. The measurement of these properties from a molecular dynamics simulation is grounded in the linear response theory of non-equilibrium thermodynamics: The fluctuation-dissipation theorem connects the small fluctuations that can be measured in a simulation to the macroscopic linear response of a system to a perturbation (in the case of diffusivity, a concentration gradient; in the case of viscosity, a velocity gradient) [26].

Although the potentials developed in this work have not yet been tested against dynamical properties, such tests will be part of the necessary course of validation of the potential for all the intended practical applications. Furthermore, the intended practical application of this potential is in predicting the viscosity, so it is useful to understand the physical basis and the methods of computation of these properties. Diffusivity simulations in methane have been conducted in the gas phase [116, 117] and in the liquid phase at a limited number of state points [51, 52], using a variety of simulation methodologies; the diffusivity of larger hydrocarbons has also been tested with a united-atom model [2]. Viscosity data for methane, on the other hand, is much more limited, probably due to the difficulty of measuring – experimentally or computationally – its extremely small viscosity.

2.3.1 Green-Kubo relations

The mathematical relations that allow the computation of transport properties from fluctuations are known as the Green-Kubo relations. They relate transport coefficients to integrals of time autocorrelation functions. For the viscosity, the coefficient to be measured is the stress (force per unit area) response to a velocity gradient; for flow in the x-direction with velocity varying in the y-direction, the linear relation is

$$\frac{F}{A} = \eta \frac{\partial v_x}{\partial y} \tag{2.1}$$

and the corresponding Green-Kubo relation is [26]

$$\eta = \frac{V}{k_B T} \int_0^\infty \mathrm{d}\tau \langle \sigma_{xy}(0) \sigma_{xy}(\tau) \rangle. \tag{2.2}$$

The off-diagonal elements of the stress tensor $\sigma_{\alpha\beta}$ can be measured by the virial tensor, $P_{\alpha\beta}$, which contains the volume averages of the corresponding elements of the stress tensor. The angle brackets strictly denote an ensemble average, though in a simulation they can be replaced by a time average by assuming ergodicity.

A similar relation exists for diffusivity:

$$D = \int_0^\infty \mathrm{d}\tau \langle v(0) v(\tau) \rangle \tag{2.3}$$

although in this case it is easier to transform the equation to use the time integrals of the velocities (i.e. the positions) instead of their time correlation functions [20, 26]. For each individual dimension:

$$\lim_{t \to \infty} \frac{1}{t} \langle (x(t) - x(0))^2 \rangle = 2D \tag{2.4}$$

Summing the vector displacements instead changes the right-hand side to $6D$, giving the canonical Einstein relation for the diffusion coefficient.

2.3.2 Practical simulation considerations

Since transport properties are sensitive to modifications of the dynamics through thermostats and barostats, the ideal approach would be to run a simulation without thermostats, i.e. run a microcanonical simulation, using the exact energy and volume that give the system the desired temperature and pressure [20, 116]. In practice, however, pressure is often extremely sensitive to both the volume and the internal energy, so a small statistical error in either variable can cause a large error in the actual simulation pressure.

For this reason, a common practice is to use a Nosé-Hoover thermostat with a large time constant to minimize interference in the dynamics while still maintaining the target temperature [3, 26]. Another option that maintains ergodic dynamics is to use a "gentle" thermostat, such as the above-mentioned NHL [19] or its subsequent "adaptive" modification [97]. Finally, a recent development has made it possible to extract dynamical information even from a simulation run using a strong stochastic coloured-noise thermostat; the effect of the GLE on the dynamics can be computed and inverted in order to obtain the dynamical properties of the unperturbed system [118].

An additional difficulty lies in computing the autocorrelation functions of Equation 2.2 from a simulation time series: The tail of the computed function is subject to increasing noise at longer (correlation) times, where the number of samples available to compute the estimate becomes smaller [3, 103]. In practice, this means accurate estimates of the autocorrelation integral require either discarding the autocorrelation beyond some point [103] or fitting the tail with a

decaying function based on some known functional form [119].

Finally, the statistical sampling error in the results is commonly reduced by averaging properties over multiple independent simulations [3, 19]. This procedure not only provides an error estimate for the computed property; it also minimizes the risk of the simulation being biased by spending too much time trapped in a small number of metastable states.

2.3.3 Alternative methods

Transport coefficients can also be calculated through non-equilibrium methods, which involve directly perturbing the simulation, driving it out of equilibrium into some other steady state, and observing the response [26]. Such simulations may be useful for systems, such as very long entangled hydrocarbon chains, where the transport processes are too slow to be obtained reliably from an equilibrium simulation. The perturbations force the transport processes to occur on a faster, more easily measurable timescale. This approach is used in e.g. [27] to model the viscosity of a variety of sizes of alkanes. While this approach has the downside that the measured viscosity is dependent on shear rate and must be extrapolated, it may be more reliable than equilibrium methods, which are especially affected by inaccuracies in the forces from a potential [66].

Another approach that works best in the limit of dilute gases is to use kinetic theory and generalized cross-sections from integrals of classical trajectories [117, 120]. For liquid-phase simulations, however, the more pragmatic approach of calculating the properties from equilibrium or non-equilibrium MD simulations is generally easier and more accurate [116].

Chapter 3

Intermolecular potential development

In order to create a potential for accurate viscosity simulations, we follow the philosophy that the most accurate potential is a systematic approximation of the entire Born-Oppenheimer potential energy surface, faithfully representing all the physical effects that contribute to the liquid's behaviour. Such a potential will give robust predictions of properties across a wide range of temperatures and pressures, since by following the detail of the surface we obtain the properties the same way as they are obtained in nature at any temperature and pressure. Note that such a potential must *not* include features that are not part of the Born-Oppenheimer potential energy surface, such as quantum nuclear effects. Our goal is first to approximate the true PES and only after to obtain additional effects arising from the dynamics in a systematic and consistent way (with path-integral MD).

In addition to delivering reliable predictions across different temperatures and pressures, the potential should also work without modification across a wide range of chemical space: In this case, across different alkane lengths and, even-

tually, branching configurations. Our approach begins with the common and successful strategy of decomposing the potential energy surface into the contributions from different physical effects. We then find out which of these effects can be represented in a way that is independent of the size and type of the molecule, and for the remaining effects, find a consistent and automatic way of capturing the variation.

This chapter begins by applying this strategy to the simplest liquid hydrocarbon system, liquid methane. We are interested in the condensed phase because we want to explore the same regions of the potential and qualitative types of molecular behaviour as the eventual full-scale viscosity simulations will encounter. The relevant physical interactions are first identified, parametrized, and, where necessary, fit with a fully flexible machine learning potential. The machine learning method is then discussed, along with ways of measuring the accuracy of the resulting potential.

3.1 Measuring accuracy

The first question when designing a systematic potential should be what level of accuracy is required. That is a difficult question, since usually we only know what accuracy we require in the predicted *properties*. The corresponcence between inaccuracies in the potental energy surface and inaccuracies in the properties predicted by simulations is still poorly understood. This gap in our understanding has gradually become more pressing with the technology to systematically fit potential energy surfaces. Early studies in this direction include the fitting of PESs of van der Waals complexes from spectroscopic data [53, 121, 122], since different types of spectra (microwave, infrared, and far-infrared) give information on different regions of the PES. Now that it is possible to fit the true physical in-

teractions (e.g. intramolecular, short-range repulsion, or long-range dispersion) with any accuracy we desire, we should ask how much accuracy we truly require for the intended application and where we require it most. We start with a preliminary investigation on a prototype problem, the united-atom model of methane where each molecule is represented by a single spherical bead, before moving on to a more detailed error measure that can be used with systematically fitted potentials for the full, all-atom system.

3.1.1 Perturbation study

The united-atom model used in this study is TraPPE-UA [9]. An explicit-hydrogen version does exist [35], albeit with scaled coordinates and rigid methane molecules. We will see later that both are relatively poor approximations of the potential energy surface itself. But the predictions these models give for the density, especially the UA version, are exceptionally accurate: Figure 3.1 shows the density predicted for compressed liquid methane at temperatures of 110 K and 188 K and pressures ranging from 0 bar to 400 bar, compared with the experimental values from [123]. The simulations were done using the LAMMPS [124] molecular dynamics package; simulation details are given in Section 4.2.

The TraPPE-UA potential was perturbed with cosine step functions:

$$\delta(r) = \begin{cases} A & r \le r_{\text{in}} \\ A\cos\left(\frac{\pi}{2}\frac{(r-r_{\text{in}})}{w}\right) & r_{\text{in}} < r \le r_{\text{in}} + w \\ 0 & r > r_{\text{in}} + w \end{cases} \tag{3.1}$$

of amplitude A and width w, starting at inner radius r_{in}. These perturbations only have effect (i.e. generate a force) over the region of width w with continuous

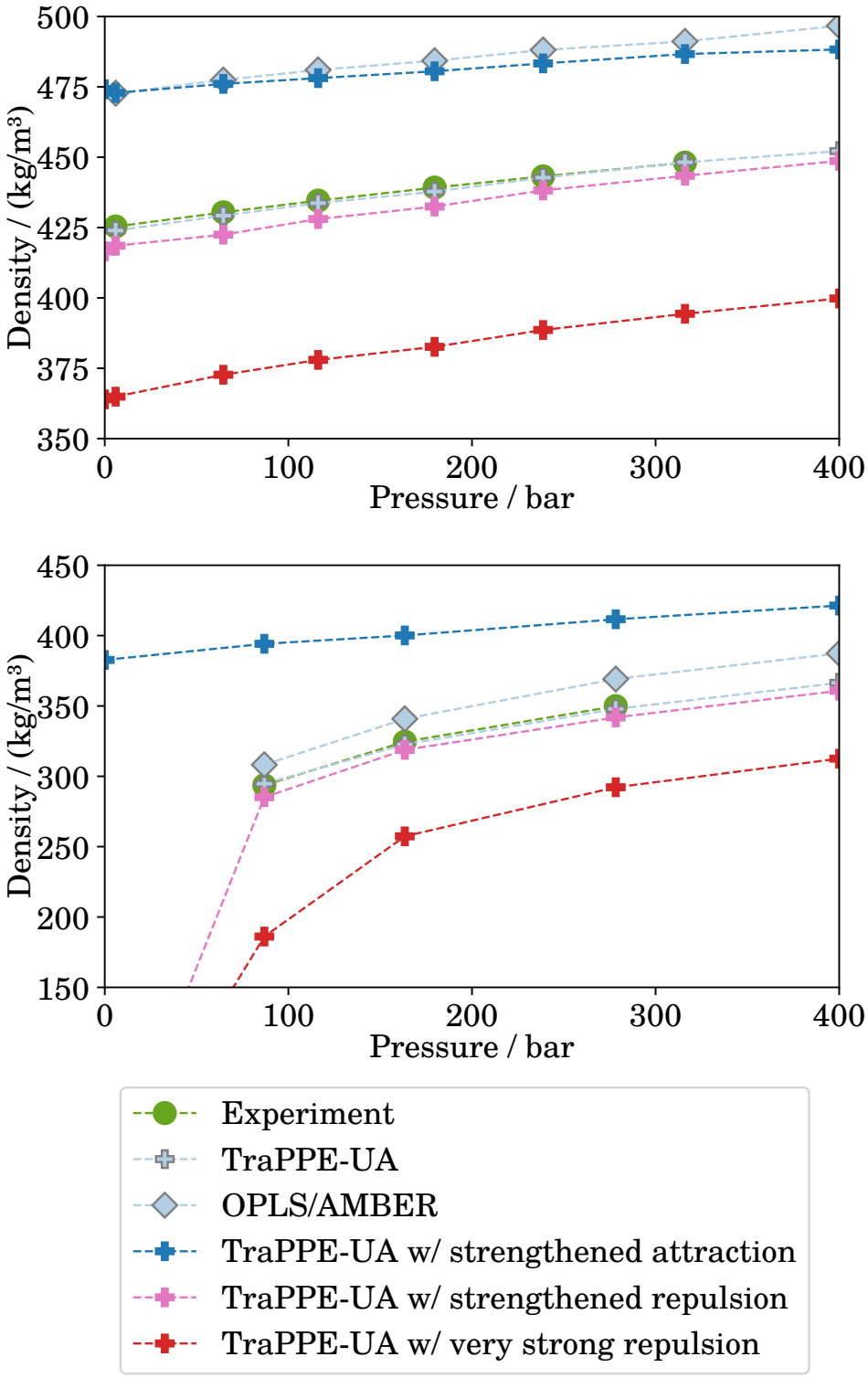


Figure 3.1: Density of liquid methane predicted by perturbed versions of TraPPE-UA at 110 K *and at* 188 K*; OPLS/AMBER is shown for comparison. Uncertainties in the density are smaller than the symbol sizes.*

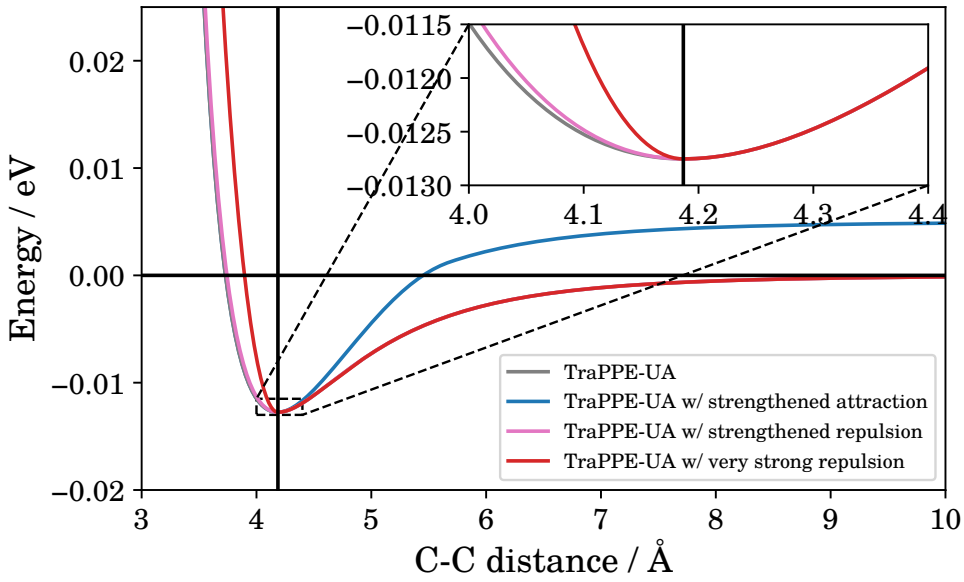


Figure 3.2: The TraPPE-UA potential for the methane dimer with the three perturbations used in this study

first derivatives at the endpoints. In this case, the perturbations were confined to either the attractive or the repulsive region of the potential. All of the perturbations used $w = 1.5\,\text{Å}$, although for a more realistic interpretation of the results these length scales should have been chosen to correspond with the length scales of the interactions being modified. One of the perturbations was confined to the attractive region (r_{in} corresponded to the minimum of the TraPPE potential) and used $A = -0.005\,\text{eV}$, thus strengthening the attraction; the other two were confined to the repulsive region ($r_{\text{in}} + w$ corresponded to the TraPPE minimum) and used $A = 0.005\,\text{eV}$ (extra repulsion) and $A = 0.1\,\text{eV}$ (much stronger repulsion). These modified (perturbed) versions of the potential are pictured in Figure 3.2. Simulation parameters for the perturbed potentials were similar, but used only 100 ps of simulation time instead of 500 ps.

First, the consistent accuracy in the density predictions of plain TraPPE-UA illustrates that a model need not be a detailed, systematic approximation to

the Born-Oppenheimer potential energy surface in order to make robust predictions of thermodynamic properties, even across a wide range of temperatures and pressures. TraPPE was fit to reproduce phase equilibria as accurately as possible [9]; evidently, fitting just two parameters is enough to reproduce the experimental equation of state of methane in this range of thermodynamic parameters. However, as the potential is put through more stringent tests, eventually the shortcomings of the approximation show through. Structural properties of alkane films, for example, cannot be accurately modelled with a united-atom potential [125, 126]. The discrepancies in dynamic properties are even more troubling: It was shown in Payal et al. [3] that united-atom models consistently underpredict the viscosity of linear alkanes. This effect can be intuitively understood, since united-atom models of molecules are much smoother than the real systems with atomistic detail and thus cannot adequately model the friction between adjacent layers of a fluid. A similar result was found in Allen and Rowley [27] where united-atom models were found to underpredict experimental viscosities, especially at high pressures and densities. An explicit-hydrogen version of TraPPE [35] was developed in part in response to a related shortcoming (united-atom models predict too large diffusivities for the same reason as just mentioned). Another approach is to use so-called asymmetric or anisotropic united-atom models where the "beads" are not spherical [2]. Such approaches introduce additional complexity, though they still fall far behind all-atom models in approximating the true potential energy surface of the methane dimer (see Section 3.1.2).

The perturbed versions gave density isotherms, shown in Figure 3.1, that were shifted from the original TraPPE isotherms, with extra attraction producing higher densities and extra repulsion producing lower densities. As expec-

ted, increasing the attraction by 0.005 eV also increased the predicted density, in this case by about 12 %, about the same as the OPLS density overprediction at 110 K. Increasing the repulsion by the same *absolute* amount of 0.005 eV had much less of an effect; only with a perturbation 20 times that magnitude does the change in density become comparable to the change due to the increase in attraction. Note that this difference can equally be attributed to the difference in length scales of the repulsion and dispersion energy components, though the interpretation of this difference is far from straightforward: For the same energy change (measured from the minimum) that the potential experiences over 1.5 Å of the dispersive region (outwards from the minimum), only about a third of that distance over the repulsive region (inwards from the minimum) is required; the repulsion energy 1.5 Å inwards from the minumum is about 250 times the well depth. In fact, the inverse-power form of the repulsion has no intrinsic length or energy scale in the same way that the attraction does. It may be illuminating to repeat this study, varying the perturbation length scale as well as the energy scale, in order to better assess the effects of both parameters on the density prediction.

Furthermore, as Figure 3.3 shows, the perturbation of 0.1 eV is of the same scale as the absolute errors OPLS-AA makes in the short repulsive range, up to the approximate minimum dimer separation of 3.25 Å. It is therefore tempting to assign the density error made by OPLS-AA entirely to the repulsive range. The main problem with such a conclusion – besides the different predictions at 188 K – is that the actual OPLS/AMBER density error is in the wrong direction: OPLS-AA appears to be *more* repulsive than the CCSD(T) reference in the short range, and yet it gives higher densities than experiment. It therefore seems that already the error in the density, even of such a simple model, cannot be under-

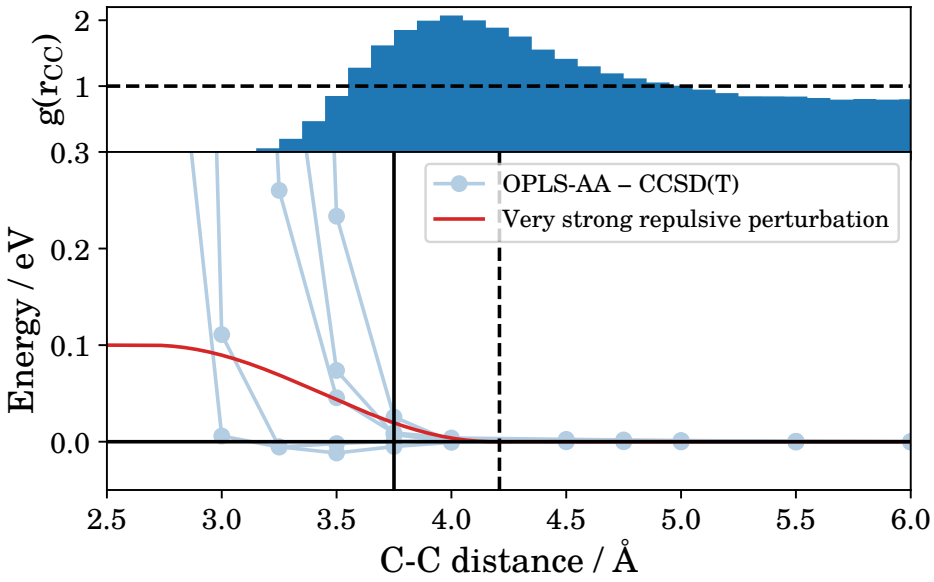


Figure 3.3: Differences of the OPLS-AA energies from CCSD(T)-F12 on a set of equally-spaced methane dimers in various orientations, compared to the "very strong" (0.1 eV) repulsive perturbation. The dashed vertical line indicates the position of the minimum of TraPPE; the solid line shows where TraPPE crosses zero. A radial distribution function is shown for reference; the distribution was taken from an OPLS/AMBER simulation at the highest temperature and pressure in the experimental density dataset, 188 K and 278 bar. Dimer orientations are the same as those in Figure 4.3; see Section 4.2 for computational details.

stood with this simple, one-dimensional analysis. Note also that the simulation results in Figure 3.1 do not include quantum nuclear effects. The TraPPE-UA model includes these implicitly thanks to their parameterization to experiment. The OPLS-AA model was also parameterized against experimental data so, in principle, it should also include these effects. On the other hand, as we will soon see, it is one of the most accurate traditional analytical potentials when compared against a quantum chemical reference.

These results do not yet help us systematically explain or attribute the errors in the density made by models more complex than the one-dimensional united-atom representation. They do provide a rough guide for what order of magnitude of accuracy to target, namely, less than 0.005 eV in the attractive region. How-

ever, the interpretation of the repulsive region is muddled by the problem of the length scale, so for now we should simply aim to keep the same *relative* error throughout the potential.

3.1.2 Dimer error measure

Measuring the accuracy of a model in approximating the underlying Born-Oppenheimer PES is likewise difficult. The question is ill-defined, in part because of the question addressed in the previous section: It is difficult to say what *part* of the surface matters most. Furthermore, the PES formally has an intractable number of dimensions: $d = 3N-6$, where N is the number of atoms in the system. This makes visualizing the differences (errors) between the true and approximate PES difficult, if not impossible, to do directly. Here we explore one way of approaching this problem, first by studying only the simplest possible system of interacting methane molecules, the dimer ($d = 24$ dimensions, or $d = 6$ if the molecules are taken to be rigid). It is much more tractable to generate samples in this space and compute average errors (root-mean-squared error, or **RMSE**) or do fits against a quantum chemical reference. In fact, this system has seen several potentials fitted to *ab initio* data before [42, 43, 51, 52, 120], but these mostly use either simple traditional or specialized analytical functional forms. Eventually we will use a fully flexible fitting method without constraint to any one functional form in order to obtain the best possible accuracy; this model is called the "6-D dimer GAP" and its development is discussed in Section 4.2.1. We can visualize the errors of each of these models along the one-dimensional coordinate of dimer separation, treating all further coordinates (describing the orientation and, if considered, flexibility of the molecules) as secondary.

Several samples of methane dimer geometries were created for this purpose,

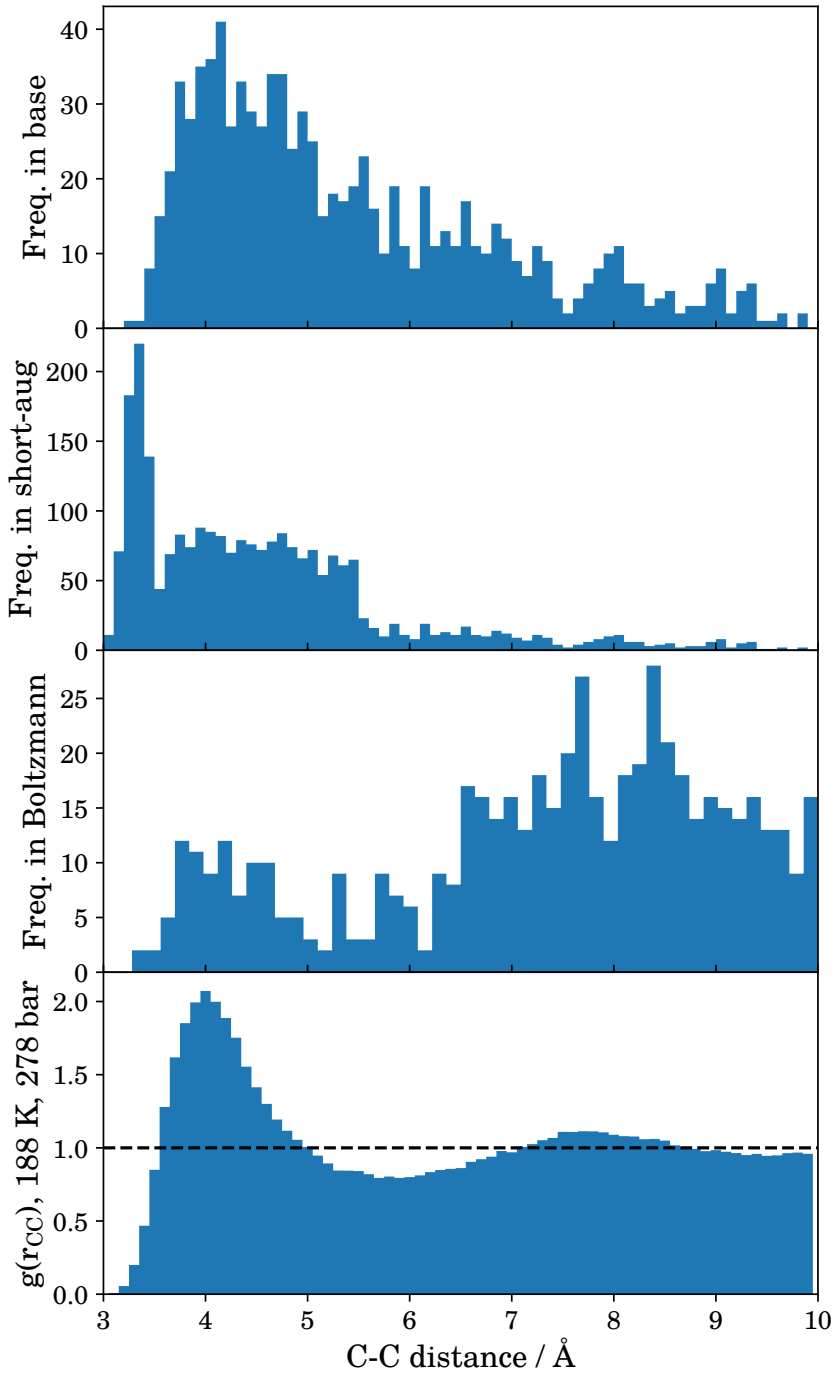


Figure 3.4: Distributions of methane dimer distances in the two dimer samples, base on top, short-augmented and unweighted (Boltzmann) below. The three samples contained 901, 2420, and 565 dimers, including geometries where quantum calculations did not converge. The bottom plot is the C-C radial pair distribution function from an OPLS/AMBER simulation at 188 K *and* 278 bar*, the highest temperature and pressure in the experimental methane density dataset.*

all taken from bulk simulations. The first two were taken from an OPLS-AA simulation at 188 K and 400 bar with the monomer geometries kept rigid, constrained to the geometry optimized with CCSD(T) (using the composite procedure described in the next section). The base sample contained 901 dimers with separations ranging from 3.0 Å to 10.0 Å. Additional samples were taken at close range to improve the accuracy in describing the repulsion, first in the range 3.0 Å to 5.5 Å and then in the range 3.0 Å to 3.5 Å; together with the base sample, this resulted in 2420 geometries heavily biased towards the short range (hence "short-augmented").

Finally, one more sample was taken from a simulation using the 6-D dimer GAP at 110 K and 316 bar, described in Section 4.2.1, this time with no weighting applied other than rejecting dimers outside the cutoff of 10 Å. This way, the sample is naturally taken from the Boltzmann distribution generated by the methane dimer potential at that temperature and pressure, hence its designation as the "Boltzmann sample". The distributions of dimer separations in these three samples – base, short-augmented, and Boltzmann – are shown in Figure 3.4.

3.1.3 Reference methods

In order to quantify the accuracy of any given approximate method, a good reference method is needed that gives as accurate a solution as practically possible to the Schrödinger equation of the dimer. Since the methane dimer is such a small system, it is possible to use a high level of quantum chemistry as a reference; here, we used coupled cluster CCSD(T) with explicitly correlated basis functions (the F12 correction [127]). The detailed procedure used a composite method, designed to control a notorious source of error in high-level quantum

chemistry calculations: the basis set incompleteness error. Since it would not have been feasible to do coupled-cluster calculations at a sufficiently large basis set, the energy is instead built up by first performing a Hartree-Fock (HF) calculation at the largest basis set possible (in this case Dunning's aug-cc-pV5Z, hereafter called AV5Z [128, 129]; the names AVQZ and AVTZ likewise refer to the aug-cc-pVQZ and aug-cc-pVTZ basis sets), then computing the MP2 correlation energy at the next smallest basis set, and finally computing the remaining coupled-cluster correlation energy at the next smallest set; this is similar to the procedure used in Gillan et al. [83]. The resulting energy expression is:

$$\begin{aligned} E_{\text{tot}} = {} & E_{\text{HF/AV5Z}} + \\ & (E_{\text{MP2/AVQZ}} - E_{\text{HF/AVQZ}}) + \\ & (E_{\text{CCSD(T)-F12/AVTZ}} - E_{\text{MP2/AVTZ}}) \end{aligned} \tag{3.2}$$

The highest level of theory in the calculations is explicity correlated coupled cluster (singles, doubles, and perturbative triples), or CCSD(T)-F12 [127, 130, 131] (hereafter referenced as just coupled cluster or CCSD(T), with the understanding that all CCSD(T) calculations done on these dimer sets were done with the F12 explicit correlation included). The dimer interaction energies were computed using the counterpoise method [132] to correct for basis set superposition error (BSSE). Using this method, the highest level of basis-set correction ($E_{\text{HF/AV5Z}} - E_{\text{HF/AVQZ}}$) was found to give a negligible improvement in the dimer *interaction* energy, although it was usually included anyway. All calculations were done using the program suite MOLPRO [133–136]. The program could compute analytical energy gradients (forces) up to the MP2 level, but not for CCSD(T) or higher.

The magnitudes of the different contributions to the energy across the sample

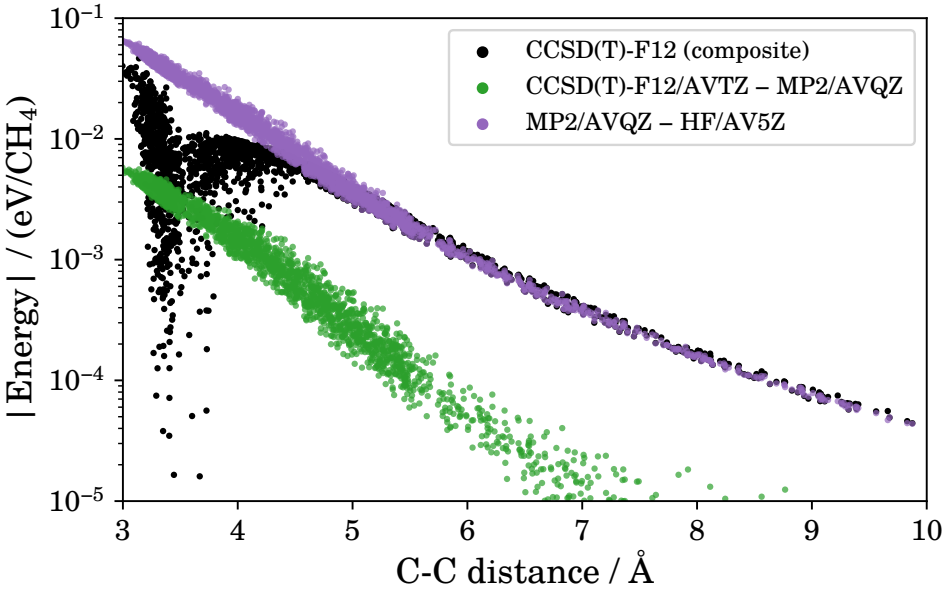


Figure 3.5: Components of the full coupled-cluster energies of the methane dimer sample shown as a function of dimer separation; the absolute values of all energies and energy differences are plotted on a log scale.

are shown in Figure 3.5. The components in the plot can be interpreted as the total energy, the MP2 *correlation energy* (measured from HF), and the CCSD(T) correction on top of MP2. Just from this plot we can see that the energies have signifiant anisotropy, especially in the close repulsive region of the potential. We also see that most of that anisotropy is coming from the Hartree-Fock level, where electron exchange is described exactly owing to the antisymmetry of the HF wavefunction [56]. This result suggests that most of the effort should be focused on describing the repulsion energy in the short to medium-short range (closer than about 6 Å C-C distance) as this is the most complicated and irregular part of the potential. Finally, the plot would also seem to suggest that the largest correction from HF is the MP2 correlation energy, which closely describes the long-range tail of the energy; the coupled-cluster correction has a comparatively small effect and can be treated separately.

Analytical model comparision

We can now use this reference to assess how well several popular or relevant analytical potentials for methane reproduce the quantum mechanical potential energy surface. The potentials chosen here are: OPLS-AA [34], TraPPE-UA [9], and TraPPE-EH [35], which were all fit to reproduce experimental properties of methane; the Li-Chao L-J fit [52], a simple pairwise Lennard-Jones fit to calculated quantum chemical energies of the methane dimer in a fixed set of orientations and shown to be successful in predicting the radial distribution functions and the self-diffusion coefficient of liquid methane; a new L-J fit to quantum chemical energies (CCSD(T)-F12) on a different set of orientations and used as the baseline for the dimer GAP (see Section 4.2.1); and the 6-D dimer GAP itself. The energies that each of these models predicts on the short-augmented set, along with the log errors against the CCSD(T)-F12 reference, are shown in Figure 3.6.

Evidently, most potentials greatly overpredict the dimer energies in the very short range (3 Å to 3.5 Å), especially TraPPE-EH, which generally follows the TraPPE-UA curve but with more scatter. The Li-Chao L-J, on the other hand, has low energies but very large scatter at medium range, much more than the reference energies. The L-J baseline also overpredicts, but with much less scatter (partly due to the use of purely repulsive C-H and H-H potentials). Finally, OPLS-AA seems to be the most accurate of the potentials tested here, with consistent accuracy across the range plotted here.

We can compute a single error number for each model by taking the root-mean-square error (RMSE) across the dimer geometries in the dataset. However, this measure still weights the error according to the distribution of dimer geometries present in the dataset; the short-augmented dimer set will weight errors

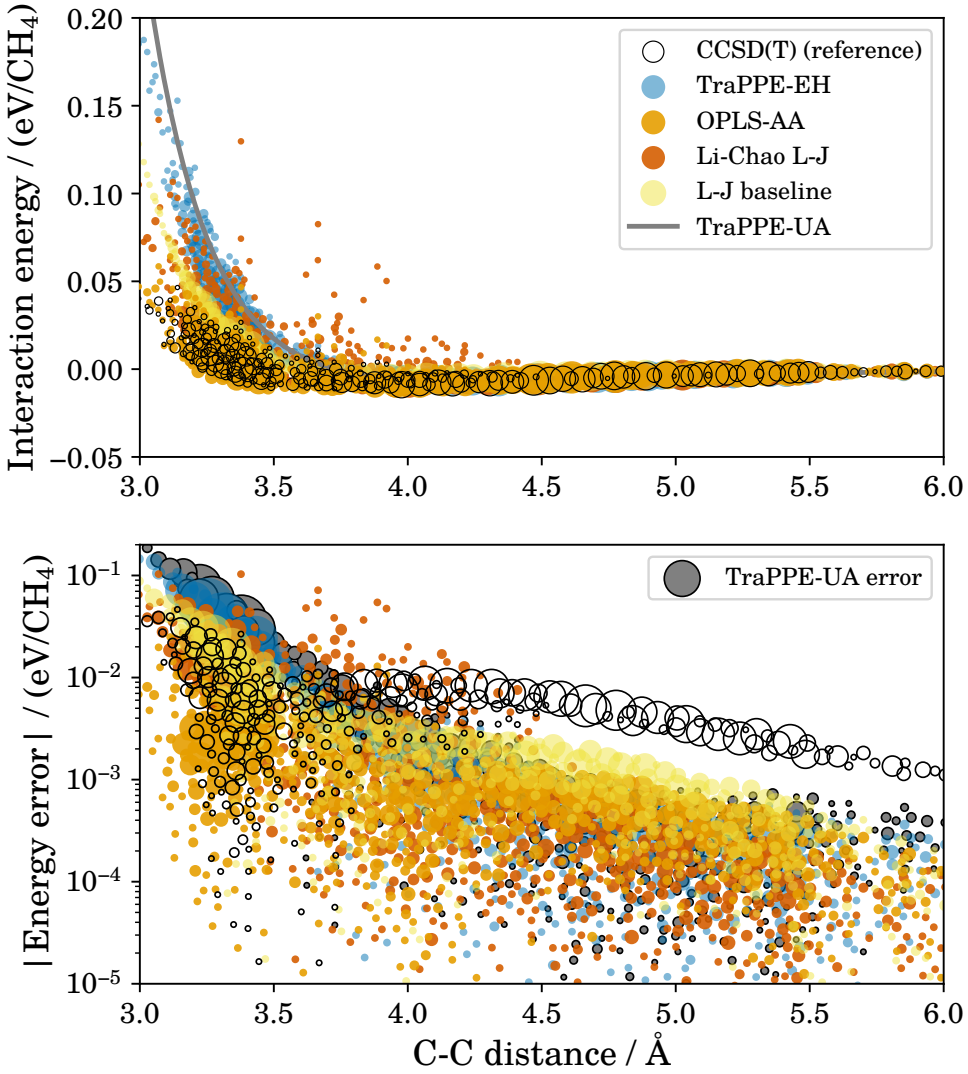


Figure 3.6: Dimer interaction energies predicted by various analytical (L-J) potentials, with CCSD(T)-F12 as a reference. Top: energies. Bottom: energy errors; the CCSD(T) energy is given for scale, everything else on the bottom plot is the error *against CCSD(T). Even though TraPPE-UA is an isotropic model, its error still depends on the dimer orientation because the error reference is the real, anisotropic CCSD(T) energy. Due to the large number of dimer geometries in the short-augmented sample, only a subset of representative points is shown with sizes proportional to the number of points they represent.*

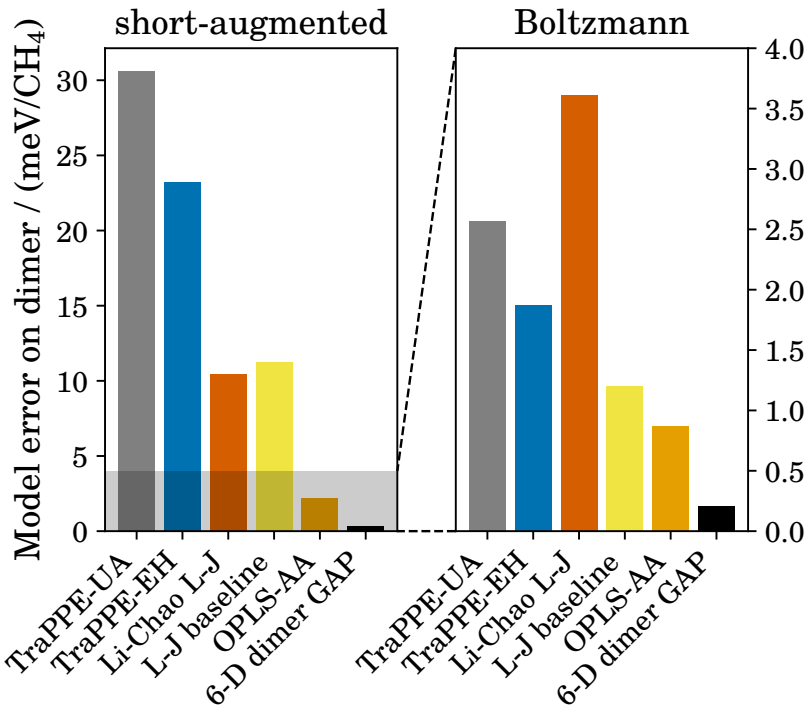


Figure 3.7: RMS errors of the dimer interaction energies predicted by the analytical (L-J) models in Figure 3.6, with the 6-D dimer GAP for comparison. Left: errors computed on the short-augmented dimer set; right: errors computed on the Boltzmann dimer set.

in the short range much more severely than the Boltzmann dimer set. Figure 3.7 shows the RMS error measures for these potentials across both dimer sets.

These error numbers confirm and quantify our conclusions from the previous figure: In the short-augmented dimer set, which heavily emphasises errors in the short range of the potential, the errors are ordered roughly as expected: The TraPPE-UA model, which ignores all anisotropy and overpredicts the repulsion energies, fares worst, with TraPPE-EH close behind; the Li-Chao and the L-J baseline fit are comparable in error, both having been fit to coupled-cluster energies on a small selection of dimer orientations; and OPLS-AA is the most accurate of the L-J potentials here – perhaps surprisingly, since it was never fit with the quantum potential energy surface of the dimer in mind. The errors on the Boltzmann sample follow approximately the same trend, though on

a smaller scale; removing the emphasis on the short range (where most potentials make large absolute errors) results in smaller error numbers. The unexpectedly large error bar of the Li-Chao L-J is likely due to its significant scatter in the medium range. Finally, the 6-D dimer GAP fit (see Section 4.2.1) is able to achieve the highest accuracy by fitting in the full six-dimensional space of dimer orientations. Its error number is comparatively consistent across the two dimer sets; on the short-augmented set (which it was fit on) it achieves 0.381 meV per methane molecule, while on the Boltzmann set that error drops to 0.204 meV/CH_4. For reference, the standard deviation of the CCSD(T) energies in the short-augmented set is 7.7 meV/CH_4 and the standard deviation on the Boltzmann set is 2.5 meV/CH_4; the depth of the dimer potential well is approximately 10 meV/CH_4.

