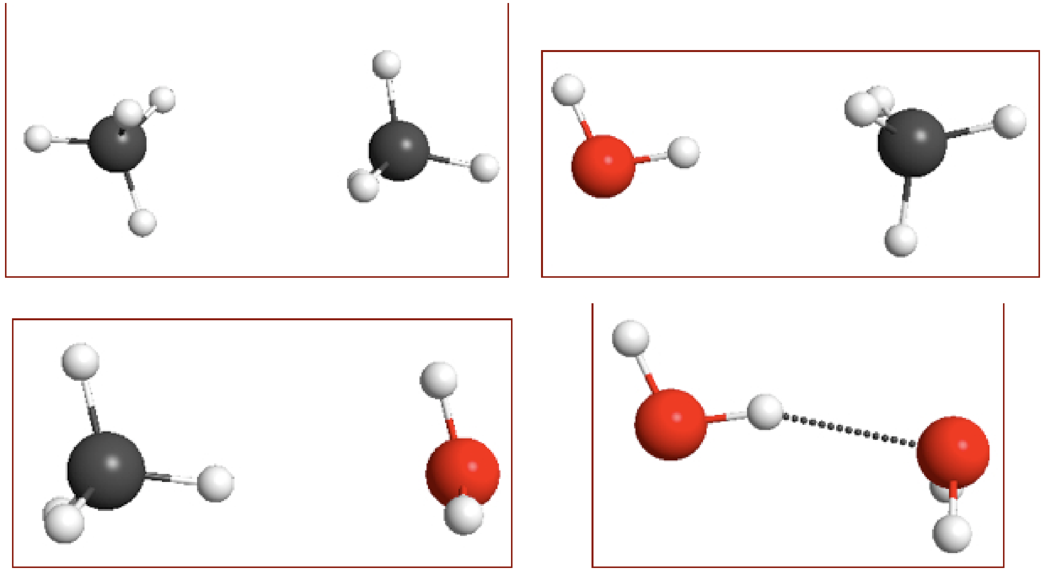

Low-Dimensional Machine Learning Potentials for Molecular Systems

SAHIL SHAH

EPFL SUPERVISORS: PROF. MICHELE CERIOTTI & DR. MAX VEIT

IMPERIAL COLLEGE LONDON SUPERVISOR: PROF. SOPHIA YALIRAKI

Pair Potentials

- Describe the potential energy between two atoms

- Used to model properties of systems, such as cohesion, thermal expansion and elastic and plastic behaviour

- Accuracy of properties measured is limited by accuracy of potential used to generate them

- Pairwise Additivity Approximation via the Body-Order Expansion

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

- Lennard-Jones potential (LJ-12-6)

$$U_{LJ}(r) = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) \tag{2}$$

Machine Learning Potentials

- Machine Learning Potentials – much more accurate due to greater functional flexibility
- Gaussian Process Regression – used to make Gaussian Approximation Potentials (GAPs)
- Distance scaling – why might it be a good idea?
- Rescaling by powers of -6 and -12, combining to give 12-6 model
- Created 4 models – Unscaled Distances, 6th-Power, 12th-Power and 12-6

$$\frac{1}{r^{6}} \qquad\qquad \frac{1}{r^{12}}$$

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.

Gaussian Process Regression

- Pair potentials can be written as a sum over basis functions, multiplied by weights.

$$\varepsilon_i = \varepsilon(d_i, \boldsymbol{w}) = \sum_h w_h \varphi_h(d_i) \tag{3}$$

- Gaussian kernels are computed, measuring the similarity between descriptors.

$$k\left(\boldsymbol{d}_i, \boldsymbol{d}_j\right) = \sum_h \varphi_h(\boldsymbol{d}_i)\varphi_h\left(\boldsymbol{d}_j\right) = exp\left(-\sum_{i,j}\left[\left(\boldsymbol{d}_i \; - \; \boldsymbol{d}_j\right)^2/2\theta^2\right]\right) \tag{4}$$

- To compute the weights, regularised loss function must be minimised and hyperparameters optimised

$$\boldsymbol{L} = \sum_{\boldsymbol{i}}(\boldsymbol{y_i} \; - \; \boldsymbol{f}(\boldsymbol{d_i}))^{\boldsymbol{2}} + \sigma^2\|\alpha\|^{\boldsymbol{2}} \tag{5}$$

$$\boldsymbol{\alpha} \; = \; (\boldsymbol{K_{NN}} \; + \; \sigma^2\boldsymbol{I_{NN}})^{\boldsymbol{-1}} \; \boldsymbol{y} \tag{6}$$

- Compute fitted energies or pair potentials

$$\varepsilon^* = \boldsymbol{\alpha}.\, k(\boldsymbol{d}, d \, *) \tag{7}$$

librascal Development

- ❖Flowchart describes the process of producing Gaussian Approximation Potentials (GAP)

