Imperial College of Science, Technology and Medicine

Master's Thesis

Low-Dimensional Machine Learning Potentials for Molecular Systems

Author: Sahil Shah Year 4 CID: 01196082

MSci Chemistry with Research Abroad

Supervisors: Prof. Michele Ceriotti, Dr. Max Veit (EPFL)

Prof. Sophia Yaliraki (Imperial College London)

Word Count: 11872 (excluding Methodology)

Research completed abroad with Computational Science and Modelling Group (COSMO) at École Polytechnique Fédérale de Lausanne (EPFL)

May 2020

Contents

1 Abstract

Interatomic pair potentials are models which describe the potential energy between two atoms, being good approximations for the potential energy surface of the system through neglecting higher body-order corrections. They have been predicted traditionally using many different models, for example, the well-known Lennard-Jones and Morse potentials. More recently, Gaussian Approximation Potentials have utilised kernel methods in the increasingly popular area of machine learning to model pair potentials. Using the code developed in *librascal,* pair distances and Gaussian kernels were computed, while the distances were able to be scaled by a predefined power. Using *librascal*, a Gaussian Approximation Potential model for methane dimers was created to generate a six-dimensional potential energy surface for regular unscaled distances. Given strong evidence in research, it was thought that scaling the distances to the powers of -6 and -12 in a Lennard-Jones fashion would lead to more accurate potentials. The 6th- and 12th-power models were combined to give a 12-6 model. Hyperparameter optimisations were carried out to minimise the root mean square error (RMSE) of the fitted total energy functions, through six-fold cross-validation. The interatomic C-C, C-H and H-H pair potentials were computed for all the models and compared. The 12-6 model produced the lowest training error and smooth pair potentials with tail forms which improved upon regular unscaled distances model, but with a degree of overfitting. The 12-6 model was compared to a Gaussian Approximation Potential produced using *QUIP* software with a shorter cut-off radius, and another classical potential model. It was found, however, that the 6th-power model generated the most accurate C-C, C-H and H-H pair potentials for a methane dimer of all the potentials made using *librascal*, through producing the lowest cross-validation error and having the most physical tail forms. This is a promising and inexpensive strategy for modelling long-range interactions in systems dominated either by dispersion or other interactions by changing the scaling power.

2 Introduction

2.1 Interatomic Potentials

Interatomic potentials describe the potential energy interaction between two or more atoms. They can be used to help model and predict properties of systems such as cohesion, thermal expansion and elastic and plastic behaviour of materials[1,2,3]. The accuracy of properties of a system obtained from molecular simulations is limited by the accuracy of the potential used to model the system.

Interatomic potentials in general, come under two main classes. Pair potentials are the focus of this project, of which a subset is repulsive potentials, such as screened Coulomb potentials[4]. The other class is many-body potentials, such as the Stillinger-Weber potential[5]. The sum of pair potentials in a system can be representative of the total energy of the system, via the pairwise additivity approximation, neglecting higher-body-order terms. Pair potentials, in particular the Lennard-Jones potential, were first developed to model the Equation of State of noble gases and are models which describe the potential energy interaction between two atoms[6,7,8]. The Lennard-Jones potential was also used from the 1960s as the long-range component of interatomic potentials for biomolecular simulation[9].

Pair potential functions can be modelled in various different ways. In addition to the Lennard-Jones potential, another well-known potential is the Morse potential. The typical Lennard-Jones potential is given by the form displayed in Equation 1[10] and takes a $1/r^{12}$ relationship for the repulsive force and a $1/r^{6}$ relationship for the attractive force. The typical Morse potential is given by the form displayed in Equation 2[11] and shows a different relationship through the use of the exponential function.

$$U_{LJ}(r) = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) \tag{1}$$

$$U_M(r) = D\left(1 - e^{-\alpha(r-R)}\right)^2 \tag{2}$$

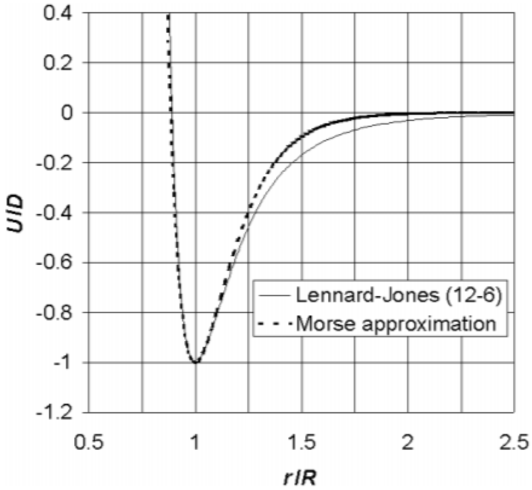


Figure 1 – Lennard-Jones and Morse Potentials Comparison[11]

For both near equilibrium states and long-range, both the Lennard-Jones and Morse potentials closely resemble each other. However, the Morse potential underestimates the repulsive energy insignificantly and overestimates the attractive energy, in comparison to the Lennard-Jones potential[11]. Overall, it can be concluded that the Morse potential is similar to the Lennard-Jones potential but with the main difference that it decays exponentially, while the Lennard-Jones potential has a 6th-power decay[12].

Models such as the Lennard-Jones and Morse potentials have been implemented in forcefields such as the COMPASS forcefield in a molecular dynamics (MD) simulation. This would provide a representation of the potential energy surface (PES) of the system. For example, an analytical interatomic potential was used to model non-equilibrium processes in a W-C-H system. In this method, Morse-like potentials were fitted to the different bond types in different unit cell structures[13]. Various parameters in the Morse potential were optimised to give the best fit.

Pair potentials for C-C, C-H and H-H pair types have been determined from *ab initio* calculations for use in MD simulations, through fitting and modelling using various Lennard-Jones and Morse potentials[14].

The development of interatomic potentials has progressed from classical models such as the Lennard-Jones and Morse potentials, to accounting for the quantum nature of a potential using Density-Functional Theory. Most recently, machine learning potentials have been developed using various machine learning techniques, such as Neural Networks[15], Gaussian Process Regression, which will be utilised in this project, and Linear Regression[16].

2.2 Body-Order Expansion of the Total Energy

The body-order expansion of a system is a way of expressing the total energy of a system, through a sum of potentials between an increasing number of bodies. These bodies can be either atoms or molecules, but in this case, bodies refer to atoms unless stated otherwise. It is extremely useful, as an approximation for the total energy can be obtained through summing over potentials, where lower-order terms are much simpler to model. It can be written as a sum over zero-, one-, two-, three-body terms and so forth. The energies of each of these terms and corrections are computed to calculate the total energy of a system of atoms, where the bodies represent atoms, as shown in Equation 3[17].

$$E = V_0 + \sum_i V^{(1)}(r_i) + \frac{1}{2}\sum_{ij} V^{(2)}(r_i, r_j) + \frac{1}{3!}\sum_{ijk} V^{(3)}(r_i, r_j, r_k) + \frac{1}{4!}\sum_{ijkl} V^{(4)}(r_i, r_j, r_k, r_l) + \ldots \quad (3)$$

From Equation 3, it can be seen that the total energy can be written as a sum of all body-order terms, where each potential is a function of the position(s), r_i, of the atom(s) included in the term. Each higher body-order term is smaller than the preceding lower body-order term. This implies that high body-order terms can be neglected, but some, for example three-body terms, are still significant.

The V_0 term is a constant that can be initialised to zero and the $V^{(1)}$ terms are the atomic potentials. In machine learning models, they are also set to the atomic potentials but have greater freedom to minimise the variance of the remaining energy to be fit, after pair potentials have been removed[18]. The $V^{(2)}$ terms are the pair potentials between atoms and $V^{(3)}$ terms are three-body order corrections.

For example, a system of three atoms, *A*, *B* and *C*, would have a total energy of: $E = V_A + V_B + V_C + V_{AB} + V_{BC} + V_{AC} + V_{ABC}$, where each of the two-body terms is the corresponding pair potential in the absence of the third atom. However, it is known that the three-body correction has a contribution to the overall energy, as it is the energy that is not accounted for when summing over all one- and two-body potentials[19].

The atomic pairwise additivity approximation for the total energy of a system can be written as a sum over all the neighbouring atoms in the system, for each atom in the system, calculating all the energies from their pair potentials and neglecting higher body-order terms[20]. Therefore, pair potentials indicate how the total energy of a system may be decomposed into pairwise contributions.

However, the limitations of this approximation appear when higher-body-order terms are required. For example, if a water molecule's energy is modelled only by two O-H pair potentials, it will be unable to distinguish between the bent and linear conformations and unable to account for hydrogen bonding. Similarly, if a propane molecule's total energy were to be modelled only by summing over all C-C pair potentials, it would be unable to distinguish between bonded and non-bonded C-C pair potentials. Therefore, in both these examples higher-body-order corrections are required to improve the accuracy of the model. Despite this, the pairwise additivity approximation provides pair potentials which are much less expensive than many-body potentials to compute, due to the lower one-dimensional scaling, and is a good approximation for systems where higher-body-order corrections have a low contribution to the total energy.

2.3 Machine Learning Potentials

In terms of selecting a machine learning model for an energy fitting function, Neural Networks, Linear Regression and Gaussian Process Regression have all been used as previously mentioned. There has been work undertaken, using atom-centred symmetry functions to construct neural network potentials, representing high-dimensional *ab initio* potential energy surfaces, since they can provide energies and forces much faster than electronic structure calculations[15]. Both Morse and Lennard-Jones potentials have been used to inspire linear least-squares polynomial fits to *ab initio* data, through computing linear coefficients to place weights on certain orders of polynomial. This method was used to compute potential energy surfaces through using permutationally invariant polynomials (PIPs)[21].

A machine learning potential is trained using total energies, forces and sometimes stresses from quantum calculations. Its accuracy can be comparable to that of a quantum calculation, in contrast with general-purpose analytical potentials. In general, they have greater accuracy while being general-purpose, but are less able to extrapolate. For example, Gaussian Approximation Potentials are more advantageous than other models such as the Lennard-Jones and Morse potentials due to the higher degree of flexibility offered by a larger number of parameters used in the fit.

2.3.1 Gaussian Process Regression

Any reasonably well-behaved function can be written as a linear combination of other functions multiplied by weights. In this case, potential energy surfaces are "well-behaved" under physically reasonable conditions, i.e. the atoms do not overlap or there are no infinite potentials. This is well known in Fourier Analysis, where a function can be approximated through being written as a sum of sine and cosine functions. In this case, the pair potential energy function is given as the sum over a set of basis functions, multiplied by the weights and is valid for linear and kernel methods only. This can be seen in Equation 4.

$$\varepsilon_i = \varepsilon(\boldsymbol{d}_i, \boldsymbol{w}) = \sum_h w_h \varphi_h(\boldsymbol{d}_i) \qquad (4)$$

Each basis function of the pair potential is a function of a descriptor, d_i, also known as a feature, which is a transformation of the coordinates between two or more bodies. In this case, the transformation corresponds to the distance between the bodies. This descriptor was chosen due to the evaluation cost of the target property, which was in this case the energy and the extent of the exploration space[22]. This set of descriptors was chosen to satisfy translational, rotational and other invariances of all atoms in the dataset.

The original Least Squares model takes the basis functions of the descriptors of distances between two bodies or atoms and map them to the energies, through multiplying by weights which would be optimised through minimising an L_2-loss function. However, with increasing complexity, the model tends to overfit, with a large variance.

Ridge Regression is similar to original Least Squares, but in this case, the loss function includes a regularisation term to prevent overfitting for high model complexities through Tikhonov regularisation. The regularisation term in Ridge Regression is the L_2-norm of the weights vector. Another possible regularisation function is the L_1-norm of the weights vector. However, Tikhonov regularisation is preferred as the L_2-norm is differentiable and so learning problems can be solved through gradient descent[23]. Both regression models, Least Squares and Ridge Regression, require the ability to compute the basis function itself.

Kernel Regression is preferred over the original Least Squares and Ridge Regression in this case due to the fact that kernels can capture the non-linear behaviour within a linear fit instead of finding a complete basis set to represent the energy. Another advantage of using kernels is that the original basis functions do not need to be known. The kernel, in this case, provides a similarity measure between two distances, being equivalent to the covariance between them.

When the probability distribution of the weights is chosen to be Gaussian with zero mean ($P(\boldsymbol{w}) = \mathrm{Normal}(\mathbf{w};\ 0, \sigma_w \boldsymbol{I})$), the covariance is shown in Equation 5[24].

$$\langle \varepsilon_i \varepsilon_j \rangle \ = \ {\sigma_w}^2 \sum_h \varphi_h(\boldsymbol{d}_i)\varphi_h(\boldsymbol{d}_j) \tag{5}$$

The product of the two basis functions in Equation 4 is equivalent to the kernel or covariance function. All kernel functions must be symmetric and positive semidefinite. A kernel is known to be symmetric if $\boldsymbol{k}(x, x') = \boldsymbol{k}(x', x)$ and is known to be positive semidefinite if $\boldsymbol{v}^T \boldsymbol{K} \boldsymbol{v} \geq 0$ for all vectors $\boldsymbol{v} \in \mathbb{R}^n$. There are many different forms of kernel function such as linear and polynomial kernel functions[25]. However, in this case, a Gaussian kernel or radial basis function (RBF) is chosen[26]. The explicit form of the kernel is shown in Equation 6[24,27] and by the squared exponential form.

$$C\left(\boldsymbol{d}_i, \boldsymbol{d}_j\right) = \sum_h \varphi_h(\boldsymbol{d}_i)\varphi_h(\boldsymbol{d}_j) = exp\left(-\sum_{i,j}\left[\left(\boldsymbol{d}_i\ -\ \boldsymbol{d}_j\right)^2/2\theta^2\right]\right) \tag{6}$$

The probability of finding a predicted value of a new energy, given previous observations, is Gaussian and is shown in Equation 7[24]. The mean of this distribution of predictions is given by Equation 8[24].

$$P(y|\boldsymbol{t}) = \frac{P(\boldsymbol{t}, y)}{P(\boldsymbol{t})} \tag{7}$$

$$\bar{y} = \boldsymbol{k}^T \boldsymbol{C}^{-1} \boldsymbol{t} \tag{8}$$

In Equation 7, ***t*** is the set of previously observed energies, i.e. the training set of energies, and *y* is the predicted energy. Equation 7 shows that the probability of predicting an energy, given a set of previously observed energies, can be calculated. In Equation 8, ***k*** is the covariance vector of function values and it shows that the prediction can be made based solely on the kernel function and previous observations, without requiring the explicit form of the basis function.

In Kernel Ridge Regression, the unknown function for the set of energies in this case, is expanded as a linear combination of radial basis functions, highlighting the relation with Gaussian processes, as shown in Equation 9[24]. The total pair potential energy is the sum over all these functions, shown in Equation 10.

$$f(\boldsymbol{d}) = \sum_{\boldsymbol{i}} \alpha_i \boldsymbol{C}(\boldsymbol{d}, \boldsymbol{d}_{\boldsymbol{i}}) \tag{9}$$

$$V = \sum_i \varphi_i(r_{ij}) \tag{10}$$

The weights, α, are optimised through minimising the regularised cost function, L, in Equation 11[28].

$$\boldsymbol{L} = \sum_i(\boldsymbol{t_i} - \boldsymbol{f}(\boldsymbol{d_i}))^2 + \sigma^2\|\alpha\|^2 \tag{11}$$

The second term is also known as the regularisation term and prevents overfitting with increasing model complexity. The norm is defined in Equation 12.

$$\|\alpha\|^2 = \alpha^T \boldsymbol{C}\alpha \tag{12}$$

The predictions made through this kernel regression are equivalent to those made through Gaussian processes, where this is a form of Gaussian Process Regression. The kernel defines the basis functions and the norm of the weights.

2.3.2 Hyperparameter Optimisation

The regularisation hyperparameter or regulariser, σ, can be selected through many different methods. One of which is called hyperparameter optimisation and is carried out to see which value minimises the cross-validation error, shown in Figure 2.

Optimisation can be done in one of two ways: Grid Search and Random Search. The Grid Search method selects a wide grid of values for the hyperparameters and computes the minimisation of the cost function for these hyperparameters. It can be effective for a rough estimate for the optimal hyperparameters but is limited by the fact that the true minimum could lie between two points on the grid. From Gaussian process analysis, it has been found that different hyperparameters are important for different datasets[29].

The Random Search method is different in that a discrete set of hyperparameters are not explored, but a continuous statistical distribution instead. A sampling distribution is initialised for the hyperparameter, therefore equal importance is not placed on each hyperparameter. An emphasis is placed on searching the hyperparameter space which is most likely to optimise the model score[30].

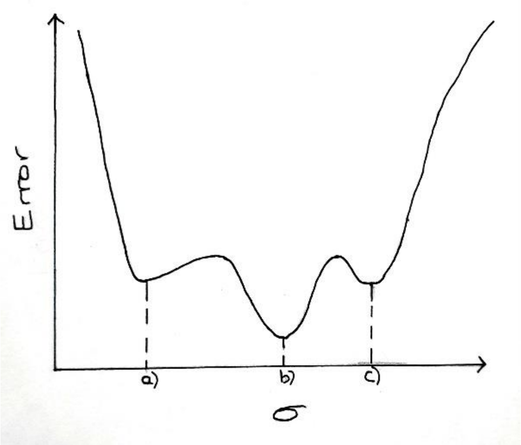


Figure 2 – Selection of Hyperparameter through Hyperparameter Optimisation

In the example of grid search cross-validation in Figure 2, it can be seen that regulariser values at a) and c) find local minima in the cross-validation error. However, the regulariser value at b) finds corresponds to a global minimum for the cross-validation error.

A large limitation of using any hyperparameter optimisation is overfitting, where the obtained hyperparameter will create a model which fits to the noise in the training data, which is not transferable to the test data[31]. Therefore, it is necessary to use k-fold cross-validation to help select a hyperparameter which avoids overfitting. In this process, the dataset is divided into *k* subsamples. From the *k* subsamples, one becomes the validation set for testing the data and the remaining *k-1* subsamples become the training set. The cross-validation process is then repeated *k* times with the average prediction error calculated and compared with other hyperparameters. The advantage of using cross-validation is that all members of the dataset are used for both training and validation[32].

2.3.3 Gaussian Approximation Potentials

In order to find the predicted total energies of the system, Equation 13[33] is used. The * indicates a new data point, which could be associated with either the train or test datasets. If the point is from the training dataset, *k* would be a train-train kernel. If the point is from the test dataset, *k* would be a test-train kernel. The training RMSE can be calculated at this point as well.

$$\varepsilon^* = \boldsymbol{\alpha}.k(\boldsymbol{d}, d^*) \tag{13}$$

Then to calculate the pair potentials, test datasets of isolated pairs of atoms must be used to compute the test-train kernels. Again, using Equation 13, the dot product of the weights, used to train the total energies, with the test-train kernel, is computed to give the pair potential. An example is shown in Figure 3. This is known as a Gaussian Approximation Potential (GAP).

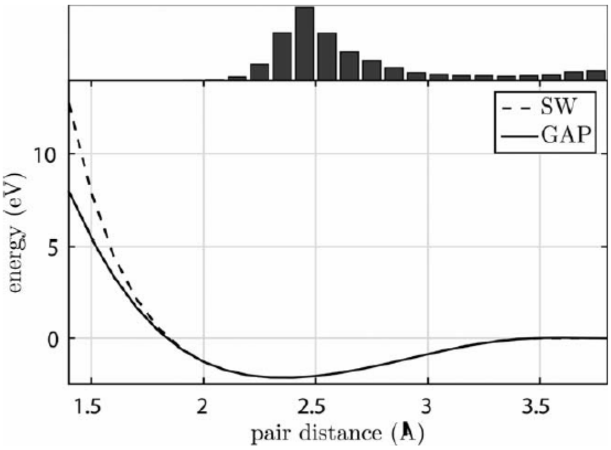


Figure 3 - GAP used to reproduce the Stillinger-Weber (SW) Potential for Silicon[24]

In this example, a two-body pair potential for silicon in silicon clusters, Si_x, was created using Gaussian Approximation Potentials (GAP) to attempt to reproduce the well-known Stillinger-Weber (SW) potential. It can be seen that the GAP provides a good fit for the SW potential, where there are distances within the dataset, as shown in the histogram above. A very strong fit is observed where there are many data points and a weak fit is observed where there are no data points. Therefore, the selected training set must contain the most distances for where the fit is desired to be most accurate.