DFT methods

While quantum chemistry is a good reference method for small systems such as the dimer, it quickly becomes intractable due to the typical N^7 scaling with the number of atoms [56] (though the less commonly used localized methods can reduce this scaling [137]). For larger systems, such as the bulk simulation cells that will later be used to train most of the GAP models, density functional theory (DFT) [138, 139] is the method of choice [57]. Two popular functionals were tested for use on this system, the generalised gradient approximation (GGA) functional PBE [140] and the hybrid GGA functional PBE0 [141]. Hybrid functionals mix in a proportion (25 % for this functional) of exact (Hartree-Fock) exchange, which may improve the description of energies in the short repulsive range.

Dispersion models

Plain density functional theory is not enough to arrive at a good approximation of the dimer interaction since, as discussed in Section 1.2.3, it neglects long-range dispersion. Here we will consider two models to correct for dispersion, both founded in physical arguments with parameters derived directly from a corresponding DFT parameters and employing a minimum of empirical parameters. The models are the pairwise correction of Tkatchenko and Scheffler [60] and its more accurate, but more complex and computationally demanding successor, MBD [62]. Both rely on some measure of atomic polarizability *relative to the free atom*.

In the pairwise T-S method, the pairwise dispersion coefficients are calculated starting from free-atom coefficients, which are calculated from first principles using dynamic polarizabilities [142]. The theoretical basis of these calculations is the Casimir-Polder integral, which expresses the coupling of fluctuations of dipoles centred on each atom; this is the lowest order of the multipole expansion commonly used to approximate the full dispersion energy [58]. The dispersion coefficients of the atoms in the molecules are then obtained through scaling by the ratios of atomic volumes between the atom in the molecule and the free atom, an idea introduced in practice by Becke and Johnson in 2006 [143] – though the underlying linear relationship (correlation) between the static polarizability of an atom in a molecule and its effective volume was known as much as 15 years earlier [144]. The effective atomic volume is computed by one of various partitioning schemes that assigns some portion of the total electron density, at any point in space, to each atom. Most of these schemes were first developed to assign partial atomic charges (along with higher multipole moments) to atoms in molecules. In this respect they are part of a large family of methods, many

of them aimed at modelling the electrostatic energy present in charged systems or systems with large charge separation. Some of these methods are reviewed in Veit [46], though for alkanes – and for methane in particular – we are neglecting the electrostatic contribution, as explained in Sections 1.2.2 and 4.1.3, and will not explore these methods further.

The pairwise T-S method, as originally proposed, uses the Hirshfeld partitioning [145] to compute relative atomic volumes. The Hirshfeld method assigns the density at each point according to the proportion of density that would come from each atom in an imaginary, non-interacting version of the molecule. This partitioning has further been re-derived using an argument from information theory: The Hirshfeld partitioning is the one that minimizes the information loss (Kullback-Leibler divergence) between the molecular and free-atom electron densities [146, 147]. A more recent, iterative extension of the Hirshfeld partitioning was proposed [148] in order to give a better description of systems with large charge separations and to strengthen the connection to the information-theoretic definition. This "iterative Hirshfeld" (HI) extension replaces the non-interacting reference system of *neutral* atoms with a system of *partially charged* atoms. In the first step, the charges are taken to be neutral, giving the regular Hirshfeld partitioning of the system. The Hirshfeld charges obtained in this step are then assigned back to the corresponding atoms of the reference system, the partitioning is recomputed with the new reference system, and the procedure is iterated to self-consistency. The density of partially charged atoms is computed by interpolation between the densities of the two closest states of integer charge.

The many-body dispersion method of Tkatchenko et al. [62] is based on the same ideas as pairwise T-S, with two key modifications that allow it to describe many-body effects in the dispersion interaction. The first modification

is to the polarizabilities of the fluctuating dipoles used to model the pairwise dispersion energy: Each dipole's polarizability is modified by the long-range electrostatic screening of its environment. This screening is solved via a classical self-consistent equation, giving rise to the TS-vdW+SCS (pairwise dispersion with self-consistent screening) model. The second modification replaces the sum of pairwise interactions between the (screened) dipoles with a true many-body model, the coupled fluctuating dipole model (CFDM) Hamiltonian. The Schrödinger equation for this Hamiltonian can be solved exactly, giving the full many-body interaction energy of the dipoles, which is the final MBD energy. Although this treatment includes dipolar interactions to arbitrary body order, it neglects the interaction of higher multipolar fluctuations, most notably the C^8 dipole-quadrupole term that D3 [61] does include. Nevertheless, MBD remains perhaps the most accurate spatially discretized (atom-centred) dispersion correction available; it predicts the binding energies of the molecules in the S22 database [149] with a mean absolute relative error of less than 5 %.

The performance on the methane dimer of the selected dispersion-corrected DFT methods is shown in Figure 3.8; both the T-S and the MBD corrections were computed* with regular (non-iterative) Hirshfeld volumes at first. A comparison of the dispersion-corrected methods using Hirshfeld vs. iterative Hirshfeld partitionings is shown in Figure 3.9. The DFT energies were computed using the Psi4 package [150], both using the AVQZ basis set. The Hirshfeld (regular and iterative) partitionings were done with the Horton software [151], itself using methods derived by Becke and Dickson for polyatomic systems [152–154]. While the iterative Hirshfeld partitioning did result in a 41 % larger average volume ratio for carbon and 19.5 % smaller average volume ratio for hydrogen than with

*The MBD code available from `http://www.fhi-berlin.mpg.de/~tkatchen/MBD/` was used here; the T-S correction was computed with my own implementation

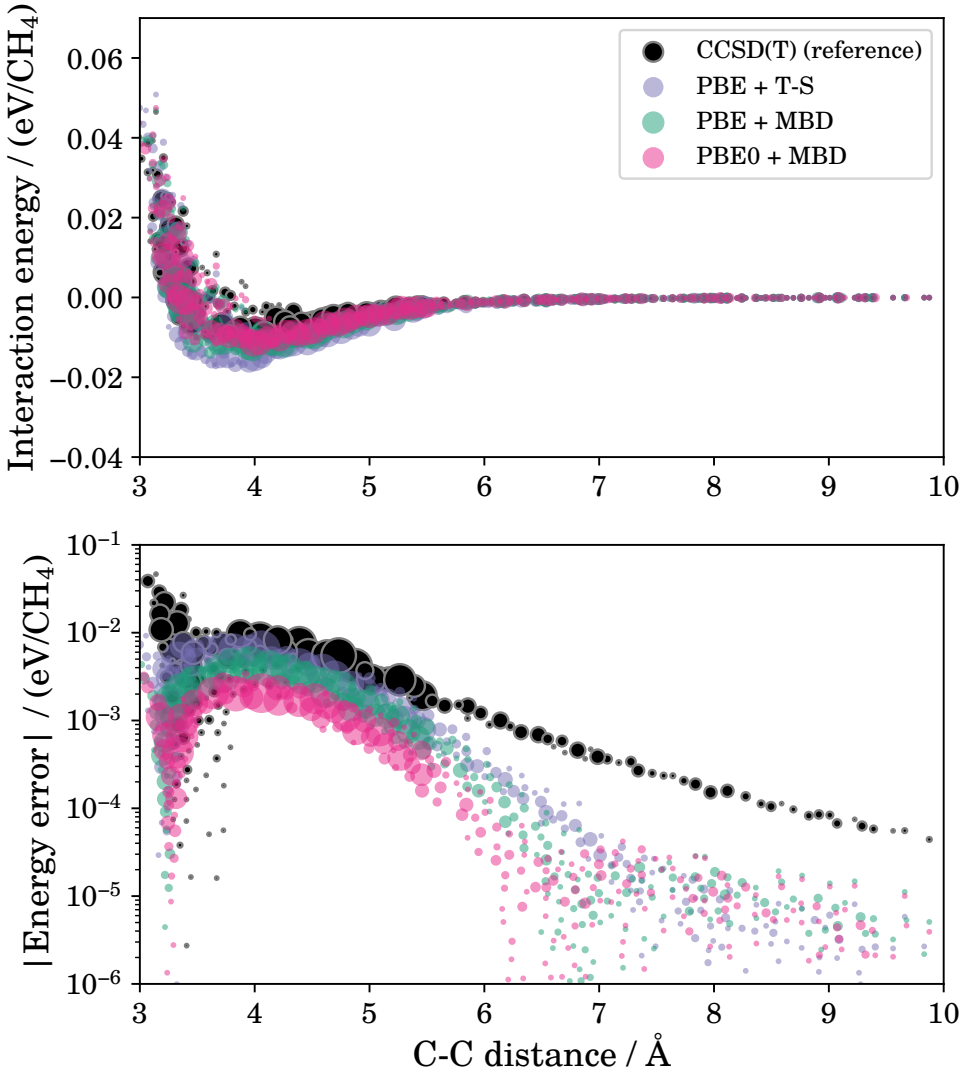


Figure 3.8: Dimer interaction energies predicted by the three dispersion-corrected DFT methods, with CCSD(T)-F12 as a reference. Energies on top, log errors against CCSD(T) (with the CCSD(T) energies themselves for scale) on the bottom. Due to the large number of dimer geometries in the short-augmented sample, only a subset of representative points is shown with sizes proportional to the number of points they represent.

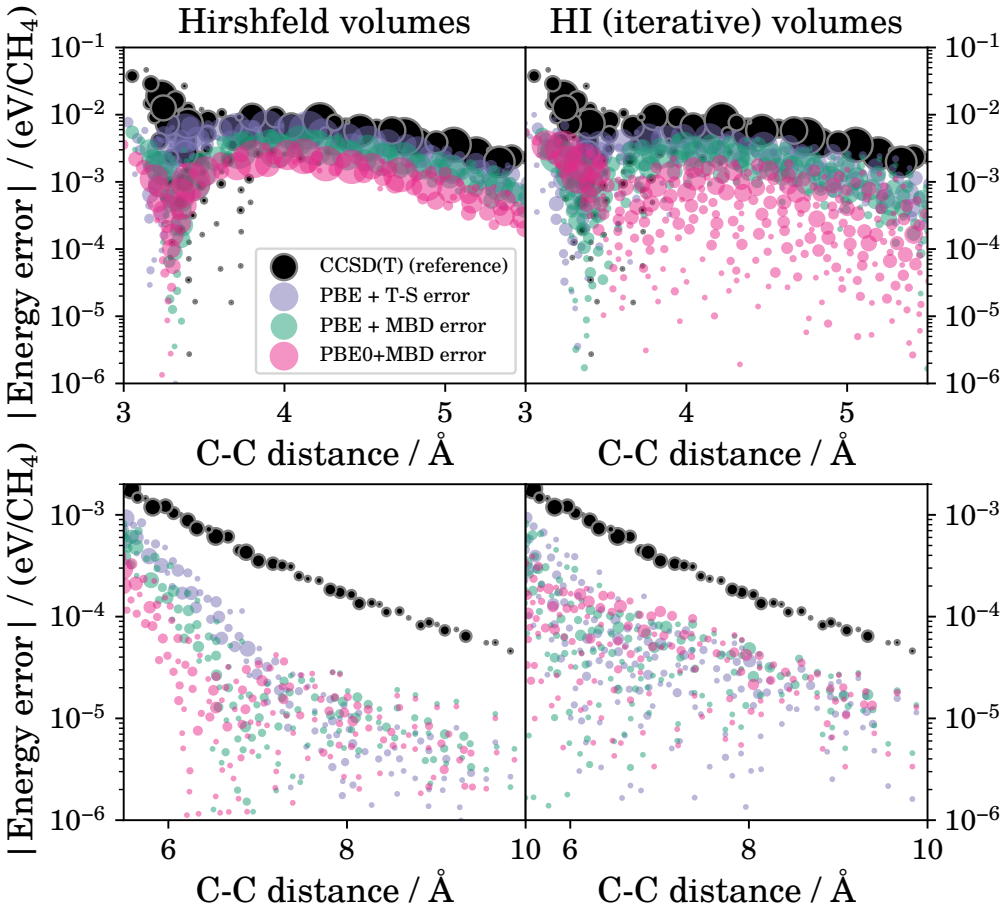


Figure 3.9: Dimer interaction energies predicted by the three dispersion-corrected DFT methods, errors against CCSD(T)-F12. Comparison of dispersion models obtained with regular Hirshfeld analysis (left) and iterative-Hirshfeld (HI) analysis (right). Top: short range, bottom: long range.

the regular Hirshfeld partitioning (both with PBE, on the short-augmented set), the average of the *total* volume ratios for each monomer (a measure of the dispersion coefficient of the entire molecule) hardly changed, being only 5.69 % smaller with the iterative Hirshfeld partitioning.

The first trend we notice from Figure 3.8 is the accuracy of the methods increasing both as the dispersion model is improved from T-S (pairwise) to MBD (many-body), and as the PBE functional is switched for PBE0. The increase in acuracy is most apparent in the medium-short range of 3 Å to 7 Å; beyond there, all three methods display errors in the same low range. Switching to iterative-Hirshfeld volumes did improve the results slightly, as expected; it was shown in Bučko et al. [155] that the T-S correction improves in accuracy for systems with large charge separation when iterative Hirshfeld is used in place of regular Hirshfeld. But the improvement seen in methane is much less pronounced: Figure 3.9 shows a modest improvement for all methods in the short range (3 Å to 5 Å), but in the long range they are even less accurate than the methods using regular Hirshfeld. Indeed, methane does not have very large charge separations: the OPLS-AA forcefield uses a charge of 0.06 e on each hydrogen atom, while the iterative Hirshfeld partitioning predicts a value* about twice as large (0.136 e). These findings are also in accordance with Bučko et al. [155], who found only small differences in the T-S correction between the iterative and non-iterative Hirshfeld versions for dispersion-dominated systems. It thus appears to be safe to simply use the regular Hirshfeld partitioning, the only type currently offered by the CASTEP code [156] that will be used for the full bulk methane cells.

The accuracy of all the methods considered here – analytical potentials as well as dispersion-corrected DFT – is summarized in Figure 3.10 on both the

*computed on the PBE0 electron density of the methane dimer, with charges averaged over the Boltzmann sample

short-augmented and Boltzmann-weighted samples. Again, note how the short-augmented set emphasises errors in the short range: if we consider only the errors on the Boltzmann sample, then the errors of the analytical potentials appear comparable to those made by dispersion-corrected DFT! This is, of course, a valid conclusion if we are only interested in the total energy in the canonical ensemble. But if we are interested in any other property, then the contributions will be weighted differently (cf. Equation (1.1)), so it is well worth considering both the short-range errors (as emphasized in the short-augmented sample) and the Boltzmann-weighted errors when assessing the accuracy of these methods. Of course, the empirical potentials make much larger errors on the short-augmented set, so only OPLS-AA remains comparable in error to dispersion-corrected DFT.

We can make several other comparisons using Figure 3.10 – for one, on the accuracy of dispersion corrections computed using regular Hirshfeld versus iterative Hirshfeld partitionings. Using iterative Hirshfeld generally results in a decreased error on the short-augmented set, though this effect is not as pronounced for the most accurate DFT-based model, PBE0 + MBD. The same trend is visible on the Boltzmann set, with the exception of PBE + T-S. It is hard to be certain why the error of this model is much *larger* on the Boltzmann set with iterative-Hirshfeld volumes than anywhere else, including with regular Hirshfeld volumes or on the short-augmented set. It is likewise puzzling why its error is so small with regular Hirshfeld volumes on the Boltzmann set, enough to upset the ordering seen in the other three cases. Perhaps this anomaly just reflects a limitation of the dimer RMSE measure, or of an insufficiently large sample size in the Boltzmann set. Otherwise, the dispersion-corrected DFT methods have fairly consistent accuracy between the Boltzmann and short-augmented

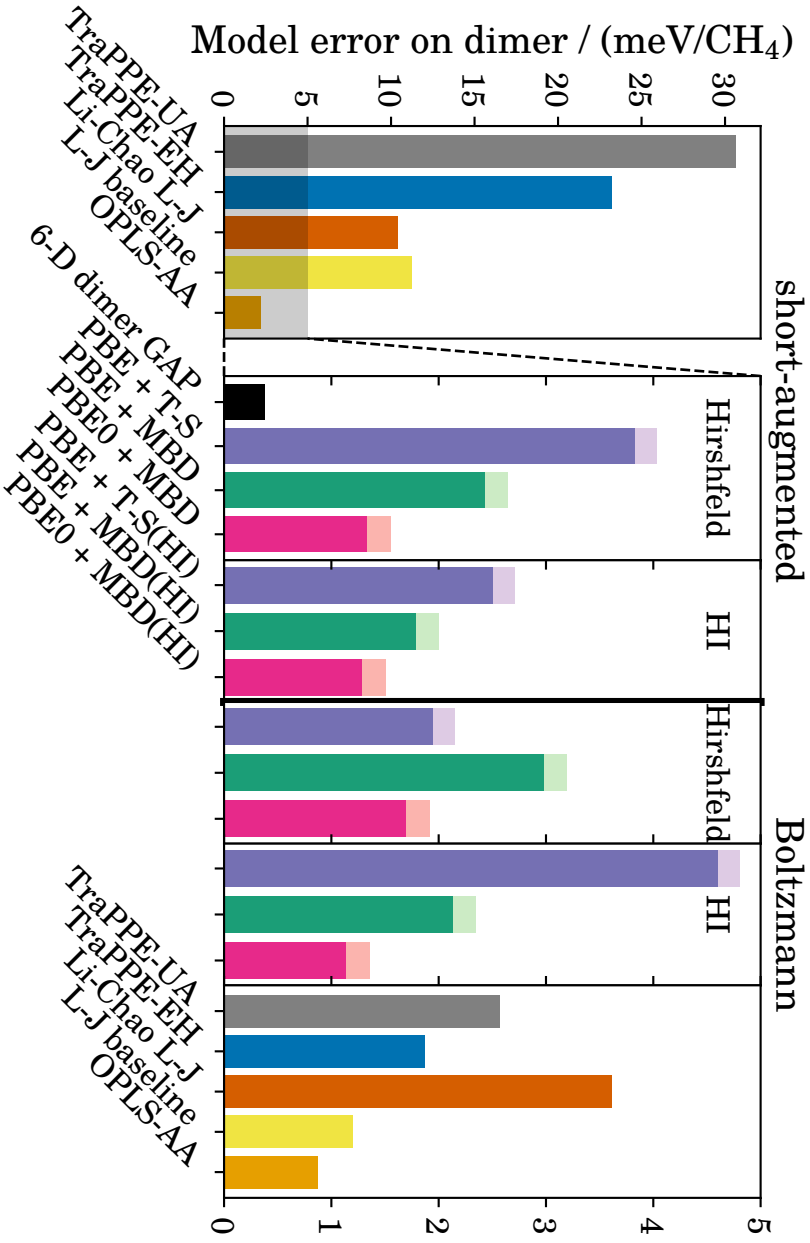


Figure 3.10: (best viewed in landscape orientation) Summary of RMS errors against CCSD(T)-F12 of various models for the methane dimer. Top (left) three plots were computed on the short-augmented dimer set, bottom (right) three were computed on the Boltzmann set. Analytical potentials on the outside, same as shown in Figure 3.7. The middle four plots are RMS errors of the DFT models, with dispersion models computed either with Hirshfeld analysis (second and fourth plots) or iterative-Hirshfeld (HI) analysis (third and fifth plots).

dimer sets and the order of accuracy of the quantum-mechanically founded methods remains clear: PBE + T-S is the least accurate, followed by PBE + MBD; PBE0 + MBD is the most accurate of the DFT-based models, and finally, the 6-D dimer GAP is the most accurate dimer model by virtue of its fitting directly to quantum chemistry in the full-dimensional space of (rigid) dimer conformations.

3.2 GAP method

The 6-D dimer GAP was the most accurate of the models considered above mainly because it relies on a robust, systematic method of fitting functions in high dimensions. The method is called Gaussian process regression and its application to potential energy surfaces (where we typically only have linear combinations and derivatives of local atomic energies) is called GAP [15, 74]. The underlying idea and previous applications were discussed in Section 1.2.4; here we go into more detail on its theoretical background, implementation, and connection to other machine learning methods.

The idea of fitting a function using a Gaussian process (GP) comes from Bayesian statistics. In a sense, the result of the fit is a *distribution* of potential energy surfaces that is updated every time we add a new quantum data point; the usual potential energy surface that we predict is the mean of this distribution [72]. This view also makes it possible to use the variance of the Gaussian process as an uncertainty on the prediction, though this variance is much less robust than the mean value that we use for prediction.

Concretely, the fitted model takes the form of a linear combination of basis functions centred on the target points, as seen here for the local energy of an

atom i as a function of some descriptor $\mathbf{d}_i$ of its local environment:

$$\varepsilon_i = f(\mathbf{d}_i) = \sum_j \alpha_j k(\mathbf{d}_j, \mathbf{d}_i). \tag{3.3}$$

The $\mathbf{d}_j$ are descriptors of the local environments in the training set and $k(\cdot,\cdot)$ is the **kernel** or covariance function. The weights $\boldsymbol{\alpha}$ are determined by a least-squares linear fit:

$$\boldsymbol{\alpha} = \mathbf{C}^{-1}\mathbf{t} \tag{3.4}$$

with $\mathbf{t}$ the vector of previous ("target") observations and $\mathbf{C}$ the **covariance matrix** of the process. In a full Gaussian process, the sum over j would need to run over *all* the environments in the training set; however, this would mean that determining the weights using Equation (3.4) would scale as N^3 with the number N of training environments due to the matrix inversion. The scaling can be greatly improved by selecting a smaller representative set of environments $\mathbf{d}_j$ and performing **sparse** GP regression, where the remaining environments are expressed as linear combinations of the representatives; this method is described in Bartók and Csányi [74] and is a form of the "subset of regressors" algorithm described in Quiñonero-Candela and Rasmussen [157].

The GAP fit is robust to noise in the data – small uncertainties due to a number of possible factors, for instance, insufficiently converged quantum calculations, finite distance cutoffs, or perhaps some other inability of the chosen descriptor to represent changes in the energy [74] – because of a step called **regularization**. This procedure smoothes and simplifies the interpolating function, removing the constraint that it pass through all of the target data points. Indeed, a model that passes exactly through all target points is said to be **overfitted**; since the extra complexity needed to represent the noise is specific to the training

dataset, overfitting limits a fit's applicability to new data.

In GP regression, the regularization is built into the covariance matrix $\mathbf{C}$:

$$C_{ij} = k(\mathbf{d}_i, \mathbf{d}_j) + \sigma_w^2 \delta_{ij} \tag{3.5}$$

with a **regularization parameter** σ_w; δ_{ij} is the Kronecker delta symbol. The covariance matrix for GAP is more complicated because it needs to account for linear combinations and derivatives of energies as well as sparse regression but the basic form is the same (see Section 5.2.1). Notice how the covariance function $k(\cdot,\cdot)$ captures essentially all the nonlinearity of the potential, with the fit itself reducing to a simple regularized linear fit. The interpretation of the kernel and the fitting parameters is the essential difference between Gaussian process regression and kernel ridge regression (KRR) with radial basis functions. In the Bayesian interpretation, the regularization parameter represents the intrinsic noise of the data, as if the data points were drawn from a normal distribution with standard deviation σ_w. In practice, it functions as a weight parameter that tells the interpolant how strictly to fit the data. KRR has a similar approach to regularization; the energy expression is identical to Equation 3.3 and the weights are determined by minmizing the loss function

$$L = \sum_i (t_i - f(\mathbf{d}_i))^2 + \lambda ||\alpha||^2_{\mathbf{C}} \tag{3.6}$$

where $||\cdot||^2_{\mathbf{C}}$ signifies the norm $\alpha^{\mathrm{T}}\mathbf{C}\alpha$ (requiring $\mathbf{C}$ to be positive definite, which is guaranteed by the conditions for the function k to be a valid kernel). If we identify λ with the regularization parameter σ_w^2, then the predictions of this KRR method become equivalent to that of Gaussian process regression [74].

3.2.1 Descriptors

There are many ways of specifying a covariance function, but most of them transform a molecular geometry into a series of geometrical parameters known as **descriptors**, then compute some similarity measure between the descriptors of the two molecular geometries. In order to be useful, a descriptor must obey the same symmetries as the energy itself: It should have the same values whether a system is translated, rotated, or atoms of the same type are permuted [74, 91, 158]. Many commonly-used descriptors are analogous to the internal coordinates used in classical force fields, which must also obey these symmetries. For example, a two-body descriptor of a local environment consists of all the distances between all pairs of atoms within a specified cutoff distance; a potential using such descriptors is analogous to the bond energy terms in local forcefields and pairwise models of the long-range energy (Lennard-Jones and fixed-charge Coulomb). By training a GAP using two-body descriptors, we can obtain a general pairwise potential that captures the physics of pairwise interactions without being constrained to any one functional form.

A GAP trained on simple descriptors, such as the two-body type, will generally use the **squared exponential** covariance function. This function captures the intuition that descriptors very close in value should correspond to very similar energies, with the correlation dropping off steeply as they become farther apart. The covariance kernel has the form [17, 72]

$$k(d_i, d_j) = \delta^2 \exp\left(-\frac{(d_j - d_i)^2}{2\theta^2}\right) \tag{3.7}$$

where in this case, atom pairs take the place of local environments, characterized by distances d_i and d_j as descriptors. The parameter θ determines the character-

istic scale over which the distances are expected to vary, while the parameter δ is an estimate for the scale of energy variation that this descriptor (e.g. the pairwise energy term) is responsible for. If multiple atom types are present, each pair type (e.g. C-C, C-H, and H-H) is represented by a separate Gaussian process; the pair potentials are added together to give the total potential.

This idea can be extended to triplets, which are characterized by three symmetric distances, and even entire dimers. In the latter case, the kernel function compares two dimer geometries [17]:

$$k(\mathbf{R}, \mathbf{R}') = \delta^2 \exp\left[-\sum_i \frac{(R_i - R_i')^2}{(2\sigma_i^2)} \right],$$

where $\mathbf{R}$ is the set of distances between all atoms in the dimer, δ is the characteristic energy scale of variation of the function, and the σ_i are the charateristic scales of variation for each distance type. This kernel must be symmetrized over the permutation group S of the dimer so that the resulting potential does not depend on the order of the atoms:

$$\tilde{k}(\mathbf{R}, \mathbf{R}') = \frac{1}{|S|} \sum_{\pi \in S} k(\pi(\mathbf{R}), \mathbf{R}').$$

This is the kernel used for the 6-D dimer GAP (see Section 4.2.1), where the kernel function is used to measure the similarity of two methane dimers. The obvious disadvantage of this kernel is that it cannot compute the covariance between a methane dimer and any other system. The fit cannot be extended to include other molecules or even beyond-dimer effects.

A more flexible approach is to decompose the energy of a molecule, dimer, or entire system into atomic contributions and fit a potential using descriptors of each atom's local environment as implied in Equation 3.3. The **SOAP** kernel

directly compares local environments by smearing out the atoms in the environment with Gaussians, resulting in the atom-centred neighbour density

$$\rho_i(\mathbf{r}) = \sum_{j \in \mathcal{N}(i)} \exp\left(-\frac{|\mathbf{r} - \mathbf{r}_{ij}|^2}{2s_{ij}^2}\right) \tag{3.8}$$

with the sum running over all atoms in the neighbourhood $\mathcal{N}(i)$ of atom i, and s_{ij} controlling the width of the atomic Gaussians (which could be permitted to vary based on the interatomic distance and atom type, though in practice it is kept constant). The similarity between two environments is computed by integrating the overlap between the two neighbour densities over all possible mutual rotations [74, 91]. The integration, which ensures the descriptor's rotational symmetry (translational and permutational symmetry are included by construction), can be done analytically by expanding each neighbour density in spherical harmonics and radial basis functions

$$\rho_i(\mathbf{r}) = \sum_{nlm} c_{nlm}^{(i)} g_n(r) Y_{lm}(\hat{\mathbf{r}}), \tag{3.9}$$

and summing the power spectrum elements:

$$p_{nn'l}^{(i)} = \frac{1}{\sqrt{2l+1}} \sum_m c_{nlm}^{(i)} (c_{n'lm}^{(i)})^\dagger \tag{3.10}$$

$$k(\rho_i, \rho_j) = \sum_{nn'l} p_{nn'l}^{(i)} p_{nn'l}^{(j)} \tag{3.11}$$

which is then normalized to obtain a proper kernel and optionally raised to some power $\zeta > 1$ to increase the sensitivity to changes in the local environment [74, 91]:

$$\tilde{k}(\rho_i, \rho_j) = \delta^2 \left(\frac{k(\rho_i, \rho_j)}{\sqrt{k(\rho_i, \rho_i), k(\rho_j, \rho_j)}}\right)^\zeta. \tag{3.12}$$

The kernel is an efficient and accurate way of representing solid-state systems

of one or two species where many-body interactions of high order are important [80]. For the new hydrocarbon potential, this kernel is ideally suited to fitting complex, many-body interactions such as the short-range repulsion (the second term of Equation (1.3)). The descriptor has recently been extended with an "alchemical" formulation that considers the neighbours' chemical species in the similarity function [77] as well as a symmetry-adapted formulation for tensorial properties [159], though neither of these modifications is needed here: GAP can already differentiate between atomic centres by fitting separate Gaussian processes for each chemical species, and fitting the potential energy surface only requires the scalar (not the tensorial) SOAP.

3.2.2 Baseline models

In practice, Gaussian process regression is best at fitting relatively smooth functions with a single length and energy scale. But real potential energy surfaces can have large, relatively steep changes (e.g. the molecular repulsive wall). Several methods are available to compensate for this difference in scales; one is to transform the descriptors with some "transfer function" in order to stretch out the regions where the potential varies fastest. But if we already have a simple model that roughly describes the energy, then an easier and conceptually simpler approach is possible: We take this model as a **baseline** and fit our GAP to the *difference* from this baseline to the true potential energy surface, effectively fitting a correction on top of the baseline. For the 6-D dimer GAP discussed in Section 4.2.1, the L-J baseline fit served the role of the baseline model (though any other simple L-J fit, e.g. OPLS-AA, might have worked just as well).

For more complex interactions (such as the molecular repulsion, or the interactions within metals and other crystalline solids) it is not always possible

to find a sensible baseline. One possible approach in this case is to first fit a GAP with simpler descriptors, e.g. a two-body GAP. The more complex contributions can then be fit using the two-body GAP as a baseline, creating a model that effectively splits the total energy into a sum of two-body contributions and higher-order effects. This approach, called hierarchical learning, may result in a more transferable potential than one using just one type of descriptor since the lower-order energy contributions could be seen as more universal and applicable to different types of systems than the complicated, higher-order contributions.

The hierarchical learning approach was not used for fitting the DFT repulsion, as it was found that a GAP fit with no baseline was accurate enough (see Chapter 4); perhaps the separation of energy components (Equation (1.3)) ensured that this component was mainly characterised by a single length and energy scale. But for the dispersion component, it was found that the MBD also needed to be represented with a GAP (the available implementation did not implement gradients and was too slow for the MD simulations planned for methane) and in this case, a ready-made baseline was already available: The pairwise T-S correction, while much simpler than the MBD method, was nevertheless based off of the same idea of polarizabilities computed from atomic volumes; the long-range tails of the two methods were also similar (see Figures 3.8 and 3.9). The MBD SOAP-GAP with the T-S correction as its baseline therefore became an integral component of the bulk methane potential from Chapter 4.

3.3 Intramolecular energy

The only remaining component of the energy from Equation (1.3) is the one-body, or intramolecular, energy. This component was not the primary focus of the potential development effort for methane, as the intramolecular potential

was thought to have a greater influence on the target properties. However, while simulations with a modified version of the AMBER potential did not find any effect of the *strength* of the intermolecular potential on the density, there was a change when the potential was substituted with COMPASS, suggesting at least a dependence on the *form* of the intramolecular potential. Furthermore, the accuracy of the intramolecular potential is especially important when quantum nuclear effects are taken into account [113, 160], so we still need a reasonably accurate intramolecular model for methane.

A systematically fitted intramolecular potential for alkanes was in fact the subject of [46], and this topic will be revisited in Chapter 5, but for now the focus is on testing the new intermolecular GAP models in simulations. Most classical, analytical potentials (see Section 1.2.2) have an intramolecular component that could be incorporated with minimal effort and computational expense; the two that were chosen were the harmonic model AMBER [32], because of its simplicity, and COMPASS [41], because of its strategy of systematically fitting to quantum data using more flexible functional forms. The comparison of these two models as part of the overall methane potential will help gauge the effect of the intramolecular component. Finally, because the intra- and intermolecular components are generally taken to be only weakly coupled [49], the option of inserting a GAP fit of the intramolecular energy of methane remains for the future.

Chapter 4

Intermolecular potential application

The following chapter details the application of the above ideas to build a series of potentials for methane in the compressed liquid to supercritical fluid state, with temperatures between 110 K and 188 K and pressures between 5 bar and 316 bar (the critical point of methane is 190.58 K and 46.04 bar [161]). It shows that a successful prediction of the density of this fluid requires taking into account the many-body dispersion effect (MBD, the highest level of dispersion correction considered here) as well as quantum nuclear effects. The majority is adapted from an article* co-authored with my supervisor and colleagues at my funding organisation, Shell; the accompanying supplementary information in Section 4.2 provides technical details, parameters, and the methodology for fitting the dimer GAP. Section 4.3 discusses some additional technical points with a view towards applying and extending the potentials to larger alkanes. The article's title page and author list is reproduced below; the rest (including page and reference numbers) has been reformatted to fit the dissertation.

*Adapted with permission from [162], in press. Copyright 2019 American Chemical Society.

Equation of state of fluid methane from first principles with machine learning potentials

Max Veit,[*,†,§] Sandeep Kumar Jain,[‡] Satyanarayana Bonakala,[‡] Indranil Rudra,[‡] Detlef Hohl,[¶] and Gábor Csányi[†]

† *Engineering Laboratory*

University of Cambridge

Trumpington Street

Cambridge, CB2 1PZ

United Kingdom

‡ *Shell India Markets Pvt. Ltd.*

Bengaluru 562149

Karnataka, India

¶ *Shell Global Solutions International BV*

Grasweg 31

1031 HW Amsterdam

The Netherlands

§ *Current address: Laboratory of Computational Science and Modeling, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

E-mail: max.veit@epfl.ch

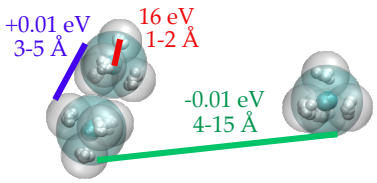


Figure 4.1: Separation of interactions in condensed-phase methane: Covalent, short-range repulsion, and dispersion.

4.0.1 Author contribution details

The bulk of the text was written, all the machine learning models were designed, and all the figures were made by me with input and comments provided by the co-authors, principally Gábor Csányi (my supervisor) and Detlef Hohl. The overall plan for the research conducted here was decided in discussions with Gábor Csányi, Detlef Hohl, and Indranil Rudra. Sandeep Kumar Jain and Satyanarayana Bonakala contributed by running some of the simulations, using the computing resources at Shell, with the (PBE0 SOAP) /COMPASS + T-S + MBD(PBE0) SOAP model detailed below. They ran both the classical (GLE) simulations as well as parts of the PIMD simulations with this model.

4.1 Introduction

This chapter is concerned with the development of a family of GAPs specifically for liquid methane, the simplest alkane, which is inherently difficult to model because its behaviour is dominated by weak dispersion interactions. It is also useful as a stepping stone towards potentials that can model larger hydrocarbons under more extreme conditions [37, 163]; such a potential would enable new research in numerous scientific and engineering applications [3–5].

There is a long history of modelling liquids at the atomistic scale with Monte Carlo (MC) or molecular dynamics (MD) methods. Section 1.2.2 outlined this history and the analytical potentials it produced, such as the venerable Lennard-Jones potential [164] and the many subsequent variations or extensions of this basic form [20, 33, 36, 41, 44, 165]. These potentials contain empirical parameters which are usually optimized until the simulations reproduce specific sections of the experimental equation of state.

Recent potentials show a trend of more closely representing the underlying quantum mechanical potential energy surface, for example by adding anharmonic and cross terms to the covalent forces to arrive at a more faithful representation [41, 48, 49] or even directly fitting the intramolecular [50] or intermolecular [10, 37, 42, 43, 52] terms to *ab initio* calculations. Such potentials, which are the type most commonly employed in simulations of liquids, have achieved high accuracy in reproducing the intramolecular potential energy. However, the restricted functional forms that they employ to describe the intermolecular interactions – typically L-J 12-6 [34, 44], 9-6 [41], or Morse [10, 37] potentials – remain too simple to represent the underlying potential energy surface faithfully. Instead, they represent thermal averages of the true potential energy surface that are useful for making predictions within a certain range of temperature and pressure. These predictions typically break down once the simulations are either taken far outside of this range, or if they are used to predict properties that were not considered in the initial fit [27, 35]. But within the "safe" temperature and pressure ranges, the traditional potentials still deliver the best predictions precisely because they have been fitted to reproduce the experimental values.

No family of potentials better exemplifies this philosophy of accurate predic-

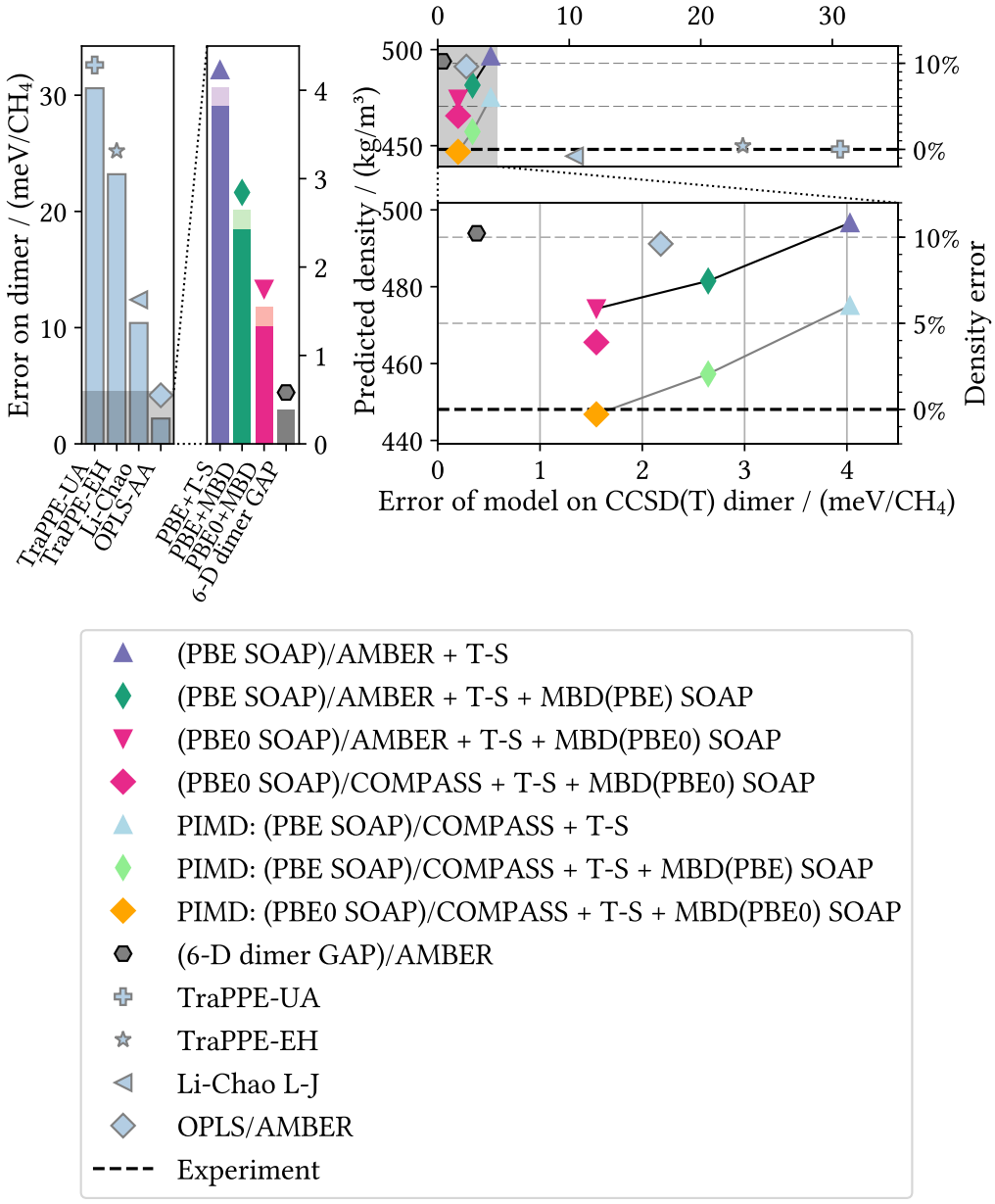


Figure 4.2: Comparison of various models for methane; the density predictions at 110 K *and* 316 bar *are compared against the error of the model's dimer potential energy surface against the CCSD(T)-F12 reference. The suffixes "/AMBER" and "/COMPASS" indicate which model was used for the intramolecular (one-body) energy (the many-body SOAP and 6-D dimer GAP models were only fitted to the beyond-one-body energy). The RMS error is computed over the sample of dimers used to train the 6-D dimer GAP. In the right-hand bar plot, solid bars represent the systematic errors due to the underlying quantum model and the pastel bars on top represent the statistical errors introduced by the GAP fit. In the left-hand bar plot, the bars represent the (systematic) error of the traditional analytical model against the same coupled-cluster reference. Density error is given relative to experiment; the uncertainties on the density are smaller than the sizes of the symbols.*

tions through thermal averaging than the TraPPE family of coarse-grained potentials. Both versions of TraPPE forcefield considered here (the coarse-grained united atom version TraPPE-UA [9] and the reduced dimensional version TraPPE-EH [35]) eliminate degrees of freedom in order to obtain a simpler description of the system. They have been fit to accurately reproduce phase equilibria, and they deliver an accurate prediction of the equation of state of liquid methane. Figure 4.2 shows the density predictions of a selection of models at one state point of liquid methane, compared with their accuracy in reproducing the interaction energy of a sample of methane dimers calculated at the explicitly correlated CCSD(T) level. We immediately see that TraPPE-UA delivers an exceptionally accurate density prediction while having the worst accuracy on the potential energy surface of the dimer (it neglects – by design – the considerable anisotropy of the dimer's potential energy surface). The TraPPE-EH version is similarly accurate in the density, though not much better than TraPPE-UA on the dimer. In contrast, OPLS-AA [34] is the most accurate empirical model of those tested here as far as the dimer potential energy surface is concerned (a tenth of the error of TraPPE-UA), but its density prediction is one of the *worst* of all of the models shown in the figure (about a hundred times worse than TraPPE-UA). Other empirical models are in between these extremes: e.g. Li and Chao's all-atom parametrization[52] is five times worse on the dimer than OPLS-AA, but ten times better in its prediction of the density.