- ❖Most of computational cost comes from computing descriptors and kernels

- ❖Pair distances and Gaussian kernels computed in *librascal*

- ❖Distance scaling implemented

- ❖Methane dimer dataset chosen to generate GAP

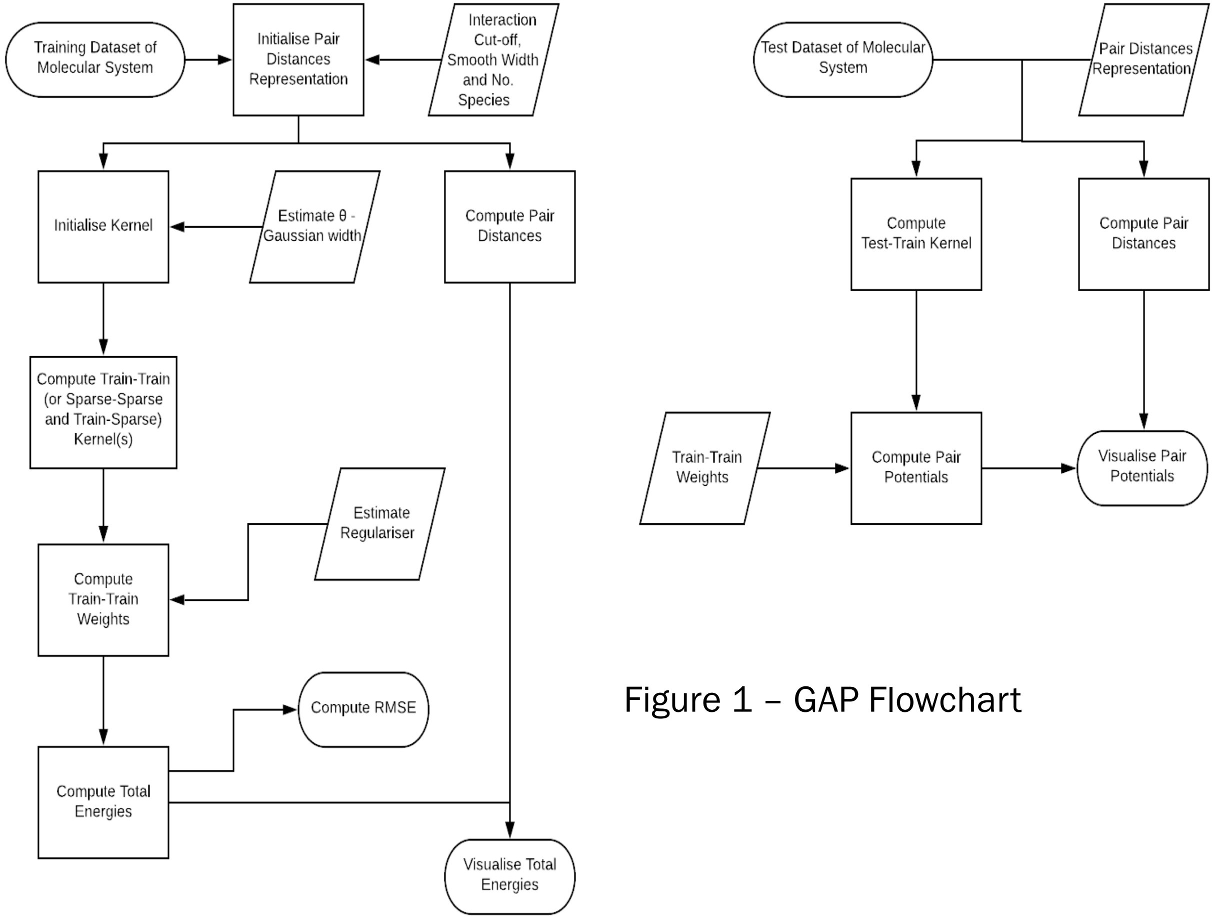
Training Dataset of Molecular System
Initialise Pair Distances Representation
Interaction Cut-off, Smooth Width and No. Species
Initialise Kernel
Estimate θ - Gaussian width
Compute Pair Distances
Compute Train-Train (or Sparse-Sparse and Train-Sparse) Kernel(s)
Estimate Regulariser
Compute Train-Train 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 1 – GAP Flowchart

Unscaled Distances Model – Hyperparameter Optimisations

- Estimation of length scale parameter at 0.8 Å

- Optimisation of regulariser using six-fold cross-validation

- Optimal and low regularisers still gave overfitted potentials

- A slightly higher regulariser was used of 0.03 with a cross-validation error of 1.15 meV per methane molecule

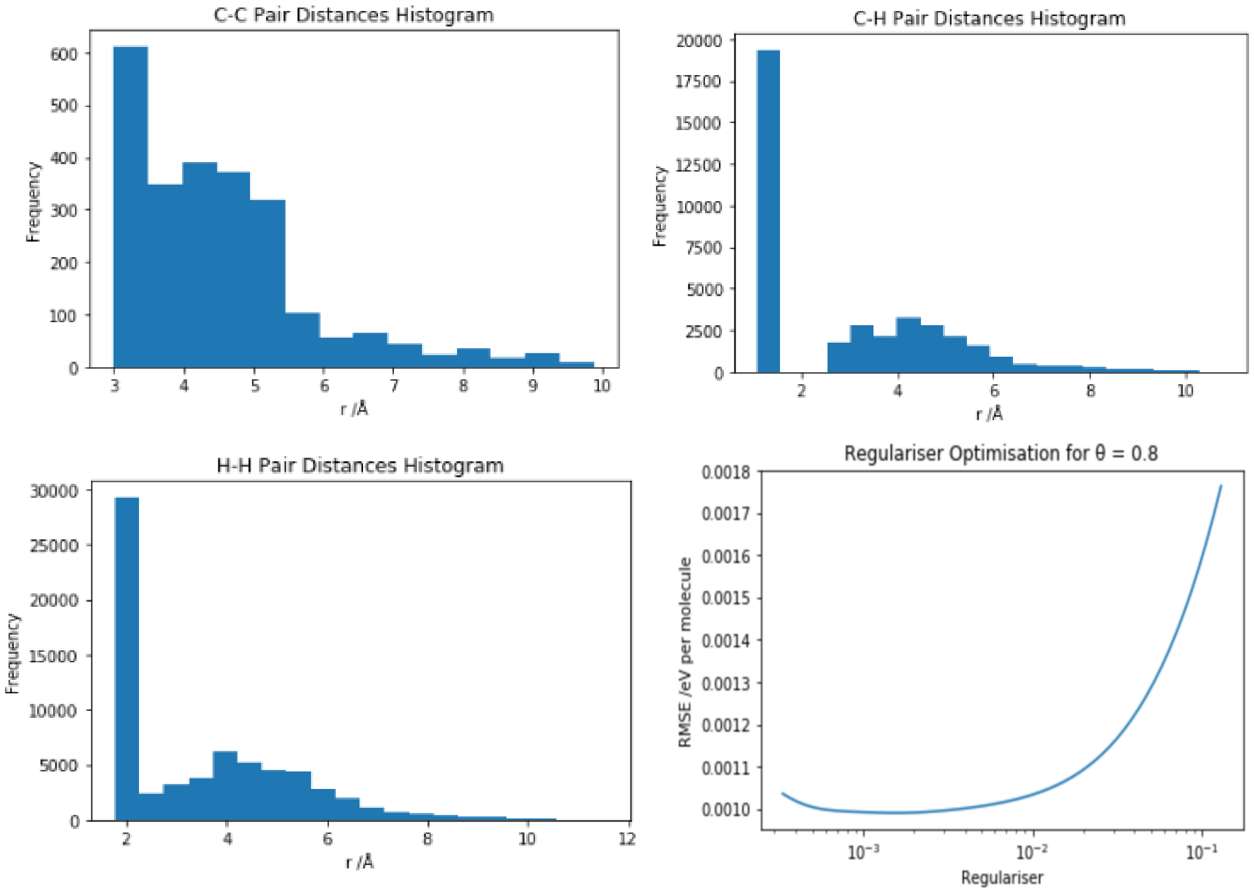

Figure 2 – Unscaled Distance Optimisations

Unscaled Distances Model

- Training error of 1.02 meV per methane molecule, lower than the standard deviation of the quantum mechanical energies
- *librascal* unable to differentiate between intramolecular and intermolecular pairs
- Larger error at short-range and smaller error at long-range
- Error oscillates at long-range

