

Gaussian Process Machine Learning of Potential Energy Hypersurfaces

by Elena Uteva, BSci

Thesis submitted to The University of Nottingham

for the degree of Doctor of Philosophy

School of Chemistry, October 2018

What's wrong? Get a bit of a scare out there? No problem. Have a seat and get comfortable. We'll both be hollow before you know it.

Crestfallen Warrior.

Contents

List of Figures **5**

Abstract **8**

Acknowledgements **10**

1 Introduction **12**

1.1 Textbook references 12

1.2 Introduction 13

1.3 Previous Research using Gaussian processes 18

1.4 Roadmap of chapters 21

2 Electronic Structure Calculation **22**

2.1 Textbook references 22

2.2 Introduction 23

2.3 Different Electronic Structure Methods 33

2.3.1 Hartree-Fock Method 33

2.3.2 Electron Correlation Methods 34

2.4 Density Functional Theory 36

3 Intermolecular Forces **38**

3.1 Textbook references 38

3.2 Introduction 39

3.2.1 Origins of intermolecular forces 41

3.2.2 Computing intermolecular forces 45

3.2.3 Supermolecule approach 47

4 Gaussian Processes and Bayesian Learning **50**

4.1 Textbook references 50

4.2 Two approaches to learning 51

4.3 Bayesian Learning 52

4.4 Gaussian Distributions 55

4.4.1 Univariate Gaussian Probability Distributions 55

4.4.2 Multivariate Gaussian Distributions 56

4.5 Gaussian Process 57

4.5.1 Mean function 58

4.5.2 Covariance Function 58

4.5.3 Matérn Covariance Function 60

4.5.4 Modifying kernels 61

4.6 Gaussian Process Learning 61

4.6.1 Gradient based optimization 65

5 Gaussian Process Interpolation of Potential Energy Surfaces **67**

5.1 Introduction 68

5.2 Methodology 70
5.2.1 Input Coordinates 70
5.2.2 Data generation 71
5.2.3 Model Construction 77
5.2.4 Symmetric Kernels 78
5.3 Results and Discussion 81
5.4 Future work 89
5.5 Conclusions 91

6 Sequential Design of Potential Energy Surfaces 92

6.1 Introduction 93
6.1.1 Molecular systems 98
6.1.2 Data sets 99
6.1.3 Co-ordinates 99
6.1.4 LHC generation for Ar_3 101
6.1.5 Active learning and GP training 101
6.1.6 Comparison to a batch space-filling design 104
6.2 Results 105
6.3 Future work 114
6.4 Conclusions 115

7 Final Discussion 118

List of Figures

2.1 A plot of GTO and STO type orbitals. The solid line is the radial decay of STO nd the dashed line is the radial decay of GTO Image sourced from Essentials of Computation Chemistry, Chris Cramer, Wiley, 2nd Edn, ... 32

3.1 Energy of interaction as a function of distance between two molecules 41

3.2 Different types of intermolecular forces 43

4.1 A 1D Gaussian distribution 56

4.2 Sample prior function of a Squared exponential covariance function with three different lengthscales (1, 0.5 and 0.05 from left to right, with variance set to 1×10^{-2}) 60

4.3 Sample prior function of a Matérn exponential covariance function with three different lengthscales (1, 0.5 and 0.05 from left to right, with variance set to 1×10^{-2}) 61

4.4 Samples drawn from the prior and posterior distribution 62

5.1 The Jacobi coordinate system for CO_2 - Ne. Figures 5.1 to 5.4 were created by Liam Goulding 72

5.2 Interatomic distance coordinate system for CO_2 - Ne. The three interatomic distances are represented by dashed lines. 73

5.3 The jacobi coordinate system of CO_2 - H_2. 74

5.4 Interatomic distance coordinate system for CO_2 - H_2. The six distances are represented by dashed lines. 75

5.5 RMSE against LHC size for CO_2-Ne. The lowest energy in the grid data is -2.90×10^{-4} E_h. 82

5.6 RMSE against LHC size for CO_2-H_2. The lowest energy in the grid data is -8.25×10^{-4} E_h. 84

5.7 RMSE against LHC size for HF-HF. The lowest energy in the grid data is -6.17×10^{-3} E_h. 85

5.8 RMSE against LHC size for HF-HF. The lowest energy in the grid data is -6.98×10^{-4} E_h. 87

5.9 CO_2-Ne Molpro calculations and the GP model, at $\cos\theta = 1$ (linear geometry), in the repulsive (a), attractive (b) and long-range (c) regions. The long-range asymptotic expansion is $E = -(0.570+0.182\cos^2\theta)r^{-6}-(1.704+7.266\cos^2\theta+1.785\cos^4\theta)r^{-8}$. 90

