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, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

E-mail: max.veit@epfl.ch

Abstract

The predictive simulation of molecular liquids requires potential energy surface (PES) models that are not only accurate, but computationally efficient enough to handle the large systems and long time scales required for reliable prediction of macroscopic properties. We present a new approach to the systematic approximation of the first-principles PES of molecular liquids using the GAP (Gaussian Approximation Potential) framework. The approach allows us to create potentials at several different levels of accuracy in reproducing the true PES, and thus to determine the level of quantum chemistry that is necessary to accurately predict macroscopic properties. We test the approach by building a series of many-body potentials for liquid methane (CH_4), which is difficult to model from first principles because its behaviour is dominated by weak dispersion interactions with a significant many-body component. The increasing accuracy of the potentials in predicting the bulk density correlates with their fidelity to the true PES, whereas the trend with the empirical potentials tested is surprisingly the opposite. We conclude that an accurate, consistent prediction of its bulk density across wide ranges of temperature and pressure requires not only many-body dispersion, but also quantum nuclear effects to be modelled accurately.

1 Introduction

The accurate simulation of molecular liquids is a problem of great scientific and industrial importance. We not only would like to be able to test the predictions of our models against experimental benchmarks to see where they need to be refined, but we also need to make predictions for new compounds or mixtures in order to identify the most promising candidates for future applications. When modelling molecular liquids one is typically obliged to trade off accuracy in the description of the potential energy surface and errors due to insufficient sampling. In this work we aim to perform simulations of *ab initio* quality but with the orders of magnitude boost in computational efficiency afforded by high dimensional regression using techniques analogous to those in machine learning. Following notable recent

success for approximating the energy of individual molecules,[1–5] here we tackle the problem of intermolecular interactions. By breaking down the total interaction potential into different components, we show explicitly that they are all modelled sufficiently accurately, and thus we obtain the right answers *for the right reasons* rather than due to uncontrolled cancellation of errors. Specifically, we create Gaussian approximation potentials (GAPs)[6–8] for liquid methane, the simplest alkane, which is inherently difficult to model because its behaviour is dominated by weak dispersion interactions. The condensed phase of methane is interesting in its own right, notably for its role in the geochemistry of Titan,[9,10] in the atmospheres of gas giants,[11] and elsewhere in the solar system.[12] Its condensed-phase mixtures with water are subjects of recent research; clathrates are the best known example, though the recently uncovered puzzle of the solubility of liquid methane in water at high pressure[13] shows there is plenty more fertile ground for investigation. This work also opens the door to potentials that can model larger hydrocarbons under extreme temperatures and pressures;[14,15] such a potential would enable new research in numerous scientific and engineering applications.[16–18]

There is a long history of modelling liquids at the atomistic scale with Monte Carlo (MC) or molecular dynamics (MD) methods. The interactions between constituent particles are often modelled using analytical potentials, which are a combination of a few simple, physically motivated expressions, such as the venerable Lennard-Jones (L-J) potential[19] and the many subsequent variations or extensions of its basic form.[20–25] 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[23,26,27] or even directly fitting the intramolecular[28] or intermolecular[15,29–32] terms to *ab initio* calculations. Such potentials, which are the type most commonly employed in simulations of liquids, have achieved

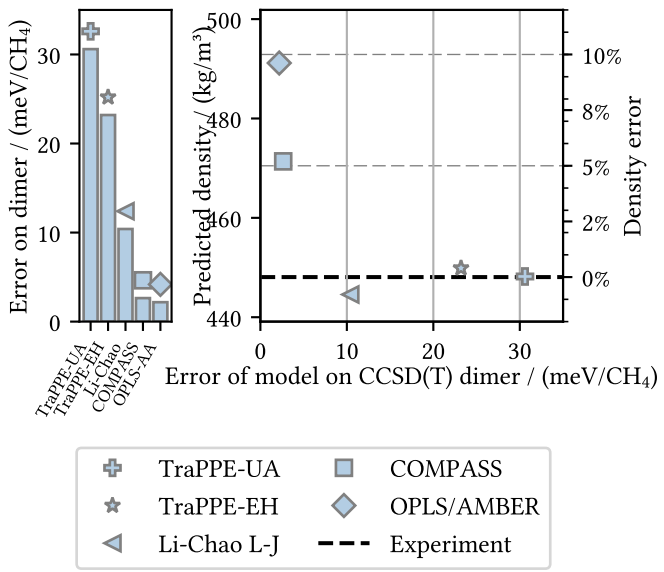


Figure 1: Accuracy of some commonly-used L-J-type empirical models for methane against the quantum mechanical potential energy surface, compared with the accuracy of the density each model predicts for bulk methane. The PES accuracy of each model is measured by the RMS error of the model's predicted energies of a sample of dimers, measured against quantum chemical (CCSD(T)-F12) reference energies; The error is computed over the sample of dimers used to train the 6-D dimer GAP (note 10 meV $\approx$ 0.96 kJ/mol $\approx$ 0.23 kcal/mol). The density predictions were done at 110 K and 316 bar. Density error is given relative to experiment; the uncertainties on the density are smaller than the sizes of the symbols.