Many studies have been carried out creating GAPs for different structures. For example, a GAP has been used to study the thermal conductivity of silicene, again using kernel regression to produce a more accurate potential than the SW potential used in the DFT calculation[34]. Transferable atomic electrostatic multipoles have also been modelled for small organic molecules again using kernel ridge regression and GAP[35]. In another study, a GAP was developed for atomistic simulations of liquid and amorphous carbon[36]. GAP models have also been used to represent the one- and two-body errors of DFT approximations of water systems, to give a better description of the energetics of small water clusters and improvements in the structure and dynamics of liquid water[18].

More specifically and relevantly, GAP has been used to study pair potentials in alkanes and their contributions to the total energy using the *Quantum Mechanics and Interatomic Potentials* (*QUIP*) software. It was found that many-body effects both within and beyond the dimer are essential in the short-range to obtain a description of the bulk density. It was found that GAPs fitted and run by *QUIP* are very computationally expensive by the standards of interatomic potentials[37].

The majority of the computational cost associated with *QUIP* comes from computing representations and kernels[38]. Therefore, through the development of *librascal*, it is hoped that the representations and kernels can be computed more efficiently, through the utilisation of parallelisation.

2.4 Long-Range Contributions

The total energy of the system is given by the sum over all local energy functionals, within a cut-off radius, r_{cut} and long-range contributions[19]. This can be seen in Equation 14[39].

$$E = \sum_{ij} \varepsilon_{ij} + \textit{long-range contributions} \tag{14}$$

In this case, a sharp cut-off with a reasonably large radius will be used to compute the distances within the structure. The local energy functionals can be easily learnt through machine learning, whereas the long-range contributions must be treated specially using a low-dimensional form, meaning that the high-dimensionality obtained from attempting to model the interactions with a many-body functional form would be limited. In the field of machine learning potentials, long-range interactions are defined as any interaction which occurs beyond a given cut-off radius, which may not necessarily correspond to intermolecular interactions. However, in this case, the long-range contributions are usually intermolecular forces such as electrostatic, induction or dispersion forces.

Electrostatic interactions are the simplest, in that they are the classical interaction, attractive or repulsive, between static charge distributions of two molecules. They are strictly pairwise additive in terms of molecules and are strongly dependent on the orientation of the molecules. The interaction remains finite unless the nuclei overlap, in which case there is strong repulsion. The dipole-dipole interaction is described as having a $1/r^3$ relationship using the multipole expansion, where r is the separation between the nuclei. However, due to the molecules being treated as point particles, there is often some error called the “penetration error”, as the molecules are not modelled as being extended in space.

Inductive effects come from distortions of a molecule induced by the electric field from all the neighbouring molecules. These interactions are always attractive and therefore negative in sign. Since the charge distribution and therefore fields of neighbouring molecules may combine either constructively or destructively, induction is strongly non-additive in terms of molecules.

Dispersion effects are purely quantum mechanical effects and so can only be understood by solving the electronic quantum mechanical equations. The motions of the electrons in the two molecules become correlated to each other, so that lower-energy configurations are favoured, and higher-energy configurations are disfavoured. The average effect is a lowering of the energy which contributes to additional stability. As the correlation effect becomes stronger as the molecules come closer together, this results in attraction between the two molecules. At large distances, the leading term in the dispersion energy follows a $1/r^6$ relationship, which comes from the quantum mechanical solution of the correlation energy of two fluctuating dipoles[40]. This is identical to the attractive tail of the Lennard-Jones potential, highlighting that this scaling is the only part of the Lennard-Jones potential which is well grounded in quantum physics. In modelling dispersion, correction terms have been added to DFT calculations for both pair potentials[41] and many-body potentials[42] to further improve their accuracies. However, a two-body GAP could outperform this correction due to its more flexible functional form, leading to greater accuracy. Furthermore, there is far more room for optimisation of computation time in the two-body GAP, making it more favourable.

Other long-range interactions include resonance and magnetic effects. Resonance effects occur when at least one of the molecules is in a degenerate (usually excited) state or when the molecules are identical, and one is excited. Therefore, this effect is not present in a set of molecules in the ground state only. Magnetic interactions overall are small in terms of their contribution to the long-range interactions. Magnetic interactions involving the electrons are present when the electrons have unpaired spins and magnetic interactions involving the nuclei are present when the nuclei have non-zero spin[19].

All three main long-range interactions remain significant over large separations between molecules. Therefore, for many-body machine learning methods which only work for short-range cut-offs, long-range contributions must be treated differently, potentially through scaling the distances by a power which is known to govern the interaction. Such logic has been implemented before, when using machine learning methods to predict molecular atomisation energies, through scaling the input features to fit to a Lennard-Jones potential[28]. In another study, descriptors were scaled and fitted to potentials of different systems[43]. In the case of methane dimers, dispersion is the governing interaction and so distance scaling using powers of -12 and -6, as described in the Lennard-Jones potential should intuitively provide a good model for the methane dimer PES, producing pair potentials similar to the Lennard-Jones potential.

3 Aims and Objectives

Pair potentials have been developed extensively for many years and are used to model properties of a system. Therefore, the accuracy of these properties is limited by the accuracy of the potentials developed to model them. Recently, machine learning potentials and in particular Gaussian Approximation Potentials have been developed to more accurately model potentials.

The software package *QUIP* is often used to develop GAPs, but the majority of the computational cost associated with computing GAPs, has been attributed to the computation of representations and kernels. Therefore, this study contains the development of code for use in the *librascal* software to be able to compute representations and kernels more efficiently.

Using the *librascal* code currently under development, the regular unscaled distances of a methane dimer set will be used to create a Gaussian Approximation Potential for a methane dimer. This will be used to compute the pair potentials for the C-C, C-H and H-H pair types.

However, the main focus of this study will be to ascertain whether scaling the distances in a Lennard-Jones fashion will lead to a more accurate potential model for methane dimers. Using *librascal*, the pair distances will be scaled to the powers -12 and -6, corresponding to the scaling seen in the Lennard-Jones potential, to create a more accurate fit. It is thought that through scaling the distances, the fit can more closely resemble the Lennard-Jones potential, which is known contain the scaling used in the governing interaction of dispersion in methane or any non-polar but polarisable substance.

4 Methodology

4.1 Methane Dimer Dataset

The methane dimer dataset was taken from a study into using GAP to model molecular liquid methane dimers. In this study, a sample of 2418 dimers was taken from a liquid MD simulation using 200 rigid geometry-optimised methane molecules. The sampling was biased towards shorter distances, since this is where the GAP was desired to be most accurate. The intermolecular energies were generated using second-order Møller–Plesset perturbation theory (MP2) and were exactly the two-body contribution to the total energy, in terms of the molecules, with the one-body methane molecular energies removed[37].

4.2 Python Prototype

A prototype for the kernel regression model was created using Python alone and trained on a methane dimer subset of 100 dimers, which was from the larger set of 2418 dimers.

The pair distances for each type were computed, giving C-C, C-H and H-H pair distances. Kernels were computed using a radial basis function of a Gaussian. The length scale parameter controlling the width of the Gaussian, θ, was set to 1 Å. From these kernels, the weights, $\boldsymbol{\alpha}$, were calculated using a regulariser, $\sigma = 1$. This was used to compute the fitted energies, which were compared with the dataset for the C-C pair type, along with its error. To measure accuracy, the training RMSE was calculated.

4.3 *QUIP* and *librascal*

QUIP is a package of software tools used to carry out molecular dynamics simulations, using different types of interatomic potentials and tight binding quantum mechanics. It is also able to call external packages and plug into other software such as LAMMPS, CP2K and the python framework ASE (Atomic Simulation Environment).

QUIP has a number of unique features such as deep access to most of the Fortran types and routines from Python and support for Gaussian Approximation Potentials (GAP). It also does not assume minimum image convention, so interatomic potentials can have cut-off distances which are larger than the periodic unit cell size. Many different interatomic potentials are coded in *QUIP*, such as Gaussian Approximation Potentials (GAP), Lennard-Jones and Morse potentials[44].

librascal is a versatile and scalable machine learning code by the Computational Science and Modelling group (*COSMO*) at EPFL, Switzerland, which is in the alpha phase of development. Its primary function is to efficiently construct representations of atomic structures which can be fed into any supervised or unsupervised machine learning algorithm.

It will be able to be interfaced with codes such as LAMMPS and PLUMED-2.0 and be able to use parallelisation over atomic structures to significantly reduce computational cost. Parallelisation can be used over atoms in a large structure, or a large collection of small structures or for representations with a large number of functions and components[45].

librascal is intended to replace the *QUIP* for the purposes of fitting and evaluating machine learning potentials, including GAP and other models, such as neural network potentials, due to its improved ability to compute descriptors. Most notably its ability to compute SOAP (Smooth Overlap of Atomic Positions) descriptors more easily through parallelisation, makes it a successor to *QUIP*. Therefore, it was necessary to enable *librascal* to have some of the same functionality as *QUIP*, while improving the speed of its computation of descriptors.

librascal uses a C++ core with Python bindings. Functionality was added to *librascal* to create and compute new Pair Distance representations using a cut-off radius. These representations were able to be used to compute the pair distances themselves and Gaussian kernels. Functionality was also added to be able to scale the pair distances by a specified power, to implement distance scaling. The distances and kernels were able to be used for Gaussian Process Regression to produce GAPs in Python, via the Python bindings. Full details of the code development carried out can be found in the Appendix.

4.4 Gaussian Approximation Potential Models using *librascal*

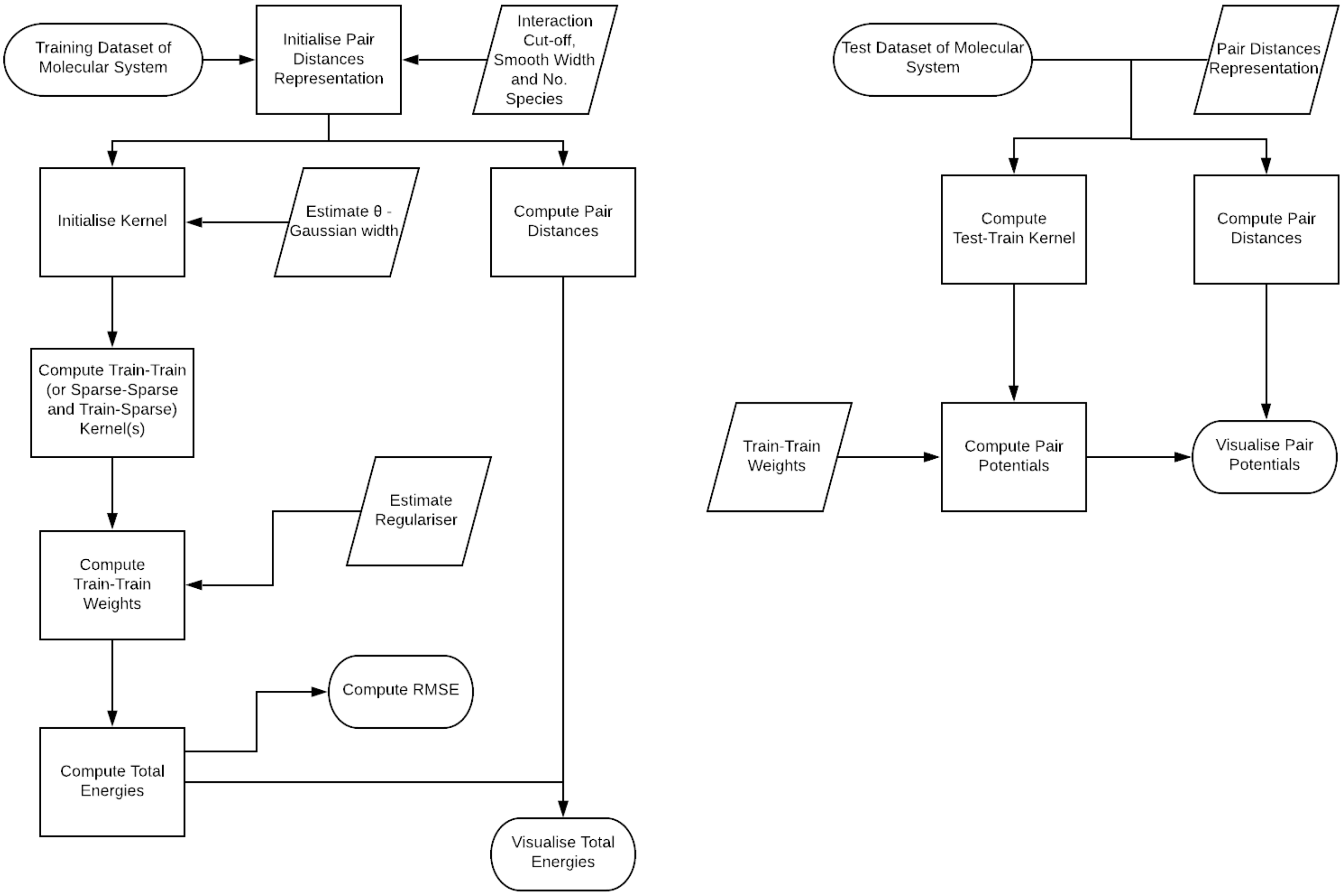

Training Dataset of Molecular System
Initialise Pair Distances Representation
Interaction Cut-off, Smooth Width and No. Species
Initialise Kernel
Estimate θ - Gaussian width
Compute Pair Distances
Compute Train-Train (or Sparse-Sparse and Train-Sparse) Kernel(s)
Estimate Regulariser
Compute Train-Train Weights
Compute RMSE
Compute Total Energies
Visualise Total Energies
Test Dataset of Molecular System
Pair Distances Representation
Compute Test-Train Kernel
Compute Pair Distances
Train-Train Weights
Compute Pair Potentials
Visualise Pair Potentials

Figure 4 - Flowcharts to calculate Pair Potentials using Kernel Regression in *librascal*

A Pair Distances representation was initialised, with a cut-off radius of 12 Å and the number of species set to 2. The pair distances were computed from the same methane dimer set used in the prototype but using all 2418 dimers as non-periodic structures, using the Pair Distances representation shown in the first flowchart of Figure 4. A Gaussian train-train kernel was initialised and computed, using a *Structure* target type, with the length scale parameter θ, set to 1 Å again, as in the prototype. The weights, α, were calculated using kernel regression on the train-train kernel matrix and using a regulariser of $\sigma=1$ *eV*. The fitted total energies of the system were compared with the dataset total energies and the error was analysed. The training RMSE of the fit for the total potential energy surface was also calculated, as demonstrated in Figure 4.

It was then thought that varying the hyperparameters could lead to an improved fit. The length scale parameter θ, used in computing the Gaussian kernel, was varied using values of 0.5 Å and 2 Å. Kernel regression was used again with a regulariser of 1 to compute the fitted total energies. The fitted total energies were compared to the dataset energies, the error analysed and the training RMSE of the two fits were also calculated. Using the Gaussian kernel with a θ value of 1 Å, the regulariser, σ, was varied, with values of 0.1, 0.5, 2 and 10 eV. Again, the weights were calculated through kernel regression and the fitted energies were computed. The fitted total energies again were compared with the dataset energies, the error analysed and the RMSE of the two fits were also calculated.

In order to understand the contribution of different pair types to the total PES, pair potentials were required to be computed, through using test datasets of single pairs of atoms, with incrementally increasing distances. Test datasets of 10,000 distances were created for C-C, C-H and H-H distances between 0.5-10 Å. From this, test-train kernels were computed using the same improved value *θ=0.5 Å* and used, in conjunction with weights computed from train-train kernels, to calculate the pair potential function for each pair type.

After observing pair potentials which oscillated at long-range, it was perceived that the pair potentials were overfitted. Therefore, histograms of the pair distances were analysed to better estimate a value for θ. A new normalised kernel was computed with the estimated value for θ. Secondly, a logarithmic grid search hyperparameter optimisation was carried out to find the value of an optimal regulariser, using six-fold cross-validation. The pair potentials for these values were then analysed.

4.5 Distance Power Models

A different model was then investigated, using the same Pair Distance representation. New train-train kernels were computed while scaling the distances by powers of -6 and -12 and using the same methane dimer set. From analysing the scaled distance histograms, estimated values of θ used in computing the kernels for 6th- and 12th-power fits were found and normalised kernels were computed. Through six-fold cross-validation, logarithmic grid search hyperparameter optimisations were carried out to find the optimal regularisers for each model. Using kernel regression, the weights were obtained, and used to compute the fitted total energy surfaces, followed by the training RMSE of the fits.

Using the same test datasets, the normalised test-train kernels were computed using the scaled distances Pair Distance representations. These kernels were used with the weights built from the train-train kernels to calculate the pair potential functions.

A 12-6 model was then created through summing the two train-train kernels and a new set of weights were computed, along with the total energies using an optimal regulariser, again calculated from six-fold cross-validation. The errors and the RMSE were computed. Using the test dataset, a test-train kernel was computed for the 12-6 model and used to compute the pair potentials again in the same fashion.

The pair potentials computed from the 12-6 model were first compared to the pair potentials from the regular unscaled distances model and then with the potentials from the 6th- and 12th-power models.

4.6 Gaussian Approximation Potential Model using QUIP

A GAP model was created using QUIP and its Potential class and trained on the set of 2418 methane dimers. The parameters of the distance computation were: C-C pair distances had a cut-off radius of 10 Å, C-H and H-H pair distances had a cut-off radius of 6 Å. The same improved hyperparameter values used in the regular unscaled distances model were used to compute the weights and therefore potential energy surface using kernel regression.

A test dataset was created with 10,000 pair distances between 0.5-10 Å for each pair type. Using the trained potential, the different pair potentials were calculated and plotted, comparing to the 12-6 model.

4.7 Classical Model

A classical potential was calculated using a COMPASS forcefield in a LAMMPS simulation, to compare the accuracy of the model with the 12-6 model using *librascal*. The total energies of the system were calculated, along with the RMSE and compared with those calculated using the 12-6 model in *librascal*. The COMPASS parameters used, were taken from a study into using a forcefield optimised for condensed phase systems[46]. The parameters for Non-Bonded Lennard-Jones 9-6 potentials were r_0=3.854 Å and ϵ=0.062 kcal/mol for the C-C, r_0=3.526 Å and ϵ=0.027 kcal/mol for the C-H and r_0=2.878 Å and ϵ=0.023 kcal/mol for the H-H pair potentials. These pair potentials were then compared to those generated by the 12-6 model using *librascal*.

5 Results and Discussion

5.1 Regular Unscaled Distances Model using *librascal*

A model was built using the regular unscaled distances, with the pair distance representations and kernels computed in *librascal*. The descriptor used in the basis function was the pair distances between given atoms or bodies in the dataset. The kernel function used was that of a Gaussian kernel as a similarity measure through the squared exponential form.

Using the built *librascal* code, a Pair Distances representation was created in Python, using a cut-off radius of 12 Å, which was large enough to encompass all pairs, and the number of species was set to 2. The model was trained on the methane dimer dataset, outlined in the first section of the Methodology, using all 2418 dimers. Unit cells were set, since the datasets were for isolated molecular systems, the unit cell was "padded" so that periodic images did not repeat within the chosen cut-off radius.

The distances for all three atom pair types were computed and a Gaussian train-train kernel, K_{NN}, was computed using Equation 7, with a value of $\theta = 1$ Å for the length scale parameter, as over this range, distances seemed to be similar in terms of the probable interaction between the atoms in a pair. In this case, *N* is the number of dimers in the dataset. The typical sparse fitting equation is given by Equation 15[24].

$$\boldsymbol{\alpha} = \left(\boldsymbol{K_{MM}} + \boldsymbol{K_{NM}}^{T}(\sigma^{2}\boldsymbol{I_{NN}})^{-1}\boldsymbol{K_{NM}}\right)^{-1}\boldsymbol{K_{NM}}^{T}(\sigma^{2}\boldsymbol{I_{NN}})^{-1}\,\boldsymbol{y} \tag{15}$$

In this fitting function, the sparse-sparse kernel, K_{MM}, is a matrix where each element corresponds to the Gaussian similarity measure between any two pair distances across all pair types and dimers. The train-sparse kernel, K_{NM}, is similar to the sparse-sparse kernel but the rows are summed across all pairs for each dimer, giving *N* rows. To simplify the sparse fitting equation, in the non-sparse case, where *N* = *M*, the equation can be reduced to the fitting equation in Equation 16[47].

$$\boldsymbol{\alpha} = (\boldsymbol{K_{NN}} + \sigma^{2}\boldsymbol{I_{NN}})^{\boldsymbol{-1}}\,\boldsymbol{y} \tag{16}$$

The vector of the weights, $\boldsymbol{\alpha}$, was computed using kernel regression, with the regulariser, $\sigma = 1$, according to Equation 16. The regulariser was selected based on the variability of the data. Using the weights, a series of fitted total energies were computed as a reduced dimensionality potential energy surface. The true potential energy surface for a system of methane dimers is six-dimensional, with one dimension for molecular separation (equivalent to C-C distance), three for the first molecular orientation and two for the second molecular orientation relative to the first. However, the fit in this case consists of only three one-dimensional potentials, highlighting the reduced dimensionality. Evidently, this is not as flexible as a full six-dimensional PES, which would come from many-body terms in the atomic body-order expansion in Equation 3. A cross-section of the fitted potential energy surface is taken along the axis of C-C pair distances and is shown in Figure 5, alongside the dataset energy values.