It is surprising and somewhat sobering that the most accurate prediction of the density of liquid methane is achieved by the *simplest* potentials (esp. TraPPE), which do not really attempt to reproduce the actual Born-Oppenheimer potential energy surface; in fact, every effort up to now to better capture the potential energy surface by a traditional analytical potential has lead to worse

predictions of the liquid density.

One might conclude that simply the OPLS-AA is still not accurate enough – and it is, of course, possible to build even more accurate models. Traditional pairwise potentials have two key limitations: First, the restricted functional form of the pairwise interaction limits its accuracy, especially when the potential must reliably model large parts of chemical space. More complex pairwise functional forms have long been used to make more accurate, physics-based potentials [39, 42, 43], though they have not been as widely applied – especially for liquid simulation and equations of state – as the simpler, traditional models. More importantly, any pairwise model neglects many-body effects. These are significant even within the dimer, giving rise to the complex, anisotropic form of the short-range potential energy surface shown in Figure 4.3. While the electrostatic and induction components are often treated within a formally many-body framework [39], other components such as the repulsion and the dispersion also exhibit significant many-body character [166] that is less commonly taken into accoun, especially in liquid simulations.

The high dimensional fitting approach of machine learning allows us to model all of this many-body character without the presumption of any particular functional form. We can explicitly fit the CCSD(T) energies with a Gaussian approximation potential (GAP) [15, 74] (more details in the supporting information) in the full six-dimensional space of mutual dimer orientations (with monomers kept rigid). The reference potential for the methane dimer that we fit with this method, which we will call the “6-D dimer GAP”, is shown along with OPLS-AA in Figure 4.3. This model achieves a consistent level of accuracy across a wide range of dimer separations and orientations. And yet, when we use it to predict the density of bulk methane (Figure 4.2), it is even farther from the experimental

value than OPLS-AA.

The goal of the present work is to resolve this apparent contradiction and develop a methodology for modelling molecular liquids that delivers more accurate predictions as we systematically increase its accuracy against the underlying quantum potential energy surface, thereby ensuring that we get accurate answers for the right reasons.

4.1.1 Quantum-mechanical energies

Several methods are available that approximate the true quantum potential energy surface; the expensive but accurate quantum methods were reviewed in Section 1.2.3 and the new generation of machine learning potentials was reviewed in Section 1.2.4 and in the recent editorial [14]. Most relevant for this chapter is the many-body dispersion correction [62]. The many-body effect has been shown to be crucial for an accurate description of many dispersion-bound systems such as supramolecular complexes [167] and organic crystals [168], though the effects on molecular liquids have not yet been extensively studied – a many-body vdW model (D3 [61]) *was* included in the water potential of [18], but it was not mentioned whether a simple pairwise model would have given different results.

4.1.2 Quantum nuclear effects

Empirical potentials have been fit to reproduce experimental equations of state, so they include quantum nuclear effects implicitly. In contrast, when simulations are done with a systematic approximation of the Born-Oppenheimer potential energy surface, it becomes necessary to account for quantum nuclear effects in an equally systematic manner [104, 115]. These effects are especially important at low temperatures and with light nuclei; their importance in liquid alkanes in

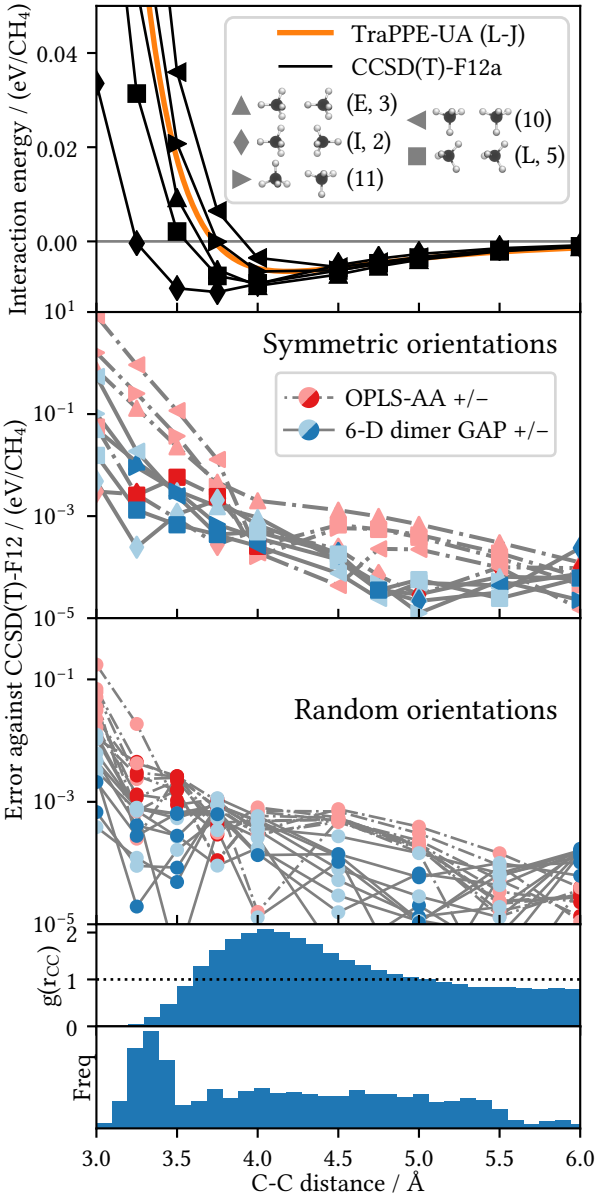


Figure 4.3: Top: Interaction energies of the rigid methane dimer in a selection of orientations. The TraPPE united-atom (and therefore isotropic) model [9] is given by the smooth line; this model gives the best overall prediction of the equation of state (Figure 4.6) even though it completely ignores the anisotropy. Configurations are labeled as in Chao et al. [51] (letters) and Hellmann et al. [42] (numbers). Middle: Errors of two models on the methane dimer energy, the OPLS-AA model [34] and a full-dimensional GAP fit, against CCSD(T)-F12 on the same orientations. Bottom: Errors with ten randomly chosen orientations. A pair correlation function at 188 K *and* 278 bar *and a histogram of the fitting database are given below for reference.*

particular has long been established [160] and was recently highlighted [169] using quantum mechanically fitted forcefields. In empirical potentials these effects are typically included in an average way because they are naturally present in the experimental data used to fit the potentials; some potentials [42] also use a semiempirical or approximate method to include these effects. But in order for a potential to systematically fit the true potential energy surface it *cannot* include quantum nuclear effects at the level of the fitting, because the true Born-Oppenheimer potential energy surface does not itself include these effects. Thus, fitting methods that include such an average contribution are not fitting the true potential energy surface and are therefore incompatible with the current strategy.

The recent developments summarized in Section 2.2 are making explicit quantum nuclear effects via PIMD practical even for large systems and expensive potentials, such as the ones employed in this work. But despite new developments in both machine learning potentials and PIMD methods, *ab initio* liquid simulation remains a challenge. The process of designing a machine learning potential for a new material, especially for amorphous or liquid simulation, is still a laborious manual process. In this work we develop a methodology that will eventually serve as a foundation for more systematic, perhaps even automated, development of potentials for more complex molecular liquids.

4.1.3 Model development methodology

Fundamental to this methodology is a strategy common to most successful potentials for molecular systems: The energy of the system is decomposed into several terms that each represent a different physical interaction, as described in Section 1.2.1. From the point of view of a physics-based analytical potential,

this decomposition is useful because the different physical interactions will typically have different functional forms, and it makes sense to parameterize them separately. From the point of view of a machine learning potential, the main advantage of an energy decomposition scheme is that it separates physical effects that take place at different length and energy scales and prevents the larger effects from overwhelming the smaller ones; while the smaller components might not be important in reproducing the *total* energy, other important observables (such as the density or the diffusivity) might well weight these contributions much higher. By controlling the accuracy of the several components separately it is possible to achieve good accuracy on any property of interest.

In a molecular liquid such as methane, the primary separation in energy scales is between the strong intramolecular (covalent) interactions and the weak intermolecular (noncovalent) interactions. These two types of interactions are easy to separate and have characteristic energy scales that are orders of magnitude apart. The second separation we will employ here is motivated by the length scales of the interactions, as machine learning potentials tend to work best for fitting functions that vary on a single length scale. In methane, the dispersion (van der Waals) interaction is very long-ranged, being still relevant at C-C distances as large as 15 Å, but the various repulsive interactions generated by electron cloud overlap die out by C-C distances of 5 Å. We can therefore represent the energy with Equation 1.3:

$$E_{\text{total}} = E_{\text{1b}} + E_{\text{repulsion}} + E_{\text{dispersion}} + E_{\text{electrostatic}} + E_{\text{induction}}$$

where the “1b” (one-body) energy is the covalent part, though it also subsumes the intramolecular contributions of exchange-repulsion, dispersion, electrostatics, and induction. The other four terms in the equation are therefore understood

to be the intermolecular (more formally, beyond one-body or "b1b") component of the corresponding energy term. The intermolecular repulsion, electrostatic, and induction terms are computed from DFT beyond-one-body interactions – electrostatics and induction, in contrast to dispersion, are handled comparatively well by DFT. The dispersion term is computed separately, as discussed above.

The electrostatic energy may be significant at short range but it decays quickly in comparison to the dispersion interaction in systems, particularly hydrocarbons, without significant charge separation (see Section 1.2.2). To illustrate for the case of methane, the electrostatic energy predicted by OPLS-AA is consistently about two orders of magnitude smaller than the other non-bonded terms; see Figure 4.7. In pure methane the molecule's symmetry additionally bounds the decay rate of the long-range electrostatic interaction: All its permanent electrostatic moments below the octupole cancel. Since the interaction energy of two octupoles decays [28] as r^{-7}, the electrostatic energy can be rigorously expected to decay more quickly than the lowest-order dispersion term, making dispersion the most important contribution for the long range – especially for the tail corrections beyond the potential's cutoff. Together, these considerations allow us to fold the electrostatic and induction energy into the short-range "repulsion" term – hereafter called $E_{\mathrm{sr,b1b}}$ (for "short-range beyond-one-body"). Future potentials could easily treat electrostatics explicitly, however, either to achieve higher accuracy or (more importantly) to be able to treat systems with significant charge separation.

Apart from separation of interaction length scales, another advantage of this energy decomposition approach is that it allows us to capture the different physical contributions and study their effects separately. Much research on *ab initio* analytical potentials follows the approach of more directly representing the un-

derlying physics by extracting forcefield parameters from fundamental physical quantities, such as the electron density, of the monomers.

Models using this approach include the *ab initio* atom-atom potential of Misquitta and Stone [170], the transferable Slater-ISA model of Van Vleet et al. [39] (including the more recent anisotropic version [166]), the Monomer Electron Density Force Field of Vandenbrande et al. [171], and the biomolecular force field of Cole et al. [172]. The IPML model of Bereau et al. [76] goes one step further by using machine learning to efficiently predict these properties across chemical compound space. The physical interpretability and systematic derivation of these models is appealing; however, they are typically applied and tested on dimers and gas-phase systems, with relatively little emphasis on condensed-phase and especially liquid systems. In one application of the *ab initio* pyridine dimer potential to the crystalline phase in Aina et al. [173], the underlying approach was transferred to condensed systems with predictive power even at high pressure; however, the authors also noted that the *ab initio* physics-based approach does not have the advantage of absorbing errors from the many-body terms (for example, from the many-body dispersion terms that the model neglects) in the same way that an empirical potential can.

Our eventual goal is to capture the best elements of both approaches: The physical rigour and interpretability of the *ab inito* approach with the full many-body character, flexibility, and ability to correct errors of many-body machine learning. The present potential represents an important step in that direction; by capturing the simple, physically motivated parts of the energy expression by simple analytical forms and fitting the complex, nonanalytical parts as corrections on top of these, we do use physics to guide our description of the interaction while maintaining complete flexibility of the functional form.

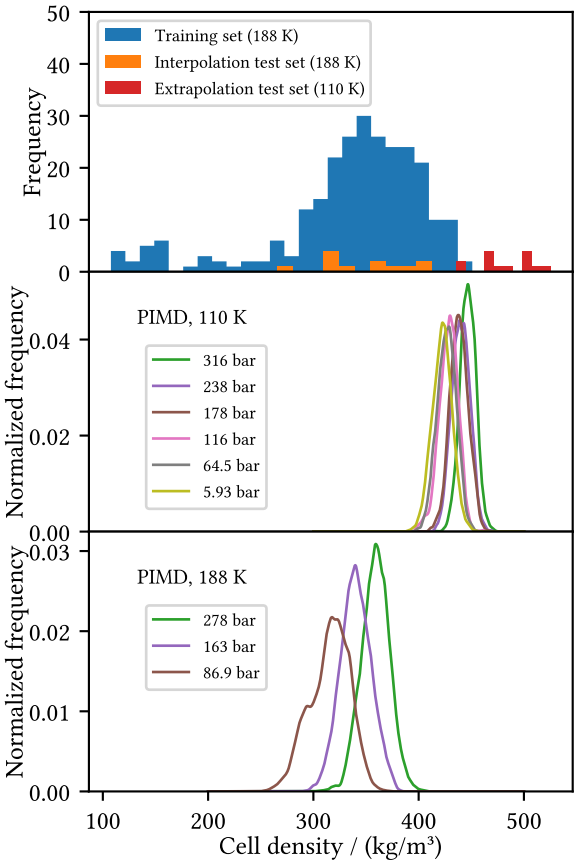


Figure 4.4: Histograms over mass density of the cells in the training and two test sets, interpolation and extrapolation. The distributions of densities encountered in the subsequent PIMD simulations with the (PBE0 SOAP)/COMPASS + T-S + MBD(PBE0) SOAP model are shown below for comparison.

4.1.4 Many-body machine learning model

The 6-D dimer GAP model introduced earlier in this section, shown in Figure 4.3, and described later in Section 4.2.1, is a good representation of the potential of the methane *dimer*. However, it has one key shortcoming for the representation of the condensed phase, namely, that it neglects *all* beyond-dimer many-body effects. Potentials following the body-expansion approach may treat such effects by including explicit trimers or by fitting to a baseline that already includes many-body effects [174]. For our condensed-phase potential, on the other hand,

we opt for a treatment that includes many-body (beyond-dimer) effects from the outset: We fit the $E_{\mathrm{sr,b1b}}$ term using the GAP method [15, 74] with the SOAP kernel [91], both developed and used by our group to fit complex, many-body potentials.

The SOAP-GAP potentials were fitted to DFT [138, 139] b1b energies and forces (beyond-one-body interactions; the monomers were computed separately and subtracted from the total cell), computed on a sample of 280 periodic unit cells of bulk methane, each containing 27 molecules, taken from MD trajectories under liquid conditions run using a classical potential (OPLS/AMBER [32, 175]) at a temperature of 188 K and five pressures ranging from 0 bar to 400 bar, thus covering the entire range of pressures encountered in the subsequent GAP MD simulations. The resulting training set consisted of a wide range of densities; see Figure 4.4. However, the typical densities encountered during a simulation at 110 K in the same pressure range fall partly outside this range, exercising both the model's interpolation and extrapolation capabilities. To validate these capabilities, independent samples were drawn from OPLS/AMBER simulations at both temperatures, with several samples taken from each of the state points where classical results are shown in Figure 4.6 below. The histogram of the densities of these test sets is also shown in Figure 4.4. Based on the position of these distributions relative to the test set, the 12 test samples taken at 188 K were labeled the "interpolation" test set and the 14 samples from 110 K were labeled the "extrapolation" test set.

The DFT calculations on all cells were done using CASTEP [156]. Two functionals were used, the pure GGA functional PBE [140] and the hybrid GGA functional PBE0 [141]. The GAP fits were done using the SOAP descriptor [91], resulting in two models called "PBE SOAP-GAP" and "PBE0 SOAP-GAP". The

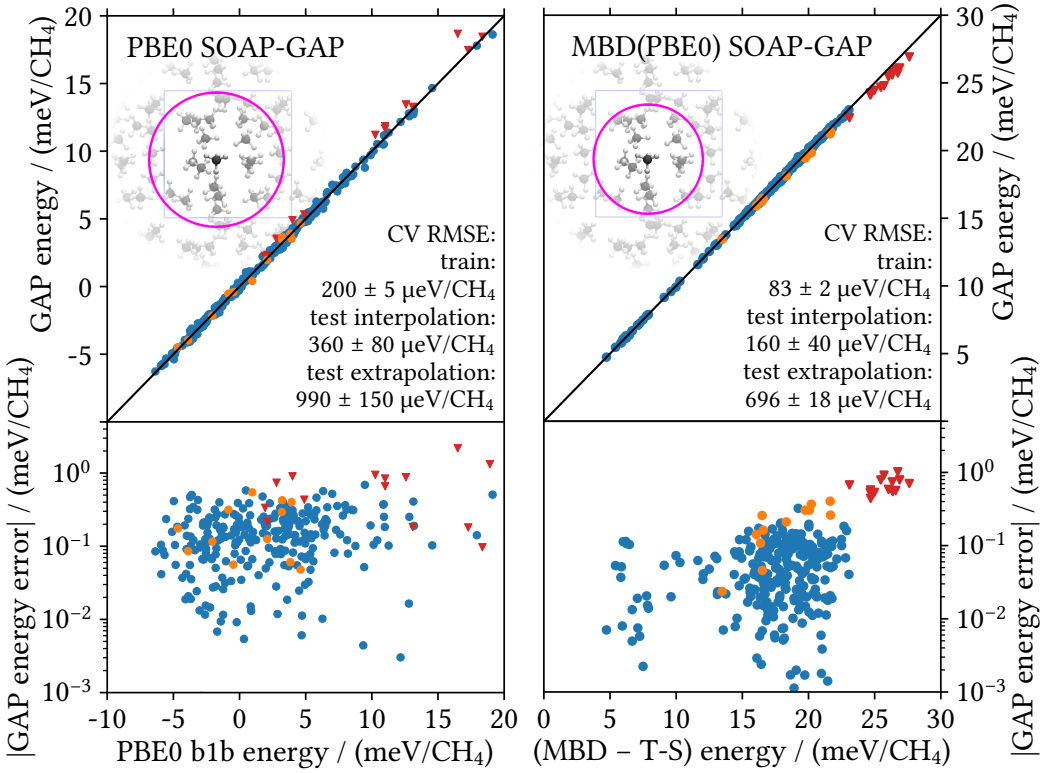


Figure 4.5: The PBE0 and MBD(PBE0) SOAP-GAP fits on 258 cell interaction (beyond one-body, "b1b") energies and (only for PBE0) corresponding forces. Top: Correlation plots with the line $y = x$ of perfect correlation. Bottom: Errors on a logarithmic scale. The blue dots represent the training set. The orange dots represent the interpolation test set and the red triangles represent the extrapolation test set, neither of which was used in training the model.

performance of the PBE0 SOAP-GAP is assessed in Figure 4.5, which indicates good reproduction of both energies and forces on the training set. With a statistical learning method such as GAP, this is usually a good measure of how the method will perform on similar geometries. The interpolation performance indicates some degree of overfitting, while the extrapolation performance is notably poorer – but the model still achieves an error of less than 1 meV per molecule under conditions that were never represented in the training set. The variability of this error measure was assessed with a cross-validation (CV) procedure: Ten disjoint sets of twelve points each were selected from the training data, and each in turn substituted with the interpolation test set to train ten additional GAP mod-

els. The numbers reported in Figure 4.5 are obtained as the mean and standard deviation of the errors across this set of eleven GAPs, with the withheld points standing in for the interpolation test set in each validation GAP. The errors on the forces show the same pattern: The training set error is (6.56 ± 0.03) meV/Å, the interpolation test set error is (6.8 ± 0.6) meV/Å, and the extrapolation test set error is (8.71 ± 0.05) meV/Å. Plots of the forces for the similar PBE SOAP-GAP, along with its energy and force errors, can be found in Section 4.2.2.

The computational effort required to generate the training database was considerable; a typical PBE calculation took 10 minutes on 24 processor cores, with the additional monomer calculations approximately doubling the total required time. The PBE0 calculations were even more expensive, taking anywhere from 50 minutes to several hours; the PBE0 database required overall about four weeks to generate using 27 nodes of 24 cores each. The fitting of the SOAP-GAPs, on the other hand, completed in less an hour on a 16-core machine, and the evaluation of the SOAP-GAP energies and forces requires less than 3 processor-seconds on a cell of 100 methanes. This illustrates a further advantage of the GAP approach, as the computational cost of evaluating the model is independent of the cost of the reference energy chosen: We can run our simulation at PBE0 accuracy without incurring additional computational cost over PBE – minus the initial cost to generate the training database, of course.

Dispersion model

The dispersion component, the third term in Equation (1.3), was accounted for using two levels of theory: The pairwise method of Tkatchenko and Scheffler [60] and the many-body extension MBD [62], both explained in Section 3.1.3.

The first level of theory is the pairwise T-S correction. One major modification

was made to this method to allow it to be used in efficient MD simulations: Recall that T-S uses the relative Hirshfeld volumes of the atoms in their molecular environments to scale the free-atom reference data appropriately. Recomputing the Hirshfeld volumes for each step of an MD simulation would be impractically expensive, however, as that would require a new DFT calculation at each step. Instead, the first level of theory only uses the per-element average of the relative Hirshfeld volumes across the sample of DFT cells. The dispersion correction can then be applied as an analytical pair potential whose form and parameters are fixed throughout the simulation, a scheme hereafter termed simply "T-S". The free-atom reference data used in this scheme was computed by Chu and Dalgarno [142] (the same used in Tkatchenko and Scheffler [60]).

The second level of theory is the MBD, or many-body dispersion, method [62, 176]. Despite the greater complexity of the MBD approach, it can still be viewed as a correction on top of the pairwise Tkatchenko-Scheffler interaction. Thus, another SOAP-GAP was fit to the difference between the MBD energies only and the (fixed) T-S term as the baseline, once each for PBE and PBE0 Hirshfeld volumes. This model, termed "MBD(PBE) SOAP-GAP" (and the corresponding "MBD(PBE0) SOAP-GAP"), accounts for relatively short-ranged many-body effects. The dispersion energy term from Equation (1.3) therefore becomes:

$$E_{\text{dispersion}} = E_{\text{T-S(fix)}} + E_{\text{MBD SOAP-GAP}}. \tag{4.1}$$

The MBD SOAP-GAP also implicitly accounts for the variability of the Hirshfeld volumes that was neglected in the fixed T-S model ($E_{\text{T-S(fix)}} - E_{\text{T-S(variable)}}$): The SOAP descriptor is sensitive to the intramolecular and short-range geometrical factors that (presumably) also account for the variability of these volumes. The MBD(PBE0) fit is likewise assessed in Figure 4.5, showing that both its interpol-

ation and extrapolation performance is similar to that of the PBE0 SOAP-GAP.

Note that the 6-D dimer GAP uses neither T-S nor the MBD SOAP-GAP to model long-range dispersion; instead, it relies on the long-range r^{-6} tail of the L-J baseline described in Section 4.2.1, which was fitted to coupled-cluster energies on various dimer spacings and orientations.

Finally, a complete model for liquid methane must also include an intramolecular component (the first term in Equation (1.3)). Two empirical potentials are considered for this purpose: AMBER [32] includes only harmonic bond and angle terms, while COMPASS [41] includes higher-order anharmonic and cross-coupling terms. Both models were tested in order to help measure the influence of such effects (anharmonic and cross-coupling) on the predicted properties, especially with the inclusion of quantum nuclear effects.

4.1.5 Results

The first test of the accuracy and applicability of any potential for liquids is how well it reproduces the experimental equation of state. While most empirical potentials (for example OPLS [175]) are fit to reproduce experimental thermodynamic data, the fitting conditions are often only a single state point per material, usually standard temperature and pressure. Some potentials, like TraPPE [9], are fit to reproduce thermodynamic data across a wide range of state points, in this case by fitting coexistence curves. Therefore, a wide range of temperature and pressure conditions were chosen to test the accuracy of the potentials considered. Two isotherms were chosen where experimental data was available (from Goodwin and Prydz [123]): At 110 K, density measurements were available at 5.93 bar, 64.5 bar, 116 bar, 179 bar, 238 bar and 316 bar*. At 188 K, density

*truncated to three significant digits; see reference for full precision

measurements were available at 86.9 bar, 163 bar and 278 bar*.

The three models chosen for testing were the "PBE SOAP-GAP" model with both fixed T-S ("+ T-S") and MBD ("+ T-S + MBD(PBE) SOAP") dispersion, and the "PBE0 SOAP-GAP + T-S + MBD(PBE0) SOAP-GAP". The 6-D dimer GAP and all of the SOAP-GAP models were first tested at the state point 110 K and 316 bar using a "smart sampling" coloured-noise thermostat for efficient equilibration [100]. The convergence of the results towards the experimental density is illustrated in Figure 4.2; for brevity, all the "SOAP-GAP" models are labeled simply with "SOAP".

The density predictions are shown against the error of the *underlying* quantum model computed on a sample of dimers (the short-augmented sample from Section 3.1.2), with CCSD(T)-F12 taken as the reference. The statistical uncertainty introduced by the fits is shown and added to the systematic uncertainty already given by the quantum model.

Evidently, the predictions for the density at both state points improve as the dispersion model is made more sophisticated, and therefore more accurate as measured on the methane dimer. Adding the MBD SOAP-GAP lowers the density by 15 kg/m^3, improving the prediction by 3.4 % with respect to experiment and further underscoring the importance of many-body, i.e. beyond-dimer, effects, discussed earlier in relation to the 6-D dimer GAP. The short-range improvement offered by switching to PBE0 gives a further 7.2 kg/m^3 (1.6 %) improvement. While the figure indicates that there are still effects not included by the dimer measure of accuracy – especially the intramolecular potential and many-body (beyond dimer) effects – it still shows a general trend of improvement of the potential's predictions as it more accurately represents the underlying potential energy surface. Crucially, this is a trait not shared by empirical potentials –

TraPPE, OPLS/AMBER and the Li-Chao L-J – which show the opposite behaviour.

The quantum nuclear effect was assessed in an explicit way, using a PIMD simulation using the PIGLET thermostat [105, 114]. With this effect included, the best model ("PBE0 SOAP + T-S + MBD(PBE0) SOAP") delivers a prediction within 0.3 % (nearly within simulation uncertainty) of the experimental density. This decrease in density is of the same order of magnitude as that reported in Pereyaslavets et al. [169], though with this potential the effect is smaller – 4.2 % instead of 9 %. Figure 4.6 shows that the size of the effect is roughly the same across the 110 K isotherm, so even at the 112 K, 1 bar state point used in that study we would expect to see a somewhat smaller effect. The decrease is evidence of the competition between two distinct effects of the zero-point vibrational motion: In the gas phase of methane, zero-point vibrational contributions *increase* the molecular C^6 (first pairwise dispersion) coefficient and hence the strength of the intermolecular attraction [42, 106, 107]. But these same effects also increase the molecular volume [160], ultimately leading to a decrease in the density of the condensed phase. The *ab initio* quality potentials presented here provide the necessary accuracy, especially in the short repulsive regime, for further study of this effect.

The performance of the models across both of the experimental isotherms is shown in Figure 4.6. For comparison, a selection of analytical potentials was tested at all the state points at 110 K and 188 K with experimental data, plus an additional point at 400 bar for each isotherm to show the high-pressure trend. In addition to the potentials shown in Figure 4.2, the figure also shows COMPASS [41]. Note in particular that the empirical all-atom potentials all shift with respect to experiment between the two isotherms. Most models, the SOAP-GAPs

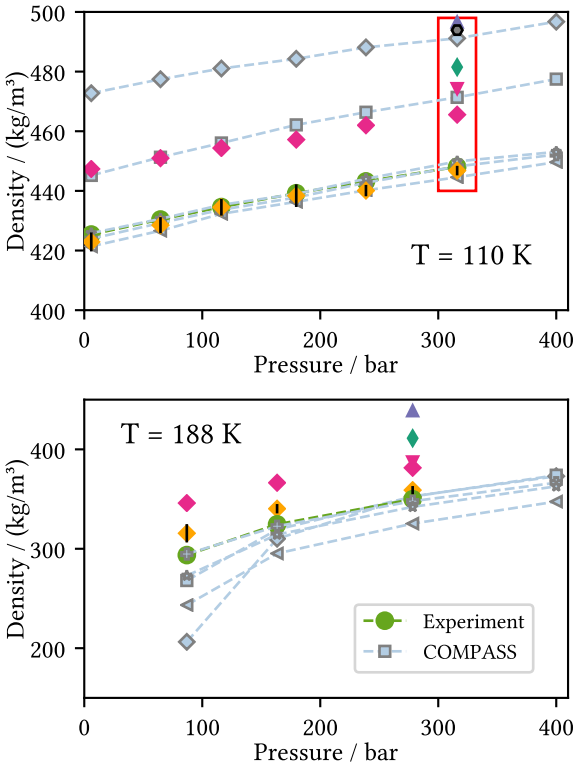


Figure 4.6: Equation of state at two temperatures, 110 K *and* 188 K*, as predicted by various atomistic models. The bulk SOAP-GAPs with different dispersion models are shown, as is the 6-D dimer GAP. All-atom empirical models are shown in grey. Experimental data from Goodwin and Prydz [123]. The small black lines are error bars on the PIMD simulations computed using the blocking method described in the supporting information. Refer to the legend of Figure 4.2 for symbols previously defined.*

included, have more trouble reproducing the density at the 188 K isotherm, perhaps because of the proximity of the lowest-pressure point to the critical point (190.58 K and 46.04 bar [161]). Only the united-atom model TraPPE-UA maintains accuracy across the whole space of conditions covered, with the explicit-hydrogen description TraPPE-EH closely following in consistency. The series of SOAP-GAP potentials delivers predictions of increasing accuracy, in correlation with the accuracy on the dimer. Despite the relatively large statistical fluctuations in the PIMD SOAP-GAP density predictions, the model is still more consistently accurate (comparing across both isotherms) than any other model fit

to the quantum PES, especially with the explicit inclusion of quantum nuclear effects. It thus appears essential to include quantum nuclear effects in order to make accurate predictions with a potential fitted to the Born-Oppenheimer quantum potential energy surface. Other potentials that achieve agreement with experiment without explicit treatment of these effects must be incorporating them into the potential energy surface itself, which is at odds with our stated goal of achieving the agreement with experiment in an *ab initio* manner by best fitting the potential energy surface.

In summary, while TraPPE potentials obtain their accuracy by fitting to experimental data across wide ranges of temperature and pressure, the SOAP-GAP potentials obtain their accuracy by fitting to the underlying quantum mechanical description of matter and systematically converge to within 0.5 % of the experimental value as their description is improved. Additionally, even the current best SOAP-GAP model still has several routes of potential improvement that would not be open to a fixed-form analytical potential, such as changing the intramolecular model for a more accurate, fitted one or improving the dimer description to the coupled-cluster dimer GAP level (which can be done using existing techniques, e.g. by adding a further two-body correction to the SOAP-GAP model [17, 83]).

While the computational cost of the SOAP-GAP potentials presented here is significant, especially including the generation of the training set, it is a tiny fraction of what the cost would be to do PIMD with the explicit PBE0+MBD method. Each PIMD datapoint required about a week on 16 nodes of 24 cores each, so the PIMD data points in Figure 4.6 required about twice as much time to generate as the PBE0 training set itself. But considering that these potentials offer a speedup of between 5000 (PBE) to 30 000 (typical PBE0) over single-

point DFT calculations, the SOAP-GAPs do more than just make simulations more efficient: They make PIMD and other expensive simulations, at the level of PBE0+MBD, possible.

4.1.6 Discussion

The fitting and testing of the SOAP-GAP and dimer potentials for liquid methane reveal three key findings for the description of molecular liquids: First, many-body effects – not only within the dimer, but also beyond-dimer effects – are essential, especially in the short range, for obtaining an accurate description of the bulk density. Second, an explicit description of quantum nuclear effects is equally important, especially at the temperatures and pressures considered here. Third, systematic measures of the accuracy of the potential (such as the dimer error measure presented here) are a good guide to improving systematically fitted potentials toward convergence with the experimental results, a goal which the best many-body GAP model (PBE0 SOAP-GAP + T-S + MBD(PBE0) SOAP-GAP) presented here comes close to achieving.

The methodology presented here represents a new, physics-based, systematic path toward creating exceptionally accurate potentials for molecular liquids. The methodology is applicable to longer hydrocarbons directly; it remains to be seen what the data requirements will be that guarantee sufficient accuracy. Furthermore, the ideas presented here could be extended to other types of long-range interactions, such as electrostatics and induction, in order to extend accurate machine learning potentials to a wider variety of molecular liquids. There is already some evidence that moderately long but finite cutoffs might be sufficient, at least for describing the liquid state [18]; if long-range contributions are required, they can be computed using machine learning of local electrostatic

properties [69, 76, 177].

Acknowledgement

M.V. acknowledges Shell Global Solutions International B.V. for funding, as well as support from the EPSRC Centre for Doctoral Training in Computational Methods for Materials Science (under grant number EP/L015552/1). This work used the ARCHER UK National Supercomputing Service (`http://www.archer.ac.uk`) under the UCKP Consortium, EPSRC grant number EP/P022596/1. We gratefully acknowledge the assistance of Venkat Kapil in preparing the GLE and PIMD simulations; and additionally thank Volker Deringer for additional useful feedback on the manuscript.

The GAP definition files and parameter files required to reproduce the MD simulations in this chapter are available at `https://doi.org/10.17863/CAM.26364`.

Finally, all the plots in this paper were made using Matplotlib [178]; the analysis was done within the Jupyter interactive computing enviroment with the IPython kernel [179], and molecular views were with VMD [180].

4.2 Dimer GAP and technical details

This section is adapted from the supporting information of the above paper; it primarily contains supporting technical information, such as DFT, MD, and GAP fitting parameters and MD trajectories. However, it also contains a more detailed account and evaluation of the 6-D dimer GAP fit mentioned before, both in the previous section and in Chapter 3. Finally, it contains a derivation of the form of tail corrections – small energy and pressure corrections due to the finite cutoff of the potential, which can be significant in constant-pressure simulations – for potentials that have a smooth, rather than sharp, cutoff. Although this type of potential and its corresponding tail correction is only used for one of the simulations in the previous section (the 6-D dimer GAP simulation), no previous derivation or mention of this type of tail correction was found in the literature.

4.2.1 Dimer energies

The binding curves of the methane dimer shown in Figure 4.3 were computed in a similar way as described in Gillan et. al. [83]: the Hartree-Fock (HF) energy was computed at the largest basis, the Dunning correlation-consistent basis set [128, 129] aug-cc-pV5Z (hereafter called AV5Z). The energy difference between MP2 and HF was computed using the smaller AVQZ basis. Finally the difference between CCSD(T) (with explicitly correlated basis functions, called CCSD(T)-F12a [127, 130, 131]) and MP2 was computed using the AVTZ basis. The corrections were successively added to the base HF energy to obtain energies at each of the HF, MP2, and CCSD(T)-F12 levels, and additionally forces at the HF and MP2 levels. Finally, all of the energies were corrected for basis-set superposition error (BSSE) using the Boys-Bernardi counterpoise procedure [132]. Calcu-

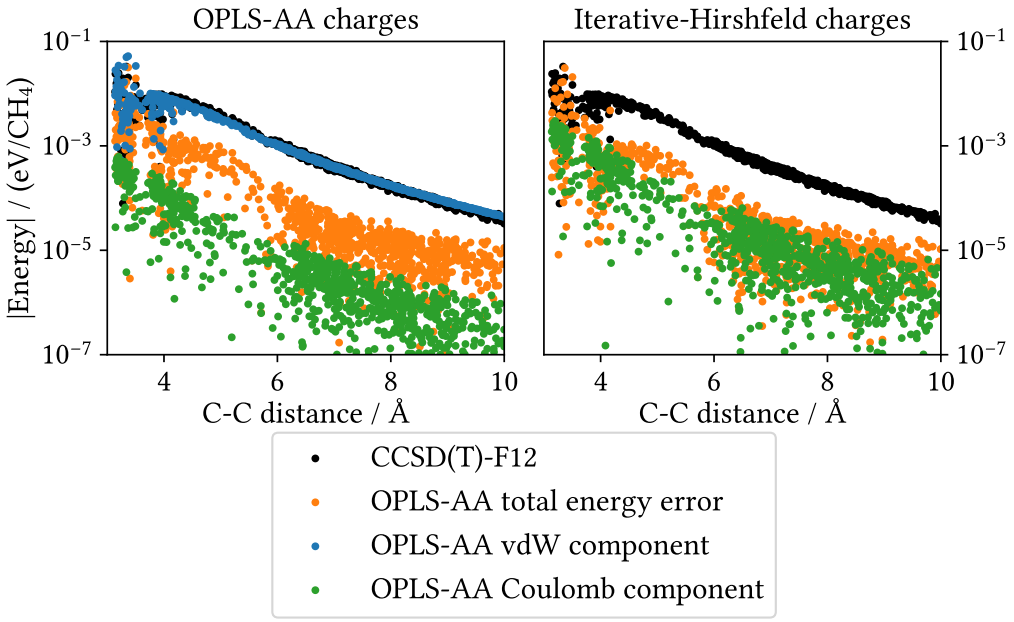


Figure 4.7: Coupled-cluster energies of the methane dimer, compared with the energy components as predicted by OPLS-AA [34] using two different sets of electrostatic parameters. Left: Original partial charges used in OPLS-AA; right: partial charges computed by iterative-Hirshfeld partitioning [148] of the electron density of the dimer predicted by PBE (carbon: 0.528 e, about twice the value of 0.24 e used by OPLS-AA). The vdW (Lennard-Jones) component is the same for both models.

lations were done using the MOLPRO suite of programs [133–136]. Figure 4.7 shows the energies compared against the predictions of the OPLS-AA model, the most accurate traditional forcefield tested against methane dimer energetics in this work.

The geometries for the symmetric orientations were generated using the Atomic Simulation Environment (ASE) [181] starting with monomers that had been optimized at the MP2/AVQZ level, resulting in a C-H bond length of 1.085 Å. The first two configurations correspond exactly with configurations used in Chao et al. [51] and Hellmann et al. [42]; the figure gives the labels each author assigned to these configurations. The other three are similar, though not exactly the same, as the corresponding labeled configurations.

A first dimer model was obtained in a similar way as in Li and Chao [52]

by fitting a pairwise L-J to the energies of the symmetric orientations shown in the main paper. The model was a standard 12-6 L-J between all atom pairs; the six coefficients for the different pair types were all optimized by a least-squares fit. The optimization produced C-H and H-H potentials that were nearly purely repulsive, so the form $\phi(r) = Ar^{-12}$ was adopted for these instead. The C-C potential has the standard L-J form: $\phi(r) = -4\epsilon((r/\sigma)^{-6} - (r/\sigma)^{-12})$. The parameters of this model are given in Table 4.1.