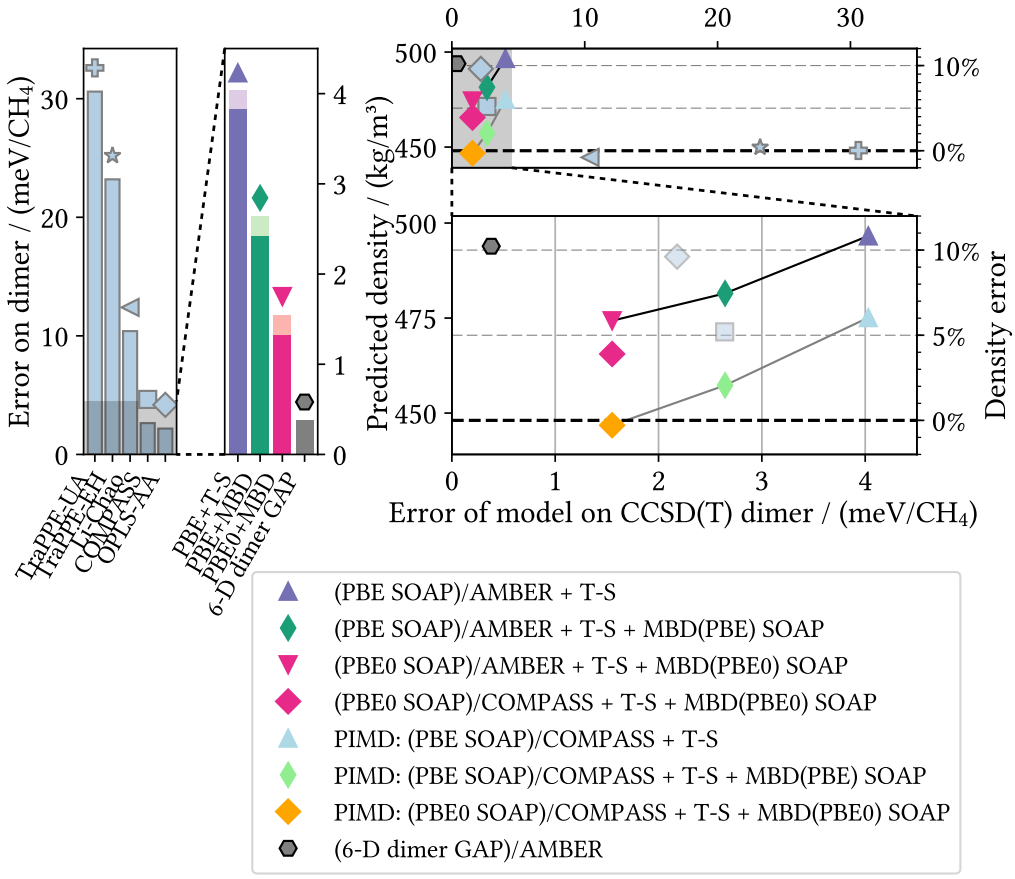


Figure 2: Comparison of PES accuracy versus density accuracy for the machine learning models for methane developed in this work; the equivalent comparison for empirical L-J-type models from Figure 1 is reproduced at the top and far left. The suffixes "/AMBER" and "/COMPASS" indicate which model was used 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). 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. Refer to the legend of Figure 1 for symbols previously defined.

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,[21,33] 9-6,[23] or Morse[15,30] potentials – remain too simple to represent the underlying potential energy surface truly 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.[34,35] 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 predictions 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[36] and the reduced dimensional version TraPPE-EH[34]) eliminate degrees of freedom in order to obtain a simpler description of the system. They have been fitted to accurately reproduce phase equilibria; they also deliver an accurate prediction of the equation of state of liquid methane. Figure 1 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[33] 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[32] is five times worse on the dimer than OPLS-AA, but ten times better in its prediction of the density. COMPASS[23] is slightly worse on the dimer than OPLS-AA, while achieving a density prediction almost twice as good as that of OPLS-AA.

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 simply conclude that 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,[29,31,37] 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 3. While the electrostatic component is often treated within a formally many-body framework,[37] other components such as the repulsion and the dispersion also exhibit significant many-body character[38] that is less commonly taken into account 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)[6,7] (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 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 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.

1.1 Quantum-mechanical energies

Several methods are available that approximate the true quantum potential energy surface. Perhaps the best known of these is density functional theory (DFT),[40] which is generally good at predicting covalent bond energies and intermolecular repulsive interactions. Standard DFT lacks dispersion interactions, however, so these must be added separately.[41] Dispersion correction schemes for DFT are generally inverse-power terms added on to the total DFT energy. They range from terms with fixed semiempirical coefficients[42] to explicitly geometry-dependent terms,[43] to terms with coefficients that use information from an existing DFT calculation.[44–46] Many of these schemes, such as DFT-D3[43] and MBD,[46] account for many-body (i.e. beyond pairwise additive) dispersion interactions. This many-body effect has been shown to be crucial for an accurate description of many dispersion-bound systems such as supramolecular complexes[47] and organic crystals,[48] though the effects on molecular liquids have not yet been extensively studied – a many-body vdW model (D3[43]) *was* included in the water potential of Morawietz *et al.*,[49] but it was not mentioned whether a simple pairwise model would have given different results.

The main drawback of quantum methods that treat electrons explicitly, such as DFT or quantum chemistry, is their computational cost: MD simulations to predict liquid properties

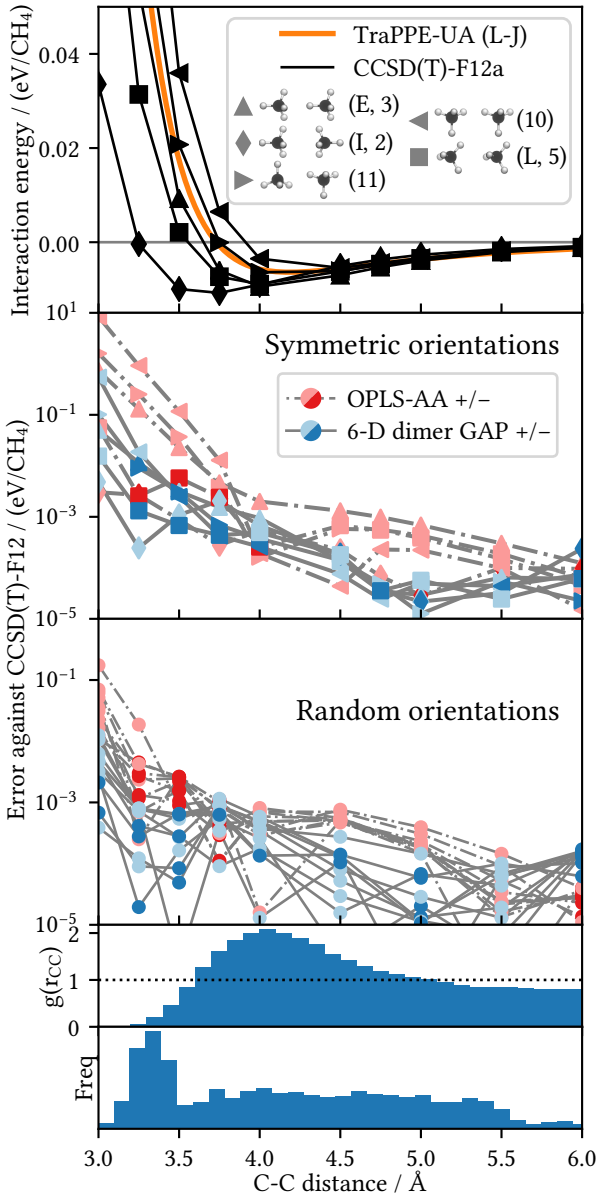


Figure 3: Top: Interaction energies of the rigid methane dimer in a selection of orientations. The TraPPE united-atom (and therefore isotropic) model[36] is given by the smooth line; this model gives the best overall prediction of the equation of state (Figure 6) even though it completely ignores the anisotropy. Configurations are labeled as in Chao *et al.*[39] (letters) and Hellmann *et al.*[31] (numbers). Middle: Errors of two models on the methane dimer energy, the OPLS-AA model[33] 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.

routinely require millions of force evaluations on thousands of atoms,[17] which would be prohibitive even for today's fastest computers using the most efficient implementations of DFT. Furthermore, MD simulations require force evaluations on many highly correlated configurations. But the Born-Oppenheimer potential energy surface is usually assumed to be smooth and regular, at least in the ordinary realm of closed-shell molecules far away from level crossings and other exotic PES irregularities. Thanks to this regularity, highly correlated (similar) configurations will also have highly correlated energies and forces. This correlation can be exploited to greatly reduce the number of force evaluations required for a molecular simulation.

1.2 Machine learning potentials

A new generation of potentials aims to exploit this correlation by using machine learning techniques to directly fit the Born-Oppenheimer potential energy surface.[6,50] These fits do not constrain the potential's functional form, relying instead on a sufficient sample of existing calculations to be able to regress (fit) these data points in the high-dimensional space of nuclear positions. Such potentials are designed to capture much of the accuracy and flexibility offered by full quantum methods but with a computational efficiency that is many orders of magnitude higher, enabling MD simulations for system sizes and timescales previously only accessible to empirical, analytical potentials.