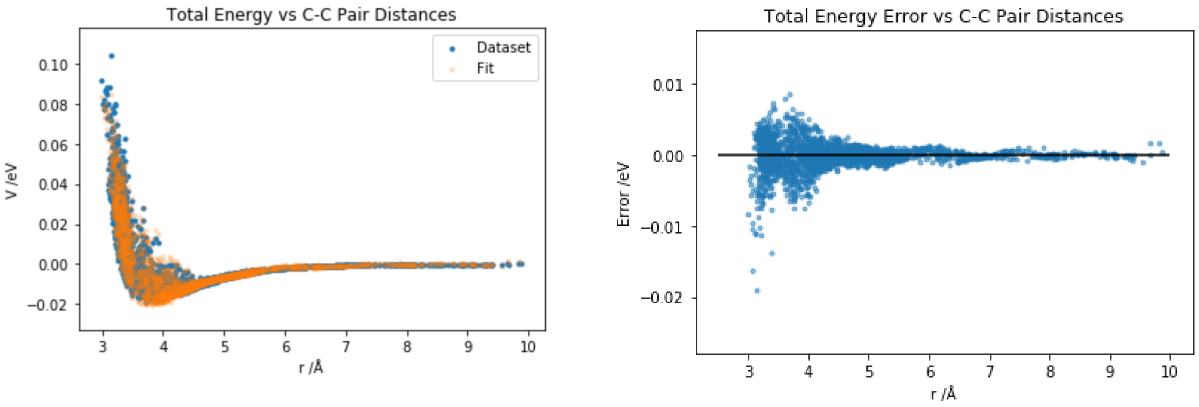


Figure 5 - Total Energy and its error as a function of C-C Predicted Pair Distances for 2418 Methane Dimers

It can be seen that the fitted total energies largely match the dataset energies. The error is larger for smaller C-C pair distances and tends to zero at larger pair distances, showing that the potential fits better when the two molecules are further apart. This is to be expected as at short range, there is strong repulsion and higher variance in the energy, leading to a greater error. However, there are small oscillations at long-range, highlighting that a shorter cut-off radius may be required. The error in this case could be reduced by optimising the hyperparameters and improving the value of σ chosen in the Gaussian kernel to create an optimal fit for the dataset. The training RMSE of the fit was calculated as being 994 µeV per methane molecule, showing a low overall error across all data points, which is much lower than the standard deviation of the quantum mechanical energies, which was 8.77 meV per methane molecule.

It was then thought that varying the length scale parameter and the regulariser would lead to an improved fit. Different parameters and hyperparameters used in building the kernel and implementing the regression were varied. The parameter ϑ, which controls the width of the Gaussian used as a similarity measure in the kernel matrix was varied, to see how the accuracy of the fit for the total potential energy surface of the methane dimers changed.

It was found that narrower Gaussian widths gave the lowest RMSEs on the training set, due to only almost identical distances qualifying as being “similar” and having any sizeable contribution to the kernel. This would potentially lead to some degree of overfitting despite the Tikhonov regularisation. Fits with too large a data “noise” (or regulariser) and length scale parameter will lead most likely to underfitting, where the potential is insensitive to variation and over-smoothed. Similarly, fits with too small a data “noise” and length scale parameter will most likely lead to overfitting, where the potential is oversensitive to fluctuations[48]. This means that larger differences in distance produced a smaller overall contribution to the kernel matrix summation and so only distances close in magnitude had the most significant if not any contribution to the summation of the kernel matrix.

In order to find an estimate for a good value for the Gaussian width in computing the kernel, histograms of the pair distances were plotted for each pair. From the histograms, differences between distances that could be classed as being similar in type, would be predicted to be good estimates for the Gaussian width.

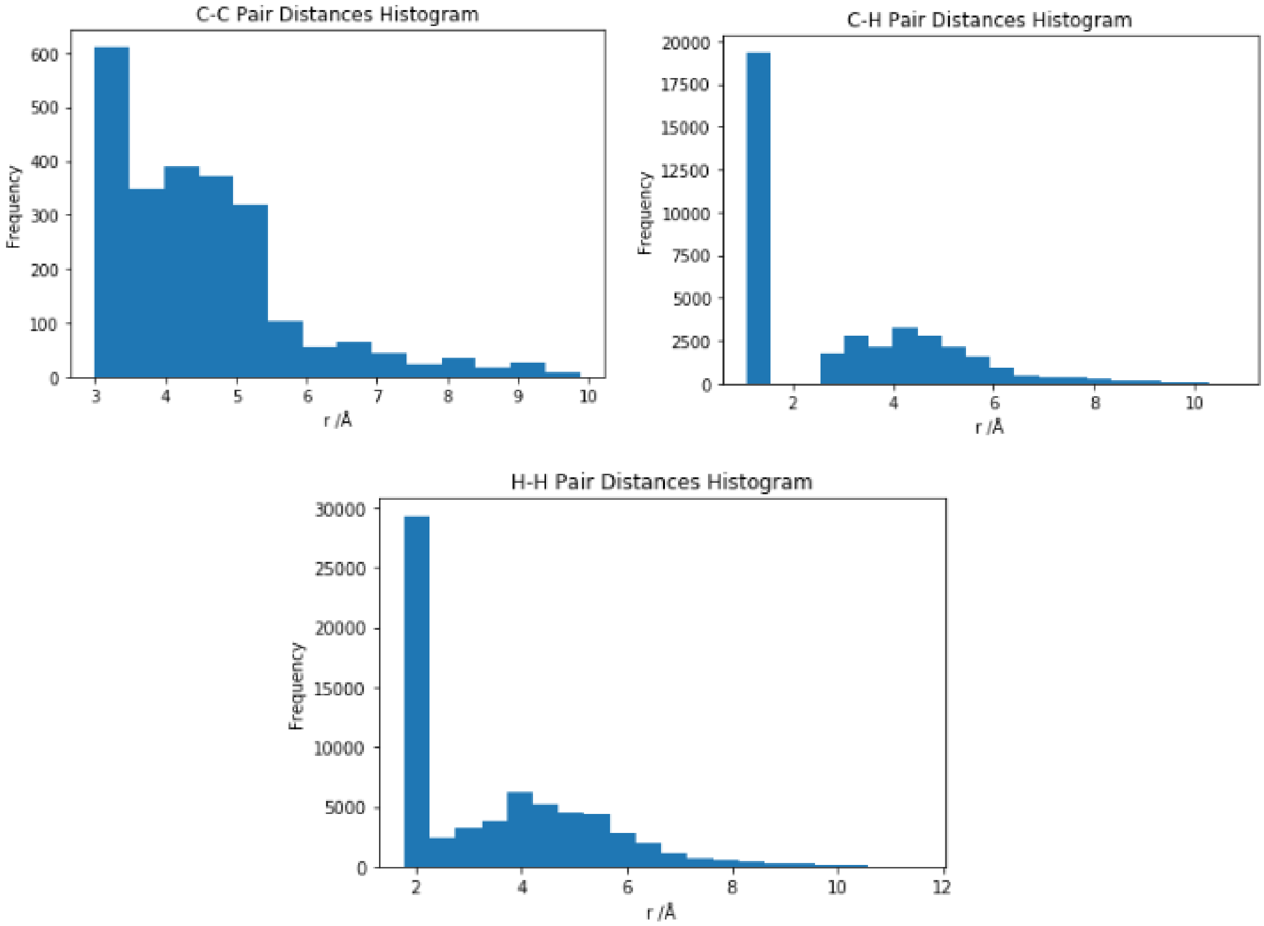


Figure 6 - C-C, C-H and H-H Pair Distances Histograms

The C-H and H-H pair distance histograms produced large bars for the intramolecular pair distances, since the methane monomers were rigid, the intramolecular distances were almost identical with extremely little variance. The intramolecular pair distances matched with the literature values of 1.094 Å for a C-H bond length[49] and 1.787 Å for the H-H distance within a methane molecule. However, in *librascal*, there was no way to distinguish between intramolecular and intermolecular distances, based on the value alone.

The parameter θ used in the Gaussian kernel calculation, governs the width of the Gaussian used as a similarity measure and should be equivalent to the distance over which two points remain correlated, and therefore over what range the function is expected to vary. It was found through observing the similarity of distances in the histograms and total PES in Figure 5, that a θ value of 0.8 Å would fit well, as it covered parts of the data which could be deemed “similar”.

The regulariser hyperparameter, σ, was varied to observe how the fit for the total potential energy surface for methane dimers changed. A lower σ value would place less emphasis on the regularisation term in the loss function. Therefore, through a loss of regularisation, lower regulariser values led to lower RMSEs on the training set and an overfitted PES. It was also seen that there was a more noticeable difference in the change in error when the order of magnitude was changed. Therefore, it was suggested that a logarithmic hyperparameter optimisation would be beneficial in finding a good regulariser, since regularisers which gave functions which were too overfitted or underfitted would give a large cross-validation error. Six-fold cross-validation was carried out to optimise the regulariser.

Furthermore, the regulariser, σ, should be representative of the noise and the variation which cannot be fitted, expressed as a ratio of the data noise or the overall standard deviation of the fitting function to the overall data scale. A logarithmic grid search regulariser optimisation was carried out and the results were plotted in Figure 7.

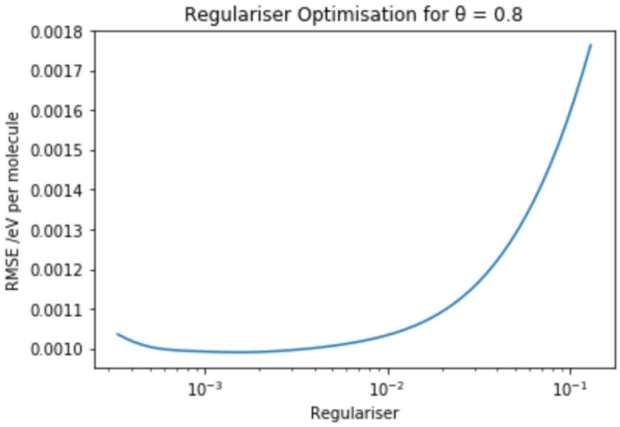


Figure 7 - Logarithmic Regulariser Optimisation

From the optimisation, it was found that very low regularisers close to zero overfitted the total energy surface, demonstrated by the fact that they produce regressions which are less affected by regularisation. This leads to higher RMSEs through cross-validation. It can also be seen that larger regularisers also have a higher RMSE, due to their underfitting of the total energy surface. Therefore, it was found that the optimal regulariser lay between 10^{-3} and $5x10^{-2}$ on the logarithmic scale. It was thought that a higher regulariser than the minimum, without heavily compromising on the RMSE, would lead to a less overfitted potential. Therefore, the regulariser value chosen was 0.03, as this did not increase the cross-validation error significantly and reduced the amplitude of oscillations of potentials at long-range. The cross-validation error for a regulariser of 0.03 was 1.15 meV per methane molecule, which was deemed low enough given that the minimum cross-validation error in the optimisation was 0.991 meV per methane molecule. Therefore, this regulariser value was used as it appeared to cover the variance of the fit well.

Using these optimised hyperparameters, the total potential energy surface was recomputed and a cross-section along the C-C distance axis was replotted in Figure 8, alongside the error.

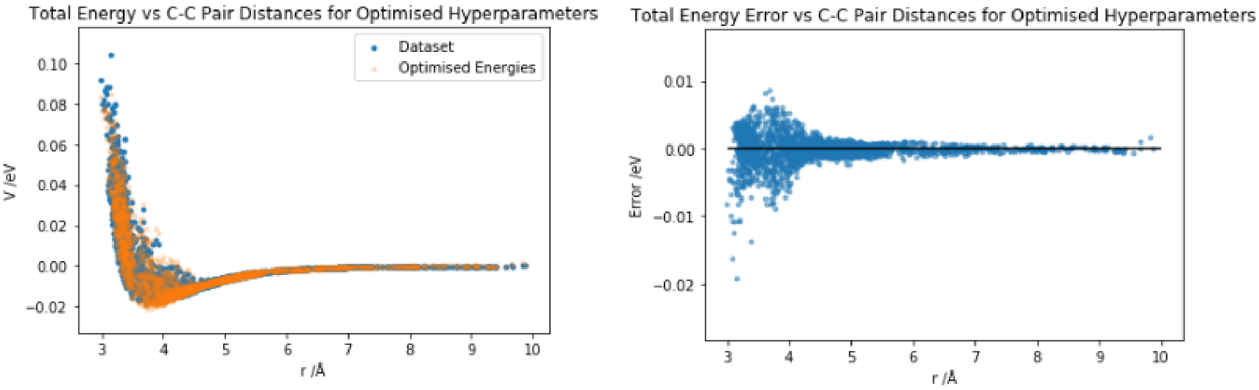


Figure 8 - Total Energy as a function C-C Pair Distances using Optimised Hyperparameters

The improved potential energy surface was shown to be a good fit, with a training RMSE of 1.02 meV per methane molecule, expectedly lower than the cross-validation error.

Finally, from the fitted total energies, the individual C-C, C-H and H-H pair potentials were calculated. Since *librascal* was unable to differentiate between bonded and non-bonded or intramolecular and intermolecular pairs, intermolecular potentials were generated only for each pair type. The improved hyperparameter values, *θ=0.8 Å* and *σ=0.03* were used.

A new test set of 10,000 distances was created for each pair type, and a new test-train kernel, K_{TN} was computed, which contained *T* rows (number of test distances) and *N* columns (number of training distances). Each test set consisted of a pair of the relevant atoms at a distance which increased incrementally between 0.5 and 10 Å.

To compute the pair potential energies, Equation 13 was used, but with the test-train kernel instead of the train-train kernel. Its dot product with the optimised training weights, $\boldsymbol{\alpha}$, obtained using the K_{NN} kernel originally, gave the pair potential energies for a specific pair type. The pair potentials are shown in Figure 9.

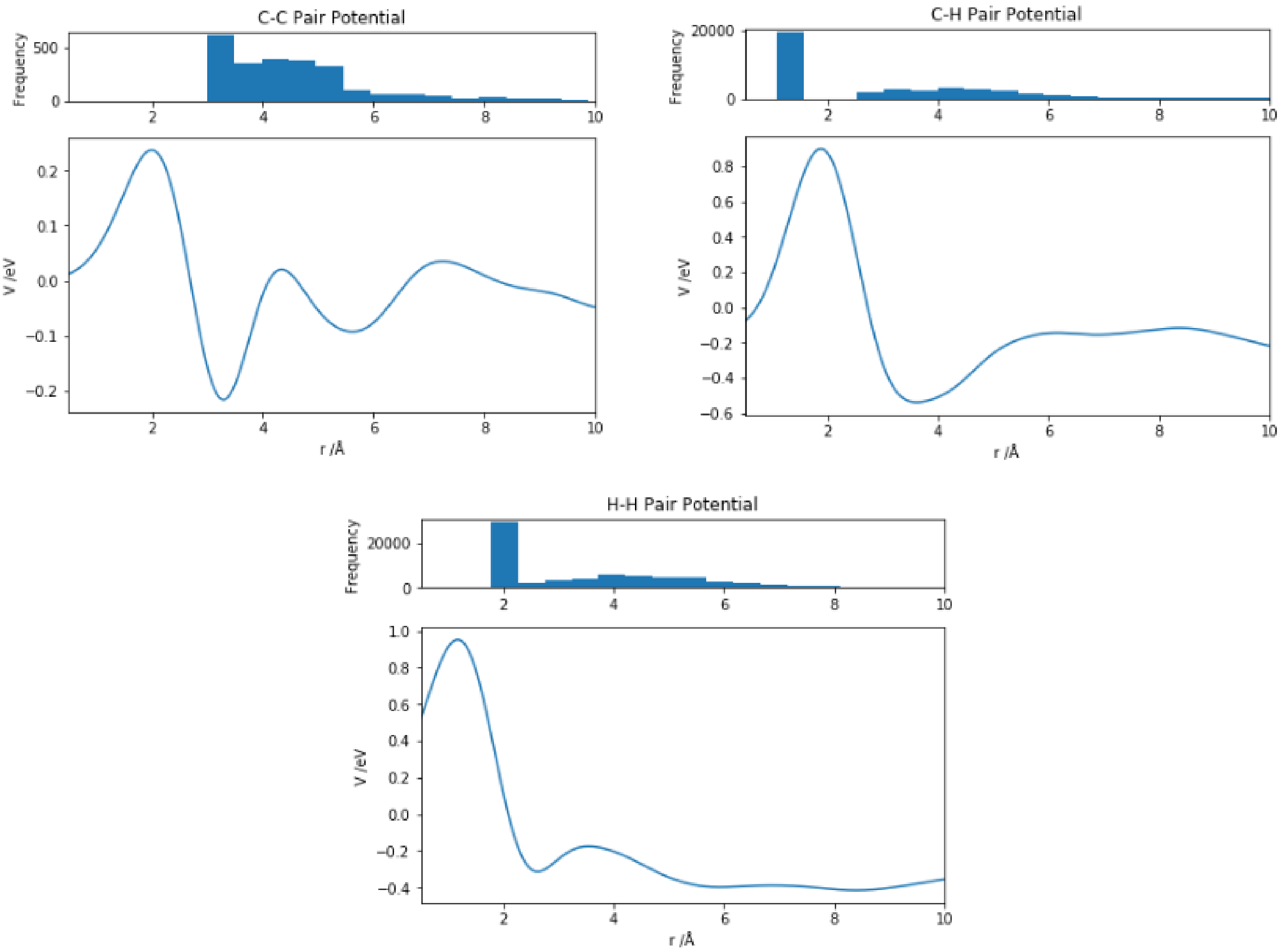


Figure 9 - Predicted Pair Potentials for C-C, C-H and H-H pair types in methane dimers

The predicted pair potentials for C-C, C-H and H-H distances in a methane dimer had differences with the Lennard-Jones potential, through showing oscillations at longer distances, suggesting that a shorter cut-off radius would lead to a better tail form. Furthermore, they had high amplitudes in the energy scale, suggesting a full optimisation for θ would be more beneficial to provide a more realistic potential. Such an optimisation for θ however, was unfeasible given the timescale of this project, but other values, 0.5, 1 and 2 Å were trialled but led to larger RMSEs on the training set.

The different pair potentials are contributions to the total energy, which is displayed in Figure 8. The C-C, C-H and H-H pair potentials' contributions to the total energies contain mostly repulsive components, shown by the large peaks which act repulsively on pairs beyond the distance corresponding to the peak. Using this model, the order of increasing equilibrium pair distance is H-H < C-C < C-H and the order of increasing well depth is C-C < H-H < C-H.

The pair distance histograms were plotted at the top of the figures, done similarly to Figure 3, in order to show the regions of the potential which are most likely to be reliable, given the concentration of data at those distances. It can be seen that predictions under 3 Å for the C-C pair potential and under 1 Å for the C-H and H-H pair potentials are likely to be less reliable as no data from the original datasets are in these regions. The greatest reliability of the pair potentials is likely to come at a distance of about 3 Å for C-C distances at the high part of the repulsive potential and about 4 Å for C-H and H-H distances at the low part of the repulsive function due to a concentration of data at those points. There is also likely to be decreasing reliability at longer distances in the tail form due to fewer data points in those regions.