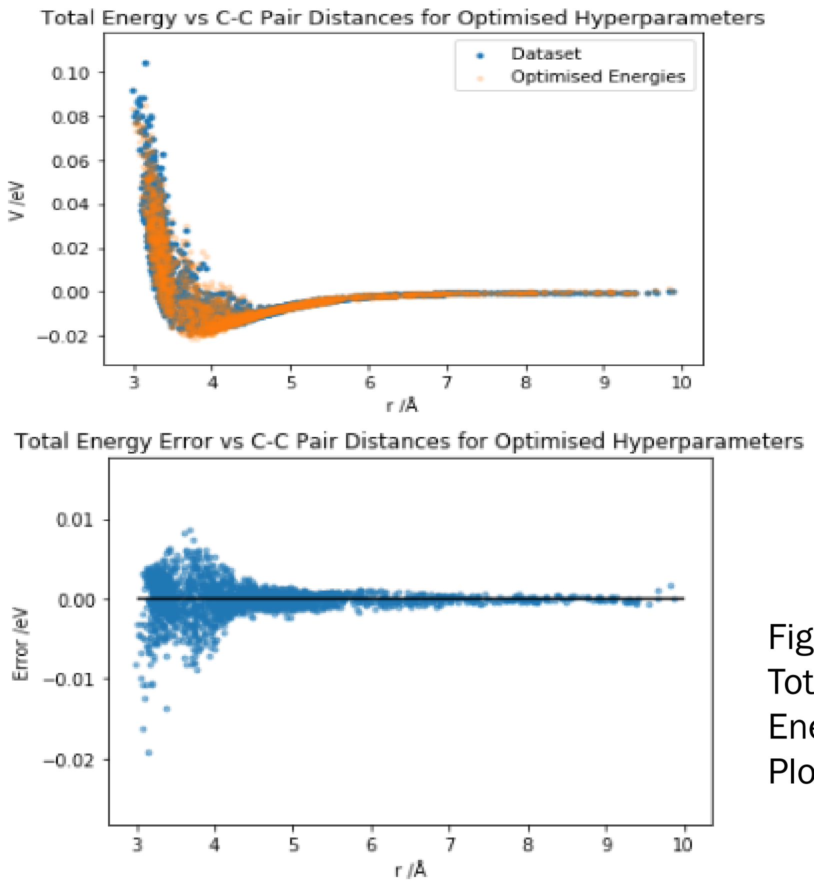

Figure 3 – Total Energy Plots

Unscaled Distance Potentials

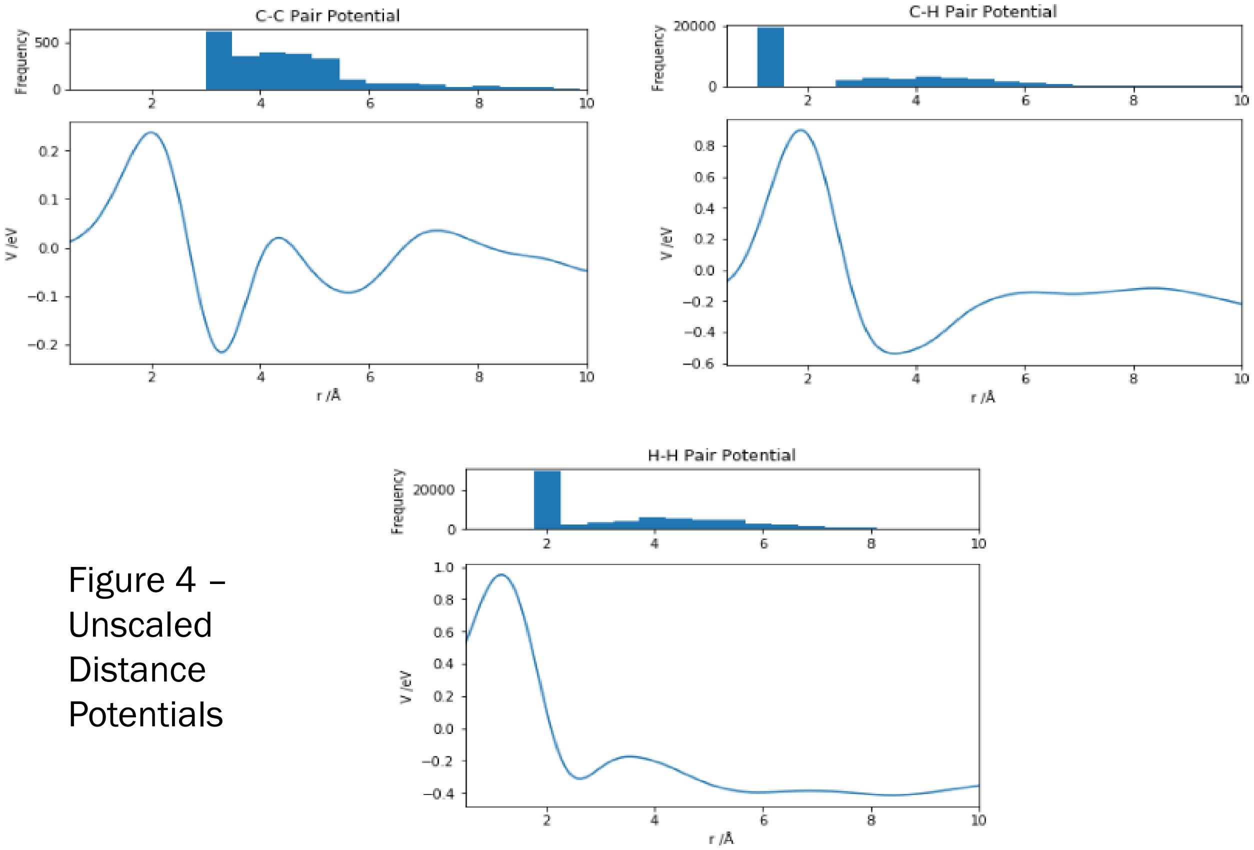

Figure 4 – Unscaled Distance Potentials

- ❖Potentials oscillate at long-range, shorter cut-off may be required
- ❖High amplitudes in the energy scale
- ❖Full optimisation of length scale parameter required
- ❖Potentials contain repulsive components to total energy at short-range