6.1 Co-ordinates used to describe the Ar_3 system 102

6.2 RMSE against training set size for CO_2-Ne. The lowest energy in the reference data is -2.90×10^{-4} E_h. 106

6.3 The placement of the first 75 points in the two-set method . . . 108

6.4 The placement of the first 75 points in the highest variance method 109

6.5 The placement of the first 75 points in the highest error method 109

6.6 LHC point location 110

6.7 RMSE against training set size for CO_2-H_2. The lowest energy in the reference data is -8.25×10^{-4} E_h. 111

6.8 RMSE against training size for Ar_3. The lowest and highest energies in the reference data are -2.10×10^{-4} E_h and 5.31×10^{-5} E_h and the root mean square over the reference data is 1.28×10^{-5} E_h. 113

Abstract

Born-Oppenheimer potential energy surfaces (PES) are a central concept in computational chemistry. Extensive research has been geared towards developing novel methods of generating accurate but computationally-feasible, global potential energy surfaces. Recent approaches have stepped away from the use of analytical methods, and have instead honed in on the application of machine learning techniques. These methods primarily focus on just mapping the input-output space of a training data set as accurately as possible, without delving too deeply into the chemistry of the system studied.

One such machine learning technique is Gaussian Processes regression - a non-parametric learning tool that has shown significant promise in being able to predict the intermolecular energies of a variety of systems to an arbitrary level of accuracy. This method is adopted in our research, producing high-accuracy global PES for several van der Waals bimolecular complexes at a fraction of the usual computational cost, when using accurate training data. The research also introduces several algorithms for sequential design, where the model selects training data points iteratively. Points are chosen based on information learnt from the previous training step. The sequential design results are pitted against 'one shot' space-filling designs (where training data are just chosen in bulk on the basis of the best coverage of space). Two of the sequential design methods significantly outperform the space filling sets in terms of predictive accuracy. Global potential energy surfaces are produced

for CO_2-H_2, CO_2 – Ne, CO_2 – CO, HF-HF, CH_4 – N_2 and Ar_3 but the algorithm can be easily applied to other molecular interactions with a only few trivial modifications.

Acknowledgements

I'd like to express a strong and heartfelt thank you to all three of my supervisors (Dr Richard Wheatley, Dr Richard Graham and Prof Richard Wilkinson), for supporting and guiding me throughout the whole course of the PhD. I owe them many thanks for showing me unlimited understanding and flexibility for the situation I was in, and for helping me grow in both my confidence and ability as a research student. Whatever next steps I take professionally, I'm indebted for their help in getting me this far.

Additionally, thank you to everyone in the Chemistry department for all their support over the course of my studies. Thank you to all the staff who helped me to secure maternity leave, to all fellow students who helped me resolve my Linux troubles and to everyone in physical undergraduate labs who made me look forward to second year demonstrating sessions.

Just as important as the professional support, I'd also like to extend a big thank you to my parents, who have both always been nothing short of my rock and have always believed in me even when I struggled to see the end of the tunnel myself.

Thank you also to my young daughter, Katia, who was born halfway into my PhD and who gave my life new meaning right at the moment it was needed. There were sleepless nights, erratic work times, research papers read as bed-

time stories and lots of moments where I doubted whether I could juggle it all, but it was without a shadow of a doubt, worth it all. One day, when she's older and per chance reading this, may she realise she was always a big reason for why I never gave up.

Last but not least, thank you to my other half, Adam Goulding, who first heard about my research on a park bench on campus, and who has since then, probably wished he got to hear a lot less of it overall. I'm forever indebted to him for showing me that I'm more than capable of finishing it, for spurring me on when my motivation was slumping and for always managing to a squeeze a laugh out of me even on the toughest of days.

"Look, I did it mum. I'm grown up now. I'm all bigger."

Katia, aged 2.5 years

Chapter 1

Introduction

Questions: *What are potential energy surfaces and how are they used? How are potential energy surfaces generated? Why is machine learning used in potential generation and which machine learning tools have been used for the task so far? What makes Gaussian Processes a good choice for modelling interatomic potential energy landscapes?*

Keywords: *Potential energy surfaces, force fields, ab initio, machine learning methods, emulation, Gaussian processes*

1.1 Textbook references

The theoretical backdrop was sourced from: Errol G Lewars' 'Computational Chemistry' (Chapter 2)[1] , David Young's 'Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems' (Chapter 20)[2]

and Behler's 'Neural network potential-energy surfaces for atomistic simulations'.[3]

1.2 Introduction

Potential Energy Surfaces (PES) are a central concept in computational chemistry and are used extensively in the theoretical branches of biology and chemistry. They are used to obtain relevant information about the structural, spectral and dynamical properties of different molecular systems. Chemically, potential energy surfaces are defined as graphical or mathematical representations of energy as a function of geometry, and arise naturally from the physical implications of the Born-Oppenheimer (BO) approximation.[4]

The Born-Oppenheimer approximation is a very important simplification in quantum chemistry. It separates electronic and nuclear motions based on their relative velocities and permits the molecular wavefunction to be defined in terms of electron and nuclear positions. Since the electron is both significantly lighter in mass and higher in speed than the nucleus, it is able to respond adiabatically to any changes in the nuclear configuration. As the nucleus is treated as a stationary body relative to the electron, the BO approximation permits the idea of a fixed nuclear geometry to exist, and allows separate potential energy surfaces to be defined for each electronic state.