5.2 Distance Powers (12-6) Model using *librascal*

The inverse power functionality added to the Pair Distances representation in *librascal* was used, to observe if the pair potentials could be better represented through scaling the distances by powers of -12 and -6. The power of -6 was chosen from the result of quantum mechanics for the phenomenon of dispersion and the power of -12 was chosen from the Lennard-Jones (12-6) potential. Then, in order to create a 12-6 model which was created using distances scaled by both the "repulsive" -12 and "attractive" -6 powers, the linear combination of kernels computed using the scaled distances would be taken. Given that studies have already been carried out in fitting potentials with scaled distances[28,43], it was thought that scaling the distances would produce a model which was more similar to the Lennard-Jones potential than the regular unscaled distances model. Therefore, the new scaled distances model was expected to remove the high error at short-range and oscillating pair potentials at long-range which were observed for the regular unscaled distances model. The new combined kernel, K_{NN}, is calculated through Equation 17.

$$(K_{NN})_{IJ} = \sum_{i,j\epsilon I} \sum_{k,l\epsilon J} w_6 k^{(6)}({r_{ij}}^{-6}, {r_{kl}}^{-6}) + w_{12} k^{(12)}({r_{ij}}^{-12}, {r_{kl}}^{-12}) \qquad (17)$$

From Equation 17, it can be seen that the new kernel is calculated through the sum of the kernel multiplied by a weight for a distance power of -6, with the kernel multiplied by a weight for a distance power of -12. The weights in this case represent the relative importance of -6 (attractive) and -12 (repulsive) kernel contributions.

Two new pair distances representations were created with same parameters as the regular distances, with a cut-off radius of 12 Å and 2 species, but with the distance powers parameter set to -6 and -12 respectively. From the original training dataset of 2418 dimers, the pair distances were appropriately scaled by a power of -6, and histograms were plotted to estimate the hyperparameter choices again. In the case of C-H and H-H pair distances, only the distances within a certain scaled range, perceived to be intermolecular, were plotted to better observe the similarity range.

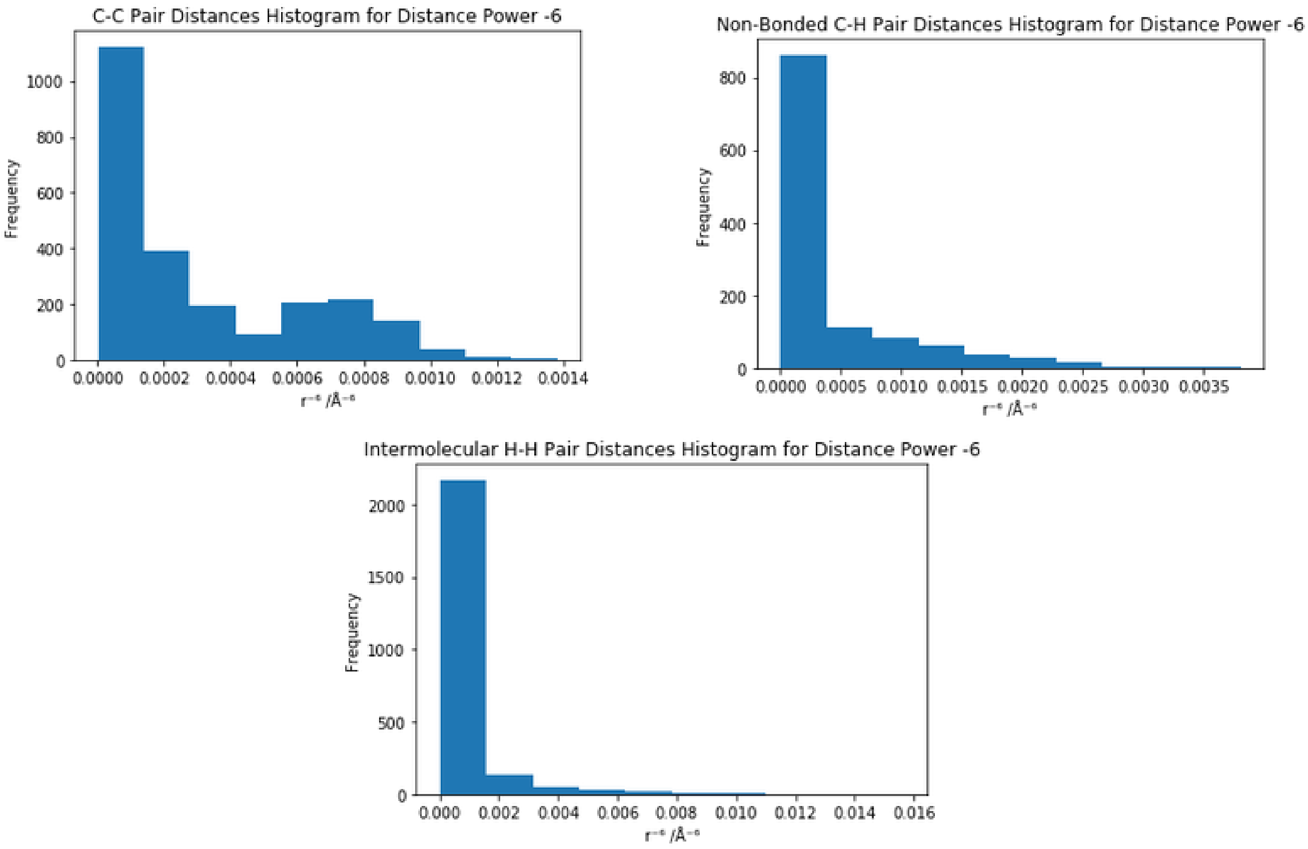


Figure 10 - C-C, C-H and H-H Pair Distances Histograms for 6th-power scaling

From the pair distances histograms for the 6th-power fit, it can be seen that the majority of the data is concentrated close to zero due to the scaling, suggesting that a good estimate for θ would be 5x10^{-4} Å^{-6} as distances within this range of each other are likely to be similar.

To estimate the regulariser for the scaled fit of -6, a logarithmic grid search hyperparameter optimisation was carried out, using six-fold cross-validation. The results are shown in Figure 11.

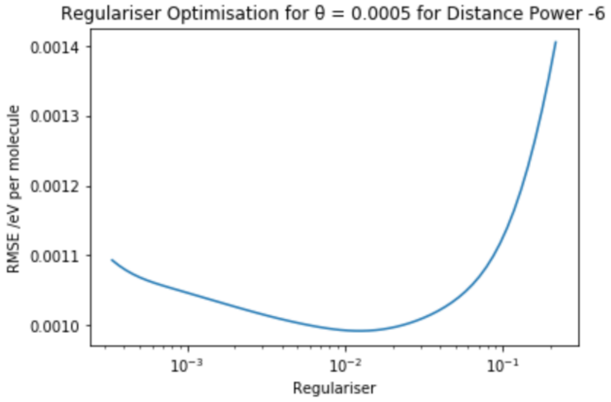


Figure 11 - Regulariser Optimisation for 6th-power fit

The logarithmic grid search gave a curved function in the RMSE and again a slightly larger regulariser of 0.07 was chosen so as not to compromise on the error and reduce overfitting in the pair potentials. For a regulariser of 0.07, the cross-validation error was 1.07 meV per methane molecule, which was reasonable given that the minimum of the cross-validation error in the optimisation was 0.992 meV per methane molecule. Since the cross-validation RMSE started to heavily increase at a regulariser of around 0.1, a choice of 0.07 for the regulariser was suitable in this case.

The pair distances were scaled by a power of -12 and scaled distance histograms were plotted to estimate the length scale parameter again.

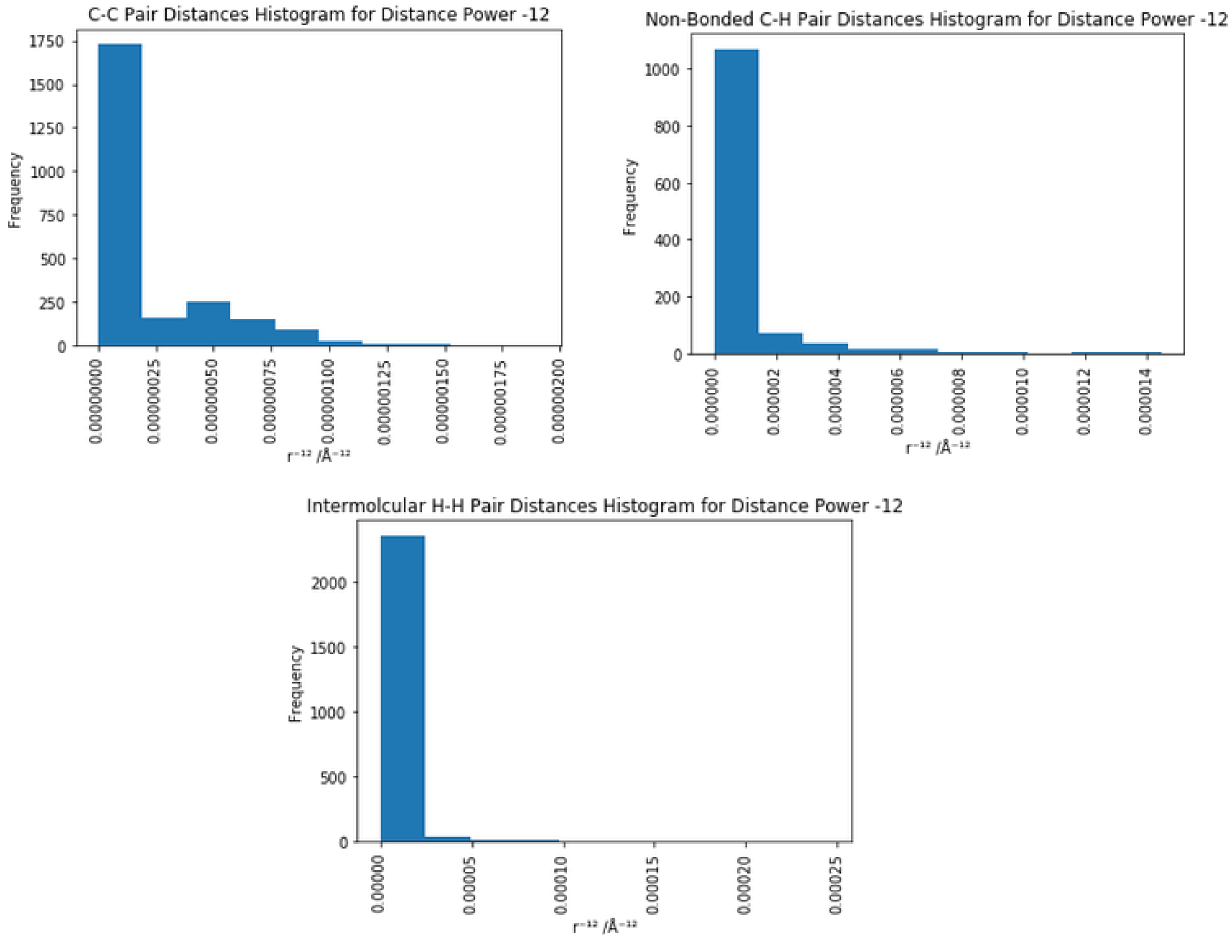


Figure 12 - C-C, C-H and H-H Pair Distances Histograms for 12th-power scaling

It can be seen again that much of the data is very close to zero, due to the strong scaling. Therefore, a value for θ of 2.5x10^{-7} Å^{-12} was chosen to be a good estimate again as distances within this range again are likely to be similar.

Another logarithmic grid search hyperparameter optimisation was carried out for 12th-power fit, using six-fold cross-validation, with the results shown in Figure 13.

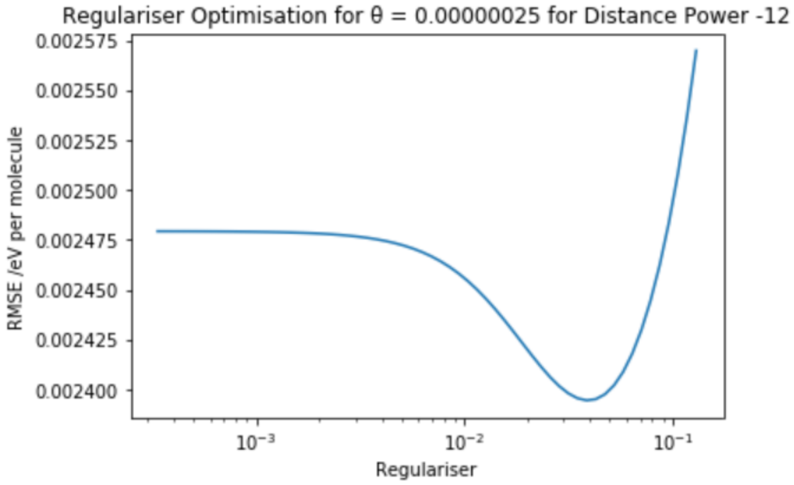
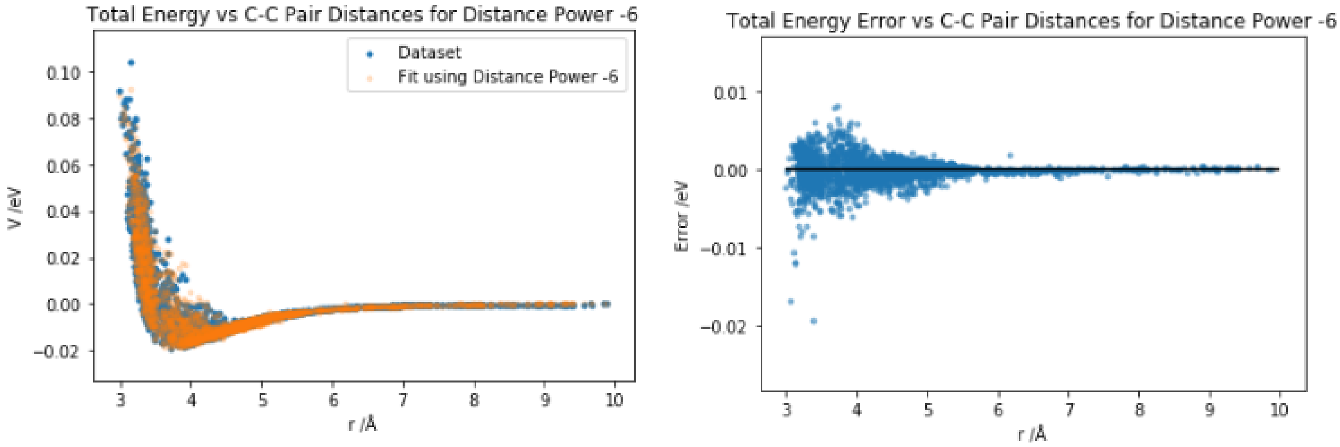


Figure 13 - Regulariser Optimisation for 12th-power fit

The logarithmic grid search again gave a curved function in the RMSE and a slightly higher regulariser could be chosen without increasing the RMSE significantly. Therefore, a regulariser value of 0.06 was chosen, since again the cross-validation error only significantly increased at a regulariser of 0.1 for the 12th-power fit. Furthermore, the cross-validation error at a regulariser of 0.06 was 2.41 meV per methane molecule, which was reasonable, given that the minimum cross-validation error for the optimisation was 2.38 meV per methane molecule.

Using these θ values, new normalised Gaussian kernels were computed. Through Kernel Regression, the weights of these kernels were computed, using the aforementioned optimal regularisers of 0.07 and 0.06 for the distance powers of -6 and -12 respectively. Using the weights, the fitted total potential energy surfaces were computed for Distance Powers of both -6 and -12 and their cross-sections are shown in Figures 14 and 15 as functions of both scaled and unscaled distances, with their errors.



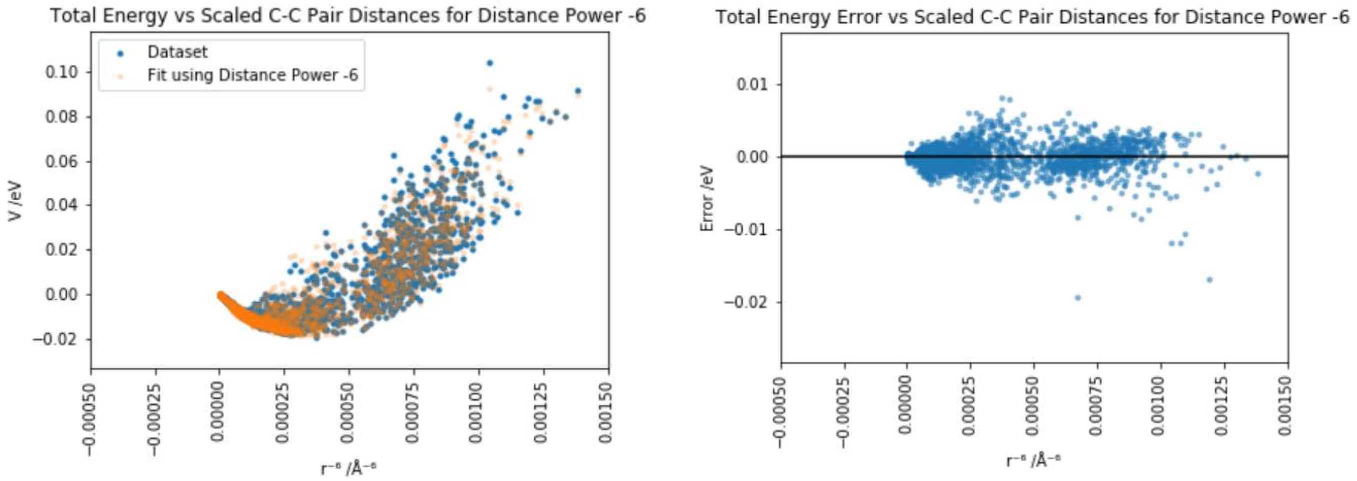


Figure 14 - Total Energy functions of Scaled and Unscaled C-C Pair Distances for 6th-power fit and their errors

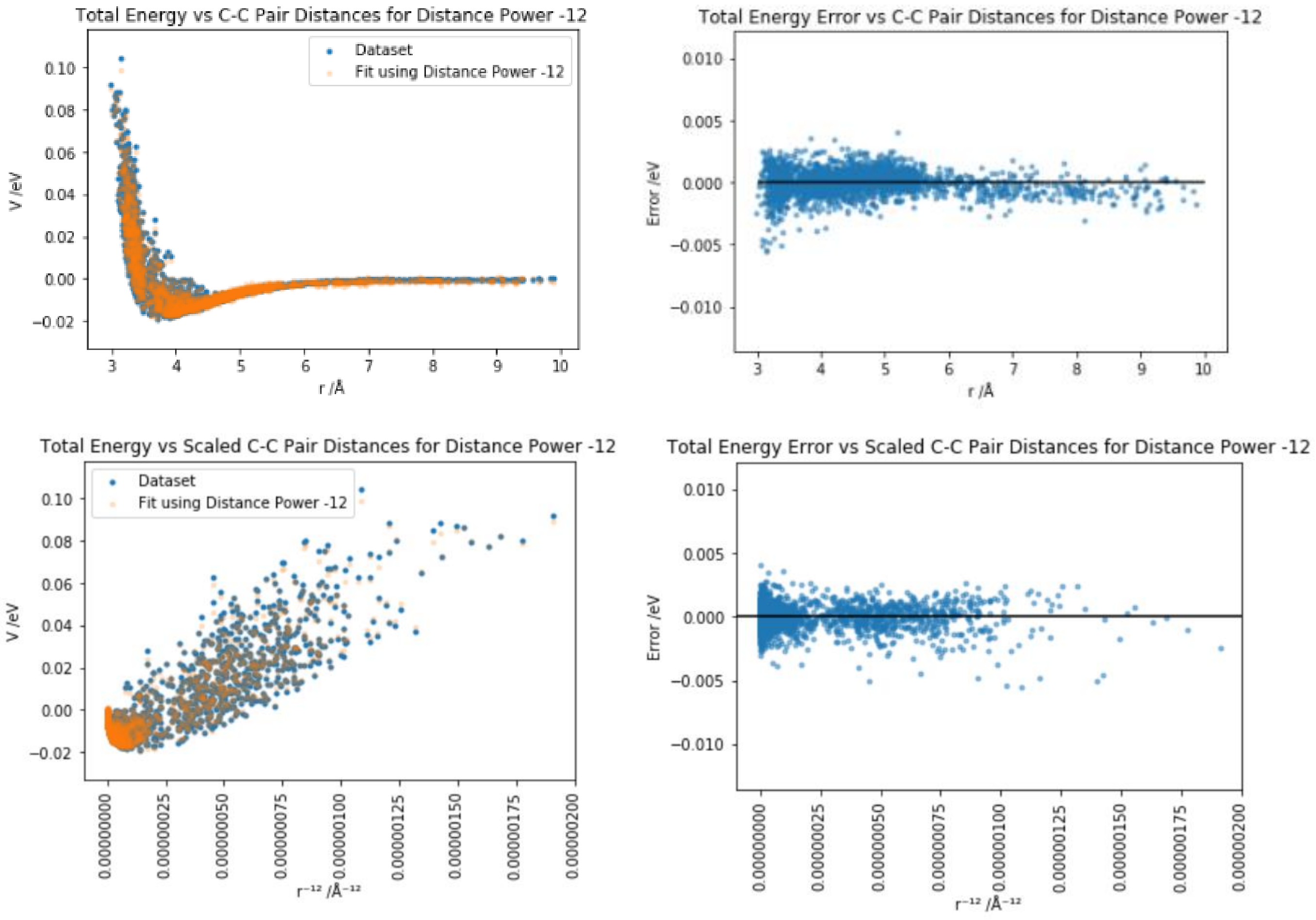


Figure 15 - Total Energy functions of Scaled and Unscaled C-C Pair Distances for 12th-power fit and their errors

From the total energy functions and error plots, it can be seen that the 12th-power fit has a lower training RMSE than the 6th-power fit. The training RMSE of the 6th- and 12th-power fits are 863 and 499 µeV per methane molecule respectively, highlighting the lower training error of the 12th-power fit. Despite the lower training error for the 12th-power fit, it has a higher cross-validation error, implying that it is overfitted despite the regulariser optimisation. Furthermore, unlike the 6th-power fit, the 12th-power fit has a bias at longer range, with error values consistently negative beyond 6-7 Å. Since no other value of the regulariser gave a significantly lower cross-validation error, it suggests that the fit could only be improved further with a full optimisation for the length scale parameter.