Machine learning potentials have been applied to a wide variety of systems.[51] The GAP method, for example, has been applied to systems ranging from the allotropes of silicon,[6] tungsten,[52] iron,[53] and boron;[54] molecular clusters[8] and liquids;[49,55] and amorphous materials.[56–58] There is also considerable interest in general, transferable molecular potentials[59] and accurate modelling of liquid water.[49] Recent progress has also been made in modelling multi-component systems,[60,61] and in interpolating between different chemical compounds[62–64] and even across different classes of materials,[2] thus approaching the level of flexibility currently offered by full quantum methods.

1.3 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.[65,66] These effects are especially important at low temperatures and with light nuclei; their importance in liquid alkanes in particular has long been established[67] and was recently highlighted[68] using quantum mechanically fitted forcefields. In empirical potentials these effects are typically included in an average way, since they are naturally present in the experimental data used to fit the potentials; some potentials[31] 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 most common and practical technique for including quantum nuclear effects (ZPVE and nuclear tunneling, but *not* the nuclear exchange) in MD simulations is via path integral molecular dynamics (PIMD), where the quantum system is represented by P replicas of the classical system, corresponding atoms being joined across the replicas by harmonic springs in a ring-polymer structure.[65,69–71] Recent techniques, including improved stochastic thermostats[72–74] and ring polymer contraction,[75] are making PIMD practical even for large systems and more expensive potentials such as the ones employed in this work.

Despite these new developments, *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, and therefore more easily automated, development of potentials for more complex molecular liquids.

2 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 represents a different physical interaction. 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. The energy equation we will use is therefore:

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

where the "1b" (one-body) energy is the covalent part and everything else makes up the intermolecular (more formally, beyond one-body or "b1b") energy. The repulsion and elec-

trostatic terms are computed from DFT beyond-one-body interactions – electrostatics, in contrast to dispersion, is handled 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.[76] To illustrate for the case of pure methane, the electrostatic energy predicted by OPLS-AA is consistently about two orders of magnitude smaller than the other non-bonded terms; see Figure 1 of the supporting information. 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[77] 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 energy along with the even smaller, shorter-ranged induction term and the strictly short-range penetration term into the short-range "repulsion" term – hereafter called $E_{\mathrm{sr,b1b}}$ (for "short-range beyond-one-body"). Future versions of this potential could easily treat electrostatics and induction 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. Some recent analytical potentials take the approach of more directly representing the underlying physics by extracting forcefield parameters from fundamental physical quantities such as the electron density. Models using this approach include the Slater-ISA model of Van Vleet *et al.*[37] (including the more recent anisotropic version[38]), the Monomer Electron Density Force Field of Vandenbrande *et al.*,[78] and the biomolecular force field of Cole *et al.*.[79] The IPML model of Bereau et. al.[80] goes one step

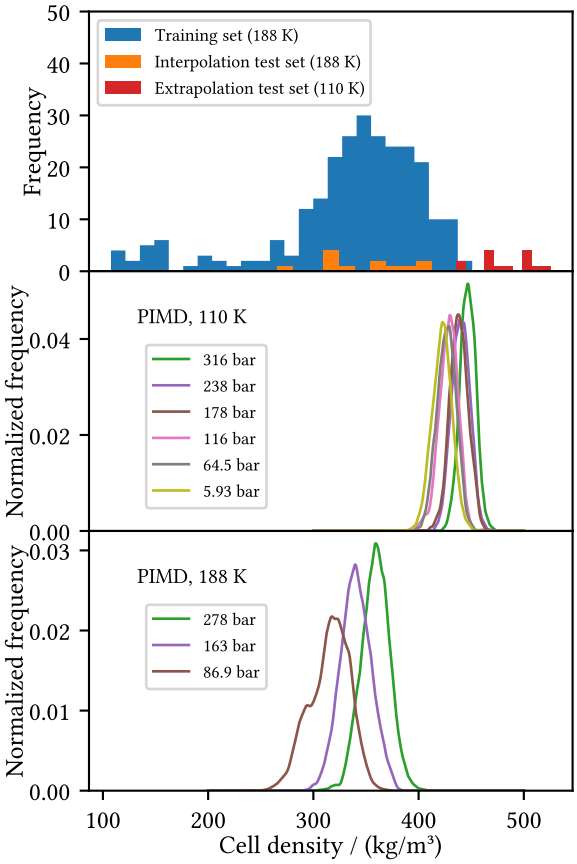


Figure 4: Histograms over mass density of the cells in the training and two test sets, interpolation and extrapolation. The distributions of densities encountered in the subsequent PIMD simulations with the (PBE0 SOAP)/COMPASS + T-S + MBD(PBE0) SOAP model are shown below for comparison.

further by using machine learning to efficiently predict these properties across chemical compound space. While the physical interpretability of these models is appealing, it comes at the cost of sacrificing a best-possible fit to the true quantum potential energy surface. In the present work, as described below, by capturing most of the dispersion energy with simple analytical form and fitting a correction on top, we do use physics to guide our description of the interaction while maintaining complete flexibility of the functional form.

2.1 Many-body machine learning model

To fit the $E_{\mathrm{sr,b1b}}$ term we use the GAP method[6,7] with the SOAP kernel,[81] both developed and used by our group to fit complex, many-body potentials. The SOAP-GAP potentials were fitted to DFT[82,83] energies and forces computed on 280 periodic unit cells representing bulk methane, each containing 27 methane molecules. The beyond-one-body components of energies and forces were obtained by separately computing them for all monomers and subtracting from the total. The samples were taken from MD trajectories performed under liquid conditions run using a classical potential (OPLS/AMBER[84,85]) at a temperature of 188 K and five pressures ranging from 0 bar to 400 bar. The resulting training set consisted of a wide range of densities; see Figure 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 6 below. The histogram of the densities of these test sets is also shown in Figure 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.[86] Two functionals were used, the pure GGA functional PBE[87] and the hybrid GGA functional PBE0.[88] The GAP fits were done using the SOAP descriptor,[81] resulting in two models called "PBE SOAP-GAP" and "PBE0 SOAP-GAP". The performance of the PBE0 SOAP-GAP is assessed in Figure 5, which indicates good reproduction of both energies and forces on the training set. Since GAP is a statistical learning method, 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