6th-Power Model – Hyperparameter Optimisations

- Pair distances scaled by power of -6 in *librascal*
- Estimated length scale parameter at 5x10^{-4} Å^{-6}
- Regulariser optimised through six-fold cross-validation
- Slightly higher regulariser was chosen at 0.07 with a cross-validation error of 1.07 meV per methane molecule

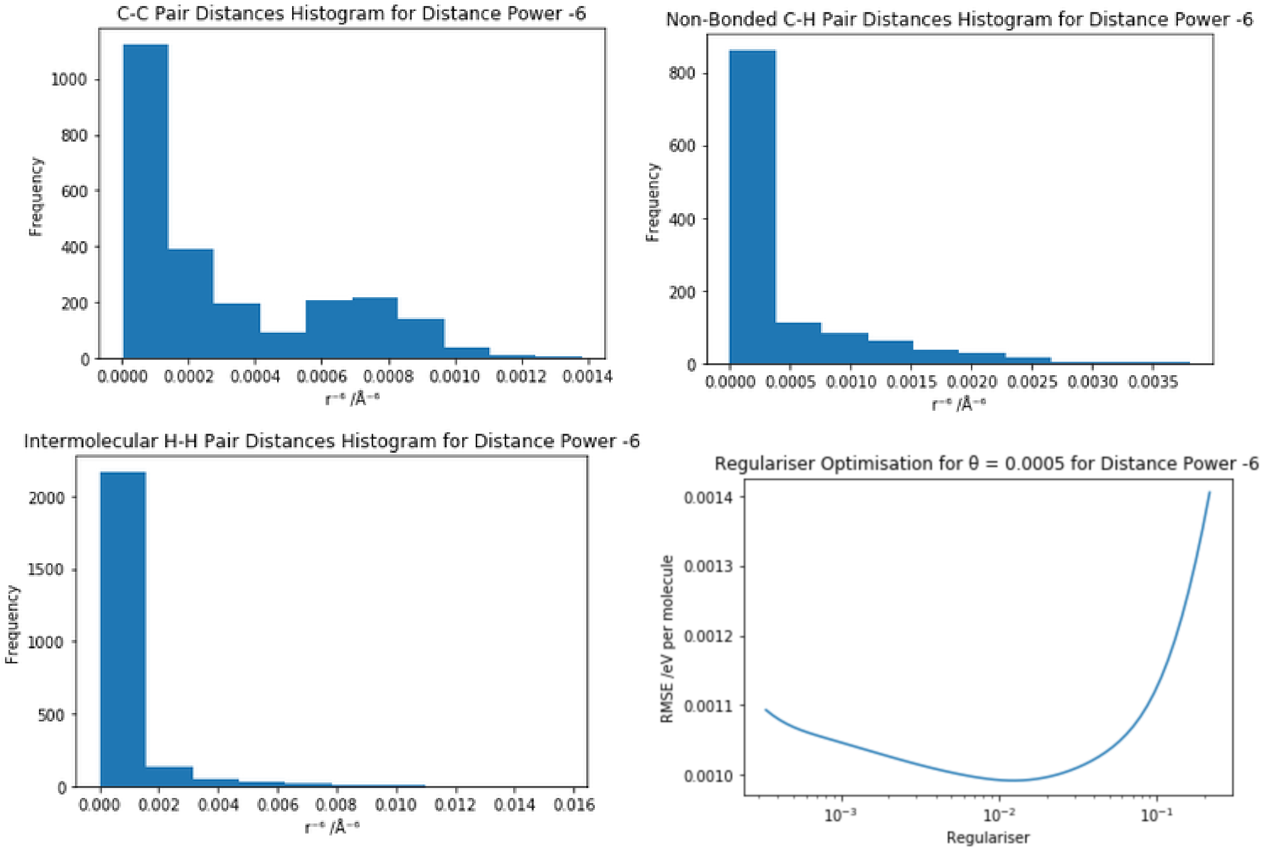

Figure 5 – 6th-power Optimisations

6th-Power Model

- ❖Training error of 863 μeV per methane molecule
- ❖Smaller error at short-range than unscaled distances fit
- ❖Oscillations at long-range removed
- ❖Good option for long-range tail of intermolecular interactions