The importance of the potential energy surface in the field of chemistry is manifold,[5] with a lot of insight to be gained from either inspecting the potential surface directly or through its subsequent application to other techniques such as statistical thermodynamics or molecular dynamics.[6] Stable molecular structures can be elucidated by identifying the local potential minima through gradient descent or second derivative methods. The global potential minimum corresponds to the most stable molecular configuration. Reaction coordinates are constructed by tracing the lowest energy pathway from reactant to products, locating transition states at the highest energies along this reaction coordinate.[7] The surface curvature is the second derivative of the potential with respect to directional distance displacement, and correlates to the force constant of various vibrational modes. This allows for the determination of vibrational energy levels,[8] corresponding IR and Raman spectra, zero-point vibrational energy, reaction rate coefficients,[9] tunnelling corrections and isotopic substitution effects. The separation between the potential energy surfaces of different electronic states yields the electronic spectrum. The application of external magnetic and electrical fields to the system can also generate information such as the dipole moment,[10] the polarisability and NMR shielding.

Since chemistry is heavily involved in dynamic processes, and the interaction between atoms and molecules is dependent on the forces acting between them, gaining insight into the dynamics of any arbitrary system through Molecu-

lar Dynamics (MD) simulations requires knowledge of all forces acting on the atoms and molecules under study. The accuracy of a MD simulation is directly dependent on the accuracy of the potential used in the calculation. Therefore obtaining a means of generating accurate potential energy surfaces is the key step to obtaining reliable theoretical calculations of chemically interesting and useful molecular properties.

Broadly speaking, the methodology for obtaining potential energy surface approximations can be divided into two categories: quantum mechanical and molecular mechanical. Quantum mechanical methods are based around solving the Schrödinger wave equation for different geometrics and the molecular mechanics focuses on producing parametrised mathematical forms based on chemical intuition and empirical data.

The molecular mechanics methods are commonly known as force fields (FF) methods. They use parametric functional forms based on the dynamic laws of classical physics and a 'ball and spring' representation of molecular systems. The free parameters include atomic charge and van der Waals coefficients, which are obtained using either empirical or high-accuracy ab initio data to calibrate the function. The FF method views the molecular system as a mechanically connected collection of atoms and only factors in quantum effects on the electron by virtue of ad hoc functional approximations. Bonding is explicitly included in the functional form with the total energy expressed

as a summation of energies for various molecular distortions. These include stretching of all bonded atom pairs, potential energy terms for bending, van der Waals interactions, electrostatic interactions, torsional energy and cross terms.

One strategy for the development of such empirical potentials begins from a standard formulation, which is expanded either by supplementing with additional terms or making the adjustable parameters more flexible to improve the accuracy of the fit. This is often tedious to accomplish for larger systems and often in the case of molecular dynamic studies, produces results that are more qualitatively than quantitatively useful.

In principle, the most accurate method for generating global potential energy surfaces in the majority of cases is through the use of ab initio electronic structure calculations. These give approximate solutions to the Schrödinger equation for many electron systems by implementing either numerical or theoretical approximations. However, despite rapid development in both the overall computational power available to research laboratories, and the algorithmic design of the various techniques employed in electronic structure calculations, this method still remains too computationally prohibitive for anything but the smallest molecular system, necessitating instead, the development of fitting methods that reduce the total number of points needed to generate the potentials while avoiding heavy penalties in the accuracy of predicted energies.

Recently, the latest methods adopted by the scientific community have focussed on using machine learning techniques to learn the potential surface through simple data mapping approaches.[11] The adoption of machine learning tools in the quantum chemistry domain marks a change in the nature of the proposed solution, stepping away from the previous heavy reliance on chemical and physical knowledge of the molecular system under study, attempting to map the input-output space as accurately as possible given a pre-existing set of data points. Examples of such machine learning approaches adopted in the generation of interatomic potentials include neural networks,[12] cubic splines,[13] kernel ridge regression[14] and Gaussian processes[15,16] . The method under focus, Gaussian processes, is a probabilistic learning tool used within a Bayesian framework. It has been established by the statistical inference community as the default machine learning method for modelling an expensive, and usually deterministic, computer simulator (a procedure commonly known as emulating). Since the nature of the problem of producing global potentials from ab initio data is an emulation one, Gaussian processes seem a natural choice when aiming to reproduce electronic structure calculations to high accuracy using a limited data set.