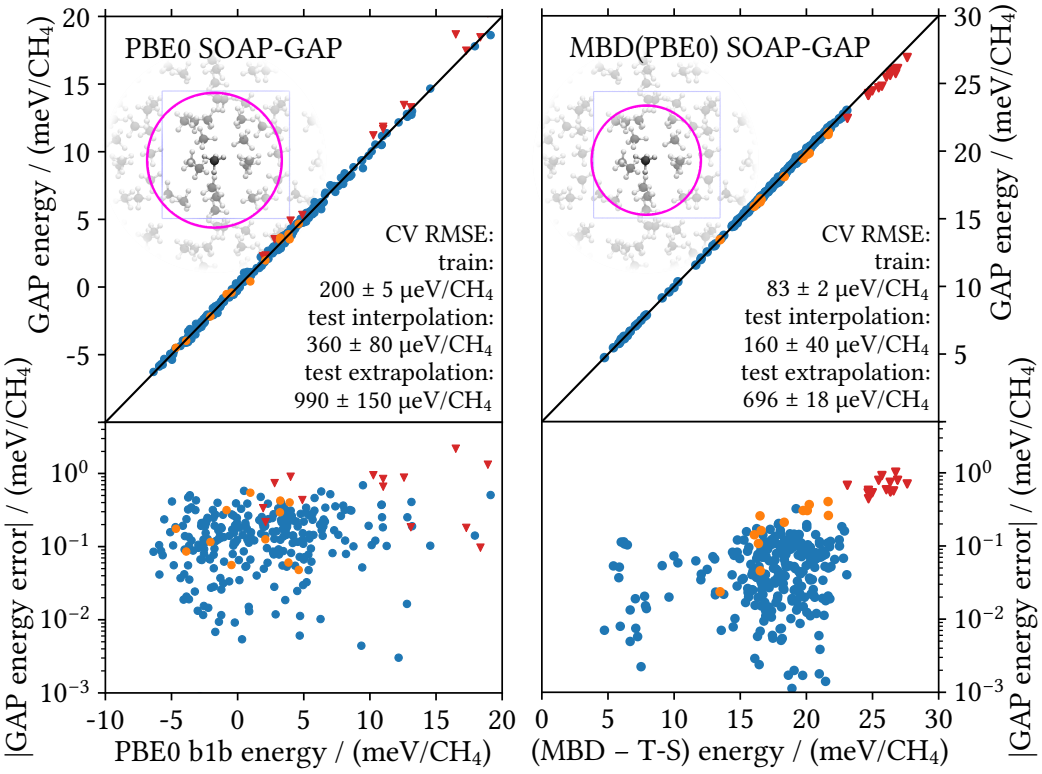


Figure 5: The PBE0 and MBD(PBE0) SOAP-GAP fits on 258 cell interaction (beyond one-body, "b1b") energies and (only for PBE0) corresponding forces. Top: Correlation plots with the line $y = x$ of perfect correlation. Bottom: Errors on a logarithmic scale. The blue dots represent the training set. The 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.

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 models. The numbers reported in Figure 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 the supporting information.

The computational effort required to generate the training database was considerable; a typical PBE calculation took 10 minutes on 24 processor cores on a Cray XC30 system,[89] 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 on the same system; 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,[90] and the evaluation of the SOAP-GAP energies and forces requires less than 3 processor-seconds on a cell of 100 methane molecules. A further advantage of the GAP approach becomes apparent here, 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. This initial computational cost is more than recovered by the subsequent savings in running the dynamics with a SOAP-GAP rather than with DFT. In fact, since doing large liquid simulations with DFT is still often beyond the capability of today's most powerful computers, the initial cost of the machine learning database and fitting serve to make the hitherto impossible possible.

2.2 Dispersion model

The dispersion component, the third term in Equation (1), was accounted for using two levels of theory. The first was the pairwise correction of Tkatchenko and Scheffler.[45] This method uses relative atomic volumes from a Hirshfeld partitioning[91] of the electron density, an idea introduced by Becke and Johnson,[44] and relates them to free-atom dispersion coefficients (those computed by Chu and Dalgarno[92] were used here). Recomputing the Hirshfeld volumes for each step of an MD simulation would be impractically expensive, 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 "T-S(fix)" or simply "T-S".

The second level of theory is the MBD, or many-body dispersion, method.[46,93] Despite the greater complexity of the MBD approach, we can still expect a large part of the total MBD energy to be captured by the pairwise Tkatchenko-Scheffler method, as evidenced by the success of the latter method in predicting dispersion energies. 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) therefore becomes:

$$E_{\text{dispersion}} = E_{\text{T-S(fix)}} + E_{\text{MBD SOAP-GAP}}. \tag{2}$$

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 5, showing that both its interpolation and extrapolation performance is similar to

that of the PBE0 SOAP-GAP.

Another, more technical motivation for fitting the dispersive interactions separately from the repulsive interactions (besides the ability to use a readily available baseline) is that analytical gradients, which significantly improve the fit, are easily available for the plain DFT energy but not for the MBD energy – at least, not in the real-space implementation used in this work. In principle, one could compensate for the lack of analytical gradient data by including more configurations, although in practice this was found not to be necessary.

Finally, a complete model for liquid methane must also include an intramolecular component (the first term in Equation (1)). Two empirical potentials are considered for this purpose: AMBER[85] includes only harmonic bond and angle terms, while COMPASS[23] 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.

3 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[84]) 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,[36] 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[94]): At 110 K, density measurements were available at 5.93 bar, 64.5 bar, 116 bar, 179 bar, 238 bar and 316 bar.[95] At 188 K, density measurements were available at 86.9 bar, 163 bar and 278 bar.[95]

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.[96] The convergence of the results towards the experimental density is illustrated in Figure 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 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.[73,74] 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

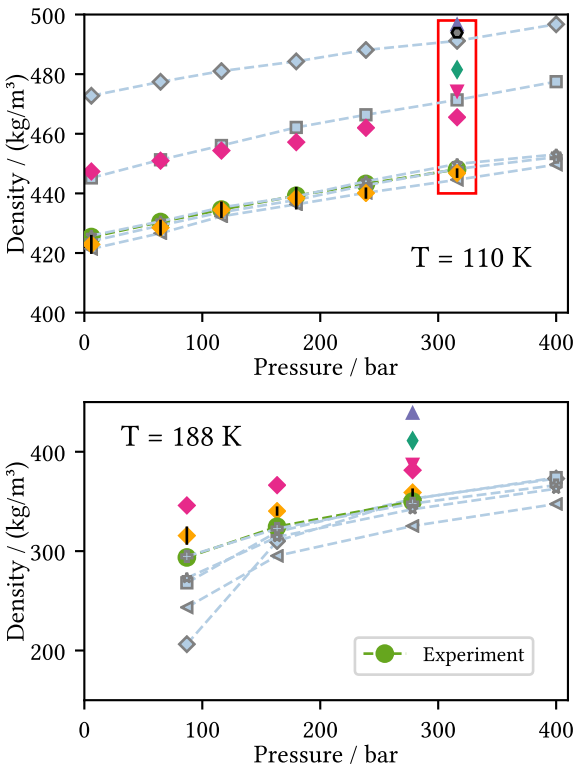


Figure 6: Equation of state at two temperatures, 110 K and 188 K, as predicted by various atomistic models. The bulk SOAP-GAPs with different dispersion models are shown, as is the 6-D dimer GAP. All-atom empirical models are shown in gray. Experimental data from Goodwin and Prydz.[94] The small black lines are error bars on the PIMD simulations computed using the blocking method described in the supporting information. Refer to the legends of Figures 1 and 2 for symbols previously defined.

magnitude as that reported in Pereyaslavets et al.[68], though with this potential the effect is smaller – 4.2 % instead of 9 %. Figure 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.[31,97,98] But these same effects also increase the molecular volume,[67] 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 6. For comparison, the L-J-type potentials from Figure 1 were 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. Note in particular that the empirical all-atom potentials all shift with respect to experiment between the two isotherms. Most models, the SOAP-GAPs 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[99]). 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[8,55]).

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 on the Cray XC30 system,[89] so the PIMD data points in Figure 6 required about twice as much time to generate as the PBE0 training set itself. Consider, however, that these potentials offer a speedup of between 5000 (PBE) to 30 000 (typical PBE0) over single-point DFT calculations on the system sizes tested; furthermore, the expensive short-range (SOAP-GAP) components of the potential scale essentially linearly with the system size thanks to their small, finite cutoffs. These improvements do more than just make simulations more efficient: They make the previously impossible – large, expensive liquid simulations, even with quantum nuclear effects, at the level of many-body dispersion-corrected DFT – possible.