The 6th-power fit has a smaller error at short-range than the unscaled distances fit, showing that the distance scaling has been successful in somewhat removing the short-range error. In addition, the oscillations observed at long-range in the unscaled distances fit have also been removed in the 6th-power fit. Therefore, it can be concluded that the 6th-power fit has slightly improved on the error of the unscaled distances fit at both short- and long-range with a lower cross-validation error and a much more physical 6th-power decay. These results also suggest that the 6th-power rescaling is a good option for fitting the long-range tail of intermolecular interactions, with the short-range potentially taken care of by another model, such as a SOAP-GAP.

In order to create the pair potentials again for the different distance powers, the test datasets of pairs created previously were used again. Test-train kernels for each of the different pair types were computed. Using the weights calculated from the train-train kernels, and the test-train kernels, the pair potentials for scaled distances were computed. The pair potentials were found to contain residual intramolecular energy, as the sum of the intramolecular energies and the intermolecular offsets was equal to zero. This residual intramolecular energy was removed from the pair potentials. The potentials are shown in Figures 16 and 17.

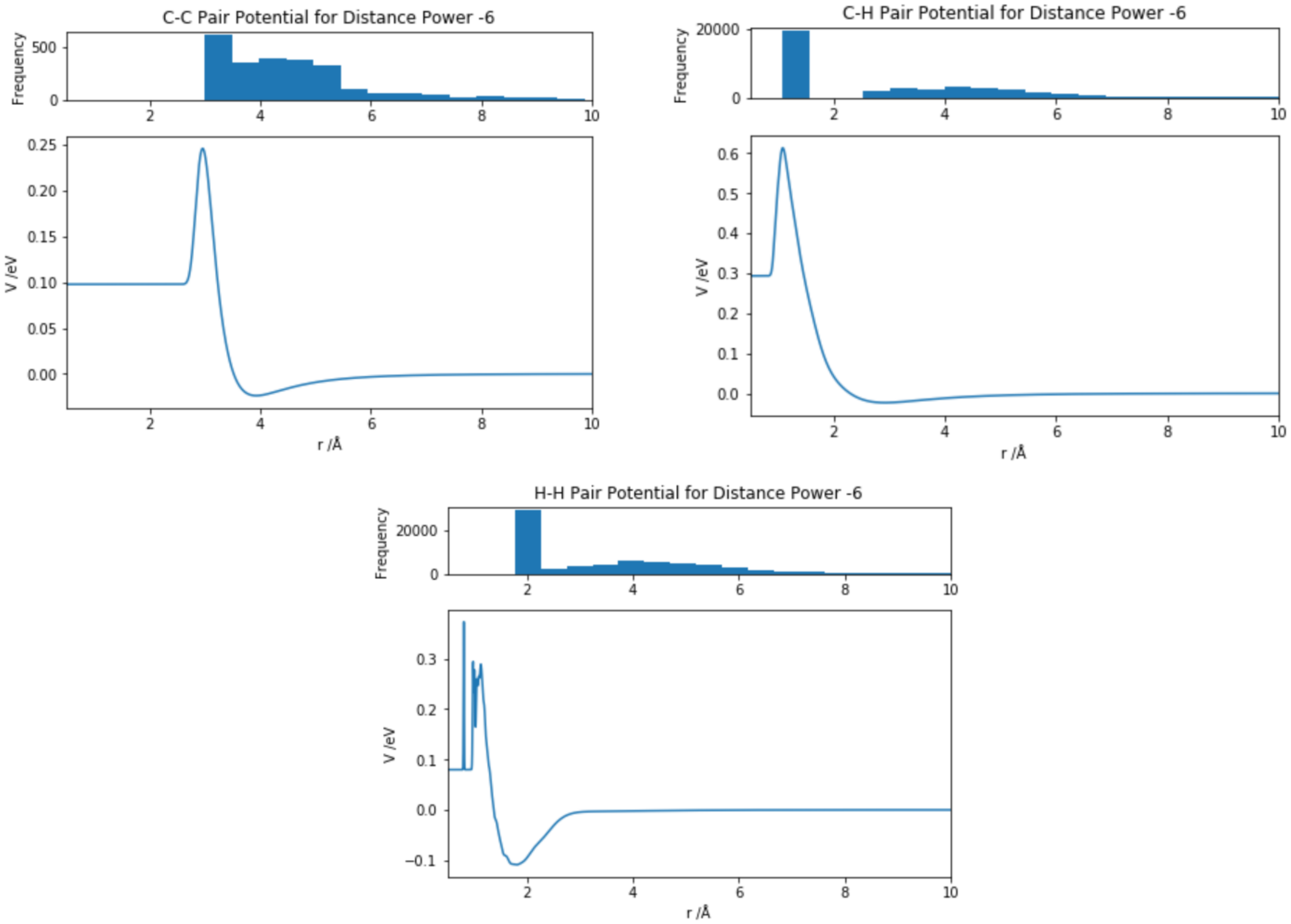


Figure 16 - Predicted Pair Potentials for C-C, C-H and H-H distances for 6th-power model

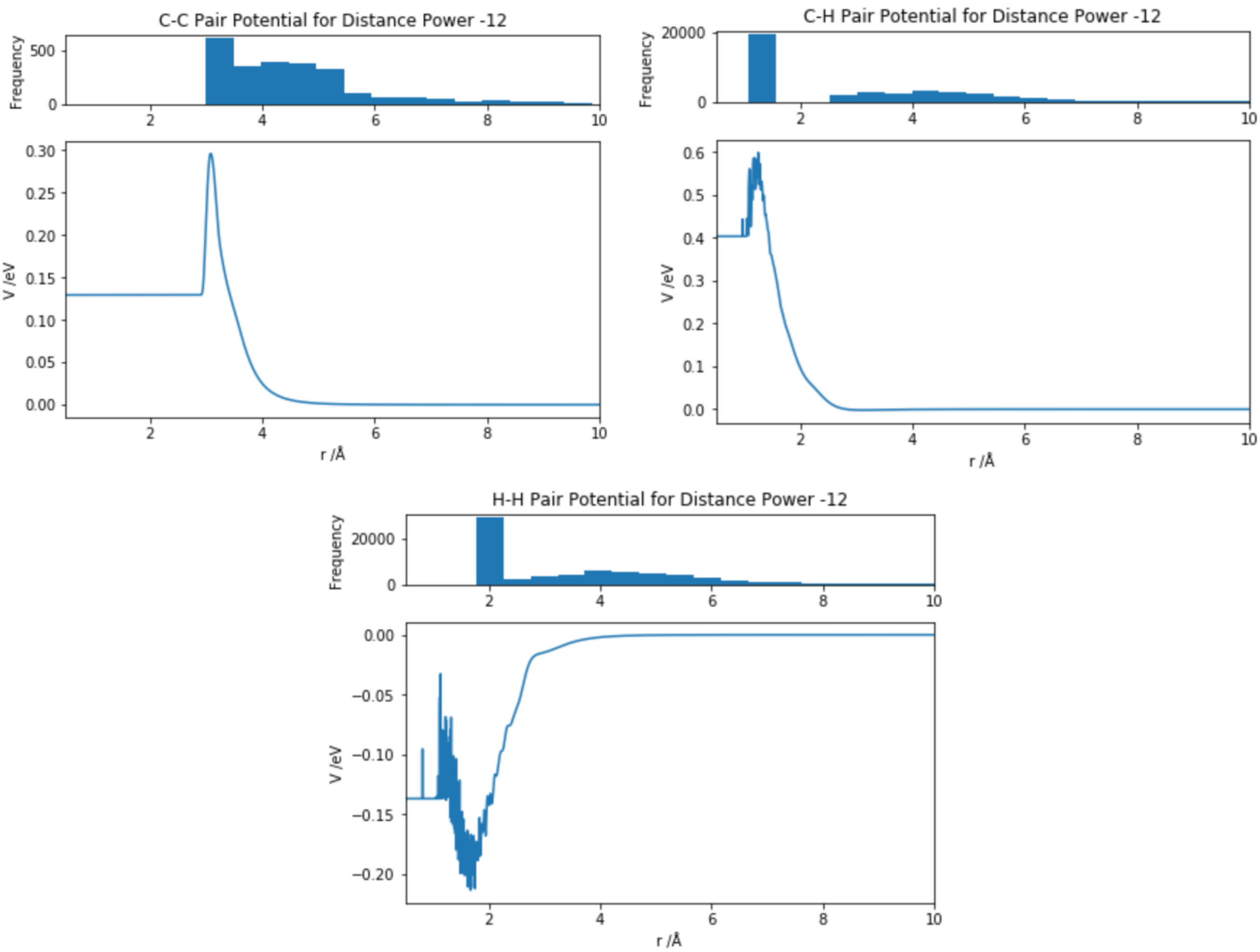


Figure 17 - Predicted Pair Potentials for C-C, C-H and H-H distances for 12th-power model

From the 6th-power potentials, all three of them show attractive and repulsive contributions to the total energy at long- and short-range respectively. On the other hand, the 12th-power potentials have only repulsive contributions to the total energy from the C-C and C-H pair potentials, while the H-H potential has both attractive and repulsive contributions. The fluctuations in the potentials for the 12th-power model can be attributed to the overfitted model, while the potentials for the 6th-power model are fairly smooth. For the 12th-power model, the only potential with a potential well was the H-H potential, with a well depth of 214 meV, a further indication of the overfitting of this model. However, for the 6th-power model, the C-C, C-H and H-H pair potentials had potential well depths of 23.6 meV, 23.3 meV and 109 meV respectively, with all the potentials holding a reasonable physical form.

Finally, the 12-6 fits were investigated. The train-train kernels, K_{NN}, from the 6th- and 12th-power models were summed to give a 12-6 train-train kernel. A logarithmic grid search hyperparameter optimisation was carried out to find the best regulariser for the 12-6 model, again using six-fold cross-validation. The results are shown in Figure 18.

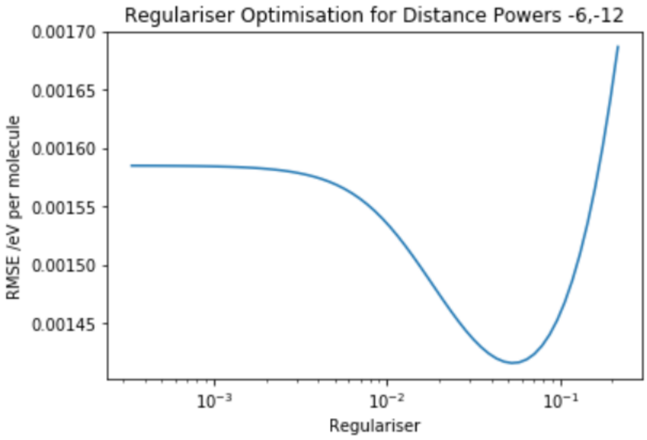


Figure 18 - Regulariser Optimisation for the 12-6 Model

The logarithmic grid search gave a minimum of 0.0523 for the regulariser in the cross-validation RMSE. Therefore, this value was chosen for the regulariser, as the curve increases significantly from a regulariser of 0.11 and this value minimised the cross-validation error, which was 1.42 meV per methane molecule.

Using the regulariser of 0.0523, kernel regression was implemented to calculate a new set of weights, which was used to calculate the fitted total energies for the 12-6 model. These energies are shown alongside their error in Figure 19.

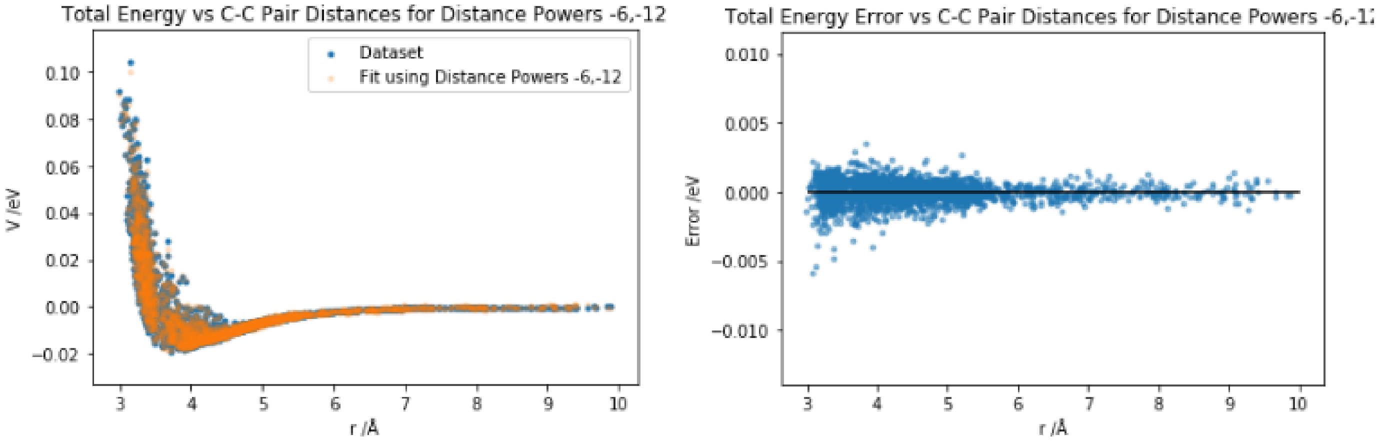


Figure 19 - Total Energy function of C-C Pair Distances using 12-6 model and its Error

From Figure 40, it can be seen that the 12-6 model has the lowest training RMSE of any model so far of 389 μeV per methane molecule, but has a cross-validation error higher than that of the 6th-power fit but lower than the 12th-power fit, making it slightly overfitted. The consequence of overfitting can be attributed to the contribution of the overfitted 12th-power fit to the 12-6 model. Furthermore, the large error displayed at short-range in the regular unscaled distances model was partially removed due to overfitting, but the bias in long-range observed in the 12th-power fit was not present, indicating another improvement on the 12th-power fit. Again, in order to improve the 12-6 model, a full optimisation of the length scale parameter in 12th-power fit ought to be carried out.

In order to create a test-train kernel for the 12-6 model, the test-train kernels from the 6th- and 12th-power models were summed.

The test-train kernels and the weights calculated from the train-train kernels were used to calculate the predicted pair potentials through kernel regression again, after which residual intramolecular energies were removed. These plots are shown in Figure 20.

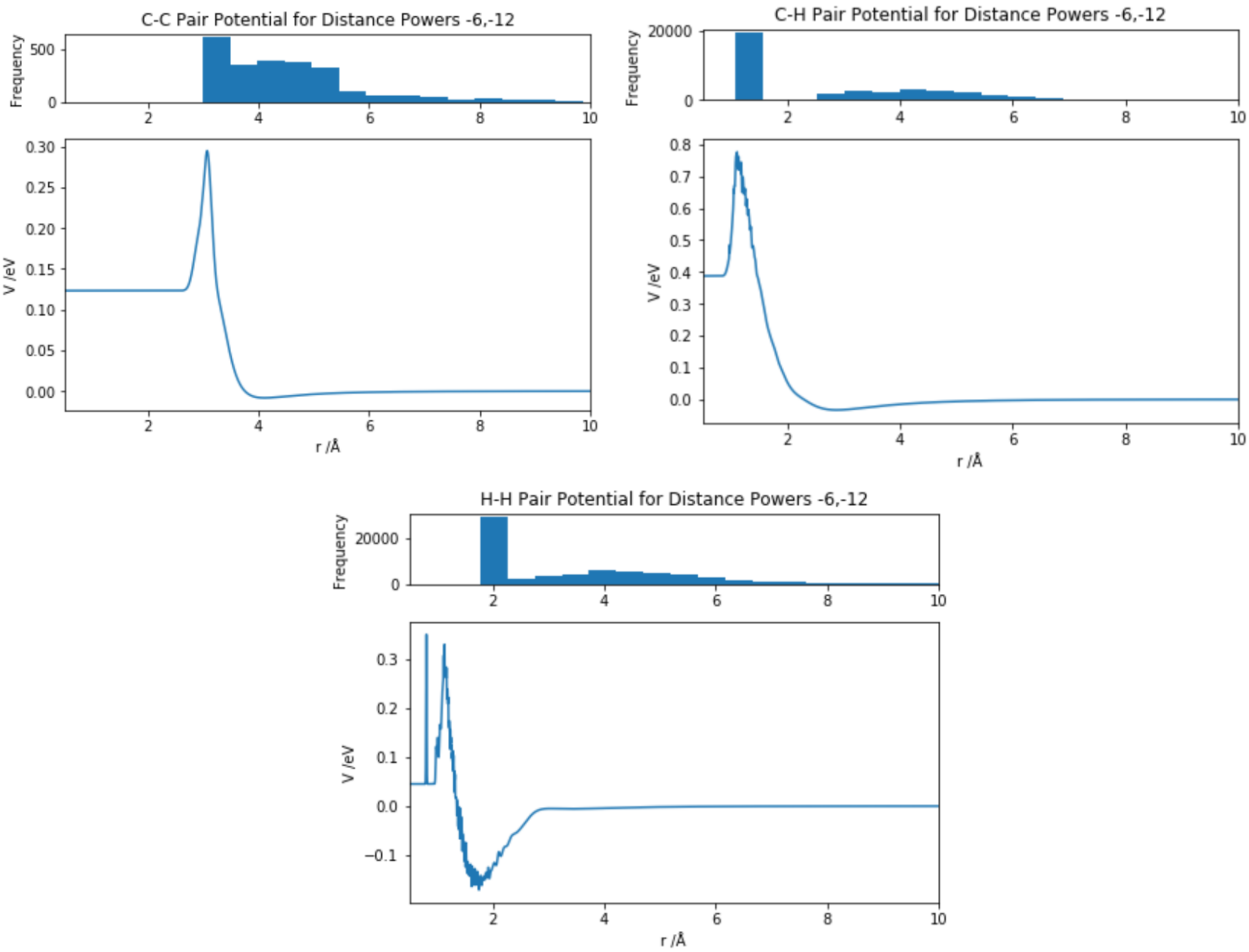


Figure 20 - Predicted Pair Potentials for C-C, C-H and H-H using 12-6 model

It can be seen that all three pair types give fairly reasonable predicted pair potentials, with all three pair potentials giving attractive and repulsive contributions to the total energy at long- and short-range respectively. All three plots show slightly high orders of magnitude for the potential wells, which could be improved with a full optimisation for the Gaussian width used to compute the kernels. For the 12-6 model, the C-C, C-H and H-H pair potentials had potential well depths of 8.17 meV, 33.1 meV and 171 meV respectively, with the potentials holding a somewhat reasonable physical form, but with some degree of overfitting.