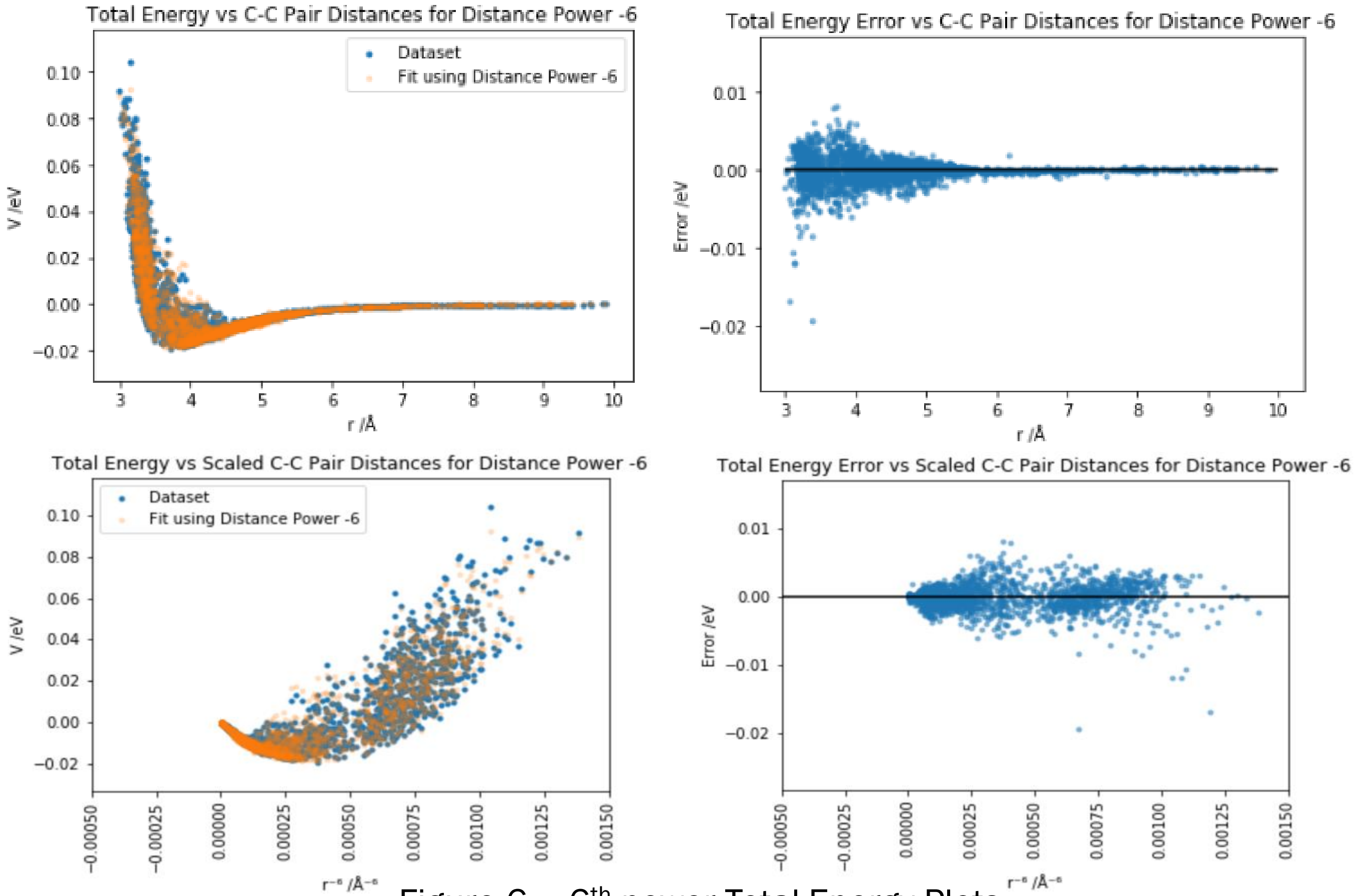

Figure 6 – 6th-power Total Energy Plots

6th-Power Potentials

❖Residual intramolecular energy removed

❖All show attractive and repulsive contributions to total energy at long- and short-range

❖Mostly fairly smooth potentials

❖Well depths of 23.6 meV, 23.3 meV and 109 meV for the C-C, C-H and H-H potentials

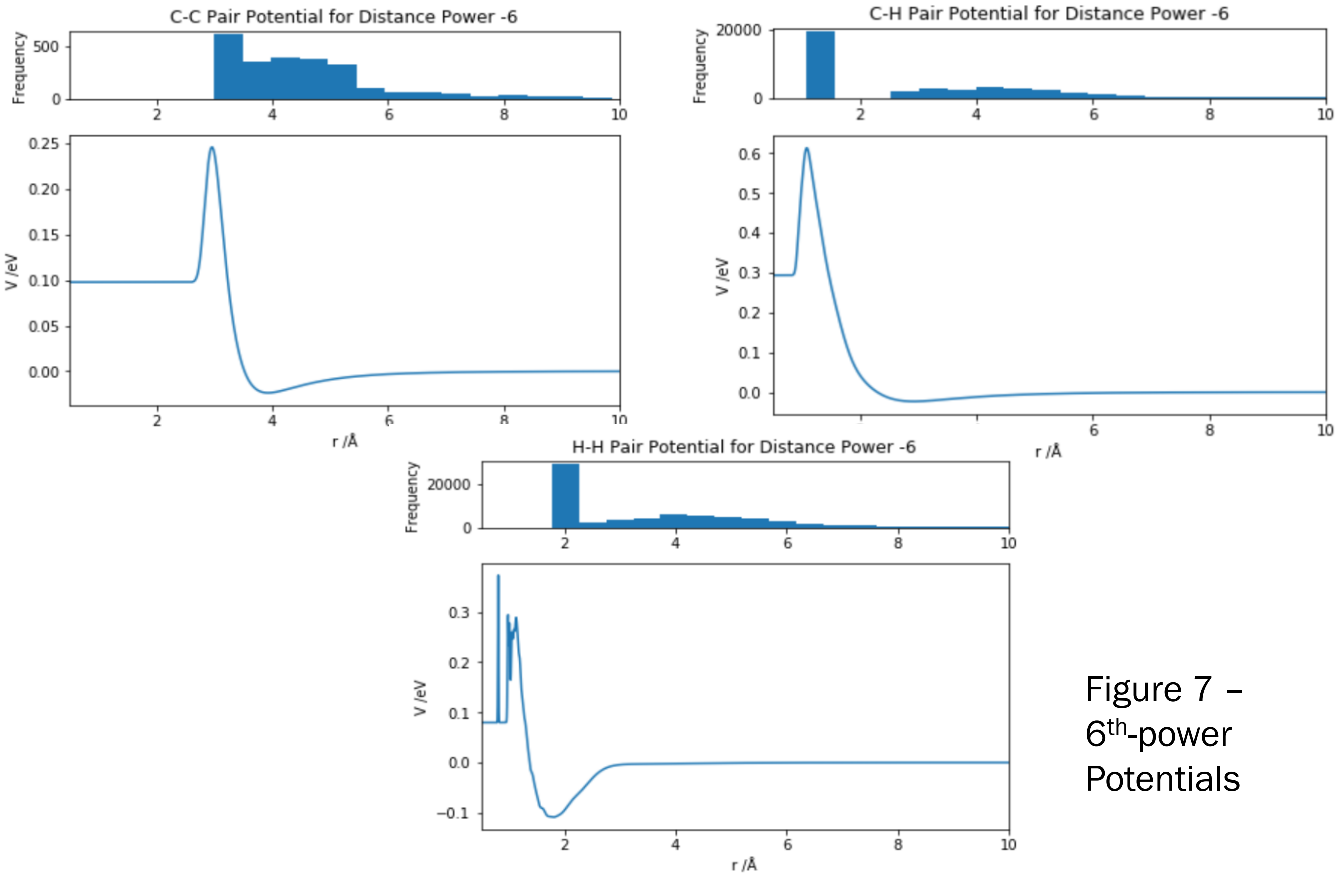

Figure 7 – 6th-power Potentials

12th-Power Model – Hyperparameter Optimisations

- Pair distances scaled by power of -12 in *librascal*
- Estimated length scale parameter at 2.5x10^{-7} Å^{-12}
- Regulariser optimised through six-fold cross-validation
- Slightly higher regulariser was chosen at 0.06 with a cross-validation error of 2.41 meV per methane molecule