Parameter	Pair type	Value
σ	C-C	3.526 08 Å
ϵ	C-C	0.001 35 eV
A	C-H	517.030 eV
A	H-H	23.4878 eV

Table 4.1: Parameters for the optimized pairwise L-J model

The fitting dataset used dimer distances from 3.5 Å to 9.5 Å in steps of 0.5 Å, with additional points at 3.25 Å and 3.75 Å. Energies larger than 0.02 eV were not used in the fit.

This model was then taken as the *baseline*, and further fits were done on the difference between this L-J baseline and the full energies in order to improve upon this model. For the new fits, a more thorough sample of the dimer configuration space was needed, so a random sample of dimers was taken from a liquid MD simulation using 200 rigid methane molecules with the monomer geometry optimized at the composite CCSD(T)/AVTZ level described below (but without the F12 correction), giving a C-H bond length of 1.088 Å, and the intermolecular interactions computed using the OPLS-AA force field [34]; the simulation was run using the LAMMPS molecular dynamics package [124]* at 188 K and 400 bar using a Langevin thermostat [182] and a Nosé-Hoover barostat[93, 94, 183–185].

*using LAMMPS stable release from 5 Oct 2015

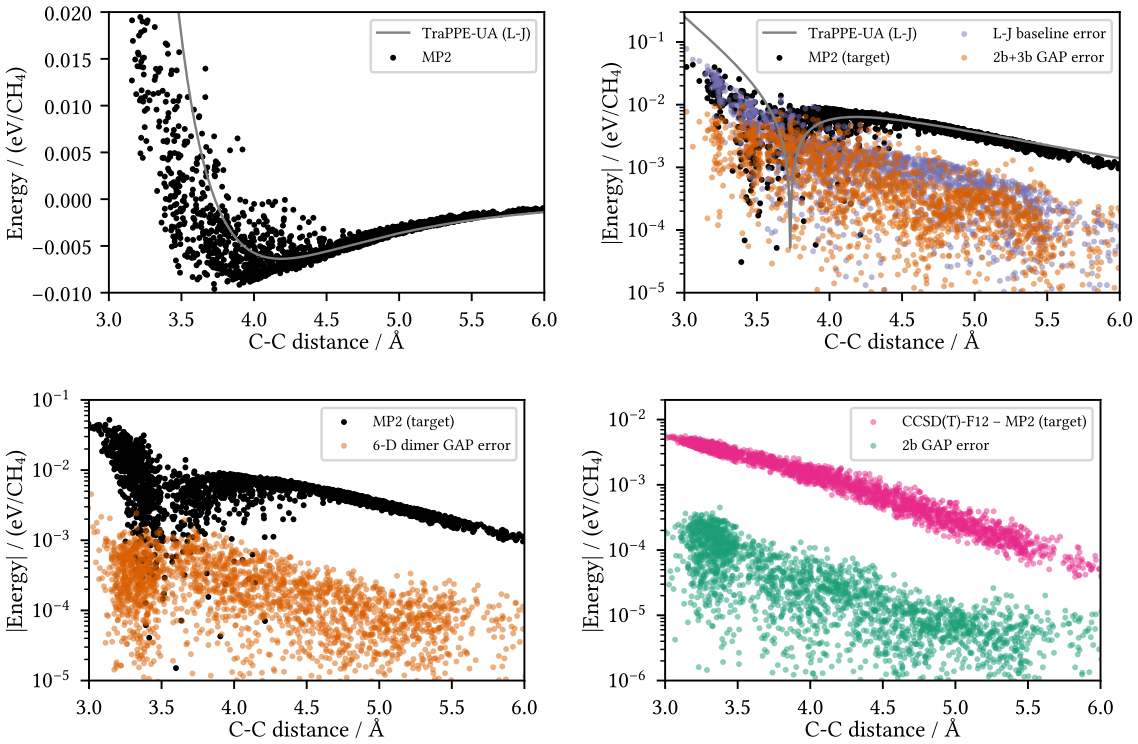


Figure 4.8: Errors of successive GAP models fit to MP2 data, shown as a function of C-C dimer separation. The baseline is a pairwise L-J model fitted to the coupled-cluster data from the symmetric orientations. The first fit uses two-body and three-body descriptors, the second uses the 6-D dimer descriptor, and the final correction to the coupled-cluster level is a simple two-body (pairwise) fit.

(The MD simulations used to generate the random orientations for the binding curves in Figure 4.3 were done the same way, except the monomers were fixed with the OPLS-AA C-H bond length of 1.09 Å.) The dimers were sampled with a C-C distance distribution from 3 Å to 10 Å, strongly favouring the short range of 3 Å to 5.5 Å and further enriched between 3 Å to 3.5 Å. A section of this distribution is pictured with the dimer binding curves in the main text; the full sample contains 2418 dimers. The interaction energies of the dimers in this dataset were computed using the same procedure as described for the fixed-orientation samples.

We first fit to MP2, since both energies and forces are available to achieve a

high-quality fit. The simplest descriptor used was the distance between pairs of atoms; each type of pair (e.g. C-C, C-H, and H-H for methane) is given a separate Gaussian process corresponding to a separate pair potential. This descriptor is called '2b' (for "atomwise two-body"). This idea can be extended to triplets of atoms, where the set of three distances is symmetrized so as to make it permutationally invariant. This descriptor is likewise called '3b'. A first fit was done using both of the above descriptors, with one Gaussian process for each pair or triplet type; the resulting potential is essentially a sum of atomwise pair and triplet potentials with fully flexible functional forms. This potential is called the '2b+3b GAP'. As Figure 4.8 shows, though, this fit offers only a modest improvement over the baseline.

We therefore attempt a fit in the full six-dimensional space of rigid dimer configurations using the dimer descriptor introduced in Section 3.2.1 . This descriptor is composed of the set of distances between all atom pairs in the dimer, symmetrized over permutations of like atoms. Concretely, the kernel or covariance function between two dimers is, as described in [17]:

$$k(\mathbf{R},\mathbf{R}') = \delta^2 \exp\left[-\sum_i \frac{(R_i - R_i')^2}{(2\sigma_i^2)}\right],$$

where $\mathbf{R}$ is the set of distances between all atoms in the dimer, δ is the characteristic energy scale of variation of the function, and the σ_i are the charateristic length scales for each distance type. This kernel must be permuationally symmetrized so that the resulting potential does not depend on the order of the atoms:

$$\tilde{k}(\mathbf{R},\mathbf{R}') = \frac{1}{|S|}\sum_{\pi\in S} k(\pi(\mathbf{R}),\mathbf{R}')$$

where S is the permutation group of the methane dimer, which – allowing both

swaps of hydrogen atoms within the monomers and swaps of whole monomers in the dimer – has order $4! \times 4! \times 2 = 1152$. The kernel is finally multiplied by a cutoff function $f_{\text{cut}}(r_{ab}) f_{\text{cut}}(r'_{ab})$, one for each dimer, which depends on the centre-of-mass separation of the monomers in the dimer. The cutoff function is designed to take the function smoothly to zero as either of the dimers approaches some cutoff distance. In our implementation, it takes the form of a half-cosine between an inner and an outer cutoff; the functional form is given in [74].

The descriptor is an overcomplete representation of the full space of mutual dimer orientations, which in the case of rigid methanes is six-dimensional. The resulting fit offers improvements of at least an order of magnitude across the entire close range (3 Å to 6 Å). The final correction is the difference from MP2 to coupled cluster CCSD(T)-F12, which is easily captured to high accuracy using an atomwise two-body (pairwise) GAP fit to the original sample of 896 dimers (a subset of the full sample, without the subsequent short-range augmentation needed for the MP2 dimer GAP). The composite model created by adding the L-J baseline, the MP2 dimer GAP, and the final coupled-cluster two-body GAP, will hereafter be referred to as the '6-D dimer GAP'. The resulting potential is a pairwise-additive two-body model and will thus miss all beyond-two-body (beyond-dimer) effects. It will still serve as a useful reference for further models, though, as it can be taken as the benchmark standard for the fictitious system of methane with only two-body interactions present. Note also that the GAP fit itself has no long-range component, leaving the long-range (tail) corrections to be handled entirely by the L-J baseline.

The new potential was evaluated on its own training set and on the dimer binding curves from the main text. It consistently achieves the level of accuracy specified in the fit, 2 meV, in the regions of the potential probed under liquid

conditions (as evidenced by the pair correlation function) and can therefore be used as a reference standard for liquid methane dimer interactions.

4.2.2 SOAP-GAP fits and evaluation

GAP is a statistical learning method and hence the quality of the SOAP fits can be evaluated by how well they reproduce the energies and forces of the training set. RMS energy and force errors are given in Table 4.2, with the 6-D dimer GAP errors given for comparison.

The fits are additionally evaluated on two test sets that were not included in the training set, as described earlier (see Figure 4.5). Additionally, the performance of the MBD SOAP-GAP fits was assessed by the finite-difference method, as gradients were not available. For this purpose, a sample of five small cells containing eight methane molecules each was taken from OPLS/AMBER NVT simulations, one each at five densities ranging from 150 kg/m^3 to 400 kg/m^3 (see Figure 4.9). Each geometry was displaced in each of five randomly selected directions for a total of 25 finite-difference forces. The results are shown in the right-hand panels of Figure 4.11.

GAP name	RMS energy error / (μeV/CH_4)	RMS force error / (meV/Å)
PBE SOAP-GAP	200	7.36
PBE0 SOAP-GAP	207	6.58
MBD(PBE) SOAP-GAP	76.1	1.3 (FD)
MBD(PBE0) SOAP-GAP	76.9	—
MP2 dimer GAP	367	5.10
CCSD(T)-F12 2b GAP	80.6	—
6-D dimer GAP	381	—

Table 4.2: RMS energy and force training errors of the GAP fits

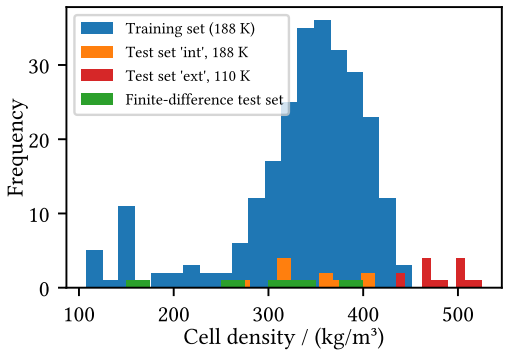


Figure 4.9: Histogram of densities in the training set and the three test sets: Interpolation, extrapolation, and the five finite-difference geometries.

The parameters for the above fits are given as command lines that can be used with the 'teach_sparse' command in the libAtoms/QUIP package [98]. The GAP code can be downloaded at http://www.libatoms.org/gap/gap_download.html, with a prepackaged version available through Docker at https://hub.docker.com/r/libatomsquip/quip/.

The parameters for the PBE SOAP-GAP are (all on one line):

```
teach_sparse at_file=mebox-minimal-nots-b1b-train.xyz gap={
      soap atom_sigma=0.5 l_max=8 n_max=8 cutoff=6.0
      cutoff_transition_width=1.0 delta=0.01
      add_species n_species=2 species_z={{1 6}}
      n_sparse=2000 covariance_type=dot_product sparse_method=cur_points
      zeta=4.0
    } default_sigma={0.0001 0.002 1.0 1.0} sparse_jitter=1e-10
    virial_parameter_name=none gp_file=gp-mebox-pbe-b1b.xml
```

(The parameters for the PBE0 SOAP-GAP are exactly the same; only the source data was computed with PBE0 instead of PBE.)

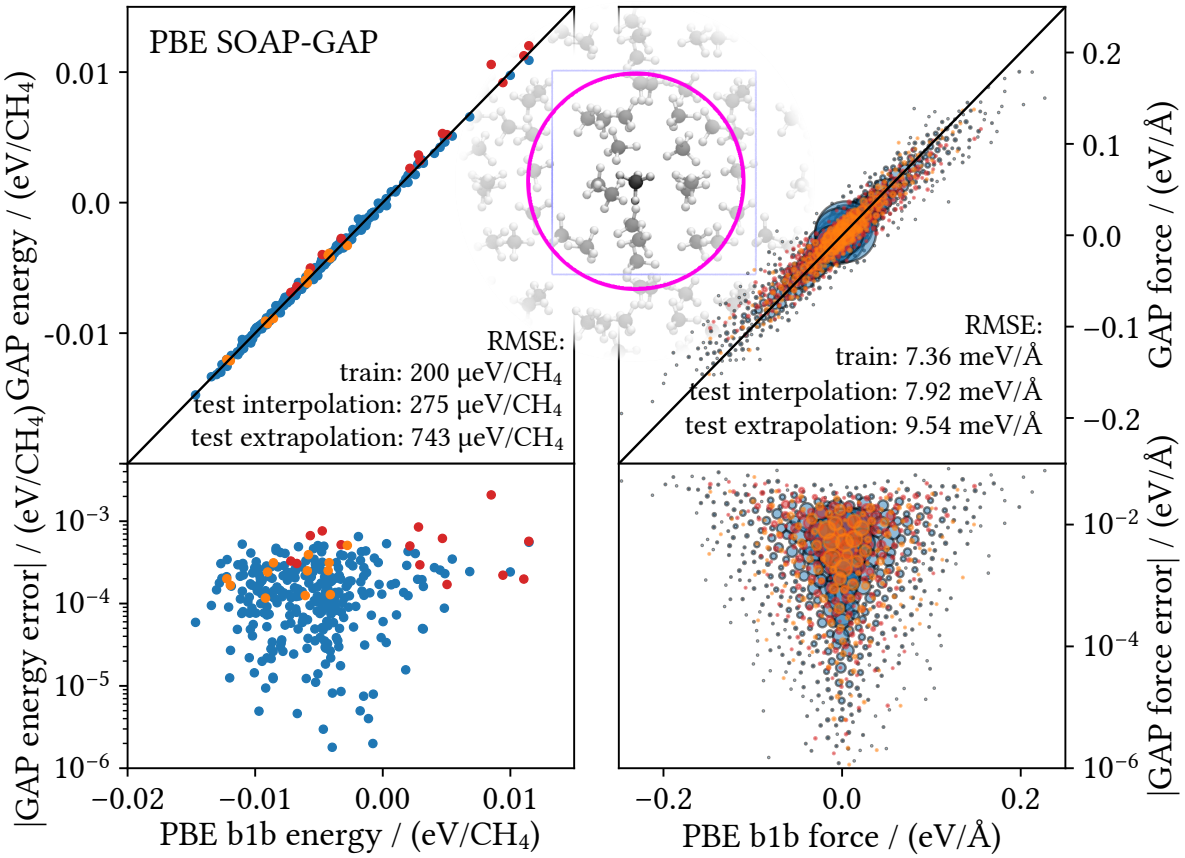


Figure 4.10: The PBE SOAP-GAP fit on 277 cell interaction (beyond one-body, 'b1b') energies and corresponding forces. Left: energies; right: Cartesian force components. In the force plots, due to the large number of points, only a subset of representative points are shown with their sizes scaled according to the 0.8 power of the number of points they represent. Top: Correlation plots with the line $y = x$ of perfect correlation. Bottom: Errors on a logarithmic scale. The blue points represent the training set. The orange points represent the interpolation test set and the red points represent the extrapolation test set, neither of which was used in training the model.

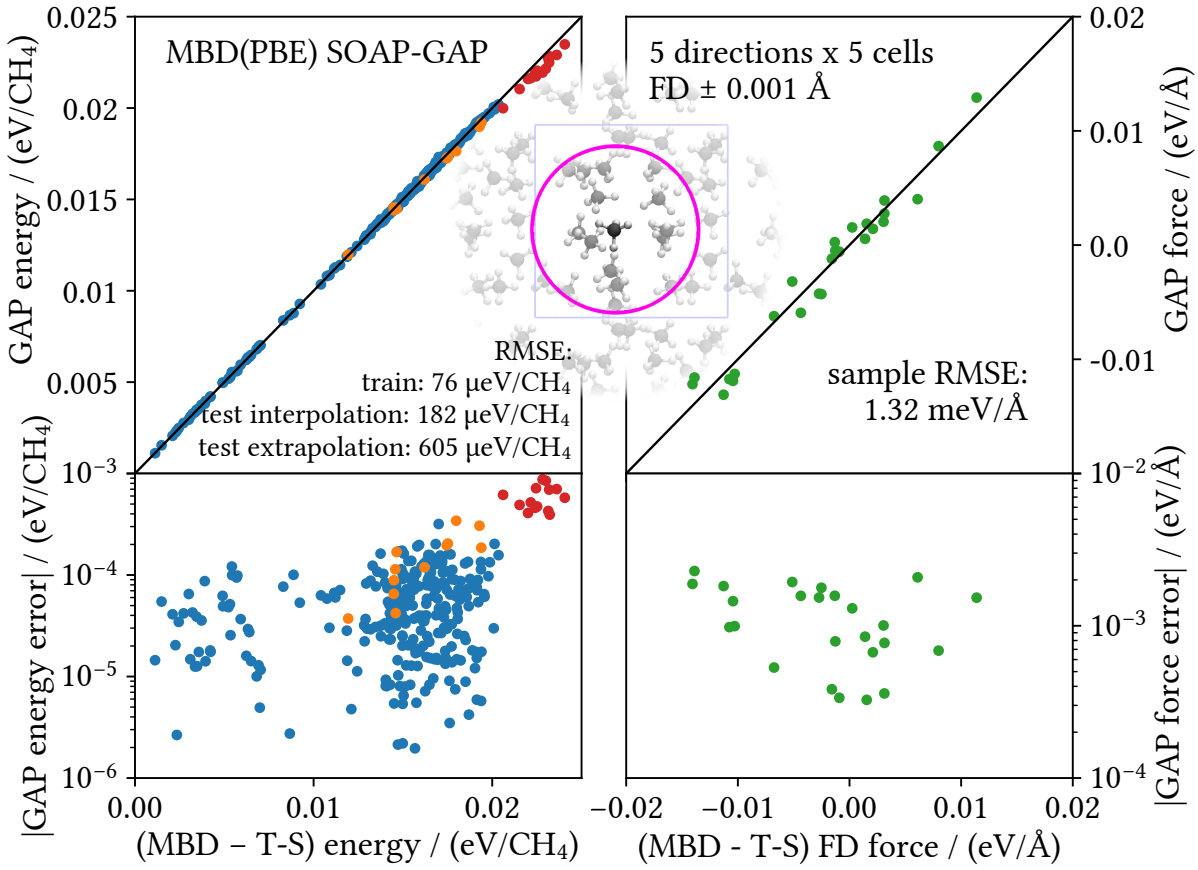


Figure 4.11: The MBD(PBE) SOAP-GAP fit to the differences between the MBD interaction energy and the T-S interaction energy, both computed using PBE-derived Hirshfeld volumes (fixed averages for T-S). Left: Energies; blue points for the training set, orange points for the interpolation test set and red points for the extrapolation test set (color online). Right: Forces estimated by finite differences on five cells of eight methanes each in five randomly chosen directions each (green points). As before, correlation plots against $y = x$ above, log errors below.

The parameters for the MBD SOAP-GAP are (again, same for both PBE and PBE0 energies):

```
teach_sparse at_file=mebox-minimal-mbdint.xyz
    core_param_file=../python/dispts_quip_params.xml core_ip_args={
        Potential xml_label=ts
        calc_args={hirshfeld_vol_name=hirshfeld_avg_volume}
    } e0=0.0 gap={
        soap atom_sigma=0.5 l_max=8 n_max=8 cutoff=5.0
        cutoff_transition_width=1.0 delta=0.001
        add_species n_species=2 species_z={{1 6}} n_sparse=2000
        covariance_type=dot_product sparse_method=cur_points zeta=4.0
    } default_sigma={0.0001 1.0 1.0 1.0}
    sparse_jitter=1e-10 gp_file=gp-mbd-soap.xml
```

The parameters for the 6-D dimer fit to MP2 are:

```
teach_sparse at_file=me-rigid-shortaug3-mp2-avqz-intnonan.xyz
    core_param_file={../empirical-pots/ljrep_quip_params.xml}
    core_ip_args={IP LJ} gap={
        general_dimer cutoff=6.0 cutoff_transition_width=1.0
        signature_one={{6 1 1 1 1}} signature_two={{6 1 1 1 1}}
        monomer_one_cutoff=1.5 monomer_two_cutoff=1.5 atom_ordercheck=F
        strict=F mpifind=T theta_uniform=1.0 covariance_type=ARD_SE
        n_sparse=2000 delta=0.02 sparse_method=CUR_COVARIANCE
    } default_sigma={0.0002 0.002 0.0 0.0} sparse_jitter=1e-10
    energy_parameter_name=energy force_parameter_name=force e0=0.0
    gp_file=gp-merig-mp2-gendim-shortaug3.xml do_copy_at_file=F
```

And for the much simpler two-body (atomwise) fit to the CCSD(T)-MP2 difference:

```
teach_sparse at_file=me-rigid-train-ljrep.xyz gap={
      distance_2b cutoff=10.0 covariance_type=ARD_SE
      n_sparse=50 sparse_method=UNIFORM Z1=6 Z2=6
      theta_fac=0.2 delta=0.0005 resid_name=resid only_inter=T
      : distance_2b cutoff=6.0 covariance_type=ARD_SE
      n_sparseX=50 sparse_method=uniform Z1=1 Z2=6
      theta_fac=0.2 delta=0.0005 resid_name=resid only_inter=T
      : distance_2b cutoff=6.0 covariance_type=ARD_SE
      n_sparseX=50 sparse_method=uniform Z1=1 Z2=1
      theta_fac=0.2 delta=0.0005 resid_name=resid only_inter=T
    } default_sigma={0.00001 0.0 0.0} sparse_jitter=1e-10
    energy_parameter_name=ediff_cc force_parameter_name=none e0=0.0
    gp_file=gp-merig-cc-ljrep-2b.xml do_copy_at_file=F sparse_separate_file=F
```

The final 6-D dimer GAP is simply the sum of the above two potentials.

These potentials, once fitted, are stored in the form of an XML file that can be read by QUIP to evaluate energies and forces on any new configuration. The XML files for the above GAPs are avalable online in the Apollo repository* as well as on our group's webpage†.

4.2.3 DFT and MBD parameters

As mentioned in Section 4.1.3, the sample for the DFT calculations was taken from MD trajectories under liquid conditions run using a classical potential

*https://doi.org/10.17863/CAM.26364
†http://www.libatoms.org/Home/DataRepository

(OPLS/AMBER [32, 175]) at a temperature of 188 K and five pressures ranging from 0 bar to 400 bar (the same ones at which OPLS/AMBER was tested in the main text, with the addition of 0 bar). There were 60 samples taken from each pressure, with the exception of the shorter 0 bar simulation, which only contributed 40 samples. Each of the 280 cells in the sample contained 27 (flexible) methane molecules; otherwise, the simulation parameters were the same as those described later for the OPLS/AMBER density simulations.

The DFT calculations were all done with the CASTEP code [156], version 8.0. The PBE calculations were done with a plane-wave cutoff of 650 eV and the default finite-basis correction. Due to the large, amorphous nature of the system, no k-point sampling was employed; calculations were only done at the Γ point. Convergence tolerances were set to 1 µeV/atom for the energies and 10 µeV/Å for the forces. The PBE0 calculations were done with a cutoff of 700 eV and no finite-basis correction, again only at the Γ point, and convergence tolerances one order of magnitude smaller (0.1 µeV/atom for the energies and 1 µeV/Å for the forces). Since computing the interaction energy requires subtracting the one-body contribution (the energy and force of each individual methane molecule in the cell) and the samples had flexible monomer geometries, an additional calculation was run on each of the 27 individual molecules in each cell, using the same periodic boundary conditions as the full cell. The energy that resulted from subtracting the sum of the monomer energies from the total cell energy is the interaction or beyond-one-body ('b1b') energy (and likewise with the interaction force). Finally, two cells were discarded because their interaction energies (both PBE and PBE0) were much higher than the rest; those cells came from the initial MD equilibration from a high-energy geometry, so they were removed to achieve a better fit for normal, equilibrium conditions. Additionally, the largest cell for PBE and

the largest 20 cells for PBE0 did not complete because the computational requirements exceeded available resources. The training set therefore comprised 277 PBE interaction energies (and $277 \times 135 \times 3 = 112185$ PBE force components), and 258 PBE0 interaction energies (and $258 \times 135 \times 3 = 104490$ PBE0 force components). These sets of interaction energies and forces were finally fit with the SOAP GAPs above, ranged at 6 Å.

The MBD energies were computed on the same sets of 277 (or 258 for PBE0) methane cells using the implementation available at `http://www.fhi-berlin.mpg.de/~tkatchen/MBD/` and interfaced with QUIP. This was done both with the PBE and PBE0 Hirshfeld volumes calculated from each geometry, as reported by CASTEP. The supercell cutoff parameter was adjusted so that a $1 \times 1 \times 1$ supercell (that is, only the unit cell) was used, in correspondence with the omission of k-point sampling in the DFT calculations. All other MBD paramters were left at their defaults. The corresponding T-S model, with fixed, per-element averaged PBE or PBE0 volumes, was then subtracted and the difference was fit with a SOAP-GAP ranged at 5 Å. The magnitude of the correction beyond this range was small enough that neglecting it was seen as safe. The given implementation did not compute gradients, so the accuracy of the GAP forces was assessed using a finite-difference scheme as described above.

The Hirshfeld volumes used to compute T-S and MBD energies on the dimer test set (computed to assess PBE+MBD and PBE0+MBD dimer model errors) were instead computed from the wavefunctions produced by the Psi4 code [150] with the HORTON post-processing functionality [151], which itself uses methods derived by Becke and Dickson for polyatomic molecules [152–154].

4.2.4 MD parameters

GAP fits

Most of the GAP MD simulations were run with i-PI [186], interfaced through LAMMPS [124]* to QUIP [98]. Only the 6-D dimer GAP simulation was run with QUIP's built-in MD functionality. It used the adaptive Langevin thermostat of Jones and Leimkuhler [97] (with a time constant of 10 fs) and a Hoover-Langevin barostat [102] (with a time constant of 100 fs and a mass factor of 100). For the SOAP simulations, the T-S correction was cut off at 15 Å, smoothed with a half-cosine curve over 1 Å. Likewise, the L-J baseline (as a component of the general dimer GAP) was cut off at 15 Å and smoothed over 1 Å (C-C potential only; the other two were simply cut off at 10 Å). Analytical tail corrections were calculated by computing the integral of the missing energy and virial outside the inner cutoff of 14 Å; see Section 4.2.5 for details. The initial configuration for these simulations was a 100-methane cell generated using Packmol [187].

The i-PI simulations used a thermostat based on the generalized Langevin equation (GLE, otherwise known as coloured-noise thermostats), namely the "smart sampling" method of Ceriotti, Bussi, and Parrinello [100]. The parameters were generated at `http://gle4md.org/` using the parameters $t_{\text{opt}} = 10\,\text{ps}$, $N_s = 6$, and $\omega_{\text{max}}/\omega_{\text{min}} = 10^4$ for the thermostat and $t_{\text{opt}} = 2\,\text{ps}$, $N_s = 6$, and $\omega_{\text{max}}/\omega_{\text{min}} = 10^3$, and a piston time constant of $\tau = 100\,\text{fs}$ for the barostat.

The PIMD simulations used the PIGLET [105, 114] thermostat to accelerate convergence to the quantum partition function. The thermostat parameters were generated at the same website, this time using the PIGLET paramters of OPT(H), $N_s = 8$, $\omega_{\text{max}} = 3000\,\text{cm}^{-1}$, $\omega_{\text{max}}/\omega_{\text{min}} = 10^4$, $\hbar\omega/k_B T = 50$, with the ap-

*using LAMMPS stable version from 11 August 2017

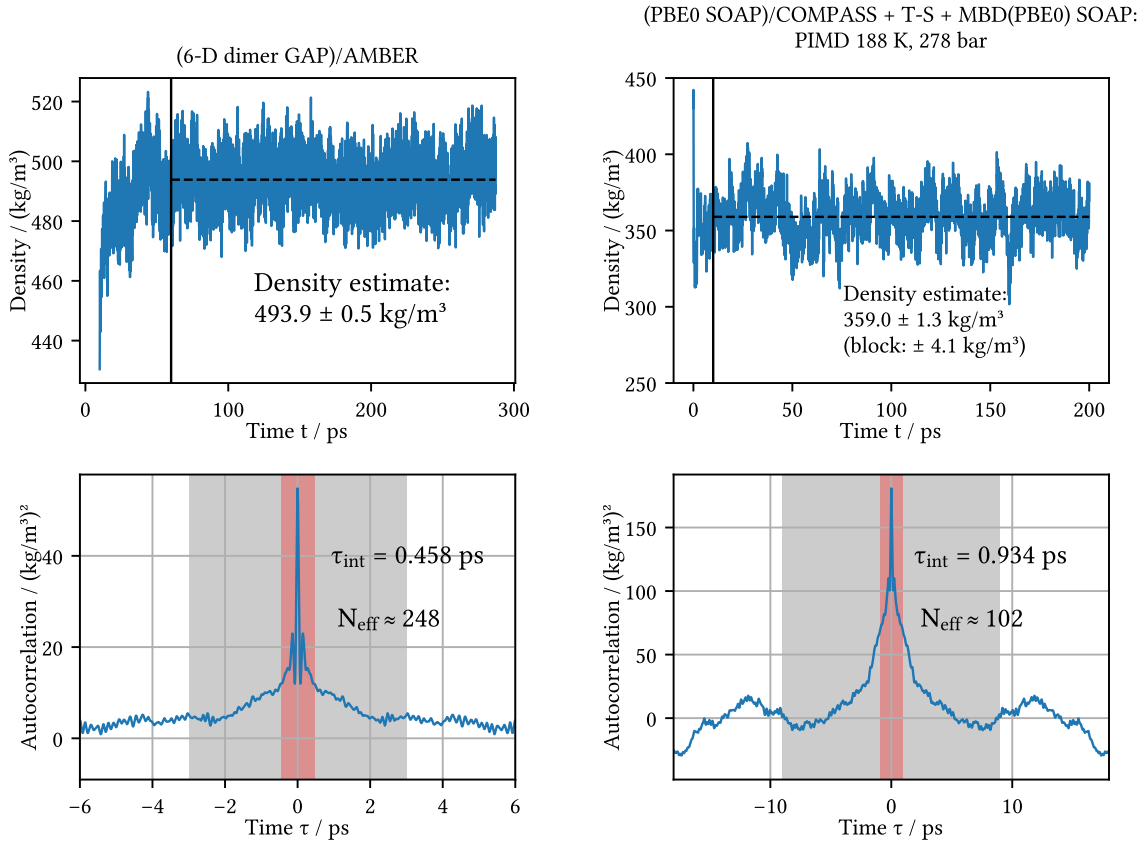


Figure 4.12: Trace of the density of the GAP NPT simulations over time. Averaging was done starting from the vertical solid line. The autocorrelation of the density timeseries (discarding the initial transient) is shown in the bottom plots. The integral τ_{int} of the normalized autocorrelation is a good estimate for the series's correlation time, which in turn can be used to estimate the number of effective *independent samples $N_{eff} = \frac{T}{2\tau_{int}}$ (T is the length of the series being averaged) and the standard error on the mean $\sigma_{corr} = \sqrt{\frac{\sigma_0^2}{2N_{eff}}}$ (where σ_0^2 is the variance of the sample being averaged). The grey region is the integration region and the red shows the correlation time.*

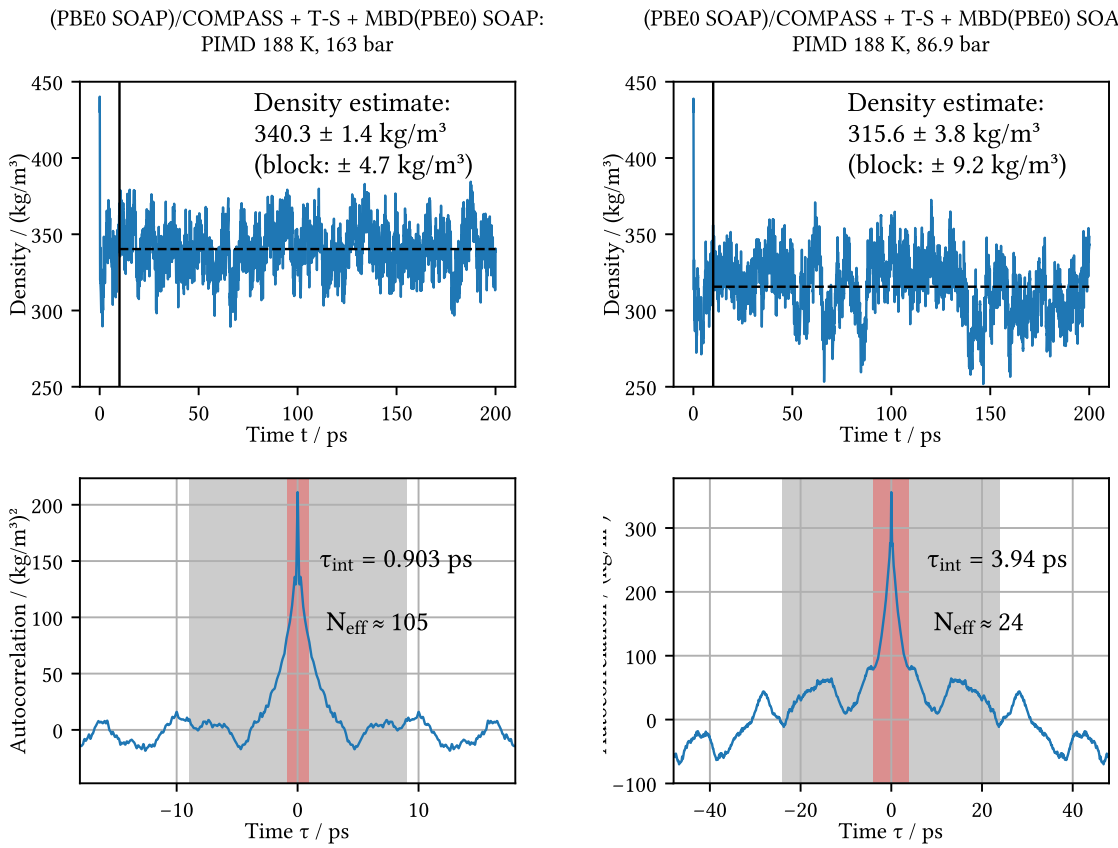


Figure 4.13: Trace of the density of the GAP NPT simulations over time: (PBE0 SOAP)/COMPASS + T-S + MBD(PBE0) SOAP at 188 K. *As in Figure 4.12, timeseries at the top, autocorrelation plots at the bottom. Most of the PIMD simulations are not yet completely equilibrated (as can be seen both in the time trace and the autocorrelation function), so an interim estimate was also computed by splitting the utilized simulation time into 10 blocks and taking the standard deviation of the individual means of the blocks.*

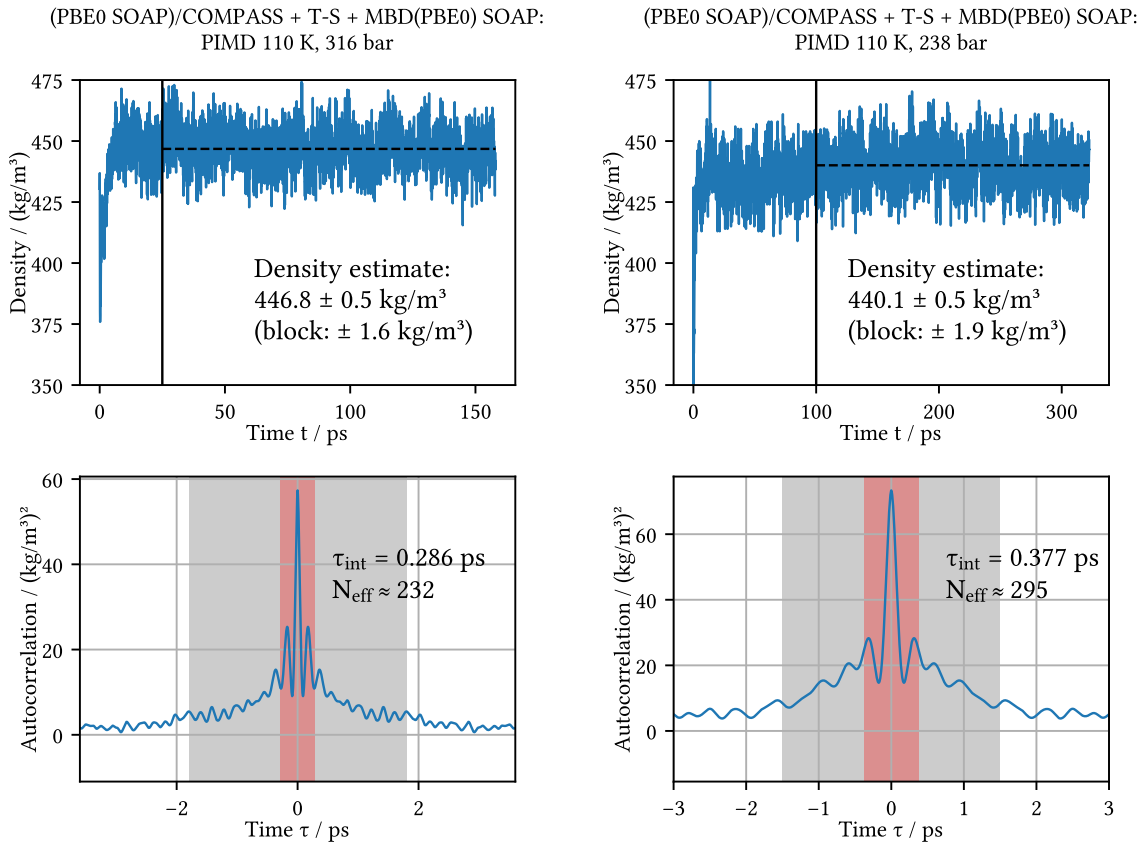


Figure 4.14: Trace of the density of the GAP NPT simulations over time: (PBE0 SOAP)/COMPASS + T-S + MBD(PBE0) SOAP at 110 K*. As in Figure 4.12, timeseries at the top, autocorrelation plots at the bottom.*

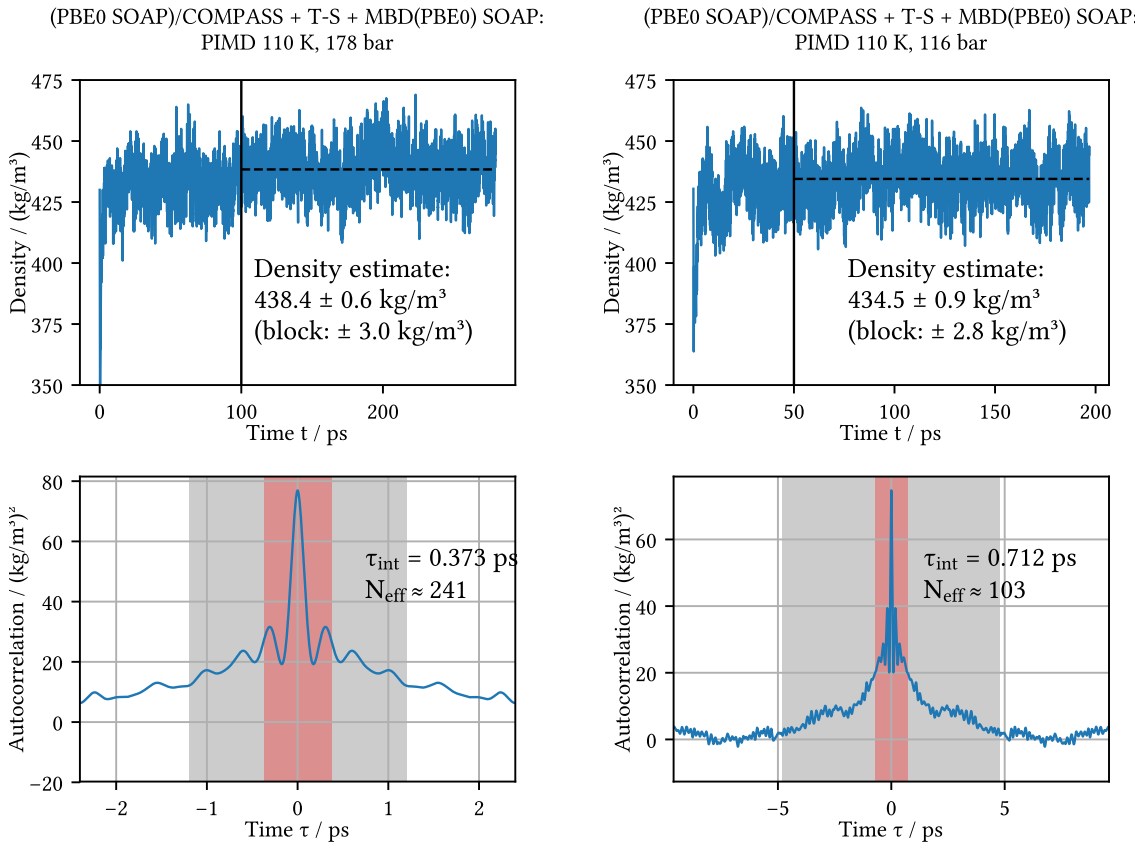


Figure 4.15: Trace of the density of the GAP NPT simulations over time: (PBE0 SOAP)/COMPASS + T-S + MBD(PBE0) SOAP at 110 K*. As in Figure 4.12, timeseries at the top, autocorrelation plots at the bottom.*

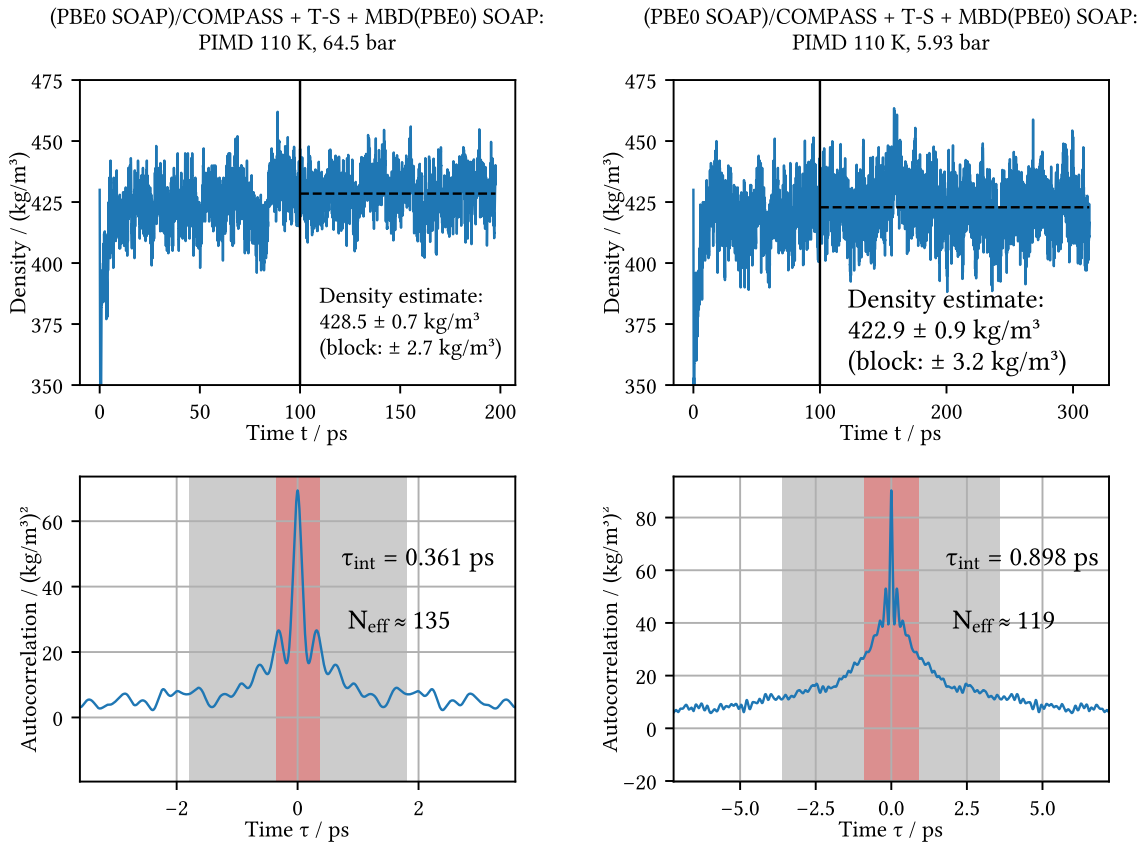


Figure 4.16: Trace of the density of the GAP NPT simulations over time: (PBE0 SOAP)/COMPASS + T-S + MBD(PBE0) SOAP at 110 K*. As in Figure 4.12, timeseries at the top, autocorrelation plots at the bottom.*

propriate temperature T (110 K or 188 K), and with both 12 and 16 beads to verify convergence to the quantum limit; the larger number was used in production simulations. The centroid barostat used the analogous "optimal sampling" method [99, 100] with the same parameters: Potential energy optimized, $N_s = 8$, $\omega_0 = 30\,\mathrm{cm}^{-1}$, and $\omega_{\max}/\omega_{\min} = 10^4$ (resulting in $\omega_{\max} = 3000\,\mathrm{cm}^{-1}$).