1.3 Previous Research using Gaussian processes

The first pioneering application of Gaussian process regression (aka kriging) techniques to potentials was carried out by the Popelier research group. Water clusters were modelled with explicit treatment of polarisation by combining fluctuating multiple moments derived from Quantum Chemical Topology (QCT) techniques.[17] Three machine learning methods (multilayer perceptrons, radial basis function neural networks and kriging) were pitted against one other to assess their relative computational speed and predictive accuracy. Kriging was found to be the most accurate interpolative method of the three, at the expense of a heightened computational demand as the dimensionality of the data set grew. Following on from their application of Gaussian processes to interatomic energies, the research group has made substantial progress in expanding the use of kriging in the field of quantum chemistry. They have made several developments that include constructing an advanced force field that accounts for anisotropy in polarisation (FFLUX).[18] FFLUX has been successfully applied to obtain the atomic properties ions in aqueous solutions,[19] the energetics of weakly bound complexes[20] and the electrostatic properties of peptide chains as a small example set of its applications.[21]

Bartok and Csanyi were one of the first groups to bridge quantum mechanical results with Gaussian Procceses in the field of material chemistry. They produced interatomic potentials for carbon using Density Functional Theory

(DFT) energies and forces as training data.[22] Total energy was seen as a summation over local energy functionals with learning achieved through the representation of local neighbourhood configurations by dot product powder spectra. Data were trained using Gaussian Process Regression. Gaussian Process Potentials of multiple models was also examined and a two- and three-body model of Si-SW (Stillinger-Weber Silicon potential) data was produced by defining total energy as a sum of pair and triplet energies, where each term was a Gaussian Process in itself and symmetry was encoded into the kernel. Using Smooth Overlap of Atomic Positions (SOAP),[23] bulk phase semiconductors of carbon, silicon and germanium were modelled, with RMSE reported as < 1 meV/ atom.[24]

Gaussian Processes have also been combined with molecular dynamics to develop accelerated techniques to form on-the-fly MD simulations. De Vita studied various crystalline and molten silicon systems,[25] and has found that the use of a growing database of ab initio calculated forces and machine learning methods to make force predictions using pre-existing data can significantly trim the computational cost of an MD simulation, allowing for smaller time steps and longer overall times when conducting dynamical simulations.

The application of Gaussian Processes to produce accurate global potential energy surfaces for polyatomic molecules has been conducted for the system N_4, where results showed that ~ 1500 points saw the same level of accuracy

as a conventional regression fit based on 16421 points.[15]

Gaussian process regression has also been carried out on the $^3Ar'$ state of SH_2: looking at the SH + H $\leftrightarrow$ S(^{2}P) + H_2 abstraction reaction and SH + H' $\leftrightarrow$ SH' + H exchange reaction. The research examined whether Gaussian Process regression required more data to accurately model reactive systems with larger energy ranges and more configurational space that needs covering and found that this was indeed the case. Gaussian process were found to be capable of providing reasonable potential energy surfaces with $\sim 1 \times 10^2$ ab initio points but required a substantially larger number of points $\sim 1 \times 10^3$ to achieve convergence of reaction probabilities.[26]

In full, the following research aims to be a systematic study of the production of global potential energy surfaces, laying the foundation to extend the algorithm to larger and more chemically complex systems in the future. Additionally it aims to fully automate the algorithm to make it easy to routinely apply to various systems. Although significant results have already been achieved for polyatomic molecular systems, this research aims to improve efficacy of the methods even further by accounting for functional non-stationarity (where non-stationarity refers to function's smoothness varying with the input), incorporating molecular symmetry and adding points sequentially to optimise the experimental design.

1.4 Roadmap of chapters

Chapter two gives a theoretical background of ab initio calculations and the different methods that exist. Chapter three describes the nature and origin of interatomic forces, and discusses the ab initio techniques that can be used to calculate them. Chapter four outlines the basic mathematical theory that the proceeding research is built upon, introducing Gaussian Processes as a stand alone concept, before embedding it into a Bayesian framework to demonstrate how it can be used to carry out the learning of highly flexible functions given a small training data set.

Chapter five outlines the basic algorithm used to conduct Gaussian Process inference on small bimolecular van der Waals algorithms using space filling designs and present results for several CO_2 complexes. Chapter six explores the use of iterative experimental design, replacing space filling designs, in a bid to improve the predictive accuracy of the training set by placing points 'informatively'.