4 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;[49] if long-range contributions are required, they can be computed using machine learning of local electrostatic properties.[80,100,101]

5 Computational Methods

5.1 Gaussian processes

The GAP machine learning method used to fit the potential energy is based on Gaussian process regression and is part of the family of kernel learning methods.[6,7] Such methods perform linear fits in a transformed data space: The nonlinearity of the function is now captured in a kernel function, also called a similarity or covariance function, which usually measures the similarity between two local atomic environments (although they can also be designed to capture long-range and global properties).

Formally, the potential energy suface is represented as a Gaussian process.[102,103] The covariance matrix of this process is formulated to use the information provided by quantum calculations, i.e. total energies and derivatives, in a natural way through linear operations on the kernel. This allows the Gaussian process to provide a smooth approximation of the potential energy surface, as sampled by the quantum data points, in the high-dimensional space of atomic or molecular environments using just a linear combination of kernels; for

example, the local energy of an atom i is given by:

$$\varepsilon_i = \sum_j \alpha_j k(\mathbf{d}_j, \mathbf{d}_i) \tag{3}$$

where the $\mathbf{d}$ are descriptors of local atomic environments, k designates the covariance or kernel function, and the weights α are determined by a regularized least-squares linear fit to the quantum mechanical training data (in this view, the predictions of Gaussian processes are the same as those given by kernel ridge regression (KRR) with a radial basis).[7] In GAP, the sum runs over a subset of *representative* configurations in the training set, allowing the fitting to scale *linearly* with the number of input data points.

The most successful kernel function for condensed-phase GAP has been the SOAP kernel,[81] which takes the similarity between local atomic environments. The environment of atom i is represented by a neighbour density $\rho_i(\mathbf{r})$, defined as a sum of Gaussians placed on each neighbouring atom, multiplied by a spherical cutoff function which smoothly takes the density to zero outside some cutoff radius. The kernel between two environments is defined as the integral over all possible mutual rotations of the square of the overlap between the two neighbour densities, thus making the kernel obey the same symmetries as the local energy: Invariance to translations (environments are atom-centred), permutations (from summing like atoms in the neighbour density), and rotations (from the rotational integration).

In practice, the integration over rotations can be done analytically by expanding each neighbour density in spherical harmonics and radial basis functions:

$$\rho_i(\mathbf{r}) = \sum_{nlm} c^{(i)}_{nlm} g_n(r) Y_{lm}(\hat{\mathbf{r}}),$$

computing the power spectrum elements

$$p^{(i)}_{nn'l} = \frac{1}{\sqrt{2l+1}} \sum_m c^{(i)}_{nlm} (c^{(i)}_{n'lm})^\dagger,$$

and summing in order to obtain the covariance function:

$$k_0(\rho_i, \rho_j) = \sum_{nn'l} p^{(i)}_{nn'l} p^{(j)}_{nn'l} \tag{4}$$

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.[7,81]

Note here that the local environment of atom i is represented by a set of numbers $\mathbf{p}^{(i)}$, which can be interpreted as a "descriptor" or even "feature vector" of the environment. Many other kernels are formulated in terms of other descriptors, such as the 6-D dimer kernel described in the supporting information.

The GAP models used in this study were all fit and evaluated using the libAtoms/QUIP package.[104] The GAP code can be downloaded at `http://www.libatoms.org/gap/gap_download.html`, with a precompiled version available through Docker at `https://hub.docker.com/r/libatomsquip/quip/`. The fitted potentials as well as all the training data are available from `http://dx.doi.org/10.17863/CAM.26364`.

5.2 MD simulations

The MD simulations were run using QUIP[104] and i-PI[105] via LAMMPS.[106,107] The former used the adaptive Langevin thermostat of Jones and Leimkuhler[108] and a Hoover-Langevin barostat[109] while the latter 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,[96] for the classical simulations and PIGLET[73,74] for the PIMD simulations. The initial configurations for all simulations were generated using Packmol.[110]

The traditional analytical potentials were run in LAMMPS[106] with a Langevin thermostat[111] and a Nosé-Hoover barostat[112–116] with the MTK correction.[117] 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.[118]

5.3 Dimer fits

The coupled-cluster CCSD(T) energies of the methane dimer were computed in a similar way as described in Gillan et. al.[8] (explained in more detail in the supporting information), up to the level of CCSD(T)-F12.[119–121] The energies were corrected for basis-set superposition error (BSSE) using the Boys-Bernardi counterpoise procedure.[122] Calculations were done using the MOLPRO suite of programs.[123–126] The Atomic Simulation Enviroment (ASE)[127] was used to generate and manipulate geometries. For the dimer error numbers used in Figure 2, energies (PBE and PBE0) were computed with Psi4[128] and the Hirshfeld partitioning[91] was done using HORTON.[129–132]

The geometries for the randomly chosen orientations were directly sampled from a liquid MD simulation (details in the supplementary information). Ten orientations were sampled and each used to produce a binding curve with regularly spaced dimer separations.

Finally, all the plots in this paper were made using Matplotlib;[133] the analysis was done within the Jupyter interactive computing enviroment with the IPython kernel,[134] and molecular views were with VMD.[135]

Acknowledgement

M.V. acknowledges Shell Global Solutions International B.V. for PhD studentship funding, as well as support from the EPSRC Centre for Doctoral Training in computational methods for materials science (CDT CMM, 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.

Supporting Information Available

Dimer fit details, parameters for the DFT and quantum chemistry calculations, GAP fitting command lines, and MD simulation trajectories and parameters.

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

References

(1) De, S.; Musil, F.; Ingram, T.; Baldauf, C.; Ceriotti, M. Mapping and classifying molecules from a high-throughput structural database. *J. Cheminf.* **2017**, *9*, 6.

(2) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modelling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, e1701816.

(3) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*, 241717.

(4) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(5) Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. Machine learning of molecular properties: Locality and active learning. *J. Chem. Phys.* **2018**, *148*, 241727.

(6) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(7) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051–1057.

(8) Gillan, M. J.; Alfé, D.; Bartók, A. P.; Csányi, G. First-principles energetics of water clusters and ice: A many-body analysis. *J. Chem. Phys.* **2013**, *139*, 244504.

(9) Atreya, S. K.; Adams, E. Y.; Niemann, H. B.; Demick-Montelara, J. E.; Owen, T. C.; Fulchignoni, M.; Ferri, F.; Wilson, E. H. Titan's methane cycle. *Planet. Space Sci.* **2006**, *54*, 1177–1187.

(10) Hayes, A. G.; Lorenz, R. D.; Lunine, J. I. A post-Cassini view of Titan's methane-based hydrologic cycle. *Nat. Geosci.* **2018**, *11*, 306–313.

(11) Benedetti, L. R.; Nguyen, J. H.; Caldwell, W. A.; Liu, H.; Kruger, M.; Jeanloz, R. Dissociation of CH4 at high pressures and temperatures: Diamond formation in giant planet interiors? *Science* **1999**, *286*, 100–102.