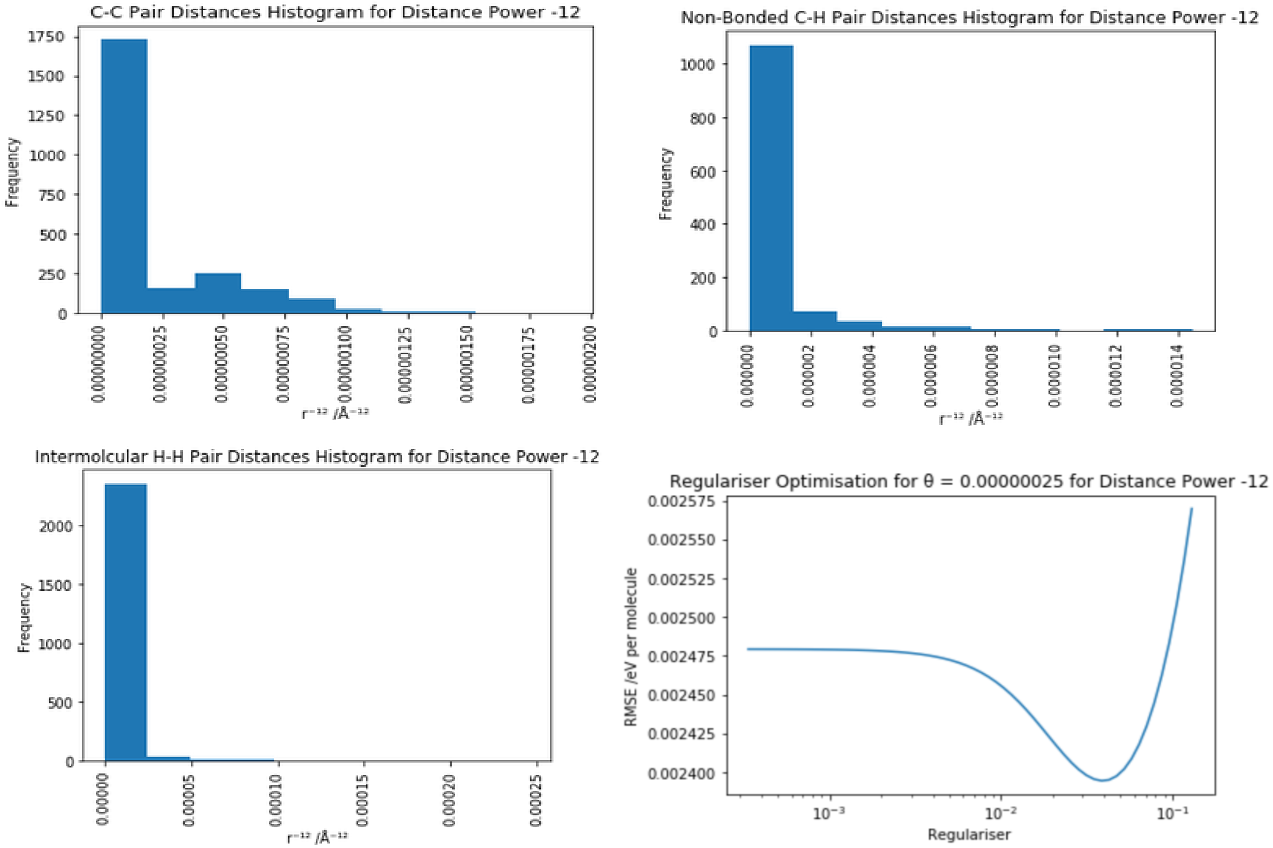

Figure 8 – 12th-power Optimisations

12th-Power Model

- Lower training error of 499 μeV per methane molecule
- Higher cross-validation error implies overfitting
- Bias at long-range, error values consistently negative
- Improve with a full optimisation of length scale parameter