Final conclusions are made in chapter seven, overviewing the research as a whole and discussing both its implications in the field of computational chemistry, and the potential it holds for future advancement.

Chapter 2

Electronic Structure Calculation

Research questions: *What are electronic structure calculations? What is a basis set? What approximations are needed to solve the Schrödinger equation for many electron systems? How can we account for electron correlation?*

Keywords: *Schrödinger equation, Born Oppenheimer Approximation, Hartree-Fock, Slater determinants, Self Consistent Field, basis set, electron correlation methods*

2.1 Textbook references

The theoretical backdrop was sourced from: Frank Jensen's 'Introduction to Computational Chemistry (chapter 3 and 4),[27] Quantum States of Atoms and

Molecules,[10] Cristopher J. Cramer 'Essentials of Computational Chemistry' (chapter 4,7,8),[28] J.M. García de la Vega's and B. Miguel's 'Basis Sets for Computational Chemistry' chapter of 'Introduction to Advanced Topics of Computational Chemistry'[29] and lecture notes including C. David Sherill's 'An Introduction to Hartree Fock Molecular Orbital Theory' and Dr Andrew Gilbert's 'Introduction to Computational Chemistry: Theory'.[30]

2.2 Introduction

Chemistry is concerned with the study of properties describing molecular systems, where a molecule can be viewed as a collection of particles (nuclei and electrons) interacting via Coulombic interactions. A Coloumb interaction between two particles is given by:

$$V_{ij} = \frac{q_i q_j}{4\pi\epsilon_o r_{ij}} \tag{2.1}$$

where q_i and q_j refer to the point charges of the two molecules, r_{ij} is the distance between the particles and ϵ_o is the dielectric constant in a vacuum.

In a molecule the electrons are in constant motion, whereas the dynamics of large bodies can be described by the deterministic laws of classical mechanics, the dynamics of electrons must be described using quantum mechanical postulates.

Broadly speaking, quantum mechanics postulates that all physical properties of a microscopic system can be described by the wavefunction, which is a mathematical formulation describing the wave characteristics of an electron. The magnitude of the wavefunction at a specific electron location can be related to the probability of finding the electron at that point in space. The physical observables can be derived by the application of the corresponding operators to the wavefunction.

In quantum chemistry, the electronic structure refers to the probability distribution of all electrons in a given chemical system. This electronic structure is obtained by solving the time-independent Schrödinger equation associated with the electronic molecular Hamiltonian. The time-independent Schrödinger equation for a system of particles is:

$$\hat{\mathcal{H}}\psi = E\psi \tag{2.2}$$

where $\hat{\mathcal{H}}$ is the Hamiltonian operator (an operator associated with observable energy), E is the total energy of the system and ψ is the wavefunction. The Hamiltonian operator includes the kinetic and potential energy parts for both the electrons, and the nuclei. The potential energy component includes electron-electron V_{ee}, electron-nucleus V_{en} and nucleus-nucleus V_{nn} parts.

$$\hat{H}_{total} = \hat{T}_n + \hat{T}_e + \hat{V}_{ne} + \hat{V}_{ee} + \hat{V}_{nn} \tag{2.3}$$

A time-independent Schrödinger equation refers to an equation that is not explicitly dependent on time and requires the time dependency of the total wavefunction, Ψ, to be separated out by a simple phase factor:

$$\Psi(r,t) = \psi(r)e^{\frac{-iEt}{\hbar}} \tag{2.4}$$

where $\hbar$ is Planck's constant, E is the energy value and t is time.

Ab initio, derived from Latin, translates to 'from first principles'. In the domain of computational chemistry, ab initio methods refer to theories that are derived from exact quantum mechanical equations and that are used in the absence of experimental data and (in most instances) in conjunction with numerical approximations. Numerical approximations are required in most instances because the Schrödinger equation can only be solved exactly for a small subset of systems: rigid rotor, harmonic oscillator, particle in a box, a hydrogen atom and a hydrogenic-like ion.

As a whole, there are three parts to solving the Schrödinger equation for many electron systems:

1. The Born-Oppenheimer approximation
2. Expansion of the wave function into Slater determinants
3. Representation of the Slater determinants by a linear combination of atomic orbitals

1. Born-Oppenheimer Approximation

The wavefunction is a function of the nuclear and electron coordinates. The motion of the two are coupled, but due to their relative speeds and masses, the nucleus is essentially 'fixed' in position compared to the moving electrons. Because of this, the electronic wavefunction (equation 2.3) can be solved for a fixed set of nuclear coordinates and the total wavefunction can be obtained by combining this with the solution of the nuclear Schrödinger equation.