(12) Guzmán Marmolejo, A.; Segura, A. Methane in the Solar System. *Bol. Soc. Geol. Mex.* **2015**, *67*, 377–385.

(13) Pruteanu, C. G.; Ackland, G. J.; Poon, W. C. K.; Loveday, J. S. When immiscible becomes miscible—Methane in water at high pressures. *Sci. Adv.* **2017**, *3*, e1700240.

(14) Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L. Free Energies of Hydration and Pure Liquid Properties of Hydrocarbons from the OPLS All-Atom Model. *J. Phys. Chem.* **1994**, *98*, 13077–13082.

(15) Hayes, J. M.; Greer, J. C.; Morton-Blake, D. A. A force-field description of short-range repulsions for high density alkane molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25*, 1953–1966.

(16) Spanu, L.; Donadio, D.; Hohl, D.; Schwegler, E.; Galli, G. Stability of hydrocarbons at deep Earth pressures and temperatures. *Proc. Natl. Acad. Sci.* **2011**, *108*, 6843–6846.

(17) Payal, R. S.; Balasubramanian, S.; Rudra, I.; Tandon, K.; Mahlke, I.; Doyle, D.; Cracknell, R. Shear viscosity of linear alkanes through molecular simulations: quantitative tests for n -decane and n -hexadecane. *Mol. Simul.* **2012**, *38*, 1234–1241.

(18) Hansen, F.; Herwig, K.; Matthies, B.; Taub, H. Intramolecular and Lattice Melting in n-Alkane Monolayers: An Analog of Melting in Lipid Bilayers. *Phys Rev Lett* **1999**, *83*, 2362.

(19) Lennard-Jones, J. E. Cohesion. *Proc. Phys. Soc.* **1931**, *43*, 461–482.

(20) Buckingham, R. A.; Corner, J. Tables of Second Virial and Low-Pressure Joule-Thomson Coefficients for Intermolecular Potentials with Exponential Repulsion. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **1947**, *189*, 118–129.

(21) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **1986**, *7*, 230–252.

(22) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. Optimized intermolecular potential functions for liquid hydrocarbons. *J. Am. Chem. Soc.* **1984**, *106*, 6638–6646.

(23) Sun, H. COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase Applications—Overview with Details on Alkane and Benzene Compounds. *J. Phys. Chem. B* **1998**, *102*, 7338–7364.

(24) Stuart, S. J.; Tutein, A. B.; Harrison, J. A. A reactive potential for hydrocarbons with intermolecular interactions. *J. Chem. Phys.* **2000**, *112*, 6472.

(25) Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Clarendon Press: Oxford, 1989.

(26) Maple, J. R.; Hwang, M.-J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. Derivation of class II force fields. I. Methodology and quantum force

field for the alkyl functional group and alkane molecules. *J. Comput. Chem.* **1994**, *15*, 162–182.

(27) Allinger, N. L.; Chen, K.; Lii, J.-H. An improved force field (MM4) for saturated hydrocarbons. *J. Comput. Chem.* **1996**, *17*, 642–668.

(28) Allen, A. E. A.; Payne, M. C.; Cole, D. J. Harmonic Force Constants for Molecular Mechanics Force Fields via Hessian Matrix Projection. *J. Chem. Theory Comput.* **2018**, *14*, 274–281.

(29) Gay, D. H.; Dai, H.; Beck, D. R. Obtaining accurate pressure second virial coefficients for methane from an ab initio pair potential. *J. Chem. Phys.* **1991**, *95*, 9106–9114.

(30) Jalkanen, J.-P.; Pakkanen, T. A.; Yang, Y.; Rowley, R. L. Interaction energy surfaces of small hydrocarbon molecules. *J. Chem. Phys.* **2003**, *118*, 5474.

(31) Hellmann, R.; Bich, E.; Vogel, E. Ab initio intermolecular potential energy surface and second pressure virial coefficients of methane. *J. Chem. Phys.* **2008**, *128*, 214303.

(32) Li, A. H.-T.; Chao, S. D. A Refined Intermolecular Interaction Potential for Methane: Spectral Analysis and Molecular Dynamics Simulations. *J. Chinese Chem. Soc.* **2016**, *63*, 282–289.

(33) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(34) Chen, B.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 3. Explicit-Hydrogen Description of Normal Alkanes. *J. Phys. Chem. B* **1999**, *103*, 5370–5379.

(35) Allen, W.; Rowley, R. L. Predicting the viscosity of alkanes using nonequilibrium molecular dynamics: Evaluation of intermolecular potential models. *J. Chem. Phys.* **1997**, *106*, 10273.

(36) Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. *J. Phys. Chem. B* **1998**, *102*, 2569–2577.

(37) Van Vleet, M. J.; Misquitta, A. J.; Stone, A. J.; Schmidt, J. R. Beyond Born–Mayer: Improved Models for Short-Range Repulsion in ab Initio Force Fields. *J. Chem. Theory Comput.* **2016**, *12*, 3851–3870.

(38) Van Vleet, M. J.; Misquitta, A. J.; Schmidt, J. R. New Angles on Standard Force Fields: Toward a General Approach for Treating Atomic-Level Anisotropy. *J. Chem. Theory Comput.* **2018**, *14*, 739–758.

(39) Chao, S.-W.; Li, A. H.-T.; Chao, S. D. Molecular dynamics simulations of fluid methane properties using ab initio intermolecular interaction potentials. *J. Comput. Chem.* **2009**, *30*, 1839–1849.

(40) Martin, R. M. *Electronic structure: Basic theory and practical methods*, 1st ed.; Cambridge University Press: Cambridge, UK, 2008.

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

(42) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

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

(44) Becke, A. D.; Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction: high-order dispersion coefficients. *J. Chem. Phys.* **2006**, *124*, 14104.

(45) Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from

Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, 073005.

(46) Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.

(47) Blood-Forsythe, M. A.; Markovich, T.; DiStasio, R. A.; Car, R.; Aspuru-Guzik, A. Analytical nuclear gradients for the range-separated many-body dispersion model of noncovalent interactions. *Chem. Sci.* **2016**, *7*, 1712–1728.

(48) Shtukenberg, A. G.; Zhu, Q.; Carter, D. J.; Vogt, L.; Hoja, J.; Schneider, E.; Song, H.; Pokroy, B.; Polishchuk, I.; Tkatchenko, A.; Oganov, A. R.; Rohl, A. L.; Tuckerman, M. E.; Kahr, B. Powder diffraction and crystal structure prediction identify four new coumarin polymorphs. *Chem. Sci.* **2017**, *8*, 4926–4940.

(49) Morawietz, T.; Singraber, A.; Dellago, C.; Behler, J. How van der Waals interactions determine the unique properties of water. *Proc. Natl. Acad. Sci.* **2016**, *113*, 8368–8373.

(50) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(51) Rupp, M.; von Lilienfeld, O. A.; Burke, K. Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. *J. Chem. Phys.* **2018**, *148*, 241401.

(52) Szlachta, W. J.; Bartók, A. P.; Csányi, G. Accuracy and transferability of Gaussian approximation potential models for tungsten. *Phys. Rev. B* **2014**, *90*, 104108.

(53) Dragoni, D.; Daff, T. D.; Csányi, G.; Marzari, N. Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron. *Phys. Rev. Mater.* **2018**, *2*, 13808.