The parameter files for the above thermostats, including the matrices used to propagate the generalized Langevin equation, are available in the Apollo repository* as well as on our group's webpage†.

No smooth cutoffs were done in the i-PI simulations due to the necessity of interfacing with QUIP through LAMMPS; analytical tail corrections were applied, though. The initial configuration was prepared with an initial 10 ps NVT equilibration using the '(PBE0 SOAP)/COMPASS + T-S + MBDGAP' potential; this configuration was used for both the classical and PIMD simulations at that temperature. All GAP MD simulations were done with a timestep of 0.5 fs.

Each run had a certain amount of initial equilibration time discarded from its trajectory, depending chiefly on the potential, the thermostat, and the temperature. The ten pairs of plots in Figures 4.12 through 4.16 show how the average density and standard error were obtained from the time evolution of the density for the 6-D dimer GAP simulation and for the '(PBE0 SOAP)/COMPASS + T-S + MBD(PBE0) SOAP' PIMD simulations. The standard error was obtained by integrating the autocorrelation of the density timeseries as described in Sokal [103]. However, many simulations showed extremely long correlation times and were not fully equilibrated within the available simulation time, rendering the autocorrelation method inapplicable. Therefore, to estimate the error incurred due to the large-scale fluctuations still observed, the simulation time utilized for

*https://doi.org/10.17863/CAM.26364
†http://www.libatoms.org/Home/DataRepository

averaging was split into ten equal blocks (corresponding approximately to the timescale of fluctuations still observed), the mean value was computed within each of those blocks, and the final error estimate computed as the standard deviation of those means. These error estimates are displayed as error bars on the '(PBE0 SOAP)/COMPASS + T-S + MBD(PBE0) SOAP' PIMD simulation results in Figure 4.6.

Analytical potentials

The analytical potentials were run in LAMMPS [124]* with a Langevin thermostat [182] and a Nosé-Hoover barostat [93, 94, 183–185] with the MTK correction [101], both using a time constant of 0.1 ps, and an initial configuration of 200 methane molecules generated using Packmol [187] and relaxed with the OPLS-AA [34] forcefield. All simulations used analytical tail corrections to account for the otherwise-neglected dispersion energy beyond their cutoffs [20, 41] (for the L-J baseline this was only done for the C-C potential). For potentials with a Coulomb component (OPLS/AMBER and COMPASS), the contributions beyond the cutoff were calculated with the particle-particle particle-mesh (PPPM) method [188]. The MD timesteps were 1 fs for TraPPE, 0.5 fs for the Li-Chao L-J and OPLS/AMBER at 110 K, and 0.1 fs for the others (OPLS/AMBER at 188 K, the L-J baseline, and COMPASS).

The potentials themselves used L-J cutoffs (and Coulomb cutoffs for OPLS/AMBER and COMPASS) of 10 Å, except for TraPPE, where the cutoff of 14 Å recommended on the website was used instead. The two pairwise L-J fits (the L-J baseline and Li-Chao) both were added to the AMBER intramolecular terms to give complete liquid methane potentials.

Equilibration and run times again varied between the potentials based on the

*using LAMMPS stable version from 5 Oct 2015

rate of convergence of the density, although the same times were used throughout an isotherm. The times are summarized in Table 4.3.

Potential	Temperature / K	Equilib. time / ps	Run time / ps
TraPPE	110	100	400
	188	100	400
OPLS/AMBER	110	25	75
	188	100	400
L-J baseline	110	50	50
	188	300	200
Li-Chao L-J	110	100	400
	188	400	100

Table 4.3: Equilibration and run times for the analytical potentials

4.2.5 Tail corrections with a smooth cutoff

For all the NPT simulations done for this work it was found important to incorporate tail corrections to account for the missing pressure neglected by cutting off the long-range dispersion potentials (the sixth-power part of L-J and T-S). These corrections can also be applied in the case where the potential is smoothed to zero before the cutoff, though the resulting integrals become more difficult to evaluate.

The expression for the missing pressure in a potential that is cut off with a smoothing function that starts at r_{in} and ends at r_{out} is, by straightforward extension of the formulae in [20, 41]:

$$\begin{aligned} p_{\text{exact}} - p_{\text{cut}} &= \frac{1}{6}\sum_{i=1}^{n_{\text{typ}}}\rho_i \sum_{j=1}^{n_{\text{typ}}}\rho_j \int_{r_{\text{in}}}^{\infty} r\frac{\mathrm{d}}{\mathrm{d}r}(\phi_{ij}(r)-\phi_{ij,\text{cut}}(r))4\pi r^2 g_{ij}(r)\,\mathrm{d}r \\ &= \frac{1}{6}\sum_{i=1}^{n_{\text{typ}}}\rho_i \sum_{j=1}^{n_{\text{typ}}}\rho_j \left[\int_{r_{\text{in}}}^{r_{\text{out}}} r\frac{\mathrm{d}}{\mathrm{d}r}\left(\phi_{ij}(r)(1-S(r))\right)4\pi r^2 g_{ij}(r)\,\mathrm{d}r \right. \\ &\qquad \left. + \int_{r_{\text{out}}}^{\infty} r\frac{\mathrm{d}\phi_{ij}(r)}{\mathrm{d}r}4\pi r^2 g_{ij}(r)\,\mathrm{d}r\right] \end{aligned} \tag{4.2}$$

where i and j run over the atom types, $g_{ij}(r)$ is the corresponding pair correlation function, ρ_i and ρ_j are the number densities of each type, and $S(r)$ is the switching function that takes the potential to zero. This function must be continuous and take values $S(r_{in}) = 1$ and $S(r_{out}) = 0$; its derivative must also be continuous and take values $S'(r_{in}) = S'(r_{out}) = 0$.

If we assume the pair correlation function $g_{ij}(r) \approx 1$ beyond $r = r_{\text{in}}$ (which is usually a good approximation for liquids at relatively large distances), the improper integral in the second term of Equation (4.2) can be evaluated analytically for simple (e.g. inverse-power) forms of the pair potential $\phi_{ij}(r)$. Using a sixth-power dispersion form $\phi_{ij}(r) = -C_{ij}^6 r^{-6}$, the improper integral becomes $\int_{r_{\text{out}}}^{\infty} 24\pi C_{ij}^6 r^{-4}\,\mathrm{d}r = 8\pi r_{\text{out}}^{-3}$ and we have:

$$p_{\text{exact}} - p_{\text{cut}} \approx p_{\text{corr}} = \frac{1}{6}\sum_{i=1}^{n_{\text{typ}}}\rho_i \sum_{j=1}^{n_{\text{typ}}}\rho_j C_{ij}^6 \left[\int_{r_{\text{in}}}^{r_{\text{out}}} -r\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1-S(r)}{r^6}\right)4\pi r^2\,\mathrm{d}r + 8\pi r_{\text{out}}^{-3}\right]. \tag{4.3}$$

Applying integration by parts to the remaining integral gives

$$p_{\text{corr}} = \frac{1}{6}\sum_{i=1}^{n_{\text{typ}}}\rho_i \sum_{j=1}^{n_{\text{typ}}}\rho_j C_{ij}^6 \left[-4\pi r_{\text{out}}^{-3} + \int_{r_{\text{in}}}^{r_{\text{out}}}\left(\frac{1-S(r)}{r^6}\right)12\pi r^2\,\mathrm{d}r + 8\pi r_{\text{out}}^{-3}\right] \tag{4.4}$$

and simplifying and rearranging leaves us with

$$
\begin{aligned}
p_{\mathrm{corr}} &= \frac{1}{6} \sum_{i=1}^{n_{\mathrm{typ}}} \rho_i \sum_{j=1}^{n_{\mathrm{typ}}} \rho_j C_{ij}^6 \left[4\pi r_{\mathrm{in}}^{-3} - \int_{r_{\mathrm{in}}}^{r_{\mathrm{out}}} 12\pi r^{-4} S(r)\,\mathrm{d}r \right] \\
&= \frac{2\pi}{3} \sum_{i=1}^{n_{\mathrm{typ}}} \rho_i \sum_{j=1}^{n_{\mathrm{typ}}} \rho_j C_{ij}^6 \left((1-\lambda) r_{\mathrm{in}}^{-3} + \lambda r_{\mathrm{out}}^{-3} \right)
\end{aligned} \tag{4.5}
$$

with

$$
\lambda = \frac{3 \int_{r_{\mathrm{in}}}^{r_{\mathrm{out}}} r^{-4} S(r)\,\mathrm{d}r}{r_{\mathrm{in}}^{-3} - r_{\mathrm{out}}^{-3}}
$$

This form isolates the problematic integral $\int_{r_{\mathrm{in}}}^{r_{\mathrm{out}}} r^{-4} S(r)\,\mathrm{d}r$ which, depending on the form of the switching function $S(r)$, may be complicated or impossible to do analytically. For practical simulations, however, λ can simply be precomputed using any suitable numerical method for a given value of r_{in}, r_{out}, and $S(r)$; this value can be used throughout the simulation.

We can also see that Equation (4.5) takes the form of a linear interpolation between the tail correction with a cutoff at r_{in} and the correction with a cutoff of r_{out}; if we choose an $S(r)$ whose values are bounded between 0 and 1 then λ will likewise be bounded between 0 and 1.

On closer inspection, however, we can see that the interpolation endpoints are *not* the values the tail correction would take with a sharp cutoff: If we start with Equation (4.3) and take $S(r) = 1$ for $r < r_{\mathrm{out}}$, we get instead:

$$
p_{\mathrm{exact}} - p_{\mathrm{cut,sharp}} \approx \frac{4\pi}{3} \sum_{i=1}^{n_{\mathrm{typ}}} \rho_i \sum_{j=1}^{n_{\mathrm{typ}}} \rho_j C_{ij}^6 r_{\mathrm{out}}^{-3}, \tag{4.6}
$$

which is the correction Tildesley and Allen [20] give for the sixth-power part of an L-J potential, and *twice* the value we would get from Equation (4.5) by letting $\lambda = 1$. This discrepancy is due to the extra $-4\pi r_{\mathrm{out}}^{-3}$ term that emerged from the integration by parts in Equation (4.4), and it can be physically interpreted

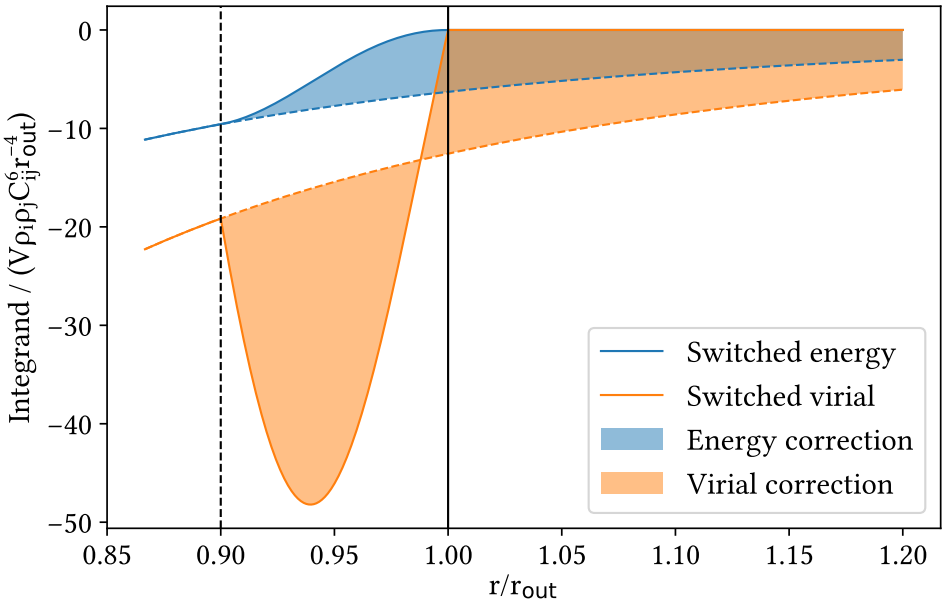


Figure 4.17: Illustration of the integrals for the energy and virial tail corrections for a potential of the form $\phi_{cut}(r_{ij}) = C^6_{ij} r^{-6}_{ij} S(r)$. The dashed line corresponds to the potential with no cutoff ($f_E(r) = -\frac{1}{2} r^{-6} \cdot 4\pi r^2$ for the energy and $f_W(r) = \frac{1}{6} r \frac{\mathrm{d}}{\mathrm{d}r} r^{-6} \cdot 4\pi r^2$ for the virial), while the solid line corresponds to the potential multiplied by the switching function $S(r)$. The shaded areas between the dashed and solid lines depict the integrals of Equations (4.3) *(times $-V$) and* (4.7)*; the area below the dotted line is negative. The energy and virial integrals (areas) are equal.*

as follows: With a sharp cutoff, an atom feels no force as it crosses the cutoff; the force just changes discontinuously from $-\phi'(r \to r^-_{\text{out}})$ to zero. With a smooth cutoff, however, the switching function provides an extra gentle inward force as the atom exits the transition region. The extra virial due to this force provides an *effective* tail correction to the system's overall pressure, albeit only about half of the difference of the pressure with the sharp cutoff to the pressure of the ideal system with an infinite cutoff; Figure 4.17 provides an illustration of this idea.

Tail corrections may also be computed, using the same method as above, for the energy. Although they do not affect the simulation dynamics in any way, they may be used for accurate bookkeeping and later analysis. The expression

is [20, 41]

$$\begin{aligned}
E_{\text{exact}} - E_{\text{cut}} &= \frac{V}{2} \sum_{i=1}^{n_{\text{typ}}} \rho_i \sum_{j=1}^{n_{\text{typ}}} \rho_j \int_{r_{\text{in}}}^{\infty} (\phi_{ij}(r) - \phi_{ij,\text{cut}}(r)) 4\pi r^2 g_{ij}(r) \, \mathrm{d}r \\
&\approx \frac{V}{2} \sum_{i=1}^{n_{\text{typ}}} \rho_i \sum_{j=1}^{n_{\text{typ}}} \rho_j C_{ij}^6 \left[\int_{r_{\text{in}}}^{r_{\text{out}}} -(1 - S(r)) 4\pi r^{-4} \, \mathrm{d}r + \int_{r_{\text{out}}}^{\infty} -4\pi r^{-4} \, \mathrm{d}r \right] \\
&= \frac{V}{2} \sum_{i=1}^{n_{\text{typ}}} \rho_i \sum_{j=1}^{n_{\text{typ}}} \rho_j C_{ij}^6 \left[-\frac{4\pi}{3} r_{\text{in}}^{-3} + \int_{r_{\text{in}}}^{r_{\text{out}}} 4\pi r^{-4} S(r) \, \mathrm{d}r \right] \\
&= -\frac{2\pi V}{3} \sum_{i=1}^{n_{\text{typ}}} \rho_i \sum_{j=1}^{n_{\text{typ}}} \rho_j C_{ij}^6 \left((1 - \lambda) r_{\text{in}}^{-3} + \lambda r_{\text{out}}^{-3} \right)
\end{aligned} \tag{4.7}$$

with λ defined as before. This is – coincidentally, with the r^{-6} potential – identical to the virial correction, i.e. the pressure correction in Equation (4.5) multiplied by $-V$.

4.3 Discussion

The results presented in this chapter essentially show us the level of detail we need to capture in order to obtain an acceptable prediction of the density from first principles, that is, from quantum mechanics rather than empirical fitting. The dimer error measure introduced in Section 3.1.2 here establishes the connection from rigorous quantum chemical theory to accurate property predictions, even when using more approximate dispersion-corrected DFT methods – but the correlation between dimer error and density error only holds when the potential energy surface is fitted in a systematic way, with GAP. Future work on systematic potentials for liquids will likely use this eror measure in combination with a few others to capture what the dimer measure misses, mainly, many-body and intramolecular effects.

The main questions determining how well this methodology will extend to other molecules are, first, how well dispersion-corrected DFT describes their properties, and second, how well these energy contributions (especially the dispersion correction) can be represented in terms of *local* parameters. As to the first question, the MBD method used in this work has been tested [176] on the S66 database [189] (an extension of the standard S22 database [149]), which includes various types of dispersion-bound alkane dimers. These databases of CCSD(T) interaction energies may serve as useful references, in combination with additional calculations on longer molecules and different orientations, for testing the applicability of the MBD method specifically to the interaction of longer alkanes.

As to the second question, the locality of the parameters underlying the MBD correction (relative Hirshfeld volumes) has not been extensively studied.

It is usually accepted, however, that properties computed directly from the electron density decay strongly with distance, a property known as "nearsightedness" [190–192]. For saturated alkanes in particular, the electron density matrix decays exponentially with distance [193]. This locality is the justification for methods that fit the electron density (or some proxy, such as partial atomic charges) based only on a local chemical environment [69, 70]. One would expect the same approach to work for the Hirshfeld volumes, being based on the atomic contribution to the electron density which itself is made of exponentially decaying components, but further study is needed to confirm and quantify this locality. It will also be necessary to test whether the locality of the MBD *correction* (MBD minus T-S energy) observed for condensed-phase methane also holds for longer molecules; if it does not hold, that would require a change in our current approach to incorporating the MBD dispersion model in our simulations.

Finally, this potential must be put through more stringent tests – namely, it must be able to predict the transport properties (diffusivity and viscosity) as described in Section 2.3. Such tests may expose weaknesses in the potential or the methodology that were not detected with this first round of tests.

4.3.1 Dimer GAP with flexible monomers

A further technical point related to the 6-D dimer GAP from Section 4.2.1 was not elaborated in the paper: This GAP was fit using a database of dimers with *rigid* monomer geometries; each methane molecule in the set was fixed to its CCSD(T)-optimized geometry. This means the training points only covered a space of dimension $d = 6$ (one for the C-C distance, five to describe the mutual dimer orientation), rather than the full $d = 24$. This makes for a more tractable fit if we can assume that monomer flexibility has a negligible impact on the dimer

PES; that is, if we can separate out the monomer potential from the dimer interaction. In simulations where monomer flexibility is required (as in the density simulations above, where the intramolecular PES *was* found to have a significant influence), an intramolecular model is added to the dimer GAP. The dimer GAP can still predict energies and forces of geometries in the extended, flexible-monomer space: Since the training samples are confined to a 6-dimensional hyperplane within the larger 24-dimensional space, the predicted value for a given flexible geometry is approximately what the GAP would pick for the closest rigid geometry to the given flexible one. It is not exactly the same as the GAP prediction for the closest rigid geometry, since the Gaussian basis functions also decay in the direction perpendicular to the "training hyperplane," but the typical variation in interatomic distances due to intramolecular distortion is so much smaller than the typical variation in these distances due to the relative movement of the monomers as they explore the dimer PES that this decay can be taken to be negligible*.

The error that the dimer GAP makes by ignoring intramolecular distortions can be quantified by taking a sample of distorted (flexible) geometries and measuring its error against the same quantum reference. Two samples of distorted geometries were taken for this purpose, both sampled from bulk methane MD simulations at 110 K and 316 bar (see above for QUIP simulation details) using the 6-D dimer GAP for intermolecular interactions. Two intramolecular potentials were used; the "flexible" one was the AMBER [32] harmonic forcefield, and the other was a stiffer version of AMBER created by multiplying both force constants (bond stretching and angle bending) by 25, hence "AMBERx25". This gives

*The rate of this decay is controlled by the typical length scale, denoted σ_i in the descriptor definition above and `theta` in the GAP parameter command lines. If the potential had been designed to include intramolecular effects in the fit, different length scales would have been used for intramolecular vs. intermolecular distances. In this case, though, the length scales were set the same for all distances.

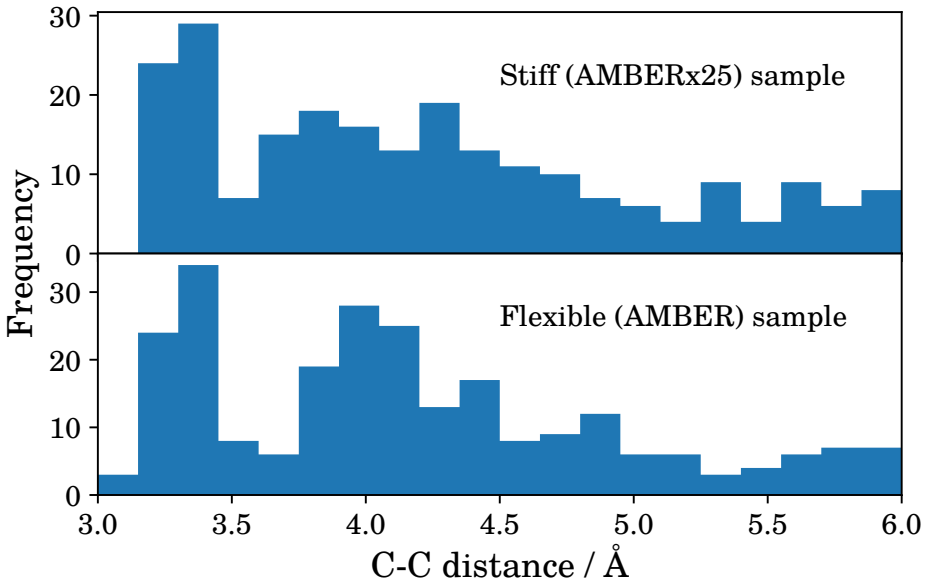


Figure 4.18: Distribution of dimer C-C distances in the two verification sets: One generated with a stiff intramolecular forcefield, one with a more flexible one.

us two degrees of distorsion of the molecule, with the flexible (AMBER) forcefield hopefully giving us distortions of the scale that will be encountered in later, more realistic simulations. The distributions of C-C distances within these samples, shown in Figure 4.18, were chosen to be similar to the distribution of the original training set.

The 6-D dimer GAP predictions (without the CCSD(T) correction) are compared in Figure 4.19 with MP2/AVQZ energies computed on each set. Compare the RMS errors of 433 µeV per molecule on the stiff set and 469 µeV per molecule on the flexible set with the original 367 µeV per molecule on the rigid training set. It therefore appears that although the GAP does slightly suffer in accuracy when going from rigid to flexible monomers, the increase in error is less than 30 %, still leaving the flexible error below 0.5 meV per molecule and consistently more than an order of magnitude below the target energies themselves. The 2b GAP for coupled-cluster correction sees a similar increase, going up to 136 meV

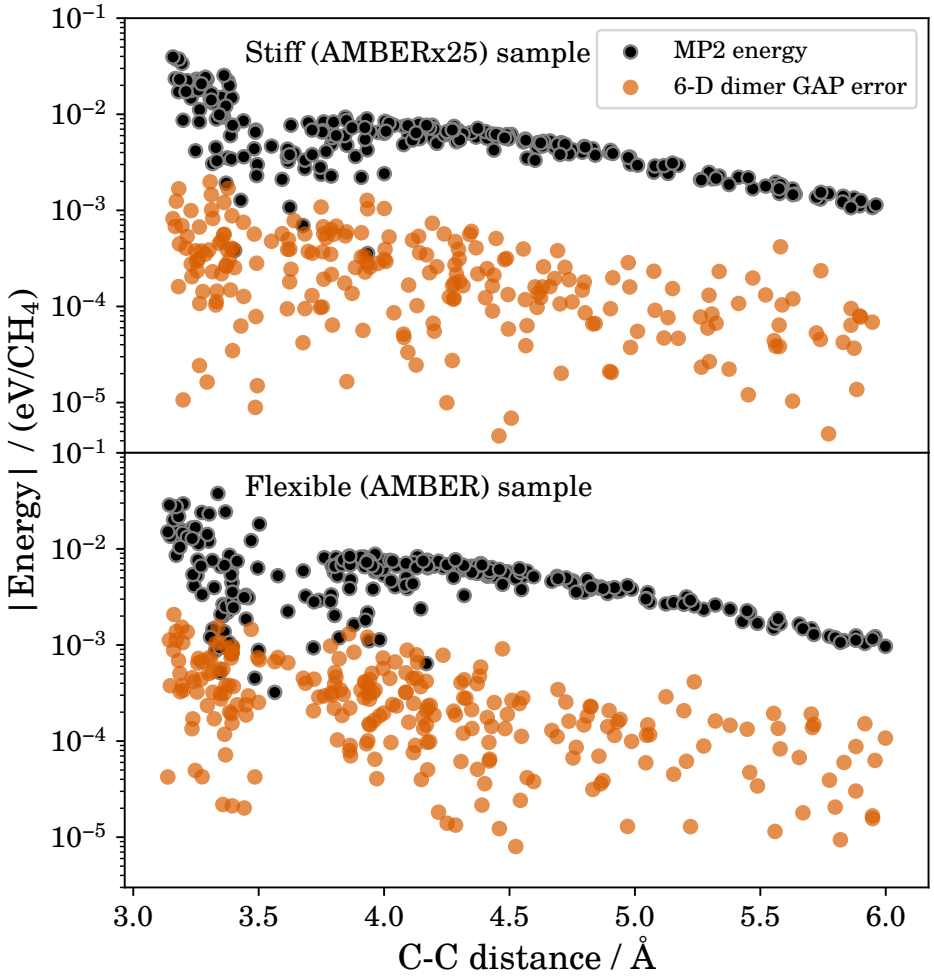


Figure 4.19: Errors made by the 6-D dimer GAP (MP2 only, without the CCSD(T) correction) on the stiff (top) and flexible (bottom) dimer test sets. RMS errors are 433 µeV *per molecule on the stiff set and* 469 µeV *per molecule on the flexible set.*

per molecule on the flexible (AMBER) test set from 80.6 meV per molecule on the rigid set, but its error remains smaller than that of the dimer GAP on either the rigid or flexible test sets.

We can therefore conclude that the current strategy of training a GAP on rigid monomers and adding the intramolecular potential as a later correction results in an acceptable level of error, although it could be worth revisiting this approximation if the dimer GAP is to be used for future applications.

Chapter 5

Intramolecular potential

In order to extend the potential to larger, more flexible molecules, we will first need to turn our focus back to the one-body, or intramolecular, term of Equation 1.3. We have already seen that the intramolecular potential has considerable influence on the bulk properties, even for a system as simple as methane (cf. the change in the predicted density of the SOAP-GAP models when COMPASS was substituted for AMBER). The accuracy of the intramolecular model becomes even more important in light of quantum nuclear effects, since important quantum effects such as the zero-point vibrational energy (ZPVE) depend almost entirely on the intramolecular potential [107, 110]. Finally, with extended alkanes, the model of intramolecular motions (especially the torsional rearrangements) has a large influence on the prediction of transport properties such as the diffusivity and viscosity [2, 27].

We will therefore extend our strategy of obtaining a systematic, accurate, best-possible fit of the quantum potential energy surface to the intramolecular component of the energy. As mentioned before in Section 1.2.2, several modern forcefields have already made progress in this direction with more flexible functional forms and the use of quantum mechanical fitting data; it is for this reason

that COMPASS [41] was used as the intramolecular component of the GAP in most of the later density simulations. However, the accuracy of forcefields like COMPASS is still limited by two major factors: First, they still use fixed functional forms (even if those forms are more flexible than their predecessors), and second, they attempt to fit a large and diverse set of chemical compounds – and, in combination with fixed functional forms, this goal necessarily limits their accuracy on any one type of molecule. Therefore, we can expect to get a much more accurate fit with a GAP fit only to alkane energies.

5.1 Random sampling GAP

The first steps toward an intramolecular hydrocarbon GAP were taken in my master's thesis [46]. First, the locality of the energy was systematically studied in long linear alkane, alkene, and ketone chains to find the maximum accuracy a model could have using only local information. Subsequently, various GAP models were fitted to DFT energies of a thermal sample of a single long alkane (and later some derived alkenes) using three different desciptors: Two-body (2b), two-body plus three-body (2b+3b), and SOAP*. The geometries for the training set were generated in much the same way as for the intermolecular SOAPs of Chapter 4, from samples of MD simulations using a simpler potential (in this case, DFTB [194], a fast approximation of DFT) at two temperatures, one slightly above room temperature (350 K) and one much higher (1500 K) to explore a larger proportion of conformational space. The GAPs trained on colder, saturated alkanes were assessed using the total-energy correlation plots in Figure 5.1: COMPASS scores somewhere in between the 2b+3b GAP and the SOAP-NN GAP. At the higher-temperature sample (Figure 5.2), COMPASS is about equal in accur-

*with a cutoff set to approximately include nearest neighbours only, hence "SOAP-NN"

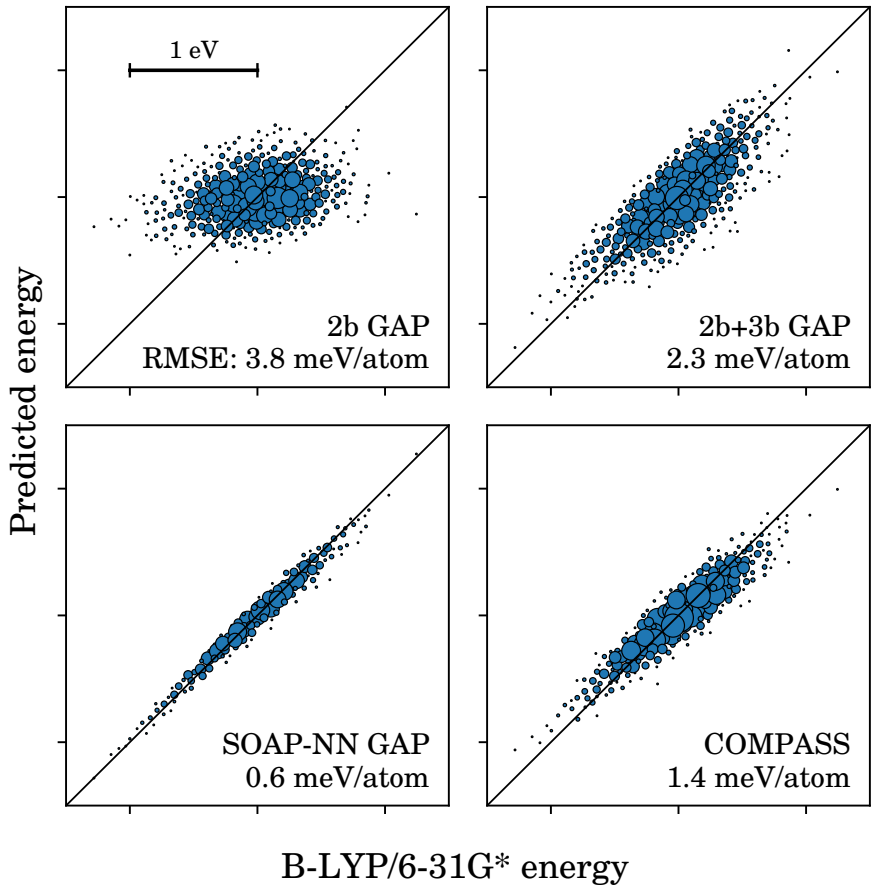


Figure 5.1: Comparison of the predictions of three random-sampling GAPs, trained on linear alkanes (saturated) with different descriptors, and the empirical model COMPASS [41]. As in Figure 3.6, only a representative sample of points is shown with sizes scaled according to the number of points they represent.

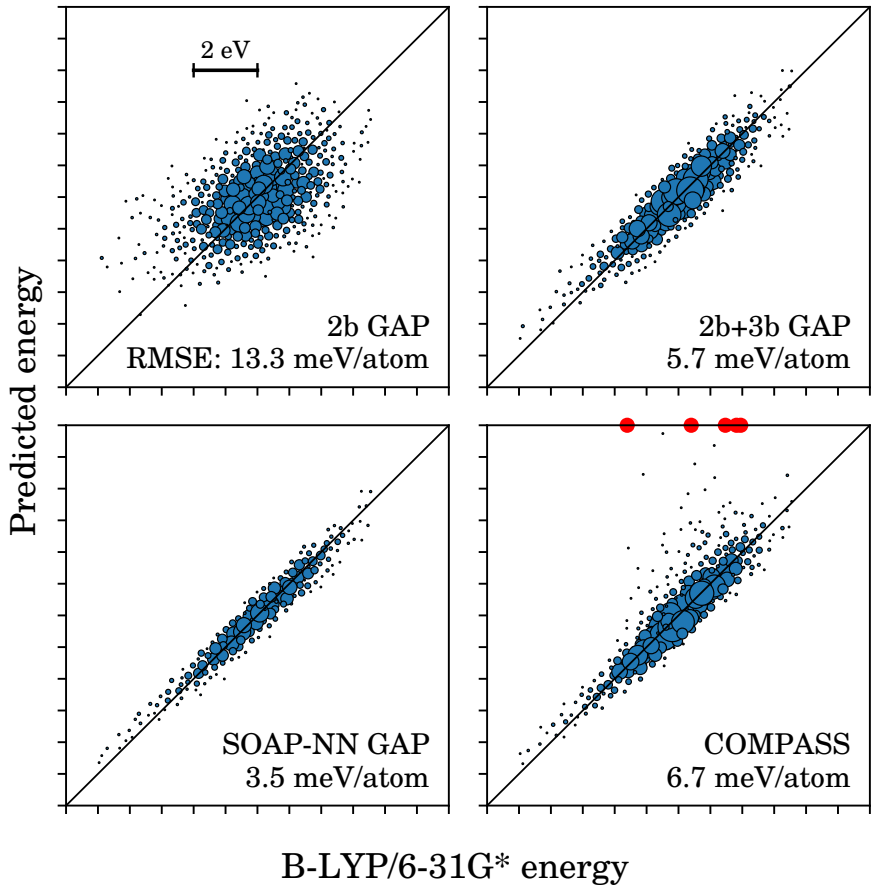


Figure 5.2: Comparison of the predictions of three random-sampling GAPs, trained on linear alkanes (saturated) sampled at a much higher temperature than in Figure 5.1, and the empirical model COMPASS. The red dots in the COMPASS plot represent points outside the plotting frame.

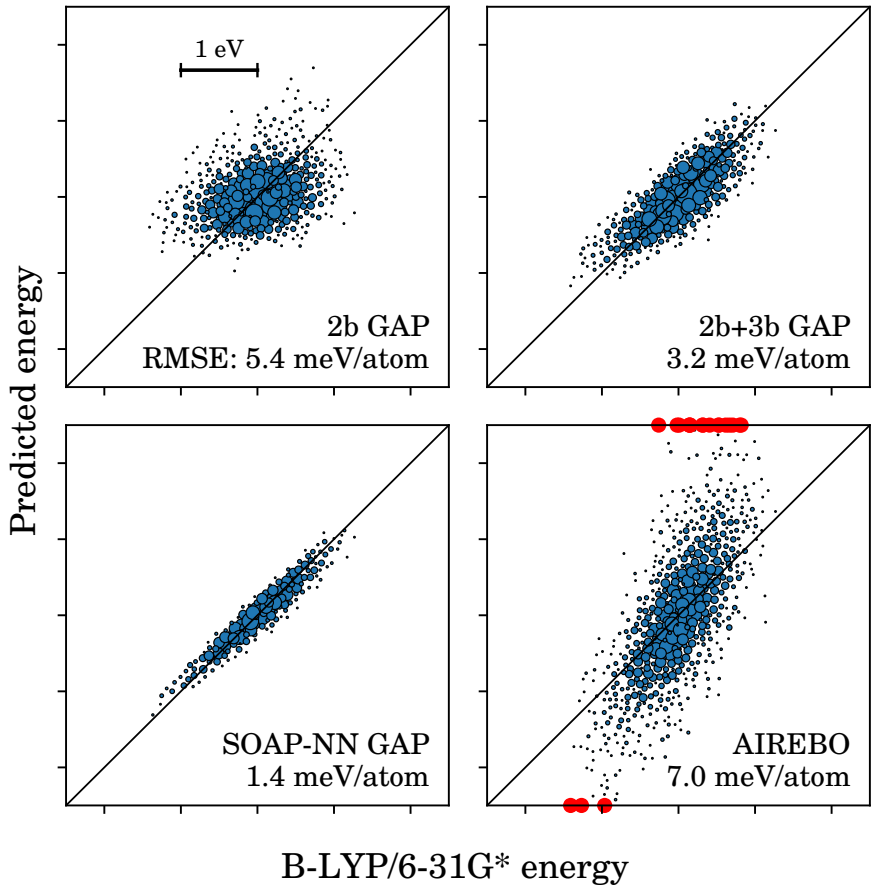


Figure 5.3: Comparison of the predictions of random-sampling GAPs trained on linear unsaturated hydrocarbon chains at the original (lower) temperature, and the empirical many-body model AIREBO [36]. The red dots in the AIREBO plot represent points outside the plotting frame. This is a corrected version of Figure 4.2 from my master's thesis [46] (the original version was affected by an error in LAMMPS that was fixed on 5 October 2015).