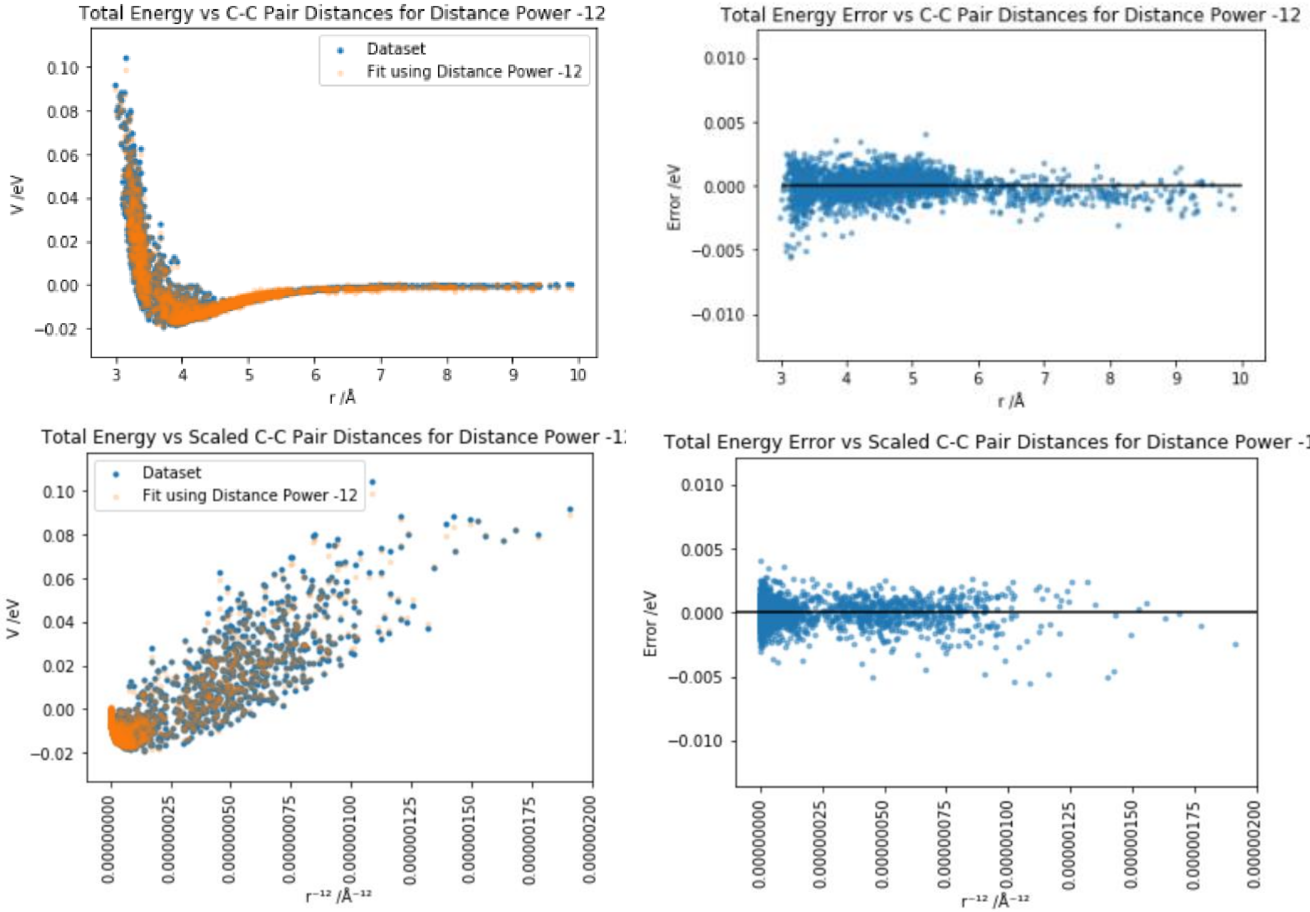

Figure 9 – 12th-power Total Energy Plots

12th-Power Potentials

- ❖C-C and C-H potentials have only repulsive contributions
- ❖H-H potential has both attractive and repulsive contributions at long- and short-range
- ❖Large fluctuations caused by overfitting
- ❖Large H-H well depth of 214 meV

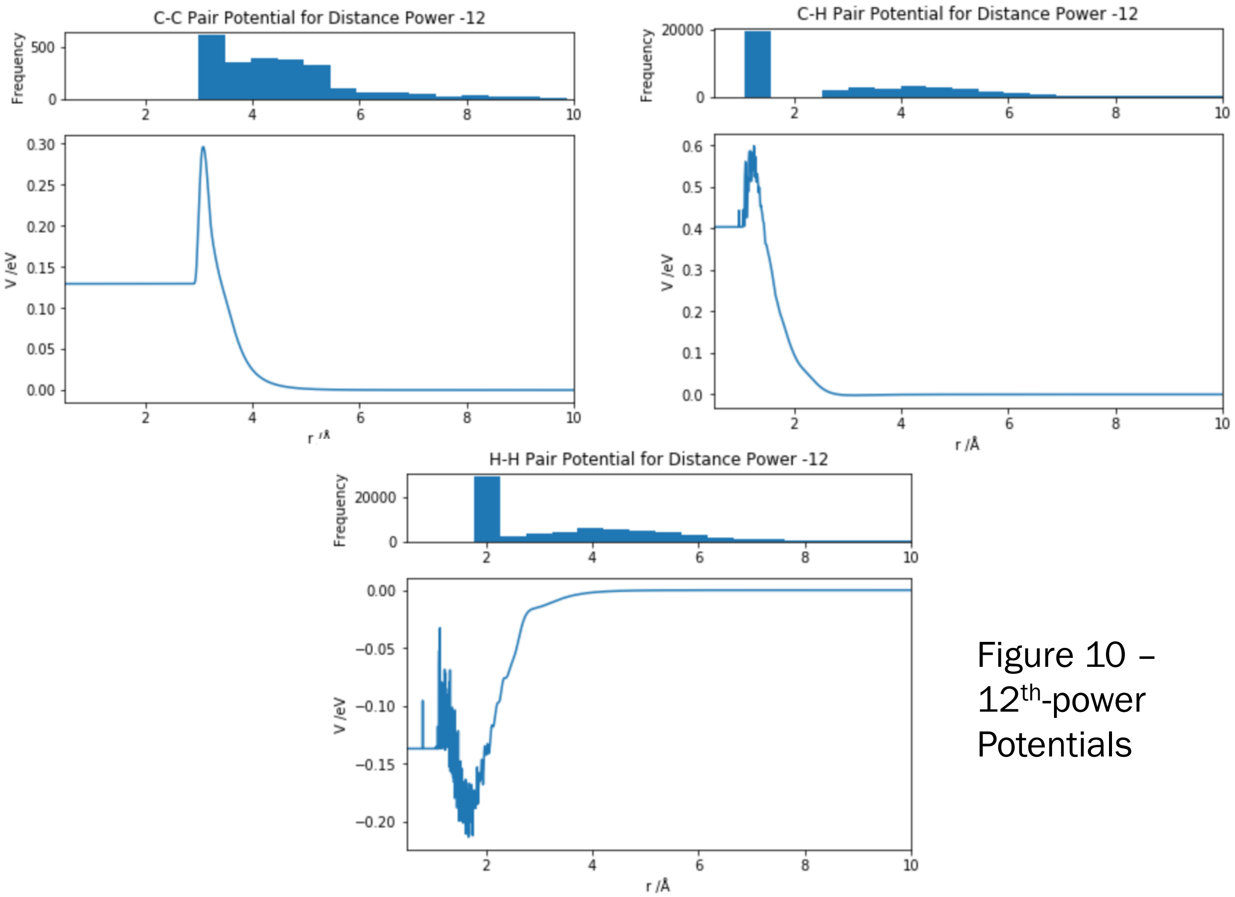

Figure 10 – 12th-power Potentials

12-6 Model

- 6th-power and 12th-power kernels summed to give 12-6 kernel

- Regulariser optimisation carried out using six-fold cross-validation, using minimum CV error of 1.42 meV per methane molecule at a regulariser of 0.0523