Figure 21 shows that the fits are more likely to be reliable in ranges of distance where there is much data and not as good in ranges of distance where there is less data. There was a removal of the oscillations in the unscaled distance potentials at long-range, without using a shorter cut-off radius, showing again the benefit of distance scaling. Given that there are more different H-H interactions than C-H interactions and more C-H interactions than C-C interactions, the gradual increased level of noise from C-C to C-H to H-H pair potentials can be somewhat rationalised, but are also indications of slight overfitting in the model.

5.3 Comparison of *librascal* potentials

The pair potentials obtained from 12-6 distance scaling model were then compared with those obtained from the regular unscaled distance model.

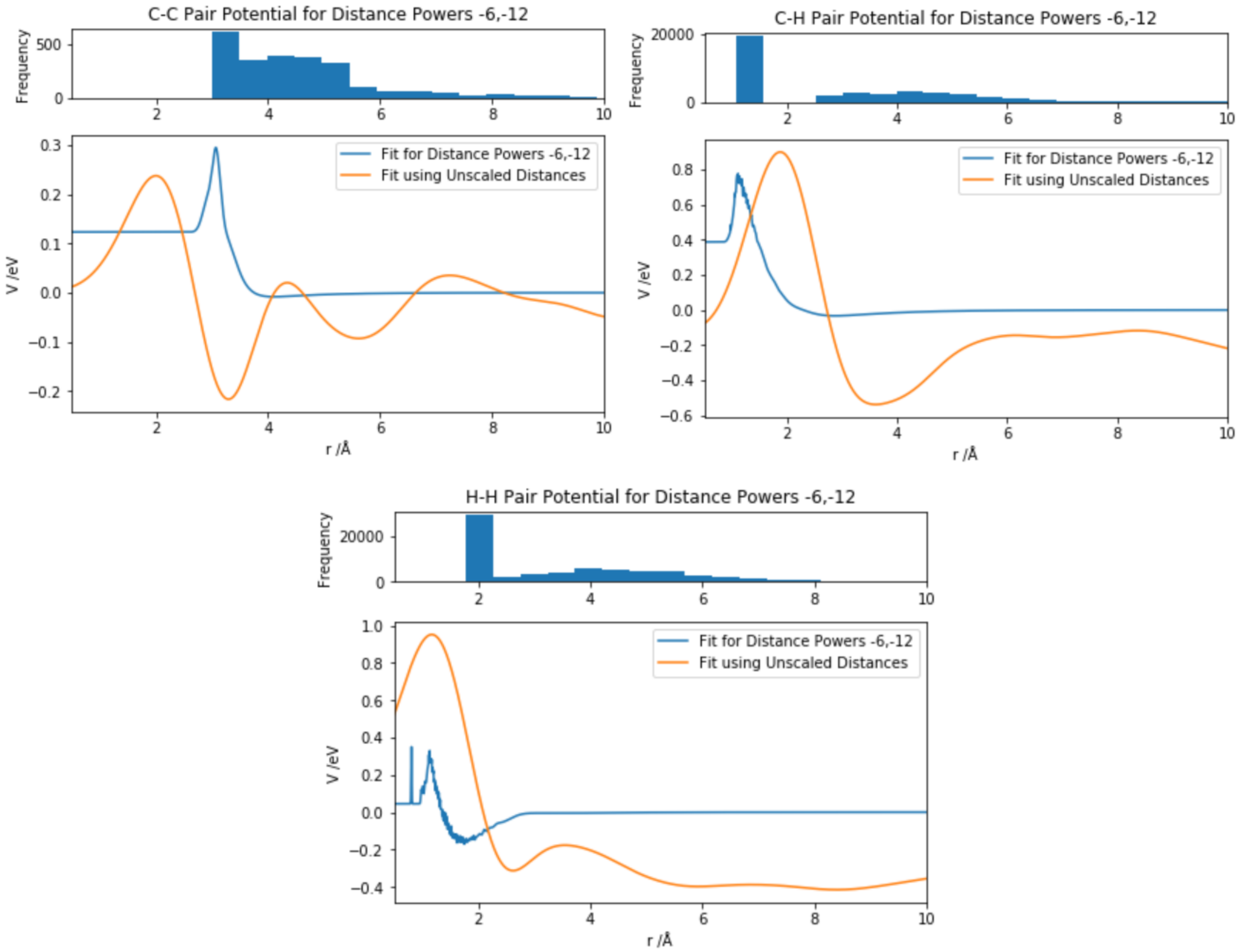


Figure 21 - Pair Potentials comparison between Unscaled Distances Model and 12-6 Model

To begin with, the C-C pair potentials are dissimilar in that the 12-6 model seems to have stronger repulsive components than the unscaled distances model, shown by the shallower nature of the repulsive part of the unscaled distances curve. They are similar in terms of the energy scale and the equilibrium distance for the 12-6 model also seems to be longer than that of the unscaled distances model, at 4.10 Å compared to 3.29 Å.

The C-H pair potentials are dissimilar, as the 12-6 model seems to be less overfitted than the unscaled distances model. The 12-6 model seems to have a smaller scale without oscillations at long-range, but some small, but large amplitude oscillations in the short-range, indicating the presence of some overfitting. The fit for unscaled distances provides a stronger attractive component at mid-range to the total energy of the system than the 12-6 model, with an equilibrium distance of 3.61 Å, while the 12-6 potential has a shorter equilibrium distance of 2.88 Å.

The H-H pair potentials are very dissimilar in terms of the scale of the function, and both provide differing contributions to the total energy of the system at long-range. The unscaled distances potential is mostly repulsive at long-range with a negative gradient, whereas the 12-6 potential is mostly attractive at long-range with a positive gradient. The fit for the unscaled distances has a much larger scale. Contrasted with the 12-6 model, which has an equilibrium distance of 1.75 Å with a shallow well, there is a sizeable difference in the range and the scale between the two fits.

Overall, it can be seen that the 12-6 model produces pair potentials which are overfitted but with a form more similar to that of the Lennard-Jones potential, with repulsive and attractive components and an equilibrium distance. Furthermore, the training RMSE is reduced further but this could link to the overfitting.

The 12-6 model was then compared with the 6th- and 12th-power fits to ascertain their contributions to the model. The three models are plotted on the same axes as shown in Figure 22.

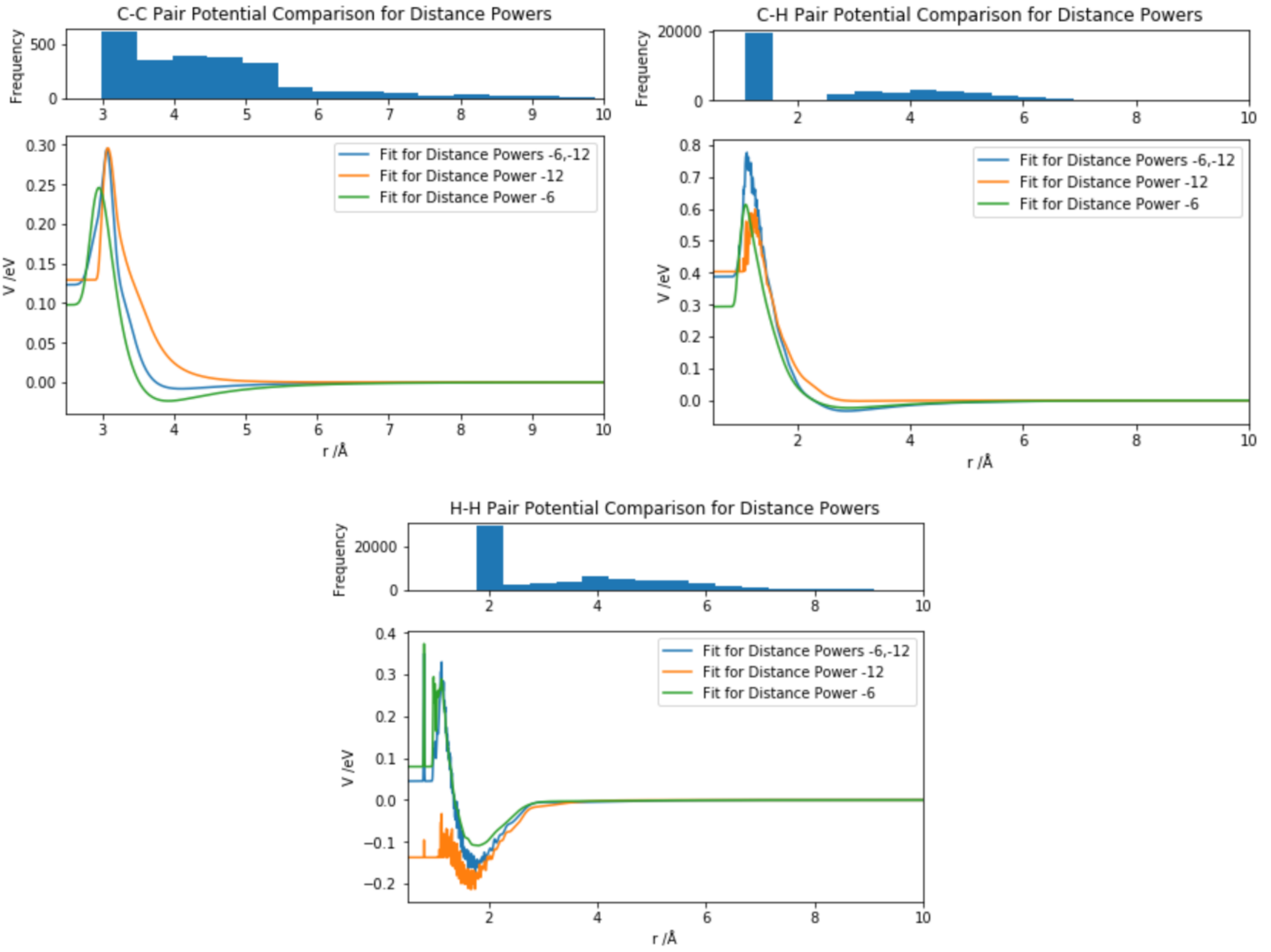


Figure 22 - Comparison between different distance power models and their pair types

For the C-C pair potentials, it can be seen that the 6th- and 12th- power models have strongly attractive and weakly repulsive contributions at long-range to their models' respective total energies, meaning that the 12-6 model, being a combination of the two, has a weakly attractive contribution at long-range to the model's total energy. The 6th-power potential has the deepest potential well at 23.6 meV, followed by the 12-6 potential with 8.17 meV.

Using the C-H pair potentials, the 6th- and 12th- power potentials have weakly attractive and repulsive contributions at long-range to their models' respective total energies, meaning that the 12-6 model has a very weakly attractive contribution at long-range to the total energy. The 12-6 potential has the deepest potential well with 33.1 meV, only just followed by the 6th-power potential with 23.3 meV.

From the H-H pair potentials, the 6th- and 12th-power potentials have strongly and weakly attractive contributions at long-range to their model's respective energies, resulting in the 12-6 model having a moderately attractive contribution at long-range to the total energy. The 12th-power potential has the deepest potential well with 214 meV, followed by the 12-6 potential with 171 meV and then the 6th-power potential with 109 meV. Almost all functions are noisy at short-range, which indicate some degree of overfitting.

The 12th-power model also produces more overfitted potentials than the 12-6 model, with a high cross-validation error and potentials that are noisy in the short-range. The 6th-power model, however, produces the best fit, with the lowest cross-validation error and with forms which are very similar to the Lennard-Jones potential with a strong 6th-power decay in the long-range of the attractive tail.

Further models were created through summing the kernels with distance powers -6 and -12 in ratios of 2:1, to give alternative 12-6 model kernels. These kernels were solved to give fitted total energies and pair potentials, but no noticeable difference was observed in the error of the total energies or shape of the pair potentials.

5.4 Gaussian Approximation Potential Model using QUIP

It was decided then to compare a GAP model using *QUIP* on the same methane dimer set of 2418 dimers, to generate the pair potentials with a shorter cut-off radius and compare the results with those generated from *librascal*.

The C-C distances calculated had a cut-off radius of 10 Å. The C-H and the H-H distances calculated each had a cut-off radius of 6 Å. The Gaussian width used to compute the kernel was 0.8 Å as used previously and the regulariser used in the kernel regression was 0.03, the same regulariser chosen from the first cross-validation for the unscaled distances. The fitted total potential energy surface was calculated and their cross-section along the C-C pair distance axis shown in Figure 23, alongside its error.

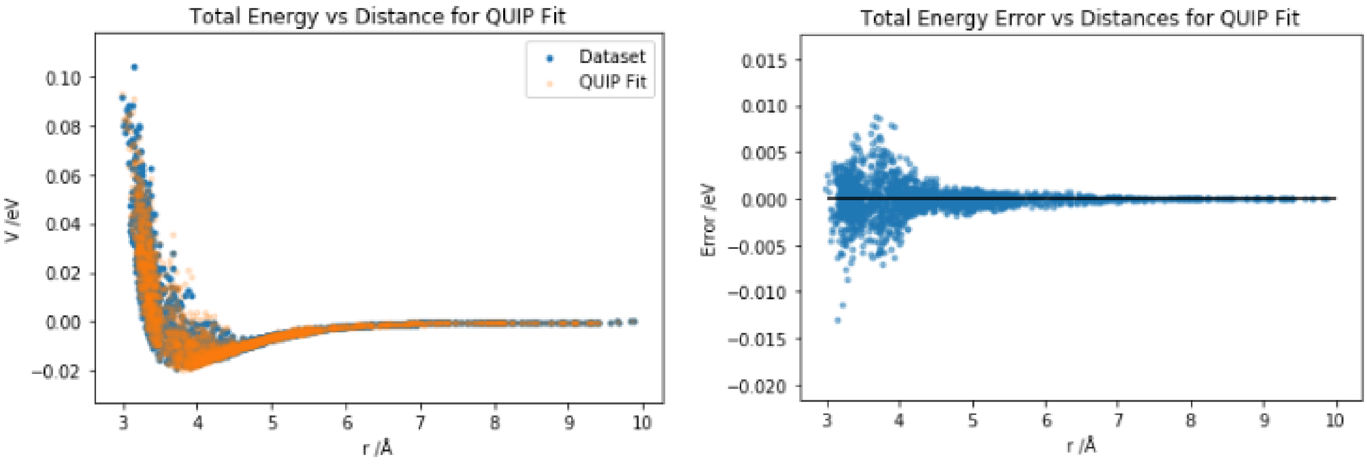


Figure 23 - Total Energy function of C-C Pair Distances using GAP and its Error

It can be seen that *QUIP* produces a good fit for the total energies, having a training RMSE of 816 µeV per methane molecule, which is understandably smaller than the cross-validation error of 1.15 meV per methane molecule. From comparing the training and cross-validation errors, it can be concluded that this fit produced by *QUIP* is comparable in quality to the 6th-power fit, with similar errors, and better than the 12th-power and 12-6 fits, which are more overfitted. Both the 6th-power fit and the fit produced by *QUIP* have smaller errors at short-range and have no oscillations in the long-range compared to the unscaled distances fit in *librascal*.

Using the test dataset, which was used to generate the pair potentials in *librascal*, the pair potentials were created using *QUIP* and are compared to the 12-6 model in *librascal*.

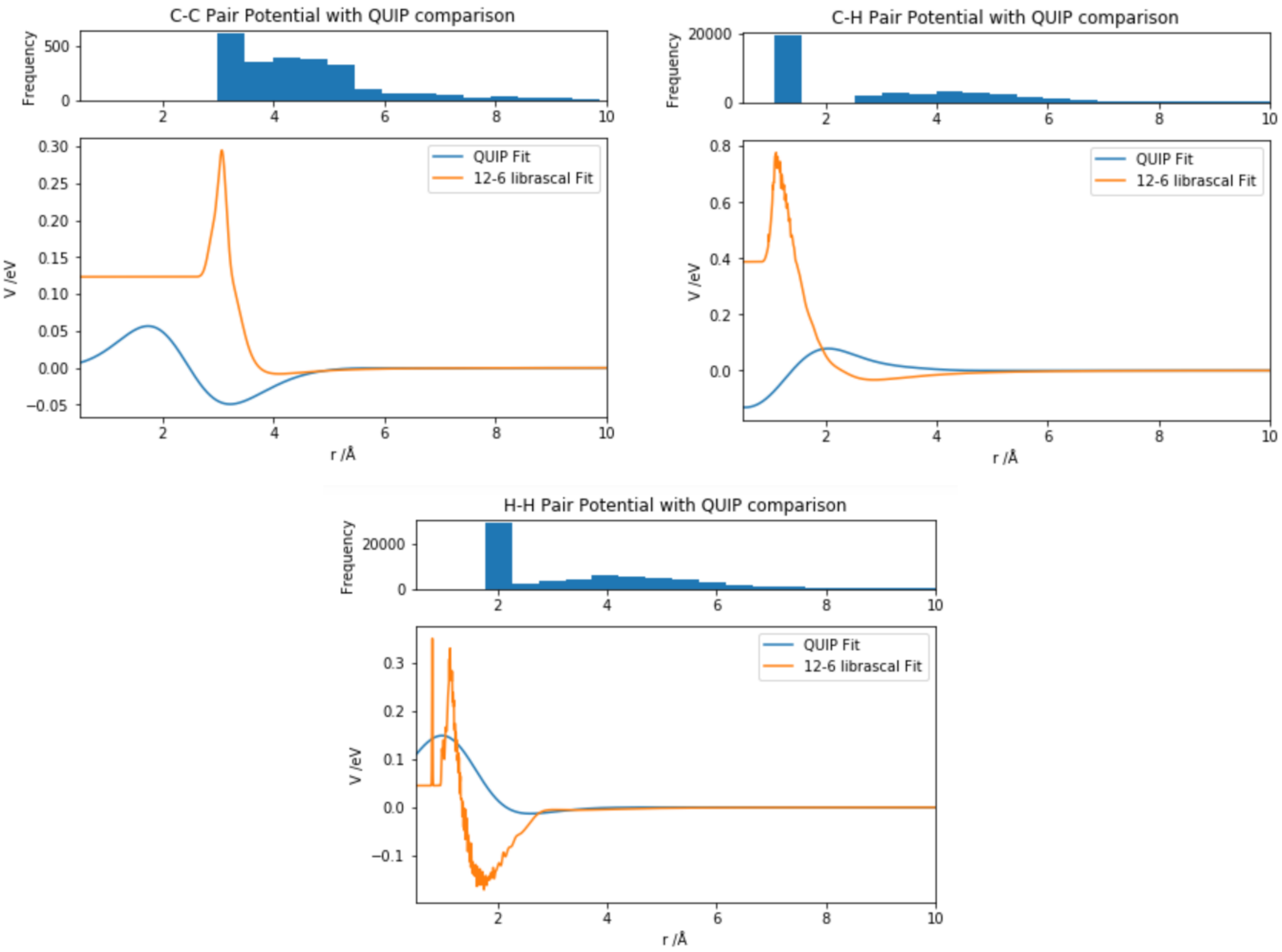


Figure 24 - C-C, C-H and H-H Pair Potentials comparison with QUIP Model

The C-C pair potential can be seen to have an attractive contribution at long-range to the total energy of the system given by the potential created using *QUIP*. This is in agreement with the shape of the C-C pair potential produced by the 12-6 model but is more attractive at long-range and less repulsive at short-range than the 12-6 model.

The C-H pair potential can be seen to have a repulsive contribution at long-range to the total energy, which is in slight disagreement with the 12-6 model created with *librascal*, as that has a weakly attractive contribution at long-range. In this case, the 12-6 potential provides a better physical form than the potential created using *QUIP*.

The H-H pair potential created using *QUIP* has an attractive contribution at long-range to the total energy, which also agrees with the shape given by the 12-6 model but has smaller attractive and repulsive contributions to the total energy than the 12-6 potential at long- and short-range respectively.

All three 12-6 potentials produced using *librascal*, have larger energy scales, indicating again a degree of overfitting, something which could be improved upon with a full optimisation of the length scale parameter.

All three potentials produced using *QUIP* do not oscillate at long-range like the unscaled distance potentials produced using *librascal*. However, the lack of oscillations at long-range in the potentials created by *QUIP*, is due to the lack of data fitted at long-range, rather than the quality of the fit itself. Therefore, the benefit of using a shorter cut-off radius, in order to produce potentials with more reasonable physical forms, can be seen evidently.