$$\Psi(\boldsymbol{r}, \boldsymbol{R}) = \Psi_{el}(\boldsymbol{r}, \boldsymbol{R})\Psi_N(\boldsymbol{R}) \tag{2.5}$$

where r denotes the electron coordinates and R, the nuclear coordinates. Since the nucleus is fixed in position under the Born-Oppenheimer approximation, the electronic wavefunction, $\Psi_{el}(r, R)$, is a function of the electron coordinates with nuclear positions (R) being the parameters that characterise the wavefunction.

$$\hat{\mathcal{H}}_{el}\Psi_{el}(\boldsymbol{r}, \boldsymbol{R}) = E_{el}\Psi_{el}(\boldsymbol{r}, \boldsymbol{R}) \tag{2.6}$$

$$\hat{\mathcal{H}}_N \Psi_N(\boldsymbol{R}) = E_N \Psi_N(\boldsymbol{R}) \tag{2.7}$$

Overall, the Born-Oppenheimer approximation is reasonable for most instances, and only begins to fail when two or more electronic states reside close in energy for a particular nuclear configuration. This requires a "non-adiabatic" wavefunction (a product of nuclear and electronic wavefunctions).[31] This approximation also ignores relativistic effects, and electron spin has to be accounted for in an ad hoc fashion, where the spin dependent terms such as spin-orbit and spin-spin coupling must be added as corrections, after a solution to the Schrödinger equation has been obtained.[32]

2. Slater Determinants

In the domain of quantum mechanics, a Slater Determinant is a determinant of one-particle spin orbitals satisfying the symmetry requirement of the Pauli exclusion principle, which states the wavefunction must change sign upon interchange of any set of electrons, or exchange of spin-spatial coordinates. The spin-spatial coordinates refer to the three spatial degrees of freedom and one intrinsic degree of freedom known as spin, denoted by either α or β.

Since electrons need to be mathematically indistinguishable, but the Pauli exclusion principle needs to be satisfied, the wavefunction can be anti-symmetrised by mathematical manipulation. This is achieved through the use of Slater detemininants, since interchanging any two rows or columns in a Slater determi-

nant leads to a change in the sign of the determinant, For a two-electron case, the symmetrised two electron wavefunction can be written in the following form:

$$\Psi(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{1}{\sqrt{2}}[\chi_1(\boldsymbol{x_1})\chi_2(\boldsymbol{x_2}) - \chi_1(\boldsymbol{x_2})\chi_2(\boldsymbol{x_1})] \tag{2.8}$$

where $\chi_i(\boldsymbol{x_i})$ refers to a spin orbital, which is a product of the spatial orbital $\phi(\boldsymbol{r})$, known as the molecular orbital, and the α or β spin functions.

The above functional form can also be written in the following form:

$$\Psi(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{1}{\sqrt{2}} \begin{vmatrix} \chi_1(\boldsymbol{x_1}) & \chi_2(\boldsymbol{x_1}) \\ \chi_1(\boldsymbol{x_2}) & \chi_2(\boldsymbol{x_2}) \end{vmatrix} \tag{2.9}$$

Generalising to a system with N electrons, the total electronic symmetrised wavefunction can be written as:

$$\Psi(\boldsymbol{x_1}, \boldsymbol{x_2}...\boldsymbol{x_N}) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(\boldsymbol{x_1}) & \chi_2(\boldsymbol{x_1}) & \dots & \chi_N(\boldsymbol{x_1}) \\ \chi_1(\boldsymbol{x_2}) & \chi_2(\boldsymbol{x_2}) & \dots & \chi_N(\boldsymbol{x_2}) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\boldsymbol{x_N}) & \chi_2(\boldsymbol{x_N}) & \dots & \chi_N(\boldsymbol{x_N}) \end{vmatrix} \tag{2.10}$$

The Slater determinants also satisfies the independent particle model, which requires the total electronic wavefunction of a molecule to be a sum of mono-electronic functions and the repulsion between electrons to be ignored. The-

one electron functions in an atom are known as the atomic orbitals. In a molecule, the one-electron functions are known as molecular orbitals.

3. Linear Combination of Atomic Orbitals

With the use of Slater determinants, there is still the imperative choice of the form of molecular orbital to be used during the calculation.

In 1951 Roothan and Hall[33 34] proposed constructing orbitals as a summation of known one-electon functions, known as basis functions. In the case of molecules, molecular orbitals ψ_i are constructed as a further linear combination of atomic orbitals:

$$\psi_i = \sum_{\mu} C_{\mu i} \chi_{\mu} \tag{2.11}$$

where χ_μ are the atomic orbitals, C_μ are the molecular orbital coefficients and μ is the number of atomic orbitals that are combined.

The number of molecular orbitals is equal to the number of atomic orbitals involved in the linear expansion. So if there are n atomic orbitals involved in the superposition, there will also be n molecular orbitals. The larger the expansion, the greater accuracy of the described molecular orbital overall and the more expensive the calculation will become.