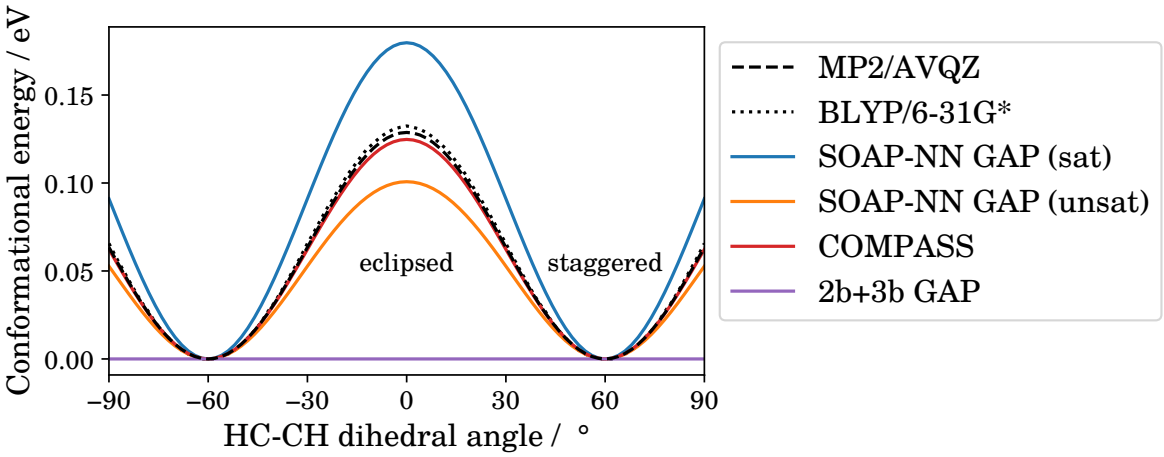


Figure 5.4: Predictions of the random-sampling GAPs (colder training set) on the torsional curve of ethane, compared with COMPASS. Note that the 2b+3b GAP has no sensitivity to four-body motions and so cannot account for any torsional energy difference.

acy with the 2b+3b GAP but with significantly more outliers.

Finally, another set of models was trained on a set that included unsaturated linear hydrocarbons (alkenes). The spread of energies (shown in Figure 5.3) was slightly larger than on the colder unsaturated set; the SOAP's accuracy was again reduced, even though it should have been able to pick up on the difference between a single-bonded and double-bonded carbon local environment.

More serious shortcomings were revealed later when these models were tested on the torsional energy curves more familiar to chemists and developers of empirical forcefields. As Figures 5.4 and 5.5 show, both the saturated and unsaturated SOAP-NN GAPs make large errors in the torsional profile, significantly under- and overpredicting the barriers with no systematic trend. One of the possible reasons these shortcomings were not found earlier is that the torsional energy makes up a relatively small component of the *total* energy, so the total energy errors shown in Figures 5.1 should be dominated by those from the the other energy components. But comparing the Boltzmann-averaged (350 K) energy error of the SOAP-NN GAP (cold saturated training set) on the butane

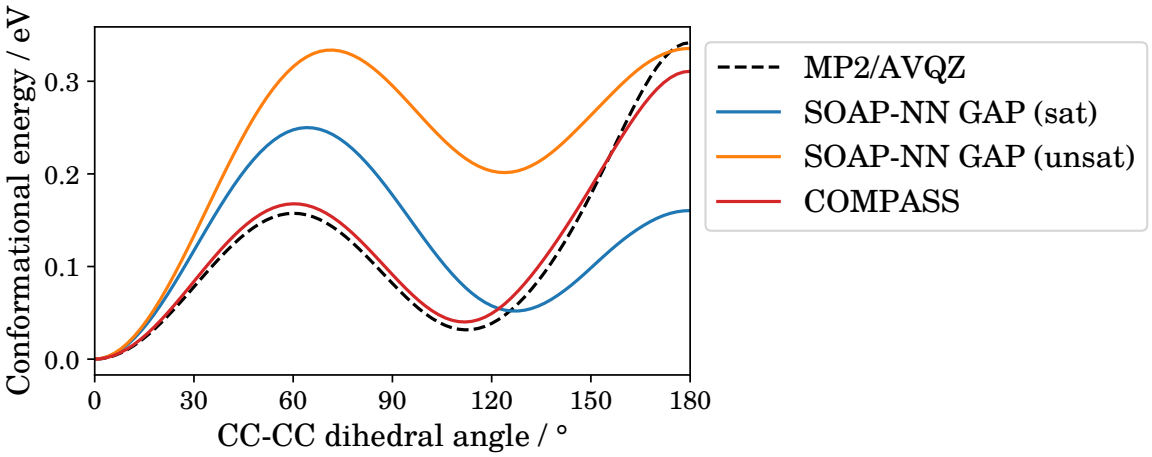


Figure 5.5: Predictions of the random-sampling GAPs (colder training set) on the CC-CC torsional curve of butane.

curve, which equates to 12 meV per atom for the 24-carbon chain in the training set, with the actual energy error of the fit, 0.6 meV per atom, starkly contradicts this explanation. It is hard to imagine that the SOAP-NN GAPs could have reached this level of accuracy without error cancellation, overfitting, or a significant devation from the expected Boltzmann distribution of torsional angles in the training set. Further study will be needed to clarify this discrepancy.

Another limitation in the random-sampling GAP is that the descriptors were not designed to capture torsional motions in the first place; the 2b+3b GAP is entirely insensitive to torsions and the SOAP-NN was only meant to capture each atom's (approximate) set of first-nearest neighbours, whereas torsions formally require knowledge of at least second-nearest neighbours. The remainder of this chapter is dedicated to efforts to fit GAPs with improved torsional energy curves, thus giving chemically more realistic properties.

5.2 Hessian GAP

Among the reasons that the random-sampling approach was so much less successful for the local GAP than for the intermolecular SOAP-GAPs of Chapter 4 was that the intramolecular potential energy surface has additional structure that should be exploited: The minima and barriers of the torsional (or, more generally, conformational) potential energy surfaces are well known, well studied [12, 13, 195], and generally occur at or near high-symmetry points. In fact, most analytical forcefields do exploit this structure, fitting directly to torsional energy curves [6, 7, 33, 48]; their use of this structure along with the locality of the energy allows them to model the enormous combinatorial array of possible alkane chain conformations with a minimal number of samples and resulting parameters. In a way, nearly all empirical forcefields are already using second-derivative information to fit harmonic bond and angle parameters; a recent study [50] has made this connection more explicit by fitting directly to the quantum-mechanically derived Hessian matrix.

We can incorporate this structure into an improved GAP by simply adding these known minima (and some maxima or transition states) into the training set. But the energies and forces at these points alone are not enough to determine the potential energy surface at intermediate points. Instead of expanding our set of training geometries in this high-dimensional space (perhaps with more random sampling), we can incorporate additional information in the form of the *curvature* of the potential energy surface at each training point. The curvature is computed as the second derivative (the Hessian matrix) of the PES, so the family of GAPs developed using this information will be called the “Hessian GAPs.”

5.2.1 Theory

The standard GAP theory already provides for training on properties obtained from linear operators on the set of local energies in the system, as long as they can be expressed as linear operators on the corresponding descriptors: Total energies and forces are both examples, with the total energy being the sum of local atomic contributions and each force component being the derivative of this sum with respect to some atomic coordinate q_α. The covariance between a total energy observation E_M and a force observation $\frac{\partial E_N}{\partial q_\alpha}$ is then [74]:

$$\mathrm{Cov}\left(\frac{\partial E_N}{\partial q_\alpha}, E_M\right) = \mathrm{Cov}\left(\frac{\partial}{\partial q_\alpha}\sum_j \varepsilon_{Nj}, \sum_k \varepsilon_{Mk}\right) = \delta^2 \sum_{j\in N}\sum_{k\in M} \nabla_{\mathbf{d}_j} k(\mathbf{d}_j, \mathbf{d}_k)\cdot\frac{\partial \mathbf{d}_j}{\partial q_\alpha} \tag{5.1}$$

(symbols used as defined in Equation (3.3), except here the function scale δ is taken outside of the covariance kernel). Note that differentiation with respect to the coordinate q_α in configuration N does not apply to the descriptor $\mathbf{d}_k$ in configuration M. The covariance between two force observations is analogous, but uses the matrix of derivatives of $k(\mathbf{d}_j, \mathbf{d}_k)$ with respect to the components of both $\mathbf{d}_j$ and $\mathbf{d}_k$; see [74] for the full formula. The covariance matrix $\mathbf{C}$ now contains the covariances between all total energy and force observations. The vector of covariances for a new environment $\mathbf{d}_i$ (which are multiplied by the weights $\boldsymbol{\alpha}$ to obtain the predicted local energy, see Equation (3.3)) becomes the vector of covariances of the new local environment with the *observed* total energies and forces.

The Hessian fitting method is a natural extension to this theory that has recently been developed in our group: In principle, to obtain the covariance of e.g. a total energy with a *second* derivative of another total energy, we could

differentiate Equation (5.1) again under application of the chain rule:

$$\mathrm{Cov}\left(\frac{\partial^2 E_N}{\partial q_\alpha q_\beta}, E_M\right) = \delta^2 \sum_{j\in N}\sum_{k\in M}\left(\left(\frac{\partial \mathbf{d}_j}{\partial q_\alpha}\right)^{\mathrm{T}}\left(\nabla_{\mathbf{d}_j} k(\mathbf{d}_j,\mathbf{d}_k)\nabla^{\mathrm{T}}_{\mathbf{d}_j}\right)\frac{\partial \mathbf{d}_j}{\partial q_\beta} + \nabla_{\mathbf{d}_j} k(\mathbf{d}_j,\mathbf{d}_k)\cdot\frac{\partial^2 \mathbf{d}_j}{\partial q_\alpha \partial q_\beta}\right). \tag{5.2}$$

where the Hessian matrix of the kernel, with components

$$\left(\nabla_{\mathbf{d}_j} k(\mathbf{d}_j,\mathbf{d}_k)\nabla^{\mathrm{T}}_{\mathbf{d}_j}\right)_{lm} = \frac{\partial^2 k(\mathbf{d}_j,\mathbf{d}_k)}{\partial (d_j)_l \partial (d_k)_m},$$

appears. Covariance functions between first- and second-derivative values, or between two second-derivative values, can be derived in the same way.

In practice, analytical-Hessian expressions such as Equation (5.2) are not used for two reasons: First, analytical second derivatives of the descriptors and kernel functions are generally not available*. Second, it is much more expensive than the equivalent first-derivative expression: Using a descriptor with L components, the kernel Hessian matrix has $\frac{L^2+L}{2}$ unique entries for each pair of descriptors (local environments); compare this with the L entries of the kernel gradient that need to be computed in Equation (5.1). Since L scales with the number of atoms in an environment (at least $3N_i - 6$ for a complete descriptor), this scaling can quickly become intractable for all but the smallest molecules or local environments.

Instead, the second differentiation with respect to q_β can be done numerically with finite differences. The natural basis for these derivatives would seem to be the eigenvectors of the Hessian matrix – similar to the normal-mode basis, only the eigenvectors are computed *without* mass-weighting the Hessian matrix.

*that is, they have not yet been derived or implemented – the situation may change if second-derivative fitting is more widely adopted

We can then take q_α and q_β to be the coordinates along these eigenvectors and compute each energy-Hessian covariance in time $2L$ rather than $\frac{L^2}{2}$. An implementation of this method has recently been added to the GAP code [98] and will be used for all the Hessian fits in this chapter.

5.2.2 Model systems

We begin with the simplest alkane systems that show torsional motion: Both ethane and propane (two and three carbons, respectively) have C-C bonds about which rotation can occur – in both cases, the motion just involves hydrogen atoms, specifically the rotation of a terminal methyl group (–CH_3). The potential energy surface along these rotations has two stationary points, called "eclipsed" and "staggered", that occur when opposing hydrogens are closest and furthest away, respectively [1]. Butane (four carbons) has a different type of torsion involving all four carbons about the central C-C bond; in this case, the special states are called *cis* (maximum, carbons same side), *trans* (global minimum, carbons opposite side), and *gauche* (an intermediate local minimum) [1, 196]. In contrast with the methyl rotations, these states are generally considered to form different **isomers** of the molecule; similar states are found in CC-CC torsions in all longer linear alkanes [12].

Quantum reference calculations were done at the MP2/AVQZ level using the MOLPRO package [133–136]. A study of relative energies of alkane conformations (minima) [13] found errors of about 10 % from MP2/pVQZ to their best CCSD(T) estimate of 25.8 meV for the relative energies of the two butane minima. The basis set incompleteness error at this level is much smaller, with a difference of 0.74 meV (3.0 %) between MP2/pVQZ and MP2/pVTZ energies. For barrier heights and overall torsional curves, another study [7] suggests MP2 is

an adequate level of theory, while my own calculations give comparable estimates for the basis set incompleteness error: The difference between the energies of the ethane barrier predicted by MP2/AVTZ and MP2/AVQZ was 3.4 meV, or 2.7 % of the barrier height.

Energies are always taken relative to the conformational global minimum; Hessian matrices were computed by finite differences of the gradients at geometries obtained by displacing each atom by 0.01 Å in each Cartesian direction. All torsional curves shown here are unrelaxed; that is, only the torsional angle was varied without changing any of the other degrees of freedom of the molecule. For that reason, it might be better to call these profiles "cut lines" of the potential energy surface, as they do not necessarily go through the true rotational barriers (saddle points). The effect of relaxation on the locations and energies of minima and barriers seems to be small, though more thorough study might eventually be called for. In any case, the GAP does not need calculations at the exact locations of the minima or saddle points; the Hessian and force together are enough to determine the exact location as long as they are computed close enough to the stationary point for the quadratic approximation to remain valid. It is also useful to compare the predictions of candidate models on cut lines other than the exact transition path, as molecules in a real simulation subject to thermal motion will deviate from this path.

Ethane

The simplest system to fit was ethane, with only one torsional degree of freedom. As Figure 5.6 shows, a GAP fitted only to information at the minimum is capable of accurately reproducing the entire torsional curve. This initial GAP was fitted to the smallest 9 non-degenerate eigenvalues and associated eigenvectors of the

Hessian matrix at the MP2/AVQZ minimum.

The fitting parameters for this GAP, in the form of a command line for use with the `teach_sparse` program distributed as part of the GAP code [98]*, are as follows:

```
teach_sparse atoms_filename=ethane-mp2opt-torsions-hesstrain-onlymin.xyz \
    gap={soap atom_sigma=0.5 cutoff=3.0 n_max=10 l_max=8 Z=6 n_species=2 \
       species_Z={{1 6}} n_sparse=84 covariance_type=DOT_PRODUCT zeta=4.0 \
       sparse_method=UNIQ delta=1.0:\
         soap atom_sigma=0.5 cutoff=4.0 n_max=10 l_max=8 Z=1 n_species=2 \
       species_Z={{1 6}} n_sparse=252 covariance_type=DOT_PRODUCT zeta=4.0 \
       sparse_method=UNIQ delta=1.0} \
    default_sigma={0.003 0.3 1.0 0.01} energy_parameter_name=Energy \
    hessian_parameter_name=hessian sigma_parameter_name=sigma \
    hessian_delta=0.01 sparse_jitter=1e-10 \
    gp_file=gp-ethane-mp2opt-hessians-onlymin.xml \
    do_copy_at_file=T sparse_separate_file=T
```

In brief, the most important parameters are the `cutoff` specifying the range of the individual SOAP descriptors (one for C environments, one for H) and the `default_sigma` giving the regularization parameters for the energies, forces, virials (unused), and Hessian eigenvalues, in that order. The rest of the parameters are explained in Bartók and Csányi [74] or in the QUIP documentation and will not be elaborated on here. This is also the command line that will be used for all other GAPs discussed in this chapter, with modifications made only to the parameters mentioned here.

The fit also included a set of 20 geometries randomly displaced from the minimum by a per-coordinate normal distribution with a standard deviation of 0.05 Å. No energies, forces, or Hessians were computed at these points; they only act as basis functions to allow a better reconstruction of the whole energy profile. This is the reverse of the typical usage of sparse GAP theory: Whereas

*available at `http://www.libatoms.org/gap/gap_download.html`, or as a precompiled binary at `https://hub.docker.com/r/libatomsquip/quip/`.

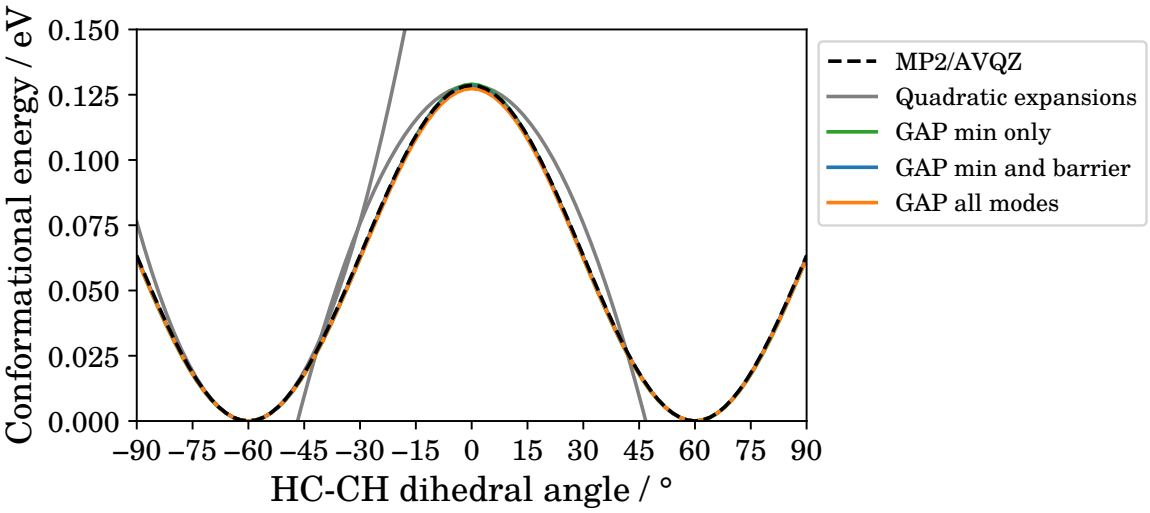


Figure 5.6: Torsional energy curve of ethane with three GAP fits, two truncated to the lowest 9 eigenvectors and one with all 19 modes (21 at the maximum) retained. Quadratic expansions of the potential along the cut line, based on the Hessian matrix at the two stationary points, are also shown. Note that all three GAP predictions are essentially identical to the MP2 reference; it is even possible to get a good fit to the whole torsional curve using only local information at the minimum.

usually it is used to fit a large amount of data with only a subset of points, here it is used to fit a large amount of data supplied at *one* point with several supplementary points. In both cases the theory is the same, using a pseudoinverse of the rectangular covariance matrix in place of $\mathbf{C}^{-1}$ in the prediction equation $\varepsilon(\mathbf{r}_*) = \mathbf{k}_*^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{t}$ [74].

Another fit was done with the Hessian computed at the maximum of the curve in Figure 5.6 (not exactly at the transition state; see above). Note that the Hessian matrix in Cartesian coordinates loses the degeneracy on three of its eigenvalues when moving away from a local minimum [197]. These eigenvalues correspond to rotation of the entire molecule; because of the way this rotation is expressed in Cartesian coordinates, they vanish only at local minima. These (small) eigenvalues and their corresponding eigenvectors are included in the fit. As Figure 5.6 shows, the new information at the maximum adds essen-

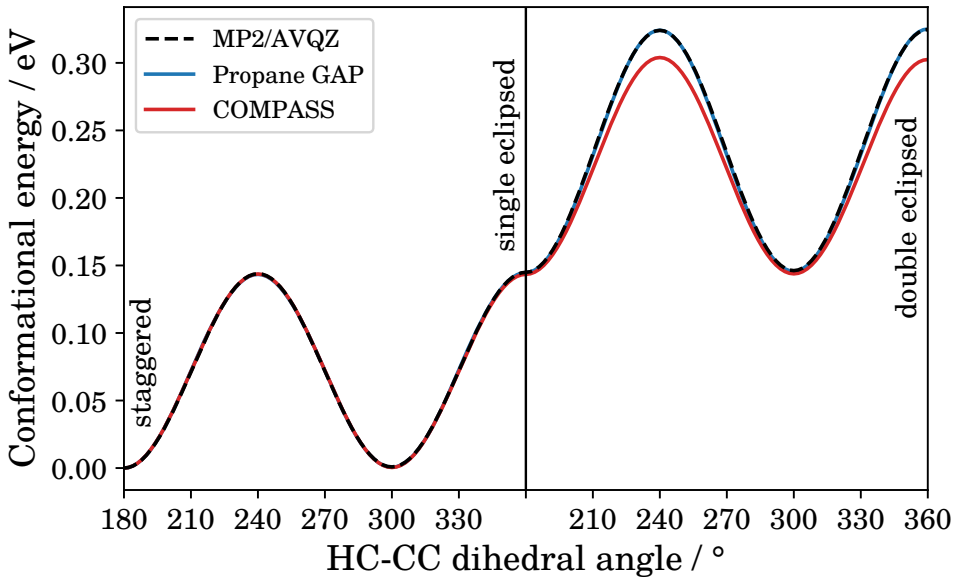


Figure 5.7: Continuous torsional profile of propane, rotating first one end methyl group through 180°*, then the other end methyl group through the same angles. The GAP was trained using three stationary points of this curve: The global minimum ("staggered"), the inflection point in the centre ("single eclipsed"), and the global maximum ("double eclipsed"); energies are relative to the staggered conformer (G2 geometry). The second rotational barrier is larger than the first, an effect that COMPASS also shows but underestimates by about* 0.02 eV*, or about half of the actual difference.*

tially nothing to the rotational profile; evidently, the minimum includes enough information to reproduce the entire cut line accurately. Sadly, this finding does not extend to more complicated systems; they need to be fitted using information both from minima and other important stationary points.

Propane

A similar fit was done for propane using three approximate stationary points. Propane has two end methyl groups that can rotate independently; the interaction between these two rotations is apparently not very well described by COMPASS. The exact stationary points are the global minimum with both bonds staggered, the geometry with one of the methyl groups rotated to the eclipsed conformation ("single eclipsed") and the geometry with both methyl groups in

the eclipsed conformation ("double eclipsed"). These geometries were approximated by three points taken from the cut line shown in Figure 5.7, which was generated from the G2 geometry [198] obtained from ASE [181]. The optimized propane geometry features slightly asymmetric methyl groups, so the three 60° segments in each half of the figure are not precisely symmetry-equivalent. Since none of the training points are minima of the MP2/AVQZ surface, the Hessian has the three non-degenerate rotational eigenvalues at each point. As with the ethane fit, this fit used a subset of modes (10 at each point, in this case) and included 20 sparse points scattered about each training point, this time displaced from a normal distribution of standard deviation 0.1 Å per coordinate. The only change in GAP parameters from the ethane fit was to tighten the force regularization parameter from 0.3 eV/Å to 0.03 eV/Å. As with ethane, the fit almost perfectly captures the entire rotational profile; it also correctly accounts for the coupling between the two methyl group torsions, while COMPASS accounts for only about half the energy difference. This is a good example of the compromises models such as COMPASS must make while fitting in a restricted subspace of the possible potential energy surfaces; the flexibility of the GAP, on the other hand, allows it to correctly account for this coupling without compromising its accuracy on any other part of the potential energy surface.

Butane

The CC-CC torsion of butane was much more difficult to fit than either of the two previous methyl rotations. This torsion is more complicated, involving the relative motion of entire methyl groups and their associated (mobile) hydrogens. The training geometries were prepared by first taking the two local minima (*trans* and *gauche*) from MP2/AVQZ geometry optimizations, then computing the tor-

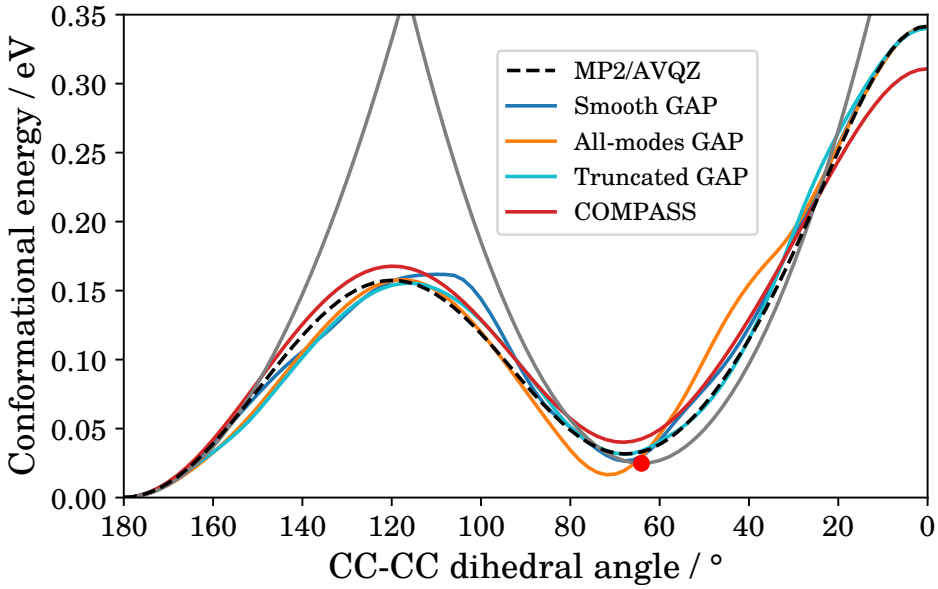


Figure 5.8: Butane torsional profile about the central C-C bond, along with several models fit to stationary points of both this curve and the two torsional curves above. As in Figure 5.6, the grey parabolas are quadratic expansions of the potential along the cut line at the two minima. The red dot shows the dihedral angle corresponding to the MP2/AVQZ local minimum.

sional curve (shown in Figure 5.8) from the global minimum and taking the two local maxima. Hessians were computed at each of the four new points as with propane, with the first 10 eigenmodes and 20 randomly-scattered sparse points used for fitting. These points were assigned the same energy and force regularization parameters as for the propane fit; the Hessian parameter was tightened from $0.01\,\mathrm{eV}^2/\text{Å}$ to $0.003\,\mathrm{eV}^2/\text{Å}$. They were then combined with the ethane and propane training data (Hessian points and sparse points), retaining the regularization parameters assigned to each configuration type (ethane, propane, or butane), to produce the first fit in Figure 5.8 (labelled "Smooth GAP"). This fit is good at the global minimum and *cis* maximum, but it misplaces the *gauche-trans* maximum as well as the *gauche* minimum; both exhibit artefacts causing them to deviate from the proper shape (which COMPASS reproduces smoothly, despite making relatively large errors in the *gauche* and *cis* energies).

Another problem was later encountered with this GAP: Since the mode truncation discarded most of the bond-length and angle vibrations, it effectively ignored those degrees of freedom. This meant it could not produce stable geometry optimizations or MD runs. To address this problem, a second version of the combined GAP was fit with all eigenmodes included*. Figure 5.6 provides proof of this concept; the third GAP in the figure, fit to the ethane maximum and minimum with a slightly larger Hessian regularization parameter, barely loses accuracy with respect to the other two, mode-truncated, GAPs.

In order to obtain an acceptable fit on the combined training set, the parameters had to be adjusted: First the cutoff of the hydrogen SOAP was taken down from 4 Å to 2 Å, in the hope that making the hydrogens more shortsighted would effectively smooth the potential. Second, the regularization parameters were tightened to better fit the available data: The energy parameter was taken down to 0.3 meV and the force parameter to 0.03 eV/Å for all configurations. The Hessian parameters were all tightened by half an order of magnitude to 0.003 eV^2/Å for ethane and propane, 0.001 eV^2/Å for butane. This fit is labelled "All-modes GAP" in Figure 5.8; while it gets better around the *trans-gauche* maximum, it introduces problematic irregularities almost everywhere else. A third fit was therefore attempted by applying the refined parameters from this GAP to the original set of truncated eigenmodes (9 for each of the ethane points, 10 each for propane and butane). The result is labelled "Truncated GAP" and achieves possibly the best fit so far to the butane dihedral: Unlike COMPASS, it gets the correct heights for both the *trans-gauche* maximum and the *cis* maximum, as well as reproducing the correct position and value of the *gauche* minimum.

The combined fits were finally tested on ethane and propane to assess whether

*except for the largest three on the butane maxima, due to a technical constraint on the number of eigenmodes

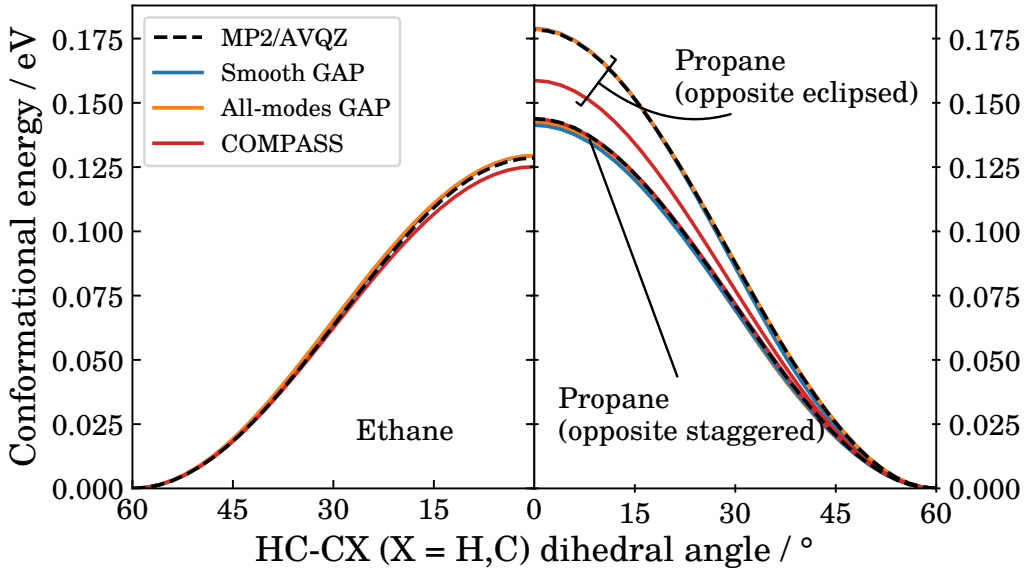


Figure 5.9: Torsional profile of ethane and propane, evaluated using two of the GAPs from Figure 5.8. The "opposite eclipsed" profile correspond to the right-hand half of Figure 5.7, with its minimum (the single eclipsed conformer) shifted to zero. Only 60° of each of the propane curves are shown for clarity; the rest is approximately equivalent. Both GAPs can reproduce, with reasonable accuracy, all the torsional curves in their training set. The "Truncated GAP" from Figure 5.8 had essentially identical predictions to the other two GAPs, so it is not shown here.

they lost any accuracy from incorporating the more complicated butane potential. Figure 5.9 shows that they did not; all three GAPs maintained meV accuracy across the methyl rotations in the training set.

5.3 Validation

While benchmarking on the training set is a useful and essential method for designing a new potential, the true test is how it performs on systems that were not originally in the training set. Since the goal of this project is to design a potential that is transferable across alkanes of varying chain lengths without explicitly including all their conformations in the training, the next step is to test on torsional profiles and molecules different from the ones considered above.

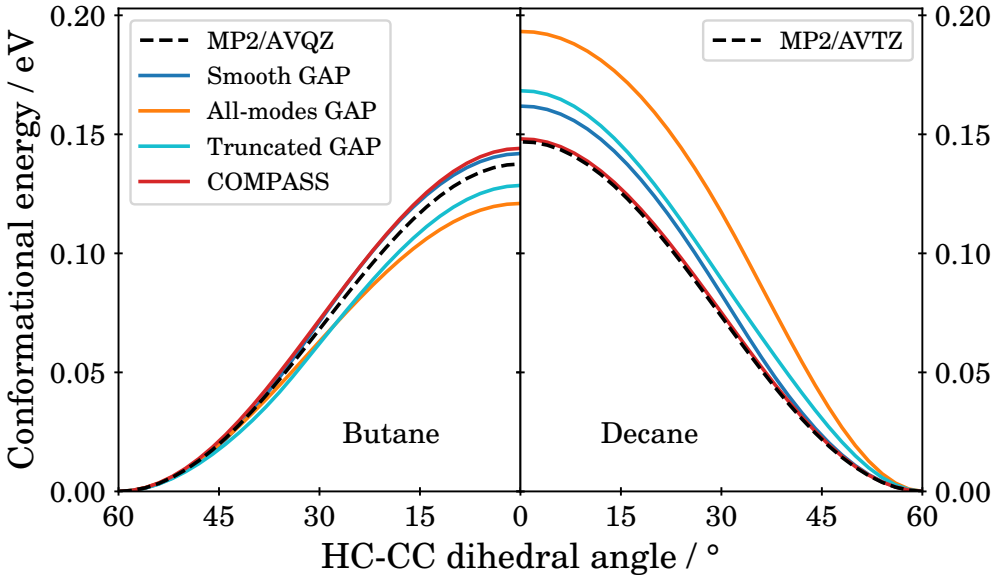


Figure 5.10: Torsional profile of the end methyl rotations of butane and decane; neither of these profiles was in the training set of the GAPs. The large size of the decane molecule required a smaller basis set (AVTZ) to be used.

One of the first validation tests is to see how well each GAP reproduces the torsional energy of the butane end methyl rotation – since the training set includes both butane Hessians and end methyl rotations, this seems like a fairly easy task. But Figure 5.10 already reveals problems indicating a possible overfitting and limited transferability of the second two GAPs (the ones fitted with tightened parameters): While the original "smooth" GAP slightly overestimates the butane maximum (by nearly the same amount as COMPASS), the other two underestimate by a relatively large amount.

5.3.1 Longer chains

The same test, done on decane, emphasizes that even the end methyl rotations are not as simple to reproduce as they originally seemed. The GAP fits all seem to be affected too strongly by changes in the geometry outside the true range of influence of the potential. The slight increase in the MP2 energy at the maximum

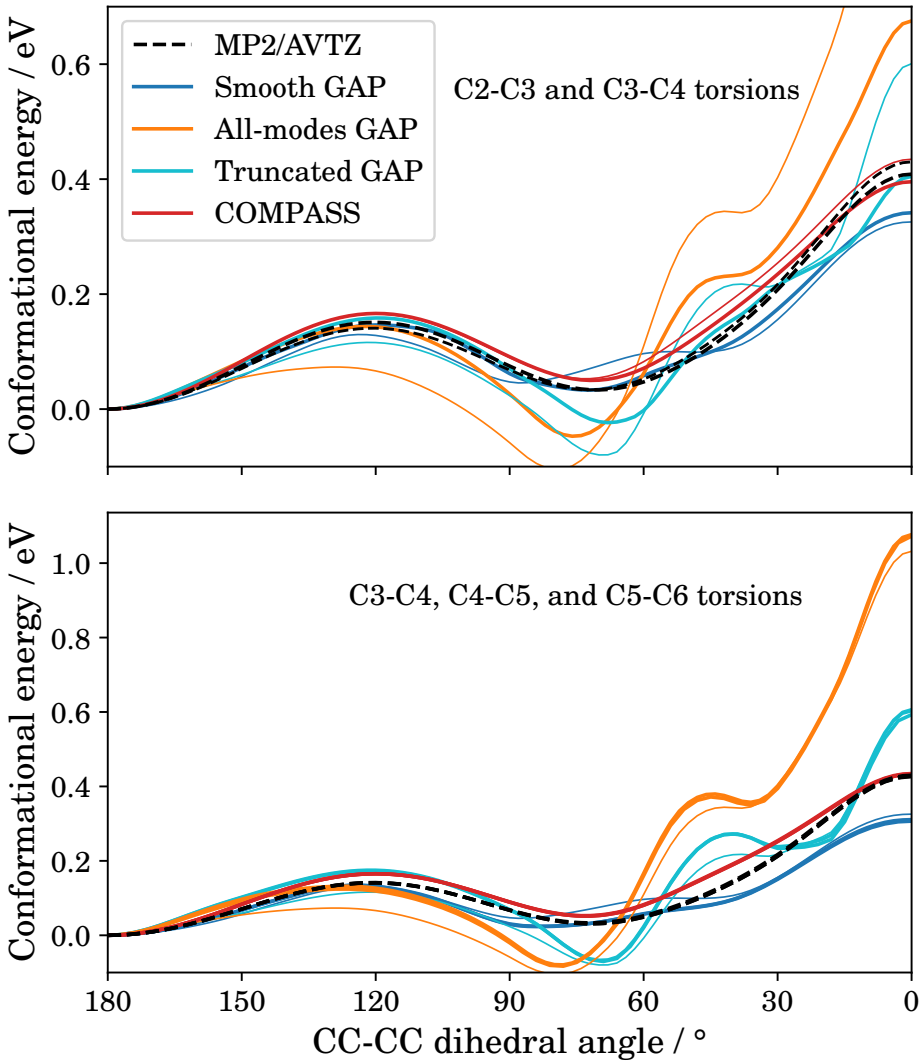


Figure 5.11: Torsional profiles about the interior C-C bonds in decane (not in training set), analogous to the butane torsion in Figure 5.8. The torsional profiles are grouped into two plots because of their similarity; the torsions about the C3-C4 bond are shown in both plots in a thinner line.

from butane to decane is probably attributable to the use of an unoptimized geometry rather than any finite-size effects, since the propane (also unoptimized) MP2 maximum is closer to that of decane than of butane. (The geometries for the larger hydrocarbons (pentane and decane) were taken from the standard fragment library of the Avogadro program [199].) COMPASS, on the other hand, undercompensates for the change – while it reproduces the decane curve perfectly, it overestimates on butane.

The decane analogues of the butane torsion from Figure 5.8 are the internal CC-CC torsions. Once again, the smooth GAP is the most regular and least affected by the change, even estimating the *trans-gauche* maximum accurately, while the other two exhibit strong overfitting irregularities. All of the models shown here have some sensitivity to end effects; that is, the energy at the CC-CC torsion at the end of the chain (the C2-C3 torsion) has a slightly higher *cis* energy and a very slightly lower *gauche-trans* maximum than the other torsions in the middle of the chain. As with the other transferability tests, the later two GAPs both greatly overestimate this effect, while the smooth GAP finds it in the wrong direction. COMPASS accounts for this effect well, despite consistently overestimating the *gauche-trans* maximum. But even though GAP makes poor predictions relative to COMPASS on these cut lines, it is encouraging that they show the same pattern of errors across the different bonds of decane, which could point to a systematic error in the training data or fitting parameters that could be corrected in the future without resorting to explicitly including decane conformers in the fit.

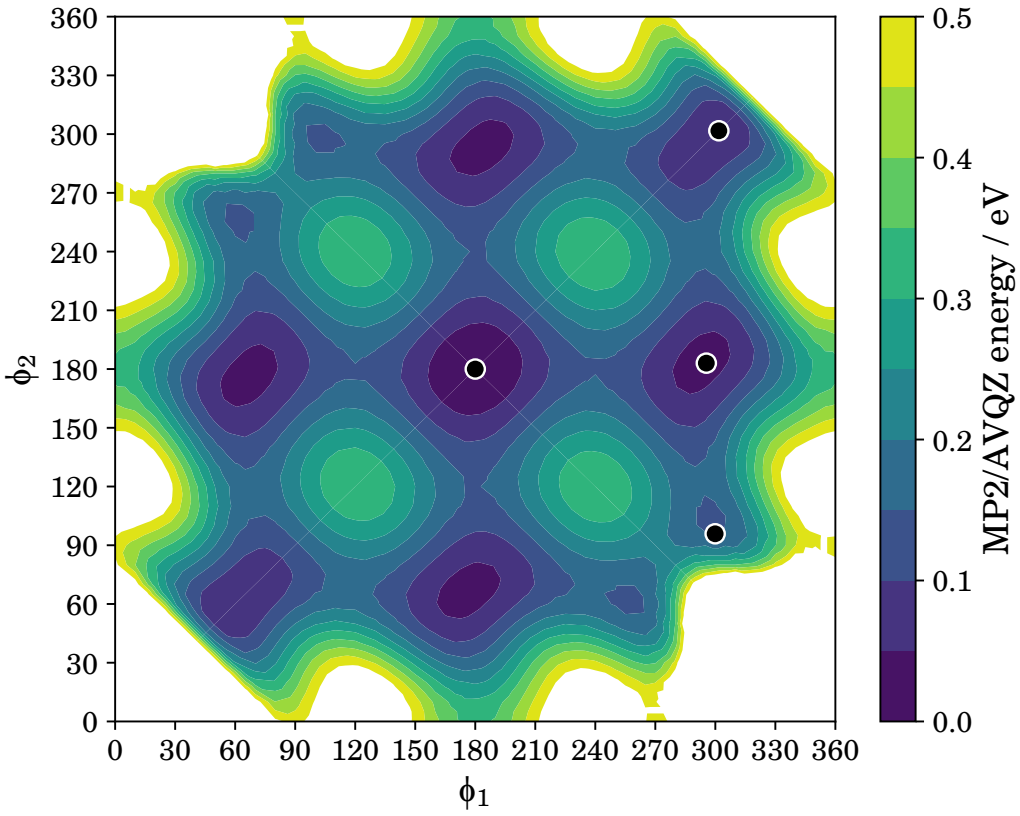


Figure 5.12: MP2/AVQZ conformational map of pentane with fixed internal degrees of freedom. Minima of this surface are shown as black dots. The map is symmetric about both diagonals. Energies larger than 0.5 eV *were removed from the plot and conformations with those large energies are shown blank.*

5.3.2 Pentane conformers

The next logical step is to consider the interaction between two neighbouring CC-CC torsions. This type of interaction has been studied for decades using pentane as a prototypical system [12, 13, 200]. For many of the same reasons, pentane is well suited to the present potential development and validation effort: It is the simplest system in which CC-CC torsions interact, and since only two torsional coordinates interact, it is easy to visualize and interpret maps of the conformational potential energy surface. Finally, the structure [12, 200] and energetics [13] of its PES have been well characterized in theoretical studies.