- Lowest training error of 389 μeV, but slightly overfitted

- Large error at short-range somewhat removed, no bias at long-range

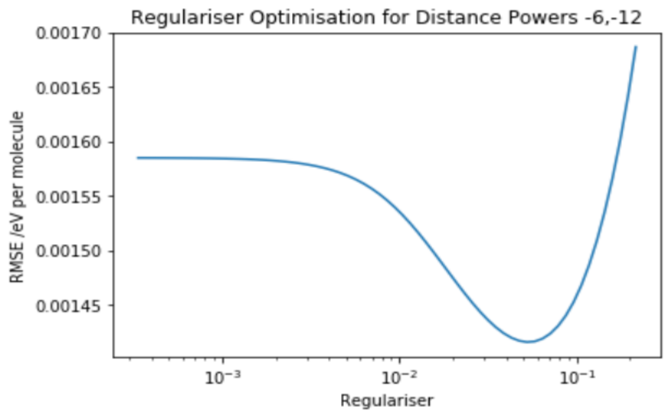

Figure 11 – 12-6 Optimisation

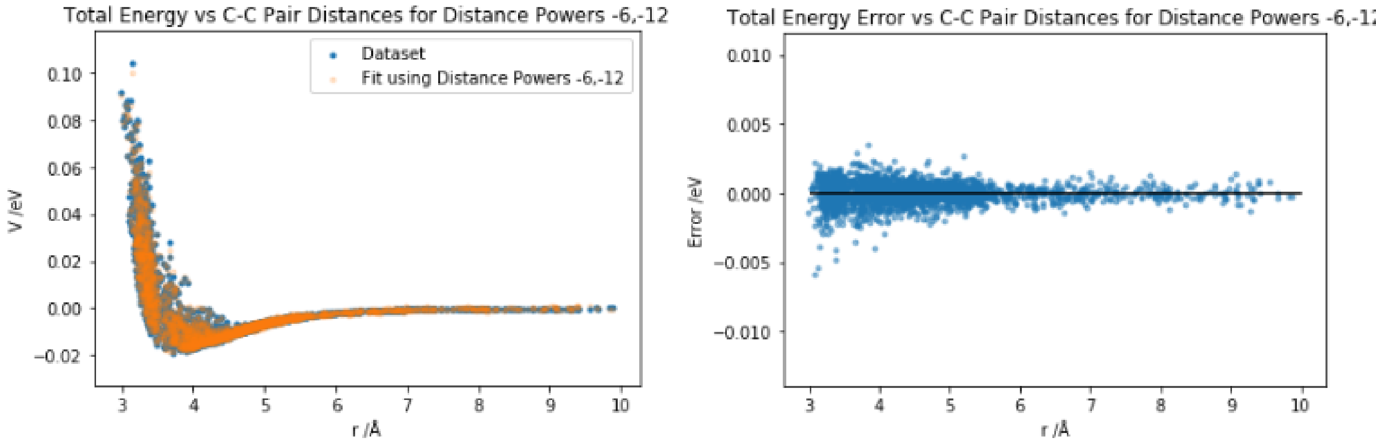

Figure 12 – 12-6 Total Energy Plots

12-6 Potentials

❖All potentials give attractive and repulsive contributions at long- and short-range

❖High orders of magnitude for potential wells – 8.17 meV for C-C, 33.1 meV for C-H and 171 meV for H-H pairs

❖Removal of oscillations at long-range without using a shorter cut-off radius

❖Noise indicative of overfitting

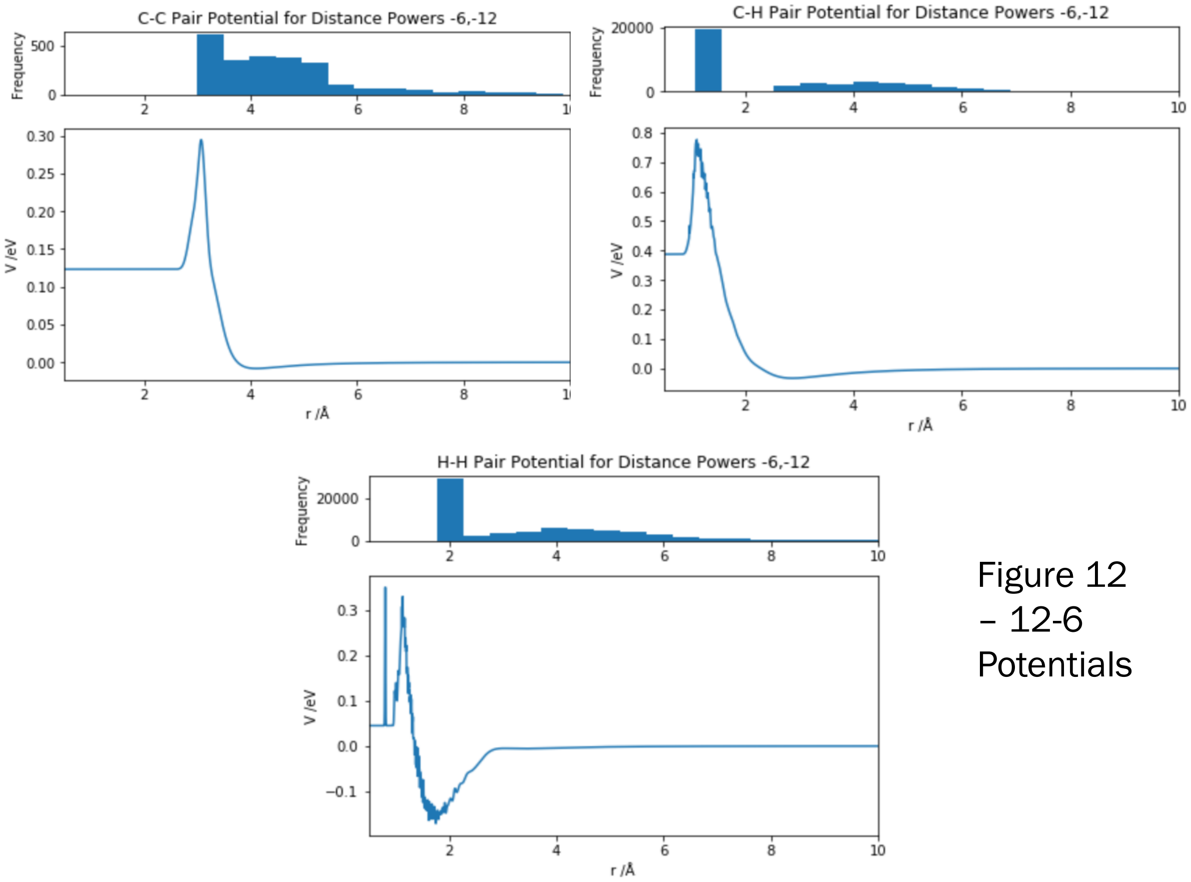

Figure 12 – 12-6 Potentials

Summary

Model	Validation Error	Training Error
Unscaled Distance	1.15 meV	1.02 meV
6th-Power	1.07 meV	863 μeV
12th-Power	2.41 meV	499 μeV
12-6	1.42 meV	389 μeV

Table 1 – Error Comparison

- Unscaled Distance Potentials computed but oscillated at long-range
- Distance scaling by powers of -6 and -12 hoped to remove this
- **6th-power potentials held a good physical form, while 12th-power potentials were overfitted**
- 12-6 model created by combining 6th- and 12th-power models
- 12-6 potentials held a reasonable physical form but with some overfitting
- *librascal* can be used to compute GAPs using distance scaling powers