(54) Deringer, V. L.; Pickard, C. J.; Csányi, G. Data-Driven Learning of Total and Local Energies in Elemental Boron. *Phys. Rev. Lett.* **2018**, *120*, 156001.

(55) Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. 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.* **2013**, *88*, 054104.

(56) Deringer, V. L.; Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **2017**, *95*, 094203.

(57) Deringer, V. L.; Bernstein, N.; Bartók, A. P.; Cliffe, M. J.; Kerber, R. N.; Marbella, L. E.; Grey, C. P.; Elliott, S. R.; Csányi, G. Realistic Atomistic Structure of Amorphous Silicon from Machine-Learning-Driven Molecular Dynamics. *J. Phys. Chem. Lett.* **2018**, *9*, 2879–2885.

(58) Caro, M. A.; Deringer, V. L.; Koskinen, J.; Laurila, T.; Csányi, G. Growth Mechanism and Origin of High s p 3 Content in Tetrahedral Amorphous Carbon. *Phys. Rev. Lett.* **2018**, *120*, 166101.

(59) Bereau, T.; Andrienko, D.; Von Lilienfeld, O. A. Transferable Atomic Multipole Machine Learning Models for Small Organic Molecules. *J. Chem. Theory Comput.* **2015**, *11*, 3225–3233.

(60) Shapeev, A. Accurate representation of formation energies of crystalline alloys with many components. *Comput. Mater. Sci.* **2017**, *139*, 26–30.

(61) Mocanu, F. C.; Konstantinou, K.; Lee, T. H.; Bernstein, N.; Deringer, V. L.; Csányi, G.; Elliott, S. R. Modeling the Phase-Change Memory Material, Ge2Sb2Te5, with a Machine-Learned Interatomic Potential. *J. Phys. Chem. B* **2018**, *122*, 8998–9006.

(62) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2015**, *18*, 13754–13769.

(63) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Anatole von Lilienfeld, O. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, 095003.

(64) Willatt, M. J.; Musil, F.; Ceriotti, M. Feature Optimization for Atomistic Machine Learning Yields a Data-Driven Construction of the Periodic Table of the Elements. *Phys. Chem. Chem. Phys.* **2018**, *20*, 29661–29668.

(65) Markland, T. E.; Ceriotti, M. Nuclear quantum effects enter the mainstream. *Nat. Rev. Chem.* **2018**, *2*, 0109.

(66) Kapil, V.; Behler, J.; Ceriotti, M. High order path integrals made easy. *J. Chem. Phys.* **2016**, *145*, 234103.

(67) Balog, E.; Hughes, A. L.; Martyna, G. J. Constant pressure path integral molecular dynamics studies of quantum effects in the liquid state properties of n-alkanes. *J. Chem. Phys.* **2000**, *112*, 870–880.

(68) Pereyaslavets, L.; Kurnikov, I.; Kamath, G.; Butin, O.; Illarionov, A.; Leontyev, I.; Olevanov, M.; Levitt, M.; Kornberg, R. D.; Fain, B. On the importance of accounting for nuclear quantum effects in ab initio calibrated force fields in biological simulations. *Proc. Natl. Acad. Sci.* **2018**, *115*, 8878–8882.

(69) Chandler, D.; Wolynes, P. G. Exploiting the isomorphism between quantum theory and classical statistical mechanics of polyatomic fluids. *J. Chem. Phys.* **1981**, *74*, 4078–4095.

(70) Habershon, S.; Manolopoulos, D. E.; Markland, T. E.; Miller, T. F. Ring-Polymer Molecular Dynamics: Quantum Effects in Chemical Dynamics from Classical Trajectories in an Extended Phase Space. *Annu. Rev. Phys. Chem.* **2013**, *64*, 387–413.

(71) Biermann, S.; Hohl, D.; Marx, D. Quantum effects in solid hydrogen at ultra-high pressure. *Solid State Commun.* **1998**, *108*, 337–341.

(72) Ceriotti, M.; Parrinello, M.; Markland, T. E.; Manolopoulos, D. E. Efficient stochastic thermostatting of path integral molecular dynamics. *J. Chem. Phys.* **2010**, *133*, 124104.

(73) Ceriotti, M.; Bussi, G.; Parrinello, M. Nuclear Quantum Effects in Solids Using a Colored-Noise Thermostat. *Phys. Rev. Lett.* **2009**, *103*, 30603.

(74) Ceriotti, M.; Manolopoulos, D. E. Efficient First-Principles Calculation of the Quantum Kinetic Energy and Momentum Distribution of Nuclei. *Phys. Rev. Lett.* **2012**, *109*, 100604.

(75) Markland, T. E.; Manolopoulos, D. E. A refined ring polymer contraction scheme for systems with electrostatic interactions. *Chem. Phys. Lett.* **2008**, *464*, 256–261.

(76) Podeszwa, R.; Bukowski, R.; Szalewicz, K. Potential energy surface for the benzene dimer and perturbational analysis of π-π interactions. *J. Phys. Chem. A* **2006**, *110*, 10345–10354.

(77) Stone, A. J. *The theory of intermolecular forces*, 2nd ed.; Oxford University Press: Oxford, 2013.

(78) Vandenbrande, S.; Waroquier, M.; Speybroeck, V. V.; Verstraelen, T. The Monomer Electron Density Force Field (MEDFF): A Physically Inspired Model for Noncovalent Interactions. *J. Chem. Theory Comput.* **2017**, *13*, 161–179, PMID: 27935712.

(79) Cole, D. J.; Vilseck, J. Z.; Tirado-Rives, J.; Payne, M. C.; Jorgensen, W. L. Biomolecular Force Field Parameterization via Atoms-in-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.* **2016**, *12*, 2312–2323.

(80) Bereau, T.; DiStasio, R. A.; Tkatchenko, A.; von Lilienfeld, O. A. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.* **2018**, *148*, 241706.

(81) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(82) Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, B864–B871.

(83) Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(84) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

(85) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(86) Clark, S. J.; Segall, M. D.; Pickard, C. J.; Hasnip, P. J.; Probert, M. J.; Refson, K.; Payne, M. C. First principles methods using CASTEP. *Z. Krist.* **2005**, *220*, 567–570.

(87) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(88) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(89) Each node contains two Intel E5-2697 v2 (Ivy Bridge) 2.7 GHz 12-core processors; see `http://archer.ac.uk/about-archer/hardware/` for details.

(90) The fitting machine has two 2.6 GHz 8-core Intel Xeon E5-2670 processors and 240 GiB of RAM.

(91) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theor. Chim. Acta* **1977**, *44*, 129–138.

(92) Chu, X.; Dalgarno, A. Linear response time-dependent density functional theory for van der Waals coefficients. *J. Chem. Phys.* **2004**, *121*, 4083–4088.

(93) Ambrosetti, A.; Reilly, A. M.; DiStasio, R. A.; Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **2014**, *140*, 18A508.

(94) Goodwin, R. D.; Prydz, R. Densities of compressed liquid methane, and the equation of state. *J. Res. Natl. Bur. Stand., Sect. A* **1972**, *76A*, 81–101.

(95) Truncated to three significant digits; see reference for full precision.