Figure 5.12 shows a conformational map calculated at the MP2/AVQZ level as a function of the two CC-CC dihedral angles, labelled ϕ_1 and ϕ_2. The map was generated by scanning through the torsional coordinates of the (unrelaxed) starting structure, without relaxing any of the other degrees of freedom. It is similar to Figure 1 in ref. [12], only their map used a different quantum chemical reference method and (presumably) relaxed the other degrees of freedom. It is interesting to note that their map shows a transition state between the "g-x+" and "x-g+" conformers ($\phi_1 \approx -60°$, $\phi_2 \approx 95°$ and its mirror image) that is lower in energy than the transition state connecting this conformer to the nearby "gt" conformers – that is, the two neighbouring "double-*gauche*"* conformers are more closely connected than to their neighbouring "*gauche-trans*" states. The unrelaxed MP2 map in Figure 5.12 shows the opposite situation, where each "double-*gauche*" conformer is more closely connected to its "*gauche-trans*" neighbour than to its mirror image. This effect is likely an artefact due to the unrelaxed torsional scan procedure; the "*gauche-trans*" conformers involve close hydrogen contacts and their interconversion involves the rotation of both end methyl groups to lower the barrier.

Unrelaxed maps for two of the GAP models, as well as for COMPASS, are shown in Figure 5.13; an alternate version with one model in each quadrant is shown in Figure 5.14 to facilitate direct comparison. We can already make useful qualitative observations about the behaviour of each GAP from Figure 5.13: The smoothest and most regular model, also the one most similar to the MP2/AVQZ map, is COMPASS, followed by the smooth GAP, which starts to display some irregularities – most notably, the splitting of the "gg" local minima on the diagonal and the shift of the "gt" minima to smaller *trans* and *gauche* angles. Finally, the all-modes GAP displays some of the same irregularities (although its minima

*_not_ the two "gg" conformers on the diagonal

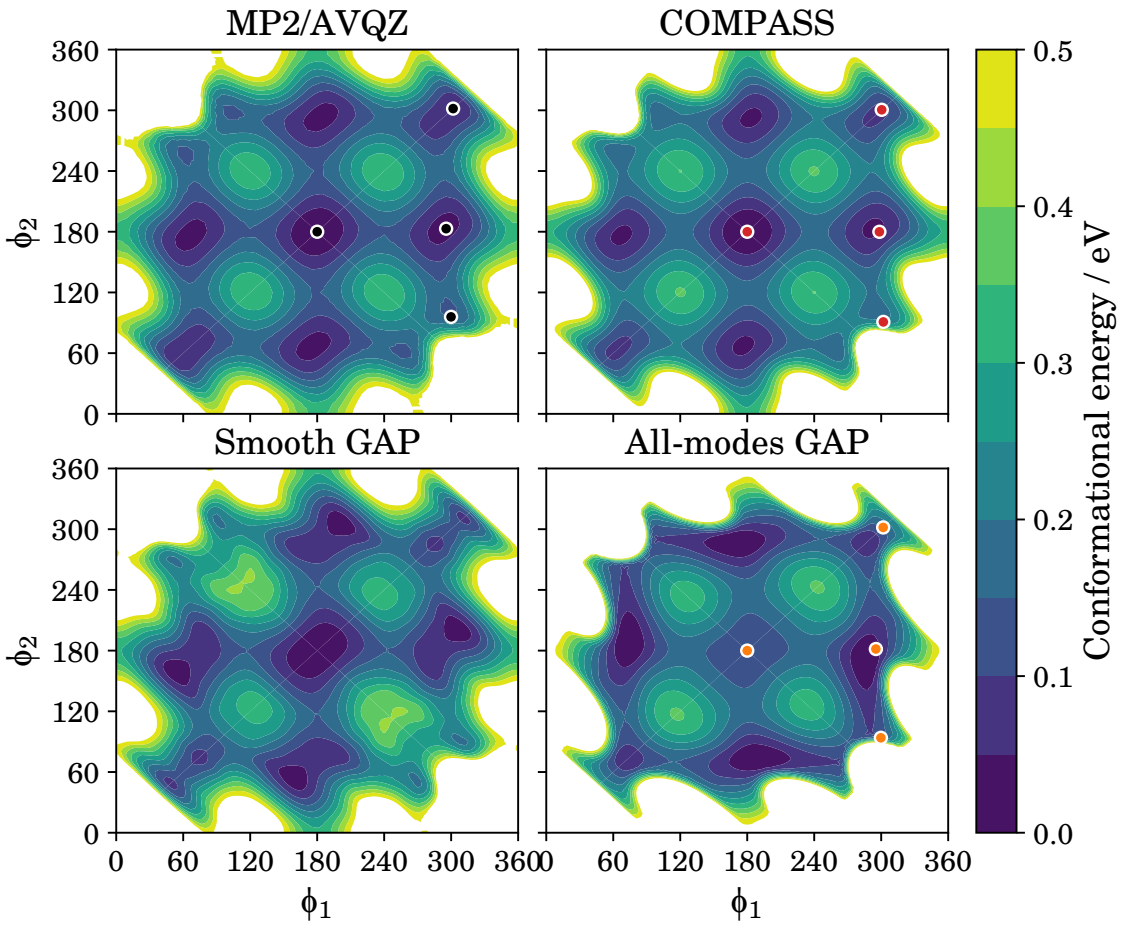


Figure 5.13: Pentane torsional maps (unrelaxed) computed with two of the combined GAPs; COMPASS is for comparison. Energies are relative to the (unrelaxed) global minimum of each model. The torsional coordinates of the fully-relaxed local minima of each model are plotted, where available, as coloured dots in white circles (colour indicates the model, not the energy of the minimum).

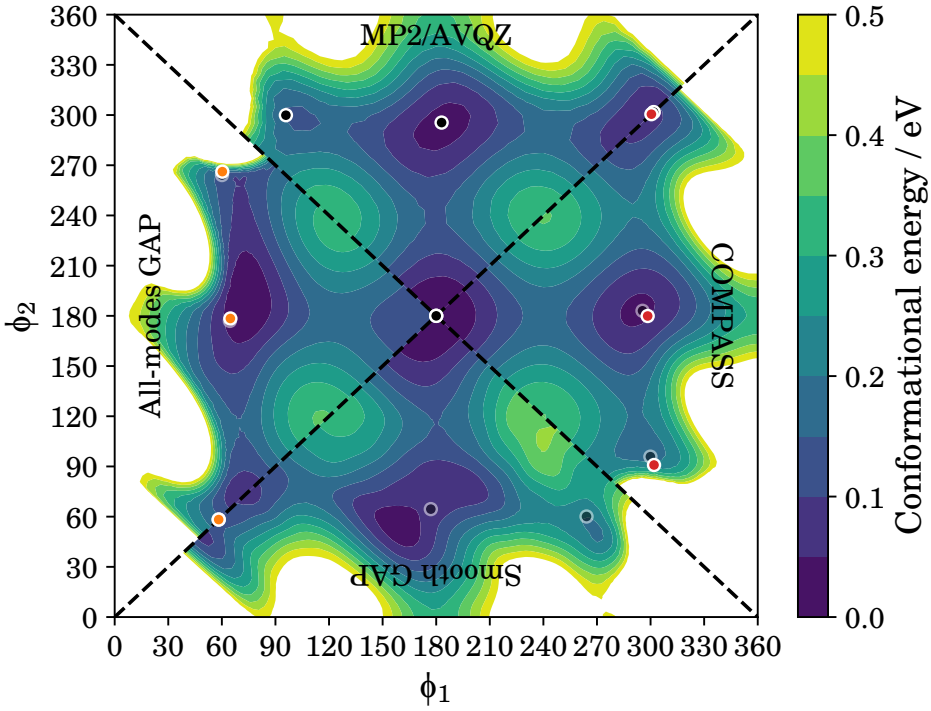


Figure 5.14: Same as Figure 5.13, but in quadrants to facilitate comparison (each quadrant contains all the unique conformations of the molecule). The MP2/AVQZ minima are also shown faintly in the other quadrants for comparison.

are in about the right place; see Figure 5.14), and in addition is much "sharper" – the energy increases more rapidly when leaving the basins of the local minima. This finding is similar to what we might have extrapolated from the decane single-torsion plots in Figure 5.11. But the most troubling finding from the all-modes GAP map is the raising of the "tt" conformer energy, which should be the most stable one. The lowering of the *gauche* minimum seen in Figure 5.11 here manifests itself as an incorrect stabilisation of the "gt" conformer with respect to the "tt" one that all other models predict as a global minimum.

The next step is to find the energies of the true, fully-relaxed local minima of each potential energy surface, as the pentane maps just shown do not include any relaxation of internal degrees of freedom that allow the molecule to adopt its most stable structure at any of the conformations. This minimization

Model	Energy/meV			
	tt	gt	gg	gx-
MP2/AVQZ	0.0	24.1	33.7	120.6
W1h-val [13]	0.0	26.6	41.7	122.0
COMPASS [41]	0.0	30.5	55.3	128.4
All-modes GAP	121.9	0.0	44.2	117.4

Table 5.1: Energies of the fully-relaxed minima of the pentane potential energy surface as predicted by the all-modes GAP and COMPASS, with two quantum chemical references, one MP2 and one extrapolated coupled-cluster CCSD(T) estimate. Energies are relative to the fully-relaxed global minimum of each model.

was carried for the MP2/AVQZ, COMPASS, and all-modes GAP. Table 5.1 shows the relative values of these local minima for each model, compared to the best quantum chemical estimate ("W1h-val") from the detailed study of alkane conformational energies [13]. The table shows that the MP2 estimate is generally in good agreement with their values, while COMPASS significantly overpredicts especially the levels of the "gg" and "gt" minima. The all-modes GAP, as previously seen, grossly overpredicts the level of the "tt" conformer, the true global minimum – or rather, it overstabilizes *gauche* torsional angles with respect to *trans*. The ordering of the minima is otherwise correct, with a good prediction of the "gg"-"gx" energy difference.

The locations of all the minima are shown in Figure 5.14, overlaid on the MP2/AVQZ minima for comparison of conformer location (in (ϕ_1, ϕ_2) space). Even though the energy values of the all-modes GAP minima are problematic, the *locations* of the minima are very close to the MP2/AVQZ conformations (while those of COMPASS deviate slightly more).

No minima were found for the smooth GAP because the optimizations were unstable to bond lengthening, angle bending, and most of the other motions accounted for in the larger Hessian eigenvalues discarded from the training sample. The same problem was found for the "truncated GAP", not shown here

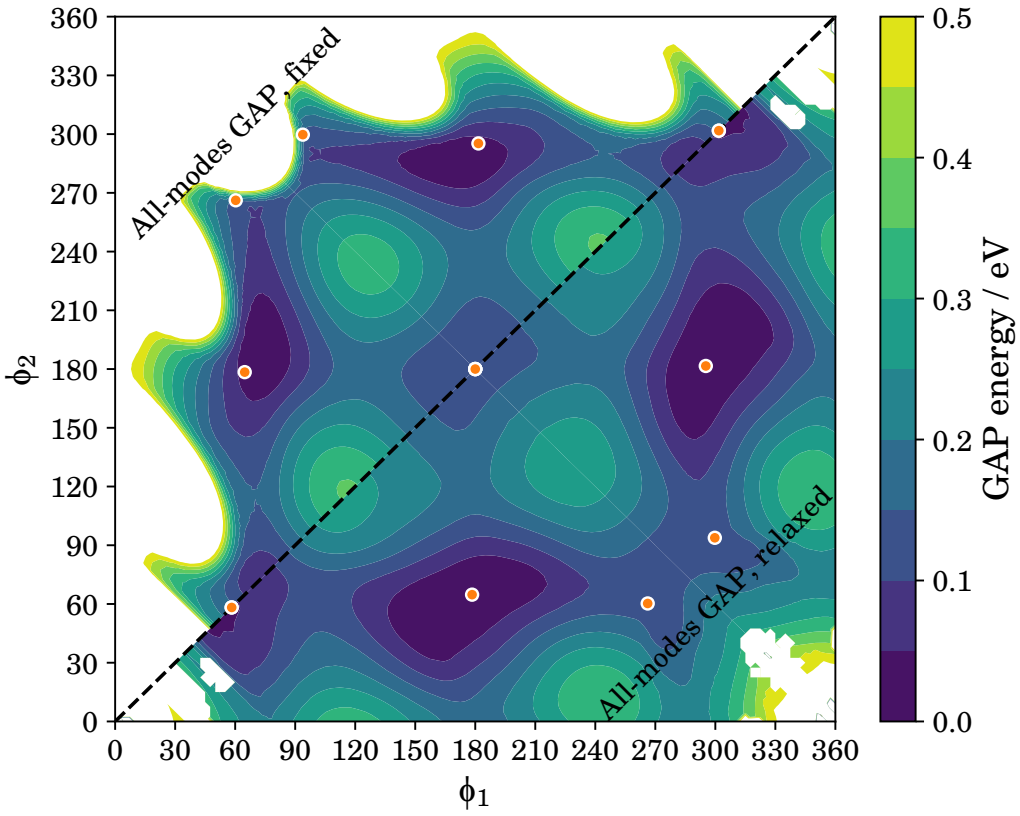


Figure 5.15: Comparison of torsional maps of pentane computed with the all-modes GAP, with fixed internal coordinates (top left) and relaxed internal coordinates (bottom right). All-modes GAP local minima are shown in each quadrant. Energies are relative to the corresponding global minimum – fixed or relaxed – of the all-modes GAP. Blank areas of the relaxed plot correspond to conformers where either the optimization was unsuccessful or the optimized energy was still larger than 0.5 eV.

because its pentane map is otherwise very similar to that of all-modes GAP.

The only stable GAP, with respect to optimization of internal coordinates, is therefore the all-modes GAP. This made it possible to produce a relaxed version of the pentane conformational map, where the torsional angles ϕ_1 and ϕ_2 are held constant but with the other degrees of freedom optimized at every (ϕ_1, ϕ_2) point on the surface. The result is shown in Figure 5.15. In fact, the surface changes drastically with relaxation: The "gx" conformers appear as local minima, along with a low-energy path connecting them. (Local minima will always appear as minima of the *relaxed* potential energy surface, while this is not gener-

ally the case with the unrelaxed maps – see e.g. the "gx" minima in Figure 5.15, which only correspond to minima on the relaxed surface.) The effect of relaxation on the GAP surface seems larger than would be expected of the real PES, especially from comparing Figure 5.12 (unrelaxed) with Figure 1 from [12] (relaxed), where the only qualitative difference is the presence of a lower-energy transition pathway between the neighbouring "gx" conformers. Nevertheless, the relaxed GAP PES shows some of the expected features of the real relaxed PES: The correct local minima (locations, at least) and lower barriers between states compared to the unrelaxed PES.

For further development of the intramolecular alkane GAP, the pentane torsional map and torsional curves presented in this section will serve a similar role as the dimer error did for the intermolecular methane GAP: It provides a systematic means of measuring the error of a fitted potential and comparing it with analytical alternatives. There are many other interesting cases that could be used for benchmarking in the future, such as the stability of long folded alkanes [11], but the pentane map offers a good balance between simplicity (ease of interpretation) and complexity (extent of model capabilities tested). In this case, COMPASS showed a level of systematic accuracy on the intramolecular potential that seemed to be impossible to reach on the intermolecular potential, for COMPASS or any other analytical potential. Its use of a relatively flexible functional form along with a quantum-mechanical fitting reference is likely responsible for this accuracy. Nevertheless, its accuracy is still limited by its fixed functional form – the overestimated pentane minimum energies and devations from the quantum torsional curves, especially in Figures 5.7 and 5.8, are subtle evidence of this.

In principle, it should be possible to achieve an better accuracy than COM-

PASS across a variety of alkane systems, but the fitting technique needs to be developed further to achieve this goal. The behaviour of the three Hessian GAPs considered in this section provide clues to what improvements are necessary: While the smooth GAP is certainly more regular and transferable than the other two, it fails in optimizations and MD because it does not account for the higher vibrational modes (a behaviour it shares with the truncated GAP). But the tightening of parameters necessary to fit all modes also make the GAP less regular and transferable, as comparison of the smooth and truncated GAPs shows (which differ only in fitting parameters, namely the regularization and cutoffs). And even the all-modes GAP suffers from failures presumably due to overfitting of certain modes: a trial MD run at room temperature using the all-modes GAP resulted in unstable C-C bond lengths, indicating that even the all-modes GAP is not properly fitting all the modes in its training set (overfitting due to too-strict regularization parameters is a likely culprit).

The ideal would be a middle ground combining the best features of the three: Smoothness and regularity due to well-chosen parameters, but a good fit due to incorporation of all the vibrational degrees of freedom. At the moment, this does not seem possible without a more advanced regularization technique, such as having one parameter for each eigenvalue (to capture the lower modes while avoiding overfitting the higher ones) or a Cartesian-basis Hessian fitting procedure (still under development) that would assign different regularizations to each atom. Together with the observations about COMPASS above, this naturally points to a combined approach where most of the energy is accounted for by COMPASS (or a similar model), while the errors and couplings COMPASS cannot account for are modelled with a Hessian GAP fit as a correction on top of a COMPASS baseline; this general approach of using baseline models to improve

transferability was described in Section 3.2.2 and used for the MBD SOAP-GAPs in Chapter 4.

Another issue to explore is the distribution of extra sparse basis points (the extra 20 geometries randomly distributed about each stationary point in the fit, with no energies or forces, to act as basis functions to support the fit away from the stationary point). It would be interesting to see how the fits depend on the number and spread of these points, and perhaps whether the fits could be improved by incorporating extra quantum energies and forces at some of them – in this sense, such a potential would be a hybrid between the random-sampling GAP discussed in the previous section and the "pure" Hessian GAPs explored in this section.

While the GAPs presented here are not yet ready for production simulations, the results in this section show significant promise for the strategy of Hessian fitting in future applications. Several avenues of improvement have been presented; it is likely that any one (or a combination) of these strategies could bring an intramolecular GAP-based model to the same level of smoothness and transferability, and an even higher level of accuracy, than that achievable with purely analytical models.

Chapter 6

Discussion and further development

This dissertation has presented progress towards an exceptionally accurate potential for alkanes obtained by a systematic fit to the entire potential energy surface. The potential development efforts have been concentrated in two domains, motivated by the separation between the intermolecular and intramolecular energies. The intermolecular efforts showed, for the first time, a machine learning potential for liquid methane with a systematic convergence towards the true Born-Oppenheimer potential energy surface. It showed that the prediction of even the density, a property previously regarded as relatively simple to reproduce (perhaps because it is easy to fit empirically), in fact requires the modelling of several different effects – including many-body dispersion and quantum nuclear effects – to reproduce accurately with the *correct* potential energy surface. The results so far indicate that this approach is robust and extensible. It will pave the way for accurate intermolecular potentials for larger and more complex systems: Longer alkanes, branched alkanes, mixtures, and conceivably even the inclusion of polar species with an appropriate polarizable electrostatic model.

It also provides tools for understanding how the different physically identifiable effects included in the potential affect other properties, especially the diffusivity and viscosity.

The intramolecular potential is further behind in development, partially because it is an application of a relatively new and unproven method, the fitting to second-derivative values. While this branch of development efforts did not yield any potentials that could be tested in an MD simulation, it did show some cases where the Hessian fitting was very successful. The results so far are encouraging for a future robust, systematic intramolecular counterpart to the intermolecular GAP with the use of second-derivative information.

6.1 Road to an integrated alkane potential

Another question that has not yet been addressed in this dissertation is how the two components of the potential – intermolecular and intramolecular – should be combined. In small, simple molecules such as methane, the split between intramolecular and intermolecular potentials is easily determined: intramolecular components involve a single connected molecule and intermolecular components are strictly the *interaction* between two or more different connected molecules. The case that has not yet been explored here is that of long, flexible molecules, where the interaction of atoms separated by many bonds behaves more as if they were governed by intramolecular than by intermolecular forces. Almost all analytical potentials switch from the intramolecular, bond-angle-torsion form to the intermolecular, pairwise form beyond 1-4 (torsion) interactions with some sort of scaling applied to the intermolecular (electrostatics and L-J) terms on the 1-4 pairs. This means that their torsional potentials (including all the COMPASS curves in Section 5.2) are in fact a combination of a true torsional potential and

some effective scaled pairwise terms. We would obviously need a more systematic way of combining the potentials, starting with a rigorous physical separation of which energy components are handled by which potential. The intramolecular potential is usually taken to be strongly local [46], so a good start would be to find some systematic separation between the local and nonlocal components of the single-molecule energy. One possibility would be to apply the intramolecular GAP to a single, large molecule and subtract the predicted energies and forces from the quantum mechanical ones, ideally leaving behind only the local component. This is only one of many possible approaches. The coupling between the intramolecular and intermolecular potentials would be described the same way as in the current model: through the intermolecular SOAP-GAP, which can equally detect changes in the intramolecular environment. For effects beyond the range of the SOAP-GAP, different approaches might be needed. Fortunately, there are several existing approaches [76, 201] to predicting locally-dependent parameters for long-range interactions; any of these could be adapted into the current framework.

In any case, the potentials that are developed will be continuously tested against experimental densities, diffusivities, and viscosities. While the best reference for *constructing* a potential is the quantum mechanical potential energy surface, the ultimate test of the *usefulness* of a model is how well and how reliably it reproduces experimental observables. This work is driven by the belief that accuracy at the microscopic, atomistic scale implies accuracy at the macroscopic scale, as long as we take care to fit the potential systematically and account for all relevant physical effects. The results for the intramolecular potential support this belief and encourage us to continue pursuing the goal of a systematically fitted potential for the entire chemical class of alkanes.

Appendix A

Technical Notes

The algorithms, scripts, and data analysis used in this dissertation were done primarily in Python version 2.7.9* using components from the SciPy ecosystem†, including the libraries NumPy 1.14.3, SciPy 1.1.0, the Matplotlib [178] plotting package (version 2.2.2), and especially the Jupyter interactive computing environment (version 1.0.0) with the IPython [179] interface (version 5.7.0). The colour schemes for many of the graphs were chosen from the ColorBrewer colourblind-friendly and print-friendly palettes‡ and distributed with Matplotlib. The random initial molecular configurations for the MD simulations were created using Packmol [187]. This project made extensive use QUIP via the `quippy` Python interface, as well as the GAP code distributed for use with QUIP§. The code, scripts, and data necessary to reproduce the results in this work will be made available at the University of Cambridge repository¶, under the same name as this dissertation.

*`http://python.org`

†`http://scipy.org/`

‡`http://colorbrewer2.org`

§currently available at `http://www.libatoms.org/gap/gap_download.html` or as a precompiled binary at `https://hub.docker.com/r/libatomsquip/quip/`

¶`https://www.repository.cam.ac.uk/`

Bibliography

[1] IUPAC. *Compendium of chemical terminology. Compiled by A. D. McNaught and A. Wilkinson*. Blackwell Scientific, Oxford, 2nd edition, 1987. ISBN 0-9678550-9-8. URL https://doi.org/10.1351/goldbook.

[2] Maurizio Mondello and Gary S. Grest. Molecular dynamics of linear and branched alkanes. *J. Chem. Phys.*, 103(16):7156, 1995. doi:10.1063/1.470344. URL http://scitation.aip.org/content/aip/journal/jcp/103/16/10.1063/1.470344.

[3] Rajdeep Singh Payal, S. Balasubramanian, Indranil Rudra, Kunj Tandon, Ingo Mahlke, David Doyle, and Roger Cracknell. Shear viscosity of linear alkanes through molecular simulations: quantitative tests for n -decane and n -hexadecane. *Mol. Simul.*, 38(14-15):1234–1241, 2012. doi:10.1080/08927022.2012.702423. URL http://dx.doi.org/10.1080/08927022.2012.702423.

[4] Leonardo Spanu, Davide Donadio, Detlef Hohl, Eric Schwegler, and Giulia Galli. Stability of hydrocarbons at deep Earth pressures and temperatures. *Proc. Natl. Acad. Sci.*, 108(17):6843–6846, 2011. doi:10.1073/pnas.1014804108. URL http://www.pnas.org/content/108/17/6843.

[5] F.Y. Hansen, K.W. Herwig, B. Matthies, and H. Taub. Intramolecular and Lattice Melting in n-Alkane Monolayers: An Analog of Melting in Lipid Bilayers. *Phys Rev Lett*, 83(12):2362, Sep 1999. doi:10.1103/PhysRevLett.83.2362. URL https://link.aps.org/doi/10.1103/PhysRevLett.83.2362.

[6] D J Tobias, K Tu, and M L Klein. Assessment of all-atom potentials for modeling membranes: molecular dynamics simulations of solid and liquid alkanes and crystals of phospholipid fragments. *J. Chim. Phys.*, 94(7-8):1482–1502, 1997. URL http://cat.inist.fr/?aModele=afficheN&cpsidt=2819547.

[7] Shirley W I Siu, Kristyna Pluhackova, and Rainer a. Böckmann. Optimization of the OPLS-AA force field for long hydrocarbons. *J. Chem. Theory Comput.*, 8(4):1459–1470, 2012. doi:10.1021/ct200908r.

[8] Norman L. Allinger, Young H. Yuh, and Jenn Huei Lii. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *J. Am. Chem. Soc.*, 111 (23):8551–8566, Nov 1989. doi:10.1021/ja00205a001. URL http://dx.doi.org/10.1021/ja00205a001.

[9] Marcus G. Martin and J. Ilja Siepmann. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. *J. Phys. Chem. B*, 102(97):2569–2577, Mar 1998. doi:10.1021/jp972543+. URL http://pubs.acs.org/doi/abs/10.1021/jp972543%2B.

[10] Jukka-Pekka Jalkanen, Tapani A. Pakkanen, Yan Yang, and Richard L. Rowley. Interaction energy surfaces of small hydrocarbon molecules. *J. Chem. Phys.*, 118(12):5474, 2003. doi:10.1063/1.1540106. URL http://scitation.aip.org/content/aip/journal/jcp/118/12/10.1063/1.1540106.

[11] Nils O B Lüttschwager, Tobias N. Wassermann, Ricardo A. Mata, and Martin A. Suhm. The last globally stable extended alkane. *Angew. Chemie - Int. Ed.*, 52(1):463–466, Jan 2013. doi:10.1002/anie.201202894. URL http://doi.wiley.com/10.1002/anie.201202894.

[12] Gyula Tasi, Fujio Mizukami, István Pálinkó, József Csontos, Werner Győrffy, Padmakumar Nair, Kazuyuki Maeda, Makoto Toba, Shu-ichi Niwa, Yoshimichi Kiyozumi, and Imre Kiricsi. Enumeration of the conformers of unbranched aliphatic alkanes. *J. Phys. Chem. A*, 102(39):7698–7703, 1998. doi:10.1021/jp981866i. URL https://pubs.acs.org/doi/10.1021/jp981866i.

[13] David Gruzman, Amir Karton, and Jan M. L. Martin. Performance of ab initio and density functional methods for conformational equilibria of C(n)H(2n+2) alkane isomers (n = 4-8). *J. Phys. Chem. A*, 113(43):11974–11983, Oct 2009. doi:10.1021/jp903640h. URL http://pubs.acs.org/doi/abs/10.1021/jp903640h.

[14] Matthias Rupp, O Anatole von Lilienfeld, and Kieron Burke. Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. *J. Chem. Phys.*, 148(24):241401, 2018. doi:10.1063/1.5043213. URL https://doi.org/10.1063/1.5043213.

[15] Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.*, 104(13):136403, Apr 2010. doi:10.1103/PhysRevLett.104.136403. URL http://link.aps.org/doi/10.1103/PhysRevLett.104.136403.

[16] Jörg Behler and Michele Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98(14):146401, Apr 2007. doi:10.1103/PhysRevLett.98.146401. URL http://link.aps.org/doi/10.1103/PhysRevLett.98.146401.

[17] Albert P. Bartók, Michael J. Gillan, Frederick R. Manby, and Gábor 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):054104, Aug 2013. doi:10.1103/PhysRevB.88.054104. URL https://link.aps.org/doi/10.1103/PhysRevB.88.054104.

[18] Tobias Morawietz, Andreas Singraber, Christoph Dellago, and Jörg Behler. How van der Waals interactions determine the unique properties of water. *Proc. Natl. Acad. Sci.*, 113(30):8368–8373, Jul 2016. doi:10.1073/pnas.1602375113.

[19] Ben Leimkuhler, Emad Noorizadeh, and Florian Theil. A gentle stochastic thermostat for molecular dynamics. *J. Stat. Phys.*, 135(2):261–277, Apr 2009. doi:10.1007/s10955-009-9734-0. URL http://link.springer.com/10.1007/s10955-009-9734-0.

[20] M P Allen and D. J. Tildesley. *Computer simulation of liquids*. Clarendon Press, Oxford, 1989. ISBN 9780198556459. URL http://search.lib.cam.ac.uk/?itemid=%7Ccambrdgedb%7C1065582.

[21] Gaurav Pranami and Monica H. Lamm. Estimating Error in Diffusion Coefficients Derived from Molecular Dynamics Simulations. *J. Chem. Theory Comput.*, 11(10):4586–4592, Oct 2015. doi:10.1021/acs.jctc.5b00574. URL http://pubs.acs.org/doi/10.1021/acs.jctc.5b00574.

[22] Burkhard Dünweg and Kurt Kremer. Molecular dynamics simulation of a polymer chain in solution. *J. Chem. Phys.*, 99(9):6983–6997, 1993. doi:10.1063/1.465445. URL https://doi.org/10.1063/1.465445.

[23] In-Chul Yeh and Gerhard Hummer. System-Size Dependence of Diffusion Coefficients and Viscosities from Molecular Dynamics Simulations with Periodic Boundary Conditions. *J. Phys. Chem. B*, 108(40):15873–15879, 2004. doi:10.1021/jp0477147. URL https://doi.org/10.1021/jp0477147.

[24] Peter J. Daivis and Denis J. Evans. Transport coefficients of liquid butane near the boiling point by equilibrium molecular dynamics. *J. Chem. Phys.*, 103(June):4261, 1995. doi:10.1063/1.470664. URL http://scitation.aip.org/content/aip/journal/jcp/103/10/10.1063/1.470664http://dx.doi.org/10.1063/1.470664.

[25] M Born and R Oppenheimer. Zur Quantentheorie der Molekeln. *Ann. Phys.*, 389(20):457–484, 1927. doi:10.1002/andp.19273892002. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19273892002.

[26] Herman J C Berendsen. *Simulating the physical world : hierarchical modeling from quantum mechanics to fluid dynamics*. Cambridge University Press, Cambridge, 2007. ISBN 9780521835275.

[27] William Allen and Richard L. Rowley. Predicting the viscosity of alkanes using nonequilibrium molecular dynamics: Evaluation of intermolecular potential models. *J. Chem. Phys.*, 106(24):10273, 1997. doi:10.1063/1.474052. URL http://scitation.aip.org/content/aip/journal/jcp/106/24/10.1063/1.474052.

[28] Anthony J. Stone. *The theory of intermolecular forces*. Oxford University Press, Oxford, 2nd ed edition, 2013. ISBN 9780199672394.

[29] Max Born and Joseph E. Mayer. Zur Gittertheorie der Ionenkristalle. *Zeitschrift für Phys.*, 75(1-2):1–18, Jan 1932. doi:10.1007/BF01340511. URL http://link.springer.com/10.1007/BF01340511.

[30] F London. Zur Theorie und Systematik der Molekularkräfte. *Zeitschrift für Phys.*, 63(3):245–279, Mar 1930. doi:10.1007/BF01421741. URL https://doi.org/10.1007/BF01421741.

[31] J. E. Jones. On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, 106 (738):463–477, Oct 1924. doi:10.1098/rspa.1924.0082. URL http://rspa.royalsocietypublishing.org/cgi/doi/10.1098/rspa.1924.0082.

[32] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, May 1995. doi:10.1021/ja00124a002. URL http://dx.doi.org/10.1021/ja00124a002.

[33] William L. Jorgensen, Jeffry D. Madura, and Carol J. Swenson. Optimized intermolecular potential functions for liquid hydrocarbons. *J. Am. Chem. Soc.*, 106(22):6638–6646, Oct 1984. doi:10.1021/ja00334a030. URL http://dx.doi.org/10.1021/ja00334a030.

[34] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, Jan 1996. doi:10.1021/ja9621760. URL http://dx.doi.org/10.1021/ja9621760.

[35] Bin Chen and J. Ilja Siepmann. Transferable Potentials for Phase Equilibria. 3. Explicit-Hydrogen Description of Normal Alkanes. *J. Phys. Chem. B*, 103(25):5370–5379, Jun 1999. doi:10.1021/jp990822m. URL http://libsta28.lib.cam.ac.uk:2327/doi/abs/10.1021/jp990822m.

[36] Steven J. Stuart, Alan B. Tutein, and Judith A. Harrison. A reactive potential for hydrocarbons with intermolecular interactions. *J. Chem. Phys.*, 112(14):6472, 2000. doi:10.1063/1.481208. URL http://scitation.aip.org/content/aip/journal/jcp/112/14/10.1063/1.481208.

[37] Joseph M. Hayes, James C. Greer, and David A. Morton-Blake. A force-field description of short-range repulsions for high density alkane molecular dynamics simulations. *J. Comput. Chem.*, 25(16):1953–1966, Dec 2004. doi:10.1002/jcc.20116. URL http://doi.wiley.com/10.1002/jcc.20116.

[38] Thomas C O'Connor, Jan Andzelm, and Mark O Robbins. AIREBO-M: a reactive model for hydrocarbons at extreme pressures. *J. Chem. Phys.*, 142(2):024903, Jan 2015. doi:10.1063/1.4905549. URL http://dx.doi.org/10.1063/1.4905549.

[39] Mary J. Van Vleet, Alston J. Misquitta, Anthony J. Stone, and J. R. Schmidt. Beyond Born–Mayer: Improved Models for Short-Range Repulsion in ab Initio Force Fields. *J. Chem. Theory Comput.*, 12(8):3851–3870, Aug 2016. doi:10.1021/acs.jctc.6b00209. URL http://pubs.acs.org/doi/abs/10.1021/acs.jctc.6b00209.

[40] J R Maple, M.-J. Hwang, T P Stockfisch, U Dinur, M Waldman, C S Ewig, and A T Hagler. Derivation of class II force fields. I. Methodology and quantum force field for the alkyl functional group and alkane molecules. *J. Comput. Chem.*, 15(2):162–182, 1994. doi:10.1002/jcc.540150207. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540150207.

[41] H. Sun. COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase Applications—Overview with Details on Alkane and Benzene Compounds. *J. Phys. Chem. B*, 102(38):7338–7364, Sep 1998. doi:10.1021/jp980939v. URL http://dx.doi.org/10.1021/jp980939v.

[42] Robert Hellmann, Eckard Bich, and Eckhard Vogel. Ab initio intermolecular potential energy surface and second pressure virial coefficients of methane. *J. Chem. Phys.*, 128(21):214303, Jun 2008. doi:10.1063/1.2932103. URL http://scitation.aip.org/content/aip/journal/jcp/128/21/10.1063/1.2932103.

[43] David H. Gay, Houfeng Dai, and Donald R. Beck. Obtaining accurate pressure second virial coefficients for methane from an ab initio pair potential. *J. Chem. Phys.*, 95(12):9106–9114, Dec 1991. doi:10.1063/1.461189. URL http://aip.scitation.org/doi/10.1063/1.461189.

[44] Scott J. Weiner, Peter A. Kollman, Dzung T. Nguyen, and David A. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.*, 7(2):230–252, Apr 1986. doi:10.1002/jcc.540070216. URL http://doi.wiley.com/10.1002/jcc.540070216.

[45] Gerardo Andrés Cisneros, Kjartan Thor Wikfeldt, Lars Ojamäe, Jibao Lu, Yao Xu, Hedieh Torabifard, Albert P. Bartók, Gábor Csányi, Valeria Molinero, and Francesco Paesani. Modeling Molecular Interactions in Water: From Pairwise to Many-Body Potential Energy Functions. *Chem. Rev.*, 116(13):7501–7528, 2016. doi:10.1021/acs.chemrev.5b00644. URL http://pubs.acs.org/doi/abs/10.1021/acs.chemrev.5b00644.

[46] Max David Veit. *Locality of forces in molecular systems*. M.Phil. thesis, University of Cambridge, 2015. URL https://doi.org/10.17863/CAM.26416.

[47] Rafał Podeszwa, Robert Bukowski, and Krzysztof Szalewicz. Potential energy surface for the benzene dimer and perturbational analysis of π-π interactions. *J. Phys. Chem. A*, 110(34):10345–10354, 2006. ISSN 10895639. doi:10.1021/jp064095o. URL https://pubs.acs.org/doi/abs/10.1021/jp064095o.

[48] J. R. Maple, M.-J. Hwang, T. P. Stockfisch, U. Dinur, M. Waldman, C. S. Ewig, and A. T. Hagler. Derivation of class II force fields. I. Methodology and quantum force field for the alkyl functional group and alkane molecules. *J. Comput. Chem.*, 15(2):162–182, Feb 1994. doi:10.1002/jcc.540150207. URL http://doi.wiley.com/10.1002/jcc.540150207.

[49] Norman L. Allinger, Kuohsiang Chen, and Jenn-Huei Lii. An improved force field (MM4) for saturated hydrocarbons. *J. Comput. Chem.*, 17(5-6):642–668, Apr 1996. doi:10.1002/(SICI)1096-987X(199604)17:5/6<642::AID-JCC6>3.0.CO;2-U. URL http://doi.wiley.com/10.1002/%28SICI%291096-987X%28199604%2917%3A5/6%3C642%3A%3AAID-JCC6%3E3.0.CO%3B2-U.

[50] Alice E. A. Allen, Michael C. Payne, and Daniel J. Cole. Harmonic Force Constants for Molecular Mechanics Force Fields via Hessian Matrix Projection. *J. Chem. Theory Comput.*, 14(1):274–281, Jan 2018. doi:10.1021/acs.jctc.7b00785. URL http://pubs.acs.org/doi/10.1021/acs.jctc.7b00785.

[51] Shih-Wei Chao, Arvin Huang-Te Li, and Sheng D. Chao. Molecular dynamics simulations of fluid methane properties using ab initio intermolecular interaction potentials. *J. Comput. Chem.*, 30(12):1839–1849,

Sep 2009. doi:10.1002/jcc.21185. URL http://doi.wiley.com/10.1002/jcc.21185.

[52] Arvin Huang-Te Li and Sheng D. Chao. A Refined Intermolecular Interaction Potential for Methane: Spectral Analysis and Molecular Dynamics Simulations. *J. Chinese Chem. Soc.*, 63(3):282–289, Mar 2016. doi:10.1002/jccs.201500358. URL http://doi.wiley.com/10.1002/jccs.201500358.

[53] A. D. Buckingham, P. W. Fowler, and Jeremy M. Hutson. Theoretical Studies of van der Waals Molecules and Intermolecular Forces. *Chem. Rev.*, 88 (6):963–988, sep 1988. ISSN 15206890. doi:10.1021/cr00088a008. URL http://pubs.acs.org/doi/abs/10.1021/cr00088a008.

[54] Donald W Brenner, Olga a Shenderova, Judith a Harrison, Steven J Stuart, Boris Ni, and Susan B Sinnott. A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons. *J. Phys. Condens. Matter*, 14(4):783–802, 2002. doi:10.1088/0953-8984/14/4/312.

[55] W A de Jong, R J Harrison, and D A Dixon. Parallel Douglas–Kroll energy and gradients in NWChem: Estimating scalar relativistic effects using Douglas–Kroll contracted basis sets. *J. Chem. Phys.*, 114(1):48–53, 2001. doi:10.1063/1.1329891. URL https://aip.scitation.org/doi/abs/10.1063/1.1329891.

[56] Tamás Veszprémi and Miklós Fehér. *Quantum chemistry : fundamentals to applications*. Kluwer Academic/Plenum, Dordrecht, 1999. ISBN 0306461641.

[57] Richard M Martin. *Electronic structure : basic theory and practical methods*. Cambridge University Press, Cambridge, UK, 1st pbk. edition, 2008. ISBN 9780521534406.

[58] Stefan Grimme, Andreas Hansen, Jan Gerit Brandenburg, and Christoph Bannwarth. Dispersion-Corrected Mean-Field Electronic Structure Methods. *Chem. Rev.*, 116(9):5105–5154, 2016. doi:10.1021/acs.chemrev.5b00533.

[59] Stefan Grimme. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.*, 27(15):1787–1799, Nov 2006. doi:10.1002/jcc.20495. URL http://dx.doi.org/10.1002/jcc.20495.