Overall, from this comparison, the GAP potentials produced by *QUIP* are somewhat similar to the 12-6 potentials produced by *librascal*, as the C-C and H-H pair potentials give a similar type of contribution to the total energy of the system. However, the most salient point learnt from the *QUIP*-produced potentials, is that the intermolecular potentials of methane dimers can have more reasonable physical forms, through utilising shorter cut-off radii.

Furthermore, in a study carried out using Gaussian Approximation Potentials in *QUIP* to model pair potentials in alkanes, the best model for methane dimers using unscaled distances was found to be 381 μeV per methane molecule[37], showing that the use of *librascal* in this case has given an equivalent result regarding training error with a similar RMSE for the 12-6 model, but could improve even more if a full optimisation of the length scale parameter is carried out.

5.5 Classical Potential Model

A classical potential of the methane dimers was computed using a COMPASS forcefield on pre-existing geometries. It was expected that increased flexibility of the GAP using *librascal* would lead to a much better fit, due to there being many (2418 in this case) parameters in the GAP fit, but just six (two per pair type) in the classical potential. The Non-Bonded Lennard-Jones 12-6 potential is known to be strongly repulsive in the short range. Therefore, the total energies were computed using the COMPASS forcefield and Non-Bonded Lennard-Jones 9-6 (LJ-9-6) potential, due to its softer nature. The Lennard-Jones 9-6 potential is given in Equation 18[50].

$$U_{LJ(9-6)}(r) \;=\; \epsilon\left(2\left(\frac{r_0}{r}\right)^9 - \;3\left(\frac{r_0}{r}\right)^6\right) \tag{18}$$

It can be seen from the functional form of the LJ-9-6 potential that more emphasis is given to the attractive part of the potential. The cross-sections of the total potential energy surfaces are shown in Figure 25, alongside their error.

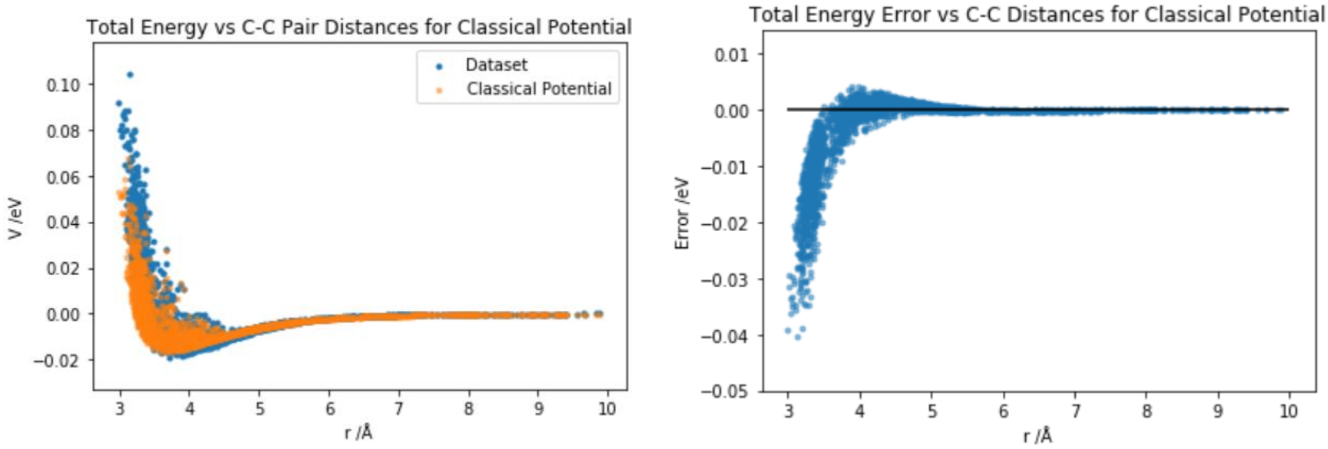


Figure 25 - Total Energy function of C-C Pair Distances for Classical Potential and its Error

From the total energy plot, it can be seen that the simulation provided a reasonable match to the dataset, but it seemed to hugely underestimate the quantum mechanical energies at short distances. It also significantly underestimates the spread of the energies around the mean total energy along the C-C distance axis, as well as placing the equilibrium distance shorter than the quantum mechanical data. This led to a larger RMSE of 4.31 meV per methane molecule, which is of an order of magnitude 10x larger than the training error calculated for the 12-6 model using *librascal*. The large difference in error can be attributed to the much greater number of parameters used in the Gaussian Approximation Potential compared to the Classical Potential.

The COMPASS parameters used were r_0=3.854 Å and ϵ=0.062 kcal/mol for the C-C, r_0=3.526 Å and ϵ=0.027 kcal/mol for the C-H and r_0=2.878 Å and ϵ=0.023 kcal/mol for the H-H pair potentials. The parameters for like atom pairs were given explicitly, but the parameters for unlike atom pairs were calculated using the 6th Order Combination Law shown in Equations 19 and 20[46].

$$r_{ij}{}^{o} = \left(\frac{(r_i{}^{o})^6+(r_j{}^{o})^6}{2}\right)^{\frac{1}{6}} \tag{19}$$

$$\epsilon_{ij} = 2\sqrt{\epsilon_i\epsilon_i}\left(\frac{(r_i{}^{o})^3(r_j{}^{o})^3}{(r_i{}^{o})^6+(r_j{}^{o})^6}\right) \tag{20}$$

All three pair potentials are shown in Figure 26.

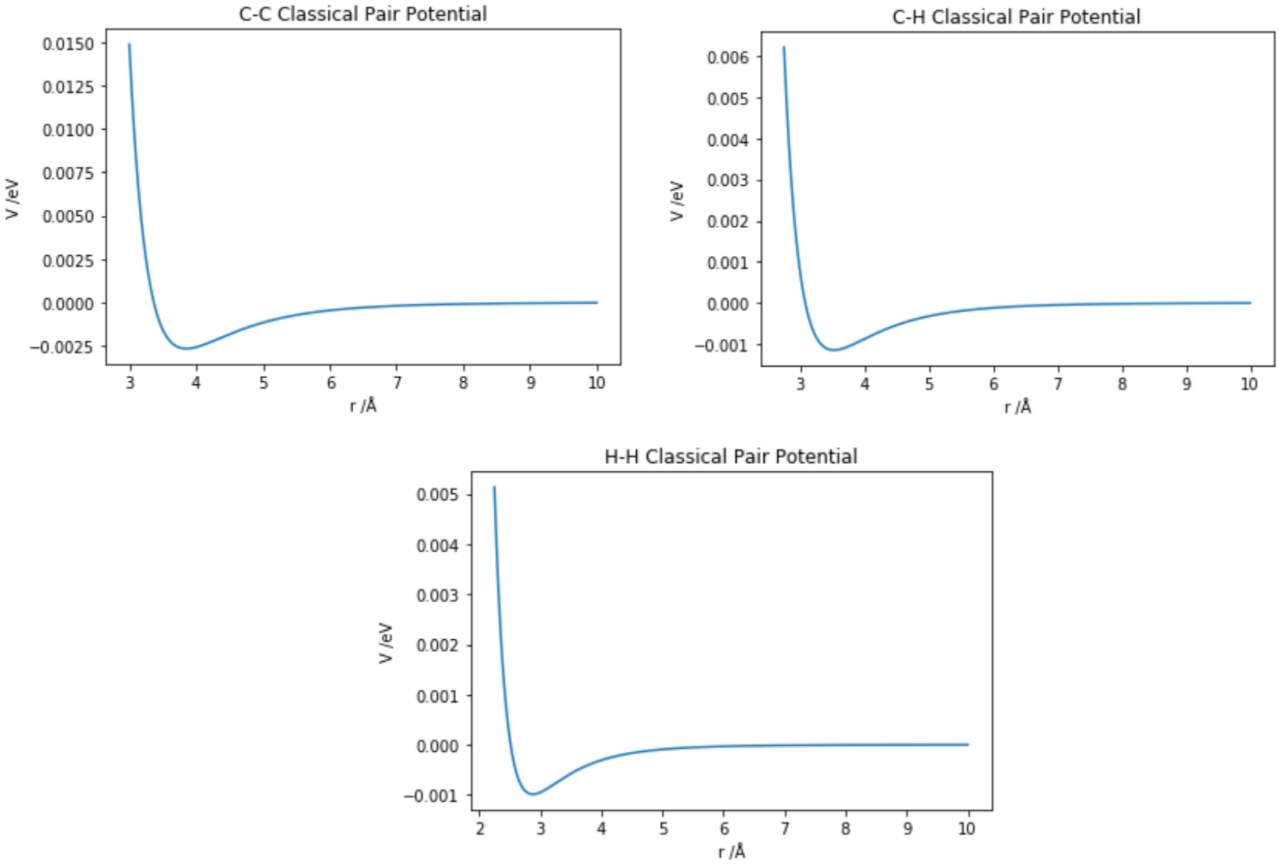


Figure 26 - C-C, C-H and H-H Pair Potentials using Non-Bonded LJ-9-6 potential

All three potentials are seen to have attractive contributions to the total energy at long-range as expected, since they are all calculated independently and since the LJ-9-6 potential places more emphasis on a smoother attractive component. Furthermore, in comparison, the C-H pair potential in the 12-6 model had a repulsive contribution at long-range to the total energy, whereas the same potential in the LJ-9-6 model has an attractive component.

The LJ-9-6 potential has a larger error than the 12-6 model computed using *librascal* and there is a difference in the type of contribution from the C-H pair potential to the total energy between the LJ-9-6 model and the 12-6 model computed using *librascal*. The energy scales for the classical potential are also more realistic than those given by the Gaussian Approximation Potentials from *QUIP* and *librascal*, due to the lack of a full length-scale parameter optimisation.

6 Conclusions

In summary, this project has shown the usefulness of producing more accurate pair potentials from fits generated from scaled distances. The distances are scaled by a known scaling power of the governing interaction. In the case of methane dimers, the dominant intermolecular force is dispersion and is known to scale in an r^{-6} fashion from the quantum mechanics and therefore being the basis for the attractive tail of the Lennard-Jones potential.

Code development was completed in the *librascal* software, where a new representation was created to compute the pair distances within each structure in a dataset and be able to scale them by a given power. Furthermore, the capacity to compute Gaussian kernels was added, to allow the user to create Gaussian Approximation Potentials through Gaussian Process Regression.

Via a process involving the pair distance histograms and the dataset energies, an estimate for an improved length-scale parameter was found to be 0.8 Å. Using six-fold cross-validation to minimise overfitting, a full regulariser optimisation was carried out and this gave a regulariser of 0.03. This regulariser had cross-validation and training errors of 1.15 and 1.02 meV per methane molecule respectively. There was found to be a larger error at short- than long-range due to the energy variance.

Interatomic pair potentials were computed, using the improved hyperparameters. All three potentials provided attractive contributions to the total energy but seemed to oscillate at long-range. Furthermore, the energy scale of the potentials was slightly too high and could have been resolved with a full optimisation of both the length scale parameter and the regulariser. However, such an optimisation was not feasible within the scope of this project, due to the computational cost associated.

Following from this, there is already good evidence to suggest that rescaling the distances by powers associated with known scaling laws of the dominant interaction, would give better fits[28]. Therefore, rescaling by powers of -12 and -6 was investigated in this relatively simple, controlled test system of methane dimers. The power of -6 was chosen from the result of quantum mechanics for the phenomenon of dispersion and the power of -12 was chosen from the Lennard-Jones (12-6) potential. The expectation was that the pair potentials would give more Lennard-Jones like fits, due to the dominant interaction between methane dimers being dispersion.

The same pair distances representation was used, and the distances were scaled as mentioned. Through a process involving both the scaled distance histograms and the dataset energies, it was found that Gaussian widths of 5×10^{-4} Å^{-6} and 2.5×10^{-7} Å^{-12} would give improved train-train kernels for the 6th- and 12th-power models respectively. Six-fold cross-validations were carried to optimise the regulariser and this provided regulariser values of 0.07 and 0.06 for the 6th- and 12th-power fits respectively. For these regulariser values, the cross-validation errors were 1.07 and 2.41 meV for the 6th- and 12th-power fits respectively, showing that the 6th-power fit was slightly more accurate than the unscaled distances fit and that the 12th-power fit was overfitted.

Gaussian Process Regression was then completed to find the training weights and thus predicted total energies using the same regulariser as for the unscaled distances. They gave training errors of 863 and 499 μeV per methane molecule for the 6th- and 12th-power fits respectively, again suggesting that the 12th-power fit was likely to be overfitted, while the 6th-power fit was more accurate.

The interatomic pair potentials were computed for the scaled distances. These were more Lennard-Jones-like, with a reduced energy scale and with better physical forms, decaying to zero over long-range, once intramolecular energies were removed. For the 6th-power model, all three potentials gave an attractive contribution to the total energy at long-range. For the 12th-power model, the C-C and C-H potentials gave repulsive contributions to the total energy at long-range, while the H-H potential gave an attractive contribution. The large amount of noise seen in the 12th-power potentials is another indication of overfitting at short-range. This overfitting would be removed with a full optimisation of the length-scale parameter. Furthermore, the 6th-power potentials gave more physical tail forms and so would be a good option for fitting the long-range tail of intermolecular interactions, with the short-range potentially taken care of by another model, such as a SOAP-GAP.

A 12-6 model was generated by combining the kernels used to compute a fit for the total potential energy surface. From a six-fold cross-validation, the optimal regulariser for this fit was found to be 0.0523. This fit had cross-validation and training errors of 1.42 meV and 389 μeV, being the model with the lowest training error, but the fairly large cross-validation error was an indicator of overfitting coming from the 12th-power model.

The interatomic pair potentials were produced in the same fashion as before. When compared to the fit for the unscaled distances, both models produced different potentials. The 12-6 C-C potential was more repulsive at short-range than the potential produced from the unscaled distances. The 12-6 C-H potential had a weaker attractive contribution at mid-range to the total energy than the unscaled distances potential. The H-H potentials were very different, with both provide differing contributions to the total energy of the system at long-range. The unscaled distances potential was mostly repulsive at long-range, whereas the 12-6 potential was mostly attractive at long-range. The energy scaling had improved to become more realistic, and the distance scaling had removed the oscillations at long-range, without requiring a shorter cut-off radius. In addition, the large error at short range was somewhat removed due to the distance scaling, resulting in a much smaller overall error, but a reasonable amount was removed due to overfitting.

The 12-6 model was then compared with a GAP produced by *QUIP* with shorter cut-off radii. The 12-6 model produced a fit with a lower error, but this is likely to have been a result of slight overfitting in the 12-6 model. Both models agreed with regard to the type of contribution each potential gave to the total energy at long-range for C-C and H-H potentials, with slight differences in the energy scale and the equilibrium distances. However, the C-H potential for the QUIP-produced model gave a repulsive contribution to the total energy at long-range, whereas the 12-6 model gave an attractive contribution at long-range.

The 12-6 model was compared with a classical potential produced using *LAMMPS*. The 12-6 model produced a better fit with much greater accuracy, due to the greater flexibility of the potential, with a much higher number of parameters involved. Both the classical and 12-6 potentials provided attractive contributions to the total energy at long-range. Furthermore, the energy scale was much smaller than the GAPs, suggesting this was the true scale of the pair potential for a methane dimer.

Finally, it was found overall that the 6th-power model was the best of those calculated, as it had the lowest cross-validation error. Furthermore, the 6th-power potentials gave more physical tail forms and so would be a good option for fitting the long-range tail of intermolecular interactions. Fitting to scaled distances is therefore a promising and inexpensive strategy for modelling long-range interactions in systems dominated either by dispersion or other interactions by changing the scaling power.

7 Future Work

Further work to this project would include a full optimisation for the length-scale parameter used in computing all train-train kernels to prevent overfitting as much as possible and give a more realistic energy scale and physical tail forms for the interatomic pair potentials. It has also been seen through creating potentials with *QUIP*, that using shorter cut-off radii can lead to more physical tail forms in the potential. More development of the *librascal* code would be required in this respect, adding the ability to have different cut-off radii for different pair types. Furthermore, the ability to compute pair distances with a smooth width given by a specific cut-off function type would lead to smoother potentials and this could be developed in *librascal*. The smooth width would remove any potential issues with a sharp discontinuous cut-off function where energies and forces calculated would suddenly go to zero. Instead, the cut-off function would smoothly and continuously go to zero over the smooth width.

A GAP model using *librascal* for a water dimer dataset could also attempt to be computed. This would be done first using the regular unscaled distances and then by scaling the distances by a power of -3, given that water is a polar molecule and that the dipole-dipole interaction is governed by a $1/r^3$ relationship with respect to the distance, via the multipole expansion. This would give interatomic O-O, O-H and H-H pair potentials, where the governing intermolecular interaction would be hydrogen bonding. This would provide different results to the methane dimer, where the dominant intermolecular interaction is the Van der Waals force.

In addition to this, the water molecular system can be scaled up to low-dimensional molecular clusters in bulk to see how well it performs, using a GAP model created using *librascal*. The same interatomic pair potentials can be computed, potentially with a different scaling power and compared to the potentials given by water dimers.

In summary, *librascal* can be used to compute pair distances and Gaussian kernels for any molecular or bulk system, using any distance scaling power, provided the intermolecular interactions for that system are known to be governed by a specific distance power law. In addition to *QUIP*, through adding pair distances to *librascal*, a greater range of use cases will be covered and GAP fits of greater accuracy will be able to be computed, due to the ability to fit to scaled distances.

8 Acknowledgements

I would like to thank Prof. Michele Ceriotti, as my supervisor at the *Computational Science and Modelling* group *(COSMO)* at *École Polytechnique Fédérale de Lausanne (EPFL),* Switzerland for his support and advice throughout the year in enabling me to carry out my research project abroad and welcoming me to his group.

I would also like to express my gratitude to Dr. Max Veit for all of his advice and guidance throughout the year. His support enabled me to carry out a project including both code development and research into Gaussian Approximation Potentials at the *COSMO* group.

Finally, I would like to thank Prof. Sophia Yaliraki, as my home supervisor at *Imperial College London* for her support and advice throughout the year.

9 References

1 R. LeSar, *Introduction to Computational Materials Science modelling*, 2013, DOI:10.1017/CBO9781139033398.

2 D. W. Brenner, *The art and science of an analytic potential, Phys. Status Solidi Basic Res.*, 2000, **217**, p. 23–40.

3 S. B. Sinnott and D. W. Brenner, *Three decades of many-body potentials in materials research, MRS Bull.*, 2012, **37**, p. 469–473.

4 K. Nordlund, N. Runeberg and D. Sundholm, *Repulsive interatomic potentials calculated using Hartree-Fock and density-functional theory methods, Nucl. Instruments Methods Phys. Res. Sect. B Beam Interact. with Mater. Atoms*, 1997, **132**, p. 45–54.

5 F. H. Stillinger and T. A. Weber, *Computer simulation of local order in condensed phases of silicon, Phys. Rev. B*, 1985, **31**, p. 5262–5271.

6 M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, 2002, p. 21–41.

7 M. W. Finnis, *Bond-order potentials through the ages*, *Prog. Mater. Sci.*, 2007, **52**, p. 133–153.

8 I. H. Umirzakov, *Van der Waals equation of state and PVT-properties of real fluid*.

9 P. N. Patrone and A. Dienstfrey, *Uncertainty Quantification for Molecular Dynamics,* 2018, p. 115–169.

10 D. Frenkel and B. Smit, *Understanding Molecular Simulation, Computational Science Series, Vol 1,* 2001, p. 664.

11 T. C. Lim, *The Relationship between Lennard-Jones (12-6) and Morse Potential Functions, Zeitschrift fur Naturforsch. - Sect. A J. Phys. Sci.*, 2003, **58**, p. 615–617.

12 L. Zhigilei, *Introduction to interatomic potentials (I),* 2013, I.

13 N. Juslin, P. Erhart, P. Träskelin, J. Nord, K. O. E. Henriksson, K. Nordlund, E. Salonen and K. Albe, *Analytical interatomic potential for modelling non-equilibrium processes in the W-C-H system, J. Appl. Phys.*, 2005, **98**, DOI:10.1063/1.2149492.