(96) Ceriotti, M.; Bussi, G.; Parrinello, M. Colored-noise thermostats à la Carte. *J. Chem. Theory Comput.* **2010**, *6*, 1170–1180.

(97) Russell, A. J.; Spackman, M. A. Vibrational averaging of electrical properties. *Mol. Phys.* **1995**, *84*, 1239–1255.

(98) Bishop, D. M.; Gu, F. L.; Cybulski, S. M. Static and dynamic polarizabilities and first hyperpolarizabilities for CH4, CF4, and CCl4. *J. Chem. Phys.* **1998**, *109*, 8407–8415.

(99) Teja, A. S.; Lee, R. J.; Rosenthal, D.; Anselme, M. Correlation of the critical properties of alkanes and alkanols. *Fluid Phase Equilib.* **1990**, *56*, 153–169.

(100) Artrith, N.; Morawietz, T.; Behler, J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B* **2011**, *83*, 153101.

(101) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; Distasio, R. A.; Ceriotti, M. Accurate molecular polarizabilities with coupled-cluster theory and machine learning. *arXiv:1809.05337* **2018**, `https://arxiv.org/abs/1809.05337` (accessed 9 Oct. 2018).

(102) MacKay, D. J. C. *Information theory, inference, and learning algorithms*; Cambridge University Press: Cambridge, 2003.

(103) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, 2006.

(104) Bartók-Pártay, A.; Bartók-Pártay, L.; Bianchini, F.; Butenuth, A.; Caccin, M.; Cereda, S.; Csányi, G.; Comisso, A.; Daff, T.; John, S.; Gattinoni, C.; Moras, G.; Kermode, J.; Mones, L.; Nichol, A.; Packwood, D.; Pastewka, L.; Peralta, G.; Solt, I.; Strickson, O.; Szlachta, W.; Varnai, C.; Veit, M.; Winfield, S. libAtoms+QUIP. 2018; `http://libatoms.org`.

(105) Ceriotti, M.; More, J.; Manolopoulos, D. E. i-PI: A Python interface for ab initio path integral molecular dynamics simulations. *Comput. Phys. Commun.* **2014**, *185*, 1019–1026.

(106) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

(107) using LAMMPS stable release from 11 Aug 2017.

(108) Jones, A.; Leimkuhler, B. Adaptive stochastic methods for sampling driven molecular systems. *J. Chem. Phys.* **2011**, *135*, 84125.

(109) Quigley, D.; Probert, M. I. J. Langevin dynamics in constant pressure extended systems. *J. Chem. Phys.* **2004**, *120*, 11432–11441.

(110) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **2009**, *30*, 2157–2164.

(111) Brünger, A.; Brooks, C. L.; Karplus, M. Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chem. Phys. Lett.* **1984**, *105*, 495–500.

(112) Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511–519.

(113) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(114) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(115) Shinoda, W.; Shiga, M.; Mikami, M. Rapid estimation of elastic constants by molecular dynamics simulation under constant stress. *Phys. Rev. B* **2004**, *69*, 134103.

(116) Tuckerman, M. E.; Alejandre, J.; López-Rendón, R.; Jochim, A. L.; Martyna, G. J. A Liouville-operator derived measure-preserving integrator for molecular dynamics simulations in the isothermal–isobaric ensemble. *J. Phys. A. Math. Gen.* **2006**, *39*, 5629–5651.

(117) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* **1994**, *101*, 4177.

(118) Hockney, R. W.; Eastwood, J. W. *Computer simulation using particles*; Hilger: Bristol, 1988.

(119) Adler, T. B.; Knizia, G.; Werner, H.-J. A simple and efficient CCSD(T)-F12 approximation. *J. Chem. Phys.* **2007**, *127*, 221106.

(120) Knizia, G.; Adler, T. B.; Werner, H.-J. Simplified CCSD(T)-F12 methods: Theory and benchmarks. *J. Chem. Phys.* **2009**, *130*, 054104.

(121) Kong, L.; Bischoff, F. A.; Valeev, E. F. Explicitly correlated R12/F12 methods for electronic structure. *Chem. Rev.* **2012**, *112*, 75–107.

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

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

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

(125) Schütz, M.; Lindh, R.; Werner, H.-J. Integral-direct electron correlation methods. *Mol. Phys.* **1999**, *96*, 719–733.

(126) Lindh, R. The reduced multiplication scheme of the Rys-Gauss quadrature for 1st order integral derivatives. *Theor. Chim. Acta* **1993**, *85*, 423–440.

(127) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.;

Jennings, P. C.; Bjerre Jensen, P.; Kermode, J.; Kitchin, J. R.; Leonhard Kolsbjerg, E.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Bergmann Maronsson, J.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiøtz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W. The atomic simulation environment - A Python library for working with atoms. *J. Phys. Condens. Matter* **2017**, *29*, 273002.

(128) Parrish, R. M.; Burns, L. A.; Smith, D. G. A.; Simmonett, A. C.; DePrince, A. E.; Hohenstein, E. G.; Bozkaya, U.; Sokolov, A. Y.; Di Remigio, R.; Richard, R. M.; Gonthier, J. F.; James, A. M.; McAlexander, H. R.; Kumar, A.; Saitow, M.; Wang, X.; Pritchard, B. P.; Verma, P.; Schaefer, H. F.; Patkowski, K.; King, R. A.; Valeev, E. F.; Evangelista, F. A.; Turney, J. M.; Crawford, T. D.; Sherrill, C. D. Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.* **2017**, *13*, 3185–3197.

(129) Verstraelen, T.; Tecmer, P.; Heidar-Zadeh, F.; González-Espinoza, C. E.; Chan, M.; Kim, T. D.; Boguslawski, K.; Fias, S.; Vandenbrande, S.; Berrocal, D.; Ayers, P. W. HORTON. 2017; `http://theochem.github.com/horton/`.

(130) Becke, A. D. A multicenter numerical integration scheme for polyatomic molecules. *J. Chem. Phys.* **1988**, *88*, 2547–2553.

(131) Becke, A. D.; Dickson, R. M. Numerical solution of Poisson's equation in polyatomic molecules. *J. Chem. Phys.* **1988**, *89*, 2993–2997.

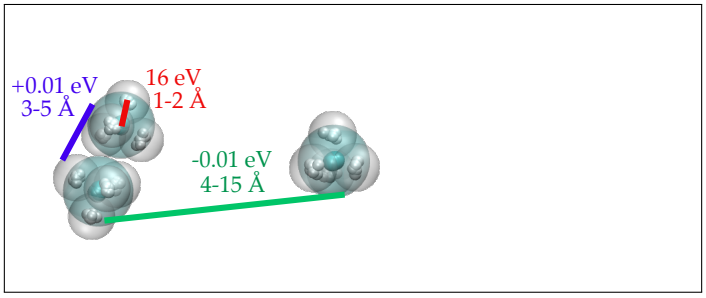
(132) Lebedev, V.; Laikov, D. A quadrature formula for the sphere of the 131st algebraic order of accuracy. *Dokl. Math.* **1999**, *59*, 477–481.

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

(134) Pérez, F.; Granger, B. E. IPython: a System for Interactive Scientific Computing. *Comput. Sci. Eng.* **2007**, *9*, 21–29.

(135) Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

Graphical TOC Entry



Separation of interactions in condensed-phase methane: Covalent, repulsion, and dispersion.