[60] Alexandre Tkatchenko and Matthias Scheffler. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.*, 102(7):073005, Feb 2009.

doi:10.1103/PhysRevLett.102.073005. URL http://link.aps.org/doi/10.1103/PhysRevLett.102.073005.

[61] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, 132(15): 154104, Apr 2010. doi:10.1063/1.3382344. URL http://aip.scitation.org/doi/10.1063/1.3382344.

[62] Alexandre Tkatchenko, Robert A. DiStasio, Roberto Car, and Matthias Scheffler. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.*, 108(23):236402, Jun 2012. doi:10.1103/PhysRevLett.108.236402. URL http://link.aps.org/doi/10.1103/PhysRevLett.108.236402.

[63] M Dion, H Rydberg, E Schröder, D C Langreth, and B I Lundqvist. van der Waals density functional for general geometries. *Phys. Rev. Lett.*, 92 (24):246401, Jun 2004. doi:10.1103/PhysRevLett.92.246401. URL http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.92.246401.

[64] Oleg A Vydrov and Troy Van Voorhis. Nonlocal van der Waals Density Functional Made Simple. *Phys. Rev. Lett.*, 103(6):63004, Aug 2009. doi:10.1103/PhysRevLett.103.063004. URL https://link.aps.org/doi/10.1103/PhysRevLett.103.063004.

[65] Jiří Klimeš, David R Bowler, and Angelos Michaelides. Chemical accuracy for the van der Waals density functional. *J. Phys. Condens. Matter*, 22(2): 22201, 2010. URL http://stacks.iop.org/0953-8984/22/i=2/a=022201.

[66] Berk Hess. Determining the shear viscosity of model liquids from molecular dynamics simulations. *J. Chem. Phys.*, 116(1):209, 2002. doi:10.1063/1.1421362. URL http://scitation.aip.org/content/aip/journal/jcp/116/1/10.1063/1.1421362.

[67] Steven Hobday, Roger Smith, and Joe Belbruno. Applications of genetic algorithms and neural networks to interatomic potentials. *Nucl. Instruments Methods Phys. Res. Sect. B Beam Interact. with Mater. Atoms*, 153 (1-4):247–263, Jun 1999. doi:10.1016/S0168-583X(99)00057-9. URL http://www.sciencedirect.com/science/article/pii/S0168583X99000579.

[68] Nils J Nilsson. *The quest for artificial intelligence : a history of ideas and achievements*. Cambridge University Press, Cambridge, 2010. ISBN 9780521116398.

[69] Nongnuch Artrith, Tobias Morawietz, and Jörg Behler. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B*, 83(15):153101, Apr 2011.

doi:10.1103/PhysRevB.83.153101. URL http://link.aps.org/doi/10.1103/PhysRevB.83.153101.

[70] S. Alireza Ghasemi, Albert Hofstetter, Santanu Saha, and Stefan Goedecker. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. *Phys. Rev. B*, 92(4):045131, Jul 2015. doi:10.1103/PhysRevB.92.045131. URL http://link.aps.org/doi/10.1103/PhysRevB.92.045131.

[71] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.*, 15(9):095003, Sep 2013. doi:10.1088/1367-2630/15/9/095003. URL http://iopscience.iop.org/article/10.1088/1367-2630/15/9/095003/meta.

[72] David J C MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, 2003. ISBN 9780521642989. URL http://www.loc.gov/catdir/description/cam032/2003055133.html.

[73] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, MA, 2006. ISBN 9780262182539.

[74] Albert P. Bartók and Gábor Csányi. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.*, 115(16):1051–1057, Apr 2015. doi:10.1002/qua.24927. URL http://doi.wiley.com/10.1002/qua.24927.

[75] Tristan Bereau, Denis Andrienko, and O. Anatole Von Lilienfeld. Transferable Atomic Multipole Machine Learning Models for Small Organic Molecules. *J. Chem. Theory Comput.*, 11(7):3225–3233, Jul 2015. doi:10.1021/acs.jctc.5b00301. URL http://pubs.acs.org/doi/abs/10.1021/acs.jctc.5b00301.

[76] Tristan Bereau, Robert A. DiStasio, Alexandre Tkatchenko, and O. Anatole von Lilienfeld. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.*, 148(24):241706, Jun 2018. doi:10.1063/1.5009502. URL http://aip.scitation.org/doi/10.1063/1.5009502http://arxiv.org/abs/1710.05871.

[77] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.*, 18(20):13754–13769, May 2015. doi:10.1039/C6CP00415F. URL http://dx.doi.org/10.1039/C6CP00415F.

[78] Albert P. Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gábor Csányi, Michele Ceriotti, Albert P. Bartok, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gabor Csanyi, and Michele Ceriotti. Machine Learning Unifies the Modelling of Materials and Molecules. *Sci. Adv.*, 3(12):e1701816, Dec 2017. doi:10.1126/sciadv.1701816. URL http://advances.sciencemag.org/lookup/doi/10.1126/sciadv.1701816.

[79] Radford M Neal. *Bayesian learning for neural networks*. Lecture notes in statistics (Springer-Verlag) ; v. 118. Springer, New York, London, 1996. ISBN 9780387947242.

[80] Wojciech J. Szlachta, Albert P. Bartók, and Gábor Csányi. Accuracy and transferability of Gaussian approximation potential models for tungsten. *Phys. Rev. B*, 90(10):104108, Sep 2014. doi:10.1103/PhysRevB.90.104108. URL http://link.aps.org/doi/10.1103/PhysRevB.90.104108.

[81] Daniele Dragoni, Thomas D Daff, Gábor Csányi, and Nicola Marzari. Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron. *Phys. Rev. Mater.*, 2(1):13808, Jan 2018. doi:10.1103/PhysRevMaterials.2.013808. URL https://link.aps.org/doi/10.1103/PhysRevMaterials.2.013808.

[82] Volker L Deringer, Chris J Pickard, and Gábor Csányi. Data-Driven Learning of Total and Local Energies in Elemental Boron. *Phys. Rev. Lett.*, 120(15):156001, Apr 2018. doi:10.1103/PhysRevLett.120.156001. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.156001.

[83] M. J. Gillan, D. Alfé, A. P. Bartók, and G. Csányi. First-principles energetics of water clusters and ice: A many-body analysis. *J. Chem. Phys.*, 139(24):244504, Dec 2013. doi:10.1063/1.4852182. URL http://aip.scitation.org/doi/10.1063/1.4852182.

[84] Volker L. Deringer and Gábor Csányi. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B*, 95(9):094203, Mar 2017. doi:10.1103/PhysRevB.95.094203. URL https://link.aps.org/doi/10.1103/PhysRevB.95.094203.

[85] Volker L. Deringer, Noam Bernstein, Albert P. Bartók, Matthew J. Cliffe, Rachel N. Kerber, Lauren E. Marbella, Clare P. Grey, Stephen R. Elliott, and Gábor Csányi. Realistic Atomistic Structure of Amorphous Silicon from Machine-Learning-Driven Molecular Dynamics. *J. Phys. Chem. Lett.*, 9(11):2879–2885, jun 2018. ISSN 19487185. doi:10.1021/acs.jpclett.8b00902. URL http://pubs.acs.org/doi/10.1021/acs.jpclett.8b00902.

[86] Miguel A. Caro, Volker L. Deringer, Jari Koskinen, Tomi Laurila, and Gábor Csányi. Growth Mechanism and Origin of High s p 3 Content in Tetrahedral Amorphous Carbon. *Phys. Rev. Lett.*, 120(16):166101, apr 2018. ISSN 0031-9007. doi:10.1103/PhysRevLett.120.166101. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.166101.

[87] Felix C. Mocanu, Konstantinos Konstantinou, Tae Hoon Lee, Noam Bernstein, Volker L. Deringer, Gábor Csányi, and Stephen R. Elliott. Modeling the Phase-Change Memory Material, Ge2Sb2Te5, with a Machine-Learned Interatomic Potential, sep 2018. ISSN 15205207. URL http://pubs.acs.org/doi/10.1021/acs.jpcb.8b06476.

[88] S T John and Gábor Csányi. Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials. *J. Phys. Chem. B*, 121(48):10934–10949, 2017. doi:10.1021/acs.jpcb.7b09636. URL https://doi.org/10.1021/acs.jpcb.7b09636.

[89] Letif Mones, Noam Bernstein, and Gábor Csányi. Exploration, Sampling, And Reconstruction of Free Energy Surfaces with Gaussian Process Regression. *J. Chem. Theory Comput.*, 12(10):5100–5110, 2016. doi:10.1021/acs.jctc.6b00553. URL https://doi.org/10.1021/acs.jctc.6b00553.

[90] A. Shapeev. Accurate representation of formation energies of crystalline alloys with many components. *Comput. Mater. Sci.*, 139:26–30, nov 2017. ISSN 0927-0256. doi:10.1016/J.COMMATSCI.2017.07.010. URL https://www.sciencedirect.com/science/article/pii/S0927025617303610.

[91] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87(18):184115, May 2013. doi:10.1103/PhysRevB.87.184115. URL http://link.aps.org/doi/10.1103/PhysRevB.87.184115.

[92] R.D. Goodwin. Apparatus for Determination for Pressure-Density- Temperature Relation and Specific Heat of Hydrogen to 350 Atmospheres at Temperatures above 14K. *J. Res.Nat. Bur. Stand*, 65C(4):231–243, 1961.

[93] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, 81(1):511–519, 1984. doi:10.1063/1.447334. URL https://doi.org/10.1063/1.447334.

[94] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695–1697, Mar 1985. doi:10.1103/PhysRevA.31.1695. URL https://link.aps.org/doi/10.1103/PhysRevA.31.1695.

[95] Frédéric Legoll, Mitchell Luskin, and Richard Moeckel. Non-ergodicity of the Nosé-Hoover thermostatted harmonic oscillator. *Arch. Ration. Mech. Anal.*, 184(3):449–463, Apr 2007. doi:10.1007/s00205-006-0029-1. URL http://link.springer.com/10.1007/s00205-006-0029-1.

[96] Glenn J Martyna, Michael L Klein, and Mark Tuckerman. Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.*, 97(4):2635–2643, 1992. doi:10.1063/1.463940. URL https://doi.org/10.1063/1.463940.

[97] Andrew Jones and Ben Leimkuhler. Adaptive stochastic methods for sampling driven molecular systems. *J. Chem. Phys.*, 135(8):84125, 2011. doi:10.1063/1.3626941. URL https://doi.org/10.1063/1.3626941.

[98] Albert Bartók-Pártay, Livia Bartók-Pártay, Federico Bianchini, Anke Butenuth, Marco Caccin, Silvia Cereda, Gábor Csányi, Alessio Comisso, Tom Daff, ST John, Chiara Gattinoni, Gianpietro Moras, James Kermode, Letif Mones, Alan Nichol, David Packwood, Lars Pastewka, Giovanni Peralta, Ivan Solt, Oliver Strickson, Wojciech Szlachta, Csilla Varnai, Max Veit, and Steven Winfield. libAtoms+QUIP, Jul 2018. URL http://libatoms.org.

[99] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. Langevin equation with colored noise for constant-temperature molecular dynamics simulations. *Phys. Rev. Lett.*, 102(2):020601, Jan 2009. doi:10.1103/PhysRevLett.102.020601. URL https://link.aps.org/doi/10.1103/PhysRevLett.102.020601.

[100] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. Colored-noise thermostats à la Carte. *J. Chem. Theory Comput.*, 6(4):1170–1180, Apr 2010. doi:10.1021/ct900563s. URL http://pubs.acs.org/doi/abs/10.1021/ct900563s.

[101] Glenn J. Martyna, Douglas J. Tobias, and Michael L. Klein. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.*, 101(5):4177, 1994. doi:10.1063/1.467468. URL http://scitation.aip.org/content/aip/journal/jcp/101/5/10.1063/1.467468.

[102] D. Quigley and M. I. J. Probert. Langevin dynamics in constant pressure extended systems. *J. Chem. Phys.*, 120(24):11432–11441, 2004. doi:10.1063/1.1755657. URL http://scitation.aip.org/content/aip/journal/jcp/120/24/10.1063/1.1755657.

[103] A. Sokal. Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms. In C. DeWitt-Morette, P. Cartier, and A. Folacci, editors, *Functional Integration. NATO ASI Series (Series B Physics)*,

pages 131–192. Springer, Boston, MA, 1997. ISBN 978-1-4899-0321-1. doi:10.1007/978-1-4899-0319-8_6. URL http://link.springer.com/10.1007/978-1-4899-0319-8_6.

[104] Thomas E. Markland and Michele Ceriotti. Nuclear quantum effects enter the mainstream. *Nat. Rev. Chem.*, 2(3):0109, Feb 2018. doi:10.1038/s41570-017-0109. URL http://www.nature.com/articles/s41570-017-0109.

[105] Michele Ceriotti and David E Manolopoulos. Efficient First-Principles Calculation of the Quantum Kinetic Energy and Momentum Distribution of Nuclei. *Phys. Rev. Lett.*, 109(10):100604, Sep 2012. doi:10.1103/PhysRevLett.109.100604. URL https://link.aps.org/doi/10.1103/PhysRevLett.109.100604.

[106] Anthony J Russell and Mark A Spackman. Vibrational averaging of electrical properties. *Mol. Phys.*, 84(6):1239–1255, 1995. doi:10.1080/00268979500100861. URL https://doi.org/10.1080/00268979500100861.

[107] David M Bishop, Feng Long Gu, and Sławomir M Cybulski. Static and dynamic polarizabilities and first hyperpolarizabilities for CH4, CF4, and CCl4. *J. Chem. Phys.*, 109(19):8407–8415, 1998. doi:10.1063/1.477503. URL https://doi.org/10.1063/1.477503.

[108] S. Biermann, D. Hohl, and D. Marx. Quantum effects in solid hydrogen at ultra-high pressure. *Solid State Commun.*, 108(6):337–341, Oct 1998. doi:10.1016/S0038-1098(98)00388-3. URL https://www.sciencedirect.com/science/article/pii/S0038109898003883.

[109] David Chandler and Peter G. Wolynes. Exploiting the isomorphism between quantum theory and classical statistical mechanics of polyatomic fluids. *J. Chem. Phys.*, 74(7):4078–4095, Apr 1981. doi:10.1063/1.441588. URL http://aip.scitation.org/doi/10.1063/1.441588.

[110] Scott Habershon, David E. Manolopoulos, Thomas E. Markland, and Thomas F. Miller. Ring-Polymer Molecular Dynamics: Quantum Effects in Chemical Dynamics from Classical Trajectories in an Extended Phase Space. *Annu. Rev. Phys. Chem.*, 64(1):387–413, Apr 2013. doi:10.1146/annurev-physchem-040412-110122. URL http://www.annualreviews.org/doi/10.1146/annurev-physchem-040412-110122.

[111] Michele Ceriotti, Michele Parrinello, Thomas E. Markland, and David E. Manolopoulos. Efficient stochastic thermostatting of path integral molecular dynamics. *J. Chem. Phys.*, 133(12):124104, Sep 2010. doi:10.1063/1.3489925. URL http://aip.scitation.org/doi/10.1063/1.3489925.

[112] Timothy J.H. Hele. On the relation between thermostatted ring-polymer molecular dynamics and exact quantum dynamics. *Mol. Phys.*, 114(9): 1461–1471, May 2016. doi:10.1080/00268976.2015.1136003. URL http://www.tandfonline.com/doi/full/10.1080/00268976.2015.1136003.

[113] Thomas E. Markland and David E. Manolopoulos. A refined ring polymer contraction scheme for systems with electrostatic interactions. *Chem. Phys. Lett.*, 464(4-6):256–261, Oct 2008. doi:10.1016/J.CPLETT.2008.09.019. URL https://www.sciencedirect.com/science/article/pii/S0009261408012542.

[114] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. Nuclear Quantum Effects in Solids Using a Colored-Noise Thermostat. *Phys. Rev. Lett.*, 103(3):30603, Jul 2009. doi:10.1103/PhysRevLett.103.030603. URL https://link.aps.org/doi/10.1103/PhysRevLett.103.030603.

[115] Venkat Kapil, Jörg Behler, and Michele Ceriotti. High order path integrals made easy. *J. Chem. Phys.*, 145(23):234103, Dec 2016. doi:10.1063/1.4971438. URL http://aip.scitation.org/doi/10.1063/1.4971438.

[116] Zhi Liang and Hai-Lung Tsai. Molecular dynamics simulations of self-diffusion coefficient and thermal conductivity of methane at low and moderate densities. *Fluid Phase Equilib.*, 297(1):40–45, Oct 2010. doi:10.1016/j.fluid.2010.06.008. URL http://www.sciencedirect.com/science/article/pii/S0378381210003079.

[117] Robert Hellmann, Eckard Bich, Eckhard Vogel, Alan S Dickinson, and Velisa Vesovic. Calculation of the transport and relaxation properties of methane. I. Shear viscosity, viscomagnetic effects, and self-diffusion. *J. Chem. Phys.*, 129(6):064302, Aug 2008. doi:10.1063/1.2958279. URL http://www.ncbi.nlm.nih.gov/pubmed/18715064.

[118] Mariana Rossi, Venkat Kapil, and Michele Ceriotti. Fine Tuning Classical and Quantum Molecular Dynamics using a Generalized Langevin Equation. *J. Chem. Phys.*, 148(10):102301, Mar 2017. doi:10.1063/1.4990536. URL http://aip.scitation.org/doi/10.1063/1.4990536.

[119] M A van der Hoef and D Frenkel. Long-time tails of the velocity autocorrelation function in two- and three-dimensional lattice-gas cellular automata: A test of mode-coupling theory. *Phys. Rev. A*, 41(8):4277–4284, Apr 1990. doi:10.1103/PhysRevA.41.4277. URL https://link.aps.org/doi/10.1103/PhysRevA.41.4277.

[120] Ebrahim Nemati-Kande and Ali Maghari. Transport properties of methane, ethane, propane, iso-butane and neo-pentane from ab initio potential energy surfaces. *J. Iran. Chem. Soc.*, pages 1–9, Mar

2016. doi:10.1007/s13738-016-0837-7. URL http://link.springer.com/10.1007/s13738-016-0837-7.

[121] J. Hutson. Intermolecular Forces From The Spectroscopy Of Van Der Waals Molecules. *Annu. Rev. Phys. Chem.*, 41(1):123–154, oct 1990. ISSN 0066426X. doi:10.1146/annurev.physchem.41.1.123. URL http://www.annualreviews.org/doi/10.1146/annurev.pc.41.100190.001011.

[122] Jeremy M. Hutson. Vibrational dependence of the anisotropic intermolecular potential of Ar-HF. *J. Chem. Phys.*, 96(9):6752–6767, may 1992. ISSN 00219606. doi:10.1063/1.462563. URL http://aip.scitation.org/doi/10.1063/1.462563.

[123] Robert D. Goodwin and Rolf Prydz. Densities of compressed liquid methane, and the equation of state. *J. Res. Natl. Bur. Stand. Sect. A Phys. Chem.*, 76A(2):81, Mar 1972. doi:10.6028/jres.076A.010. URL http://nvlpubs.nist.gov/nistpubs/jres/76A/jresv76An2p81_A1b.pdf.

[124] Steve Plimpton. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.*, 117(1):1–19, Mar 1995. doi:10.1006/jcph.1995.1039. URL http://www.sciencedirect.com/science/article/pii/S002199918571039X.

[125] James P Bareman and Michael L Klein. Collective tilt behavior in dense, substrate-supported monolayers of long-chain molecules: a molecular dynamics study. *J. Phys. Chem.*, 94(13):5202–5205, 1990. doi:10.1021/j100376a003. URL https://doi.org/10.1021/j100376a003.

[126] M A Moller, D J Tildesley, K S Kim, and N Quirke. Molecular dynamics simulation of a Langmuir–Blodgett film. *J. Chem. Phys.*, 94(12):8390–8401, 1991. doi:10.1063/1.460071. URL https://doi.org/10.1063/1.460071.

[127] Liguo Kong, Florian A Bischoff, and Edward F Valeev. Explicitly correlated R12/F12 methods for electronic structure. *Chem. Rev.*, 112(1):75–107, Jan 2012. doi:10.1021/cr200204r. URL http://dx.doi.org/10.1021/cr200204r.

[128] Thom H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.*, 90(2):1007–1023, Jan 1989. doi:10.1063/1.456153. URL http://aip.scitation.org/doi/10.1063/1.456153.

[129] Rick A. Kendall, Thom H. Dunning, and Robert J. Harrison. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.*, 96(9):6796–6806, May 1992.

doi:10.1063/1.462569. URL http://aip.scitation.org/doi/10.1063/1.462569.

[130] Thomas B. Adler, Gerald Knizia, and Hans-Joachim Werner. A simple and efficient CCSD(T)-F12 approximation. *J. Chem. Phys.*, 127(22):221106, Dec 2007. doi:10.1063/1.2817618. URL http://aip.scitation.org/doi/10.1063/1.2817618.

[131] Gerald Knizia, Thomas B. Adler, and Hans-Joachim Werner. Simplified CCSD(T)-F12 methods: Theory and benchmarks. *J. Chem. Phys.*, 130(5):054104, Feb 2009. doi:10.1063/1.3054300. URL http://aip.scitation.org/doi/10.1063/1.3054300.

[132] S.F. Boys and F. Bernardi. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.*, 19(4):553–566, Oct 1970. doi:10.1080/00268977000101561. URL http://www.tandfonline.com/doi/full/10.1080/00268977000101561.

[133] 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, and M Wang. MOLPRO, version 2012.1, a package of ab initio programs, 2012.

[134] H.-J. Werner, P J Knowles, G Knizia, F R Manby, and M Schütz. Molpro: a general-purpose quantum chemistry program package. *WIREs Comput Mol Sci*, 2:242–253, 2012.

[135] Martin Schütz, Roland Lindh, and Hans-JOACHIM Werner. Integral-direct electron correlation methods. *Mol. Phys.*, 96(4):719–733, 1999. doi:10.1080/00268979909483008. URL http://dx.doi.org/10.1080/00268979909483008.

[136] Roland Lindh. The reduced multiplication scheme of the Rys-Gauss quadrature for 1st order integral derivatives. *Theor. Chim. Acta*, 85(6):423–440, 1993. doi:10.1007/BF01112982. URL http://dx.doi.org/10.1007/BF01112982.

[137] Claudia Hampel and Hans-Joachim Werner. Local treatment of electron correlation in coupled cluster theory. *J. Chem. Phys.*, 104(16):6286–6297, 1996. doi:10.1063/1.471289. URL https://doi.org/10.1063/1.471289.

[138] P Hohenberg and W Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136: B864–B871, 1964.

[139] W Kohn and L J Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, 1965.

[140] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, 77(18):3865–3868, Oct 1996. doi:10.1103/PhysRevLett.77.3865. URL https://link.aps.org/doi/10.1103/PhysRevLett.77.3865.

[141] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.*, 110(13):6158–6170, 1999. doi:10.1063/1.478522. URL https://doi.org/10.1063/1.478522.

[142] X. Chu and A. Dalgarno. Linear response time-dependent density functional theory for van der Waals coefficients. *J. Chem. Phys.*, 121(9):4083–4088, Sep 2004. doi:10.1063/1.1779576. URL http://aip.scitation.org/doi/10.1063/1.1779576.

[143] Axel D Becke and Erin R Johnson. Exchange-hole dipole moment and the dispersion interaction: high-order dispersion coefficients. *J. Chem. Phys.*, 124(1):14104, Jan 2006. doi:10.1063/1.2139668. URL http://dx.doi.org/10.1063/1.2139668.

[144] Tore Brinck, Jane S. Murray, and Peter Politzer. Polarizability and volume. *The Journal of Chemical Physics*, 98(5):4305–4306, 1993. doi:10.1063/1.465038. URL https://doi.org/10.1063/1.465038.

[145] F L Hirshfeld. Bonded-atom fragments for describing molecular charge densities. *Theor. Chim. Acta*, 44(2):129–138, 1977. doi:10.1007/BF00549096. URL http://dx.doi.org/10.1007/BF00549096.

[146] R. F. Nalewajski and R. G. Parr. Information theory, atoms in molecules, and molecular similarity. *Proc. Natl. Acad. Sci.*, 97(16):8879–8882, Aug 2000. doi:10.1073/pnas.97.16.8879. URL http://www.pnas.org/cgi/content/long/97/16/8879.

[147] Paul W. Ayers. Information Theory, the Shape Function, and the Hirshfeld Atom. *Theor. Chem. Acc.*, 115(5):370–378, Feb 2006. doi:10.1007/s00214-006-0121-5. URL http://link.springer.com/10.1007/s00214-006-0121-5.

[148] Patrick Bultinck, Christian Van Alsenoy, Paul W. Ayers, and Ramon Carbó-Dorca. Critical analysis and extension of the Hirshfeld atoms in molecules. *J. Chem. Phys.*, 126(14):144111, Apr 2007.

doi:10.1063/1.2715563. URL http://scitation.aip.org/content/aip/journal/jcp/126/14/10.1063/1.2715563.

[149] Petr Jurečka, Jiří Šponer, Jiří Černý, and Pavel Hobza. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.*, 8(17):1985–1993, 2006. doi:10.1039/B600027D. URL http://dx.doi.org/10.1039/B600027D.

[150] Robert M Parrish, Lori A Burns, Daniel G A Smith, Andrew C Simmonett, A Eugene DePrince, Edward G Hohenstein, Uğur Bozkaya, Alexander Yu. Sokolov, Roberto Di Remigio, Ryan M Richard, Jérôme F Gonthier, Andrew M James, Harley R McAlexander, Ashutosh Kumar, Masaaki Saitow, Xiao Wang, Benjamin P Pritchard, Prakash Verma, Henry F Schaefer, Konrad Patkowski, Rollin A King, Edward F Valeev, Francesco A Evangelista, Justin M Turney, T Daniel Crawford, and C David Sherrill. Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.*, 13(7):3185–3197, 2017. doi:10.1021/acs.jctc.7b00174. URL https://doi.org/10.1021/acs.jctc.7b00174.

[151] Toon Verstraelen, Pawel Tecmer, Farnaz Heidar-Zadeh, Cristina E. González-Espinoza, Matthew Chan, Taewon D. Kim, Katharina Boguslawski, Stijn Fias, Steven Vandenbrande, Diego Berrocal, and Paul W. Ayers. HORTON, 2017. URL http://theochem.github.com/horton/.

[152] A. D. Becke. A multicenter numerical integration scheme for polyatomic molecules. *J. Chem. Phys.*, 88(4):2547–2553, 1988. doi:10.1063/1.454033.

[153] A. D. Becke and R. M. Dickson. Numerical solution of poisson's equation in polyatomic molecules. *J. Chem. Phys.*, 89(5):2993–2997, 1988. doi:10.1063/1.455005.

[154] V.I. Lebedev and D.N. Laikov. A quadrature formula for the sphere of the 131st algebraic order of accuracy. *Dokl. Math.*, 59(3):477–481, 1999.

[155] Tomáš Bučko, Sébastien Lebègue, János G Ángyán, and Jürgen Hafner. Extending the applicability of the Tkatchenko-Scheffler dispersion correction via iterative Hirshfeld partitioning. *J. Chem. Phys.*, 141(3):034114, Jul 2014. doi:10.1063/1.4890003. URL https://aip.scitation.org/doi/10.1063/1.4890003.

[156] S J Clark, M D Segall, C J Pickard, P J Hasnip, M J Probert, K Refson, and M C Payne. First principles methods using CASTEP. *Z. Krist.*, 220:567–570, 2005.

[157] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A Unifying View of Sparse Approximate Gaussian Process Regression. *J. Mach. Learn. Res.*, 6:1939–1959, dec 2005. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1046920.1194909.

[158] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.*, 134(7):074106, Feb 2011. doi:10.1063/1.3553717. URL http://scitation.aip.org/content/aip/journal/jcp/134/7/10.1063/1.3553717.

[159] Andrea Grisafi, David M. Wilkins, Gábor Csányi, and Michele Ceriotti. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.*, 120(3):036002, jan 2018. ISSN 10797114. doi:10.1103/PhysRevLett.120.036002. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.036002.

[160] E Balog, A L Hughes, and G J Martyna. Constant pressure path integral molecular dynamics studies of quantum effects in the liquid state properties of n-alkanes. *J. Chem. Phys.*, 112(2):870–880, 2000. doi:10.1063/1.480614. URL https://doi.org/10.1063/1.480614.

[161] A S Teja, R J Lee, D Rosenthal, and M Anselme. Correlation of the critical properties of alkanes and alkanols. *Fluid Phase Equilib.*, 56:153–169, 1990. doi:https://doi.org/10.1016/0378-3812(90)85100-O. URL http://www.sciencedirect.com/science/article/pii/037838129085100O.

[162] Max Veit, Sandeep Kumar Jain, Satyanarayana Bonakala, Indranil Rudra, Detlef Hohl, and Gábor Csányi. Equation of state of fluid methane from first principles with machine learning potentials. *Journal of Chemical Theory and Computation*, 2019. URL https://doi.org/10.1021/acs.jctc.8b01242.

[163] George Kaminski, Erin M. Duffy, Tooru Matsui, and William L. Jorgensen. Free Energies of Hydration and Pure Liquid Properties of Hydrocarbons from the OPLS All-Atom Model. *J. Phys. Chem.*, 98(49):13077–13082, Dec 1994. doi:10.1021/j100100a043. URL http://dx.doi.org/10.1021/j100100a043.

[164] J E Lennard-Jones. Cohesion. *Proc. Phys. Soc.*, 43(5):461–482, Sep 1931. doi:10.1088/0959-5309/43/5/301. URL http://stacks.iop.org/0959-5309/43/i=5/a=301?key=crossref.81d7fd40429f59f086896015181ad062.

[165] R. A. Buckingham and J. Corner. Tables of Second Virial and Low-Pressure Joule-Thomson Coefficients for Intermolecular Potentials with Exponential Repulsion. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, 189(1016):

118–129, Mar 1947. doi:10.1098/rspa.1947.0032. URL http://rspa.royalsocietypublishing.org/content/189/1016/118.

[166] Mary J Van Vleet, Alston J Misquitta, and J R Schmidt. New Angles on Standard Force Fields: Toward a General Approach for Treating Atomic-Level Anisotropy. *J. Chem. Theory Comput.*, 14(2):739–758, 2018. doi:10.1021/acs.jctc.7b00851. URL https://doi.org/10.1021/acs.jctc.7b00851.

[167] Martin A. Blood-Forsythe, Thomas Markovich, Robert A. DiStasio, Roberto Car, and Alán Aspuru-Guzik. Analytical nuclear gradients for the range-separated many-body dispersion model of noncovalent interactions. *Chem. Sci.*, 7(3):1712–1728, Feb 2016. doi:10.1039/C5SC03234B. URL http://pubs.rsc.org/en/content/articlehtml/2016/sc/c5sc03234b.

[168] Alexander G. Shtukenberg, Qiang Zhu, Damien J. Carter, Leslie Vogt, Johannes Hoja, Elia Schneider, Hongxing Song, Boaz Pokroy, Iryna Polishchuk, Alexandre Tkatchenko, Artem R. Oganov, Andrew L. Rohl, Mark E. Tuckerman, and Bart Kahr. Powder diffraction and crystal structure prediction identify four new coumarin polymorphs. *Chem. Sci.*, 8(7):4926–4940, Jun 2017. doi:10.1039/C7SC00168A. URL http://xlink.rsc.org/?DOI=C7SC00168A.

[169] Leonid Pereyaslavets, Igor Kurnikov, Ganesh Kamath, Oleg Butin, Alexey Illarionov, Igor Leontyev, Michael Olevanov, Michael Levitt, Roger D Kornberg, and Boris Fain. On the importance of accounting for nuclear quantum effects in ab initio calibrated force fields in biological simulations. *Proc. Natl. Acad. Sci.*, 115(36):8878–8882, 2018. doi:10.1073/pnas.1806064115. URL http://www.pnas.org/content/115/36/8878.

[170] Alston J. Misquitta and Anthony J. Stone. Ab Initio Atom-Atom Potentials Using CamCASP: Theory and Application to Many-Body Models for the Pyridine Dimer. *J. Chem. Theory Comput.*, 12(9):4184–4208, sep 2016. ISSN 15499626. doi:10.1021/acs.jctc.5b01241. URL http://pubs.acs.org/doi/10.1021/acs.jctc.5b01241.

[171] Steven Vandenbrande, Michel Waroquier, Veronique Van Speybroeck, and Toon Verstraelen. The monomer electron density force field (medff): A physically inspired model for noncovalent interactions. *Journal of Chemical Theory and Computation*, 13(1):161–179, 2017. doi:10.1021/acs.jctc.6b00969. URL https://doi.org/10.1021/acs.jctc.6b00969. PMID: 27935712.

[172] Daniel J. Cole, Jonah Z. Vilseck, Julian Tirado-Rives, Mike C. Payne, and William L. Jorgensen. Biomolecular Force Field Parameterization

via Atoms-in-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.*, 12(5):2312–2323, May 2016. doi:10.1021/acs.jctc.6b00027. URL http://pubs.acs.org/doi/10.1021/acs.jctc.6b00027.

[173] Alexander A. Aina, Alston J. Misquitta, and Sarah L. Price. From dimers to the solid-state: Distributed intermolecular force-fields for pyridine. *J. Chem. Phys.*, 147(16):161722, oct 2017. ISSN 0021-9606. doi:10.1063/1.4999789. URL http://aip.scitation.org/doi/10.1063/1.4999789.

[174] Thuong T. Nguyen, Eszter Székely, Giulio Imbalzano, Jörg Behler, Gábor Csányi, Michele Ceriotti, Andreas W. Götz, and Francesco 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):241725, jun 2018. ISSN 0021-9606. doi:10.1063/1.5024577. URL http://aip.scitation.org/doi/10.1063/1.5024577.

[175] William L. Jorgensen and Julian. Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110(6):1657–1666, Mar 1988. doi:10.1021/ja00214a001. URL http://dx.doi.org/10.1021/ja00214a001.

[176] Alberto Ambrosetti, Anthony M. Reilly, Robert A. DiStasio, and Alexandre Tkatchenko. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.*, 140(18):18A508, May 2014. doi:10.1063/1.4865104. URL http://aip.scitation.org/doi/10.1063/1.4865104.

[177] David M Wilkins, Andrea Grisafi, Yang Yang, Ka Un Lao, Robert A Distasio, and Michele Ceriotti. Accurate molecular polarizabilities with coupled-cluster theory and machine learning. *arXiv:1809.05337*, 2018. URL https://arxiv.org/pdf/1809.05337.pdf.

[178] J D Hunter. Matplotlib: A 2D graphics environment. *Comput. Sci. & Eng.*, 9(3):90–95, 2007.

[179] Fernando Pérez and Brian E Granger. IPython: a System for Interactive Scientific Computing. *Comput. Sci. Eng.*, 9(3):21–29, May 2007. doi:10.1109/MCSE.2007.53. URL http://ipython.org.

[180] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *J. Mol. Graph.*, 14:33–38, 1996.

[181] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N.

Groves, Bjørk Hammer, Cory Hargus, Eric D. Hermes, Paul C. Jennings, Peter Bjerre Jensen, James Kermode, John R. Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S. Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W. Jacobsen. The atomic simulation environment - A Python library for working with atoms. *J. Phys. Condens. Matter*, 29:273002, 2017. doi:10.1088/1361-648X/aa680e.

[182] Axel Brünger, Charles L Brooks, and Martin Karplus. Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chem. Phys. Lett.*, 105(5):495–500, 1984. doi:https://doi.org/10.1016/0009-2614(84)80098-6. URL http://www.sciencedirect.com/science/article/pii/0009261484800986.

[183] M Parrinello and A Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981. doi:10.1063/1.328693. URL https://doi.org/10.1063/1.328693.

[184] Wataru Shinoda, Motoyuki Shiga, and Masuhiro Mikami. Rapid estimation of elastic constants by molecular dynamics simulation under constant stress. *Phys. Rev. B*, 69(13):134103, Apr 2004. doi:10.1103/PhysRevB.69.134103. URL https://link.aps.org/doi/10.1103/PhysRevB.69.134103.

[185] Mark E Tuckerman, José Alejandre, Roberto López-Rendón, Andrea L Jochim, and Glenn J Martyna. A Liouville-operator derived measure-preserving integrator for molecular dynamics simulations in the isothermal–isobaric ensemble. *J. Phys. A. Math. Gen.*, 39 (19):5629–5651, May 2006. doi:10.1088/0305-4470/39/19/S18. URL http://stacks.iop.org/0305-4470/39/i=19/a=S18?key=crossref.a2cc46a8f51a0e19c9850cf3f0d88954.

[186] Michele Ceriotti, Joshua More, and David E. Manolopoulos. i-PI: A Python interface for ab initio path integral molecular dynamics simulations. *Comput. Phys. Commun.*, 185(3):1019–1026, Mar 2014. doi:10.1016/J.CPC.2013.10.027. URL https://www.sciencedirect.com/science/article/pii/S001046551300372X.

[187] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.*, 30(13):2157–2164, Oct 2009. doi:10.1002/jcc.21224. URL http://doi.wiley.com/10.1002/jcc.21224.

[188] Roger W Hockney and James W Eastwood. *Computer simulation using particles*. Hilger, Bristol, 1988. ISBN 0852743920.

[189] Jan Řezáč, Kevin E Riley, and Pavel Hobza. S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.*, 7(8):2427–2438, 2011. doi:10.1021/ct2002946. URL https://doi.org/10.1021/ct2002946.

[190] W Kohn. Density Functional and Density Matrix Method Scaling Linearly with the Number of Atoms. *Phys. Rev. Lett.*, 76(17):3168–3171, Apr 1996. doi:10.1103/PhysRevLett.76.3168. URL https://link.aps.org/doi/10.1103/PhysRevLett.76.3168.

[191] Roi Baer and Martin Head-Gordon. Sparsity of the Density Matrix in Kohn-Sham Density Functional Theory and an Assessment of Linear System-Size Scaling Methods. *Phys. Rev. Lett.*, 79(20):3962–3965, Nov 1997. doi:10.1103/PhysRevLett.79.3962. URL http://link.aps.org/doi/10.1103/PhysRevLett.79.3962.

[192] E Prodan and W Kohn. Nearsightedness of electronic matter. *Proc. Natl. Acad. Sci.*, 102(33):11635–11638, 2005. doi:10.1073/pnas.0505436102. URL http://www.pnas.org/content/102/33/11635.

[193] P. E. Maslen, C. Ochsenfeld, C. A. White, M. S. Lee, and M. Head-Gordon. Locality and Sparsity of Ab Initio One-Particle Density Matrices and Localized Orbitals. *J. Phys. Chem. A*, 102(12):2215–2222, Mar 1998. doi:10.1021/jp972919j. URL http://dx.doi.org/10.1021/jp972919j.

[194] M Elstner. The SCC-DFTB method and its application to biological systems. *Theor. Chem. Acc.*, 116(1-3):316–325, 2006. doi:10.1007/s00214-005-0066-0. URL http://dx.doi.org/10.1007/s00214-005-0066-0.

[195] Alfred Karpfen, Cheol Ho Choi, and Miklos Kertesz. Single-Bond Torsional Potentials in Conjugated Systems: A Comparison of ab Initio and Density Functional Results. *J. Phys. Chem. A*, 101(40):7426–7433, 1997. doi:10.1021/jp971606l. URL http://pubs.acs.org/doi/abs/10.1021/jp971606l.

[196] Dóra Barna, Balázs Nagy, József Csontos, Attila G. Császár, and Gyula Tasi. Benchmarking Experimental and Computational Thermochemical Data: A Case Study of the Butane Conformers. *J. Chem. Theory Comput.*, 8(2):479–486, Feb 2012. doi:10.1021/ct2007956. URL http://pubs.acs.org/doi/10.1021/ct2007956.

[197] István Kolossváry and Colin McMartin. On the degeneracy of the Hessian matrix. *J. Math. Chem.*, 9(4):359–367, Dec 1992. doi:10.1007/BF01166099. URL https://doi.org/10.1007/BF01166099.

[198] Larry A Curtiss, Krishnan Raghavachari, Paul C Redfern, and John A Pople. Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. *J. Chem. Phys.*, 106(3):1063–1079, 1997. doi:10.1063/1.473182. URL https://doi.org/10.1063/1.473182.

[199] Avogadro: an open-source molecular builder and visualization tool. Version 1.1.1, 2017. URL http://avogadro.openmolecules.net/.

[200] Seiji Tsuzuki, Lothar Schafer, Hitoshi Goto, Eluvathingal D Jemmis, Haruo Hosoya, Khamis Siam, Kazutoshi Tanabe, and Eiji Osawa. Investigation of intramolecular interactions in n-alkanes. Cooperative energy increments associated with GG and GTG' [G = gauche, T = trans] sequences. *J. Am. Chem. Soc.*, 113(12):4665–4671, 1991. doi:10.1021/ja00012a040. URL https://doi.org/10.1021/ja00012a040.

[201] Federico M. Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Emsley. Chemical shifts in molecular solids by machine learning. *Nat. Commun.*, 9(1):4501, dec 2018. ISSN 2041-1723. doi:10.1038/s41467-018-06972-x. URL http://www.nature.com/articles/s41467-018-06972-x.