14 R. L. Rowley, Y. Yang and T. A. Pakkanen, *Determination of an ethane intermolecular potential model for use in molecular simulations from ab initio calculations, J. Chem. Phys.*, 2001, **114**, p. 6058–6067.

15 J. Behler, *Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys.* , 2012, DOI:10.1063/1.3553717.

16 A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles and G. J. Tucker, *Spectral neighbour analysis method for automated generation of quantum-accurate interatomic potentials, J. Comput. Phys.*, 2015, **285**, p. 316–330.

17 R. Drautz, *Atomic cluster expansion for accurate and transferable interatomic potentials, Phys. Rev. B,* 2019, **014104**, p. 1–15.

18 A. P. Bartók, M. J. Gillan and F. R. Manby, *Machine-Learning approach for one- and two-body corrections to density functional theory : Applications to molecular and condensed water* 2013, **054104**, p. 1–12.

19 A. J. Stone, *The Theory of Intermolecular Forces*, *2^nd^ edition*.

20 S. H. Walmsley, *Lattice vibrations and elastic constants of molecular crystals in the pair potential approximation, J. Chem. Phys.*, 1968, **48**, p. 1438–1444.

21 C. Qu, Q. Yu and J. M. Bowman, *Permutationally Invariant Potential Energy Surfaces, Ann. Rev. Phys. Chem.,* 2018.

22 A. Seko and I. Tanaka, *Descriptors for Machine Learning,* Ch. 1.

23 T. A. Johansen, *On Tikhonov Regularisation, Bias and Variance in Nonlinear System Identification, Automatica*, 1997, **33**, p. 441–446.

24 A. P. Bartók and G. Csányi, *Gaussian Approximation Potentials: A Brief Tutorial Introduction* 2015, p. 1051–1057.

25 M. Jaggi and R. Urbanke, *Kernel Ridge Regression and the Kernel Trick*, *Machine Learning*, October 2019.

26 C. E. Rasmussen, C. K. I. Williams, G. Processes, M. I. T. Press and M. I. Jordan, *Gaussian Processes for Machine Learning*, 2006.

27 T. Stecher, N. Bernstein and G. Csányi, *Free energy surface reconstruction from umbrella examples using Gaussian process regression, J. Chem. Theory Comput.*, 2014, **10**, p. 4079–4097.

28 K. Hansen, F. Biegler, S. Fazli, M. Rupp, M. Sche, O. A. Von Lilienfeld, A. Tkatchenko and K. Mu, *Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies*, 2013, DOI:10.1021/ct400195d.

29 J. Bergstra, J. B. Ca and Y. B. Ca, *Random Search for Hyper-Parameter Optimization*, *J. Machine Learning* Research, 2012, vol. 13, p. 281-305.

30 Hyperparameter Tuning in Python | Towards Data Science, https://towardsdatascience.com/hyperparameter-tuning-c5619e7e6624, (accessed 15 February 2020).

31 D. J. Leinweber, *Stupid Data Miner Tricks, J. Invest.*, 2007, **16**, p. 15–22.

32 K. A. Ross, C. S. Jensen, R. Snodgrass et al., *Cross-Validation,* in *Encyclopedia of Database Systems*, Springer US, 2009, p. 532–538.

33 A. P. Bartók, *On representing chemical environments,* 2013, **184115**, p. 1–16.

34 C. Zhang and Q. Sun, *Gaussian Appoximation Potential for studying the thermal conductivity of silicene, J. Appl. Phys.*, 2019, **126**, DOI:10.1063/1.5119281.

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

36 V. L. Deringer, *Machine Learning based interatomic potential for amorphous carbon,* 2017, **094203**, p. 1–15.

37 M. Veit, S. K. Jain, S. Bonakala, I. Rudra, D. Hohl and G. Csányi, *Equation of State of Fluid Methane from First Principles using Machine Learning Potentials, J. Chem. Theory Comput.*, 2019, **15**, p. 2574–2586.

38 M. D. Veit, *Designing a machine learning potential for molecular simulation of liquid alkanes*, DOI:10.17863/CAM.37522.

39 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Gaussian approximation potentials: The accuracy of quantum mechanics without the electrons, Phys. Rev. Lett.*, 2010, **104**, p. 1–4.

40 F. London, *Zur Theorie und Systematik der Molekularkräfte, Zeitschrift für Phys.*, 1930, **63**, p. 245–279.

41 A. Tkatchenko and M. Scheffler, *Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data*, 2008, DOI:10.1103/PhysRevLett.102.073005.

42 A. Tkatchenko, R. A. Distasio, R. Car and M. Scheffler, *Accurate and Efficient Method for Many-Body van der Waals Interactions*, DOI:10.1103/PhysRevLett.108.236402.

43 F. A. Faber, A. S. Christensen, B. Huang and O. A. Von Lilienfeld, *Alchemical and structural distribution based representation for universal quantum machine learning, J. Chem. Phys.*, 2018, **148**, p. 241717.

44 GitHub - libAtoms/QUIP: libAtoms/QUIP molecular dynamics framework: http://www.libatoms.org, https://github.com/libAtoms/QUIP, (accessed 28 March 2020).

45 GitHub - cosmo-epfl/librascal: A scalable and versatile library to generate representations for atomic-scale learning, https://github.com/cosmo-epfl/librascal, (accessed 28 March 2020).

46 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*, 1998, **102**, p. 7338–7364.

47 F. Musil, *Practical Introduction to Rascal: Build a simple machine learning interatomic potential*, February 2020.

48 1.7. Gaussian Processes — scikit-learn 0.22.2 documentation, https://scikit-learn.org/stable/modules/gaussian_process.html, (accessed 2 May 2020).

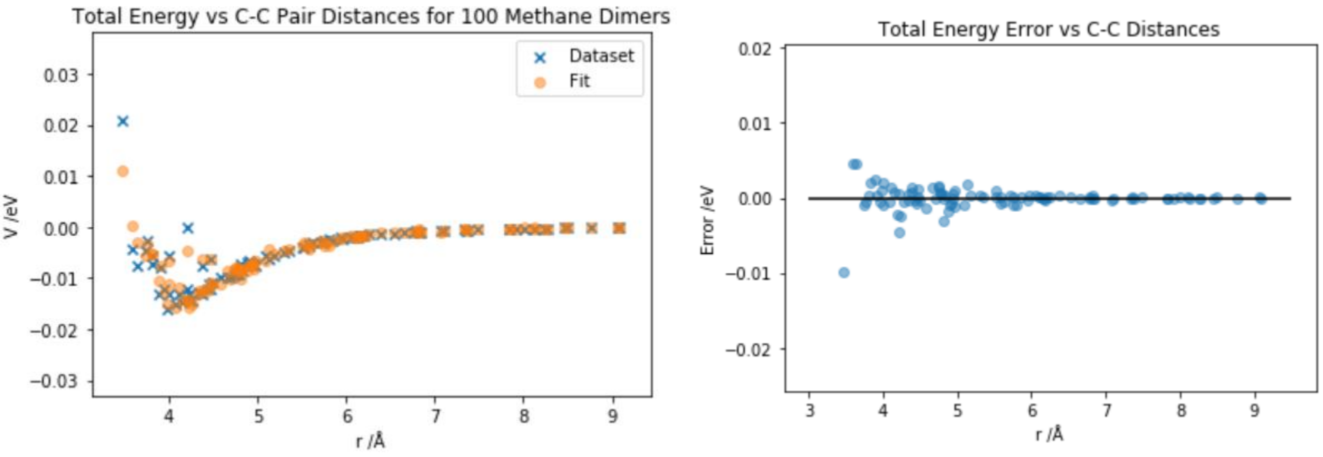
49 L. S. Bartell, K. Kuchitsu and R. J. DeNeui, *Equilibrium bond lengths in methane and deuteromethane as determined by electron diffraction and spectroscopic methods, J. Chem. Phys.*, 1960, **33**, p. 1254–1255.

50 T. C. Lim, *Alignment of buckingham parameters to generalized Lennard-Jones potential functions, Zeitschrift fur Naturforsch. - Sect. A J. Phys. Sci.*, 2009, **64**, p. 200–204.

10 Appendix

10.1 GAP Prototype using Python

Total Energy against C-C Pair Distances using Dataset for 100 Methane Dimers and its Error



Training RMSE: 758 μeV per methane molecule

10.2 *librascal* Code Development

Within *librascal*, a new representation header file was built called *calculator_pair_distances.hh*, using C++. The function of this file was to take a dataset and compute the pair distances themselves. Functionality was added to use a cut-off radius as a parameter to avoid having to compute all long-range distances. The ability to scale the pair distances by a power was also added, with the default power being 1.

A Python binding was built to complement *calculator_pair_distances.hh*, which was called *pair_distances.py*. This file gave the ability to call a *PairDistances* object in Python, in order to compute all the pair distances for a given dataset. The parameters included a cut-off distance and the number of species.

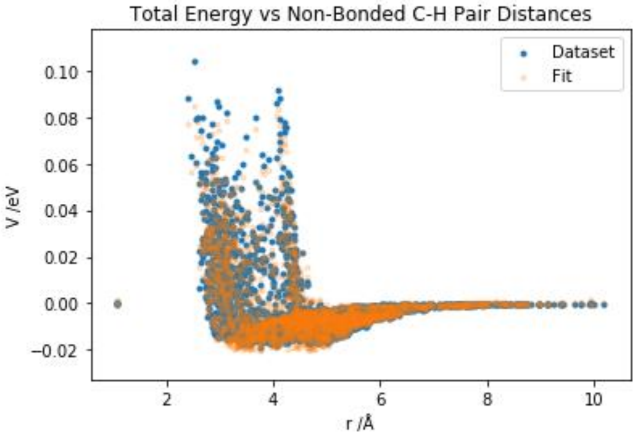
The Kernels header file, *kernels.hh*, was adapted to include the ability to compute a Gaussian kernel in addition to a Cosine kernel. For the Gaussian kernel, there were two different *TargetType*, *Structure* and *Atom*. Selecting *Structure* involved summing over all rows and columns for each system, giving a K_{NN} train-train kernel matrix, where N is the number of molecular systems. Selecting *Atom* involved either giving all computed values in a K_{MM} sparse-sparse kernel matrix or summing over the rows only to give a K_{NM} train-sparse kernel matrix, where M is the number of pair distances across all molecular systems within the dataset.

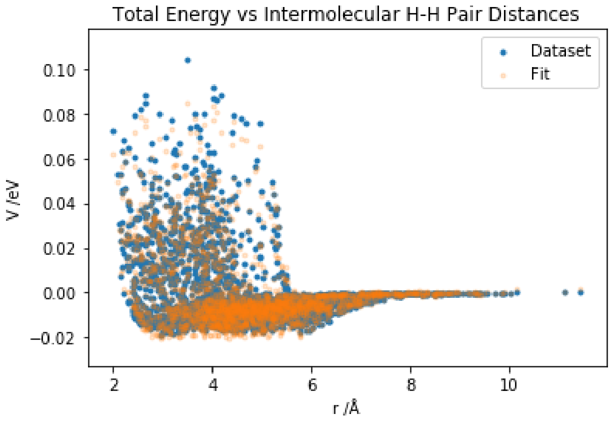
The Python binding *kernels.py* was adapted to fully complement *kernels.hh*, to enable the ability to call a *Kernel* object. The function of the binding was to compute a Gaussian kernel from the pair distances of a dataset in Python, either using a target type of *Structure* or *Atom*.

The tests *test_calculator.cc* and *test_calculator.hh* were updated to check for runtime errors upon computing pair distance representations.

10.3 Total Energy function of Intermolecular C-H and H-H Distances

Using the first *librascal* model created, intermolecular C-H and H-H distances were sampled with a consistent distance for each dimer, which appeared to be intermolecular based on the distance value being much greater than the known intramolecular distances. The total energy was then plotted as a function of these distances to observe how the distribution of points changes with pair type.

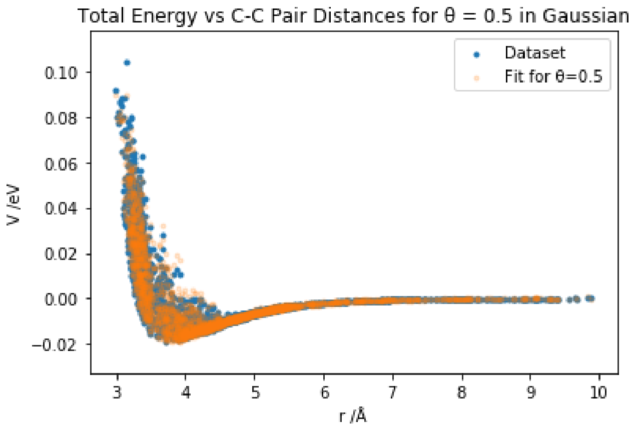


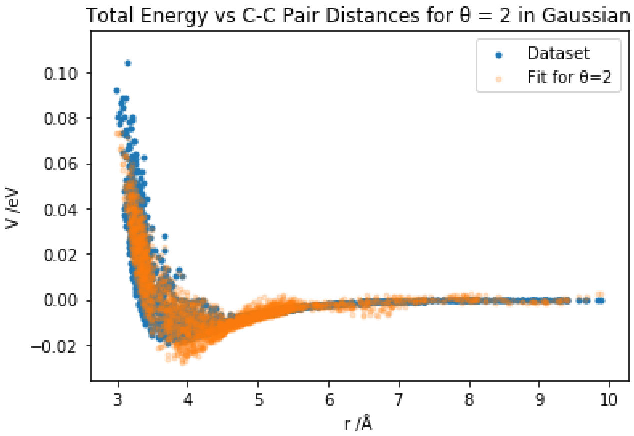


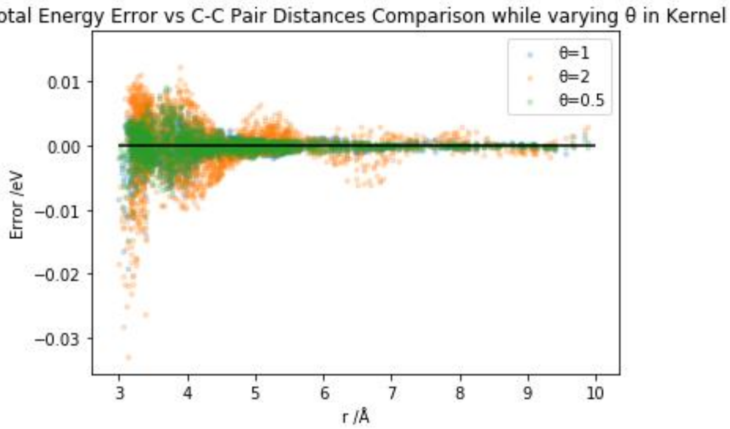
There is a wider spread in the repulsive part of the potential than there is in the C-C distance plot. This can be rationalised through the variety of intermolecular C-H distances possible, as for each carbon atom, there will be distances with four other hydrogen atoms in the neighbouring molecule.

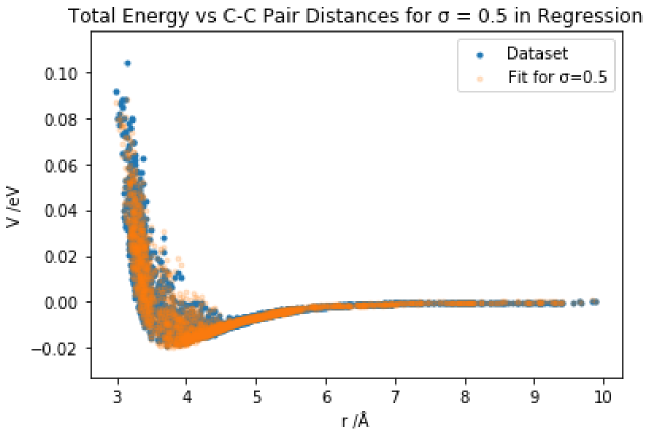
There is an even wider spread of data and predictions in the repulsive part of the potential than the intermolecular C-H distance cross-section. This can again be rationalised by the fact that there are sixteen different H-H distances, compared to four different C-H distances, leading to greater variation and spread.

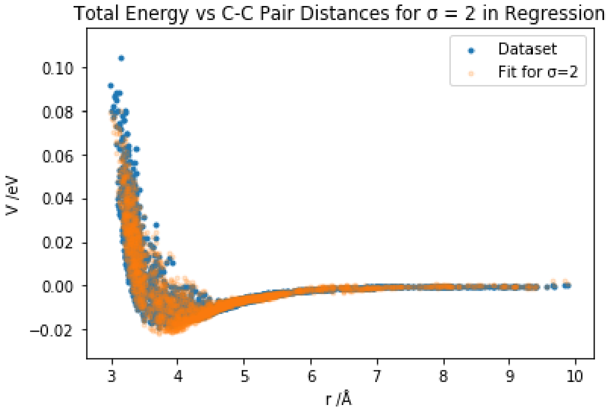
10.4 Plots for varying hyperparameters with Non-Normalised Kernels



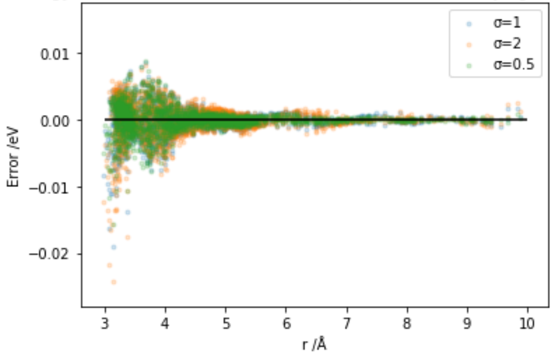




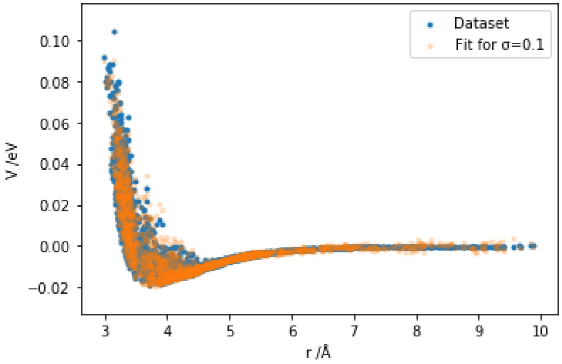




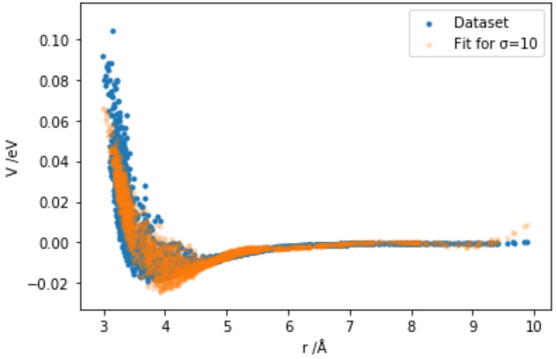
Total Energy Error vs C-C Pair Distances Comparison while varying σ in Regression



Total Energy vs C-C Pair Distances for σ = 0.1 in Regression



Total Energy vs C-C Pair Distances for σ = 10 in Regression



Total Energy Error vs C-C Pair Distances Comparison while varying σ in Regression

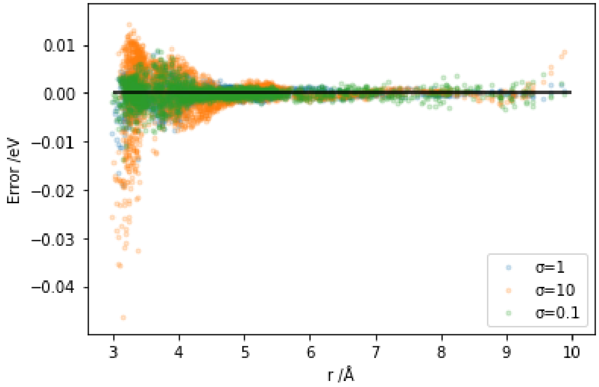


Table of RMSE for fits with Non-Normalised Kernels:

Gaussian Width, θ /Å	Regulariser, σ	RMSE /eV per methane molecule
1	1	$9.94\text{x}10^{-4}$
0.5	1	$8.49\text{x}10^{-4}$
2	1	$2.12\text{x}10^{-3}$
1	0.5	$8.91\text{x}10^{-4}$
1	2	$1.22\text{x}10^{-3}$
1	0.1	$8.90\text{x}10^{-4}$
1	10	$2.60\text{x}10^{-3}$