If an electronic structure calculation method is variational, then the approximate energy will always be of higher or equal value to the true energy of the system. Therefore any variation in the wavefunction that lowers the expectation value of energy, brings the approximate energy closer to the actual value. This allows for a systematic iterative improvement of the wavefunction by varying the MO coefficients in the linear combination of atomic orbitals (LCAO) equation until the energy is minimised.

2.2.0.1 Basis Sets

The basis functions are atom-centred functions which act as approximate representations of atomic orbitals. The basis set is a collection of these functions, and the selection of which basis set to use is especially important as it needs to approximate the wavefunction well enough to give chemically meaningful results. The calculations also need to be computationally reasonable in cost, so there must be a balance between the speed of integral evaluation and accuracy of results, with a choice of basis set that is right for the molecular system being modelled.

An obvious choice for a mathematical description of an atomic orbital is the exact Schrödinger solution for a hydrogen atom. The solution is a Slater-type orbital (STO) and has both the correct long and short term behaviour. However due to their problematic usage in computation, Gaussian-type orbitals (GTO) are often used instead (see figure 2.1 for orbital types). Single GTOs

are too flat near the nucleus and fall off too quickly at longer distances. Such deviations from the STO is usually accounted for by the use of several GTOs linearly combined, known as contracted GTOs. Contracted Gaussian orbitals are usually used for inner orbitals and simple GTOs are usually used for outer shells.

There are two categories of a basis set: minimal and extended. Minimal basis sets only offer the most basic description of the basis function (e.g. one function for each angular momentum component) whereas extended basis sets offer a greater description and typically include multiple functions for each angular momentum component. The larger the basis sets, the more accurate the approximation of the molecular orbital, as fewer restrictions are imposed on the location of the electron in space.

Single (minimal) basis sets contain only one basis function for each occupied orbital. Double zeta and triple zeta basis sets use two and three basis functions for each orbital, respectively. Due to the higher chemical importance of the valence electrons relative core ones, split-valence basis sets are often employed and valence electron orbitals are defined by more basis functions than the core orbitals.

Polarised basis sets allow orbitals to become polarised by permitting more asymmetry around the nucleus. As one atom approaches another, an atom's

orbital may shift to one side. This can be represented by the mixing of atomic orbitals with an orbital of a higher momentum (e.g. mixing an s orbital with a p orbital or mixing a p orbital with a d orbital). Generally speaking, to polarise a basis function with angular momentum l, it is mixed with a basis function with angular momentum $l + 1$. Polarised basis functions are important for accurately describing bonding between atoms.

Diffuse functions can be used to augment a basis set by using Gaussian functions with a small exponent. They aim to describe the "tail component" of the atomic orbital by decaying less quickly with distance. Diffuse basis functions are an important choice for anions and Rydberg excited electronic states. They are also a valid choice for electronegative anions and van der Waals complexes.

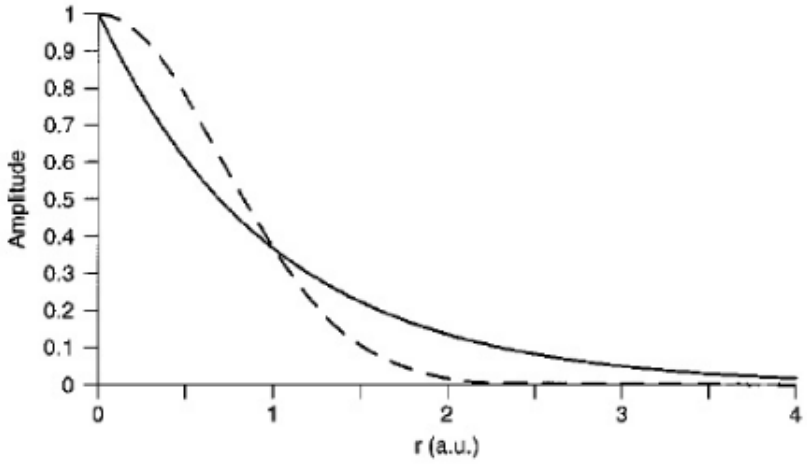


Figure 2.1: A plot of GTO and STO type orbitals. The solid line is the radial decay of STO nd the dashed line is the radial decay of GTO Image sourced from Essentials of Computation Chemistry, Chris Cramer, Wiley, 2nd Edn,

2.3 Different Electronic Structure Methods

2.3.1 Hartree-Fock Method

The Hartree Fock (HF) method is a fundamental method in calculating electronic structure and forms the starting step for other more accurate methods that describe many-body potentials.

The HF method relies on a conceptual approximation, which models the electrons as moving independent of each other and only perceiving an average interaction over all other electrons (the central field approximation). Under this approximation, the Schrödinger equation may be alternatively written as:

$$(K + V_{e-n} + V_i^{eff})\psi_i = E_i\psi_i \tag{2.12}$$

where K is the kinetic energy operator of the electrons, V_{e-n} is the potential energy operator for electron - nucleus interaction and V^{eff} is the effective field of all other electrons.

The HF theory describes each electron by an atomic orbital and describes total wavefunction as a product of orbitals, breaking down the many-electron Schrödinger equation into many simpler one-electron equations. The wavefunction is approximated by a single Slater determinant consisting of one spin

orbital per electron. The accuracy of the results depends on the size of the basis set used and the choice of basis functions. However, due to the central field approximation, the Hartree Fock energy will always be greater than the exact energy. The calculated energy will tend to the basis set limit with increasing basis set size, allowing for systematic improvement of results using an iterative procedure known as the Self Consistent Field (SCF) procedure.

The SCF procedure is an iterative procedure used to deal with the effective field's dependency on the wavefunction. An initial set of one electron wavefunctions is guessed. This is used to calculate the effective potential, which in turn, produces a new set of wavefunctions that forms a new effective potential. The cycle is repeated until convergence is reached and the set of wavefunctions no longer changes.

2.3.2 Electron Correlation Methods

One limitation of the Hartree-Fock Method is that electron-electron interactions are instead replaced by an average interaction over all electrons, with no explicit treatment of electron correlation. This introduces an error into both the wavefunction and calculated energy, and is especially problematic for van der Waals complexes, since dispersion forces are a direct result of electron correlation effects.

Electron correlation methods account for instantaneous electron-electron re-

pulsion. The electron correlation energy, E_{corr} is the difference between the calculated energy using an exact ψ and using the Hartree-Fock model ψ_{HF}:

$$E_{corr} = E_{exact} - E_{HF} \tag{2.13}$$

Electron correlation methods are known as post-Hartree Fock methods and go beyond the *ansatz* of a single determinant wavefunction. The three most mainstream methods include pertubation theory, coupled cluster techniques and configuration interaction.

2.3.2.1 Perturbation Methods

The perturbation methods are built from the assumption that the problem to be solved (correlated ψ and E) is similar to one already solved (the Hartree-Fock ψ and E). Electron correlation is therefore treated as a small perturbation from the Hartree-Fock energy. In terms of the perturbation theory formulation, the Hartree-Fock wavefunction becomes the first-order perturbation and MPn pertubations (where n is the order ranging from 2 to 5) are the succeeding order perturbations, with progressively more corrections and computational cost, the greater the order.

2.3.2.2 Configuration Interaction

The configuration interaction (CI) approach expresses the wavefunction as a multiple-determinant wavefunction (a Hartree-Fock determinant and other de-

terminants produced by the promotion of electrons into unoccupied orbitals and whose coefficients are found using the variational principle). The CI method is classed in terms of the number of excitations used to produce the determinants. CIS, CISD, CISDT and CISDTQ refer to configuration interactions with single, double, triple and quadruple excitations respectively

2.3.2.3 Coupled Cluster

The coupled cluster correlation method expresses the wavefunction as an exponential product. The method stems from the idea that accounting for the interaction of one electron with more than one electron is unlikely to be of importance, and focuses on simultaneous "pair interactions" known as "disconnected clusters".

2.4 Density Functional Theory

Density Functional Theory (DFT) is often classed as an ab initio method, but utilises functionals (functions of a function) that are derived using experimental data or higher level ab initio calculations. The DFT method sidesteps wavefunctions in full, and instead expresses energy in terms of total electron density. As with the other ab initio methods, the method starts with the Born Oppenheimer approximation and then expresses the Hamiltonian as a sum of an external field, $\hat{\mathcal{V}}_{ext}$ and $\hat{\mathcal{F}}$, denoting the kinetic energy of the electrons and electron-electron interactions.

$$\hat{\mathcal{H}} = \hat{\mathcal{F}} + \hat{\mathcal{V}}_{ext} \tag{2.14}$$

The external potential is a unique functional of electron density and has a trivial derivation:

$$V_{ext[p]} = \int \hat{\mathcal{V}}_{ext} p(r) dr \tag{2.15}$$

However, the kinetic energy and electron-electron functionals are not known. If a good appproximation for the two functionals could be found, then direct energy minimisation is possible. The chemical interpretation of DFT was simplified through the Kohn-Sham formulation[35] where non interacting fictitious molecular orbitals are introduced. This procedure is the most common DFT approach but does not work well for large systems due to the extensive computational cost.

DFT is a popular method in computational chemistry but struggles to describe dispersion forces due to the functionals not capturing the long range correlation. Therefore DFT is not the best choice for molecular systems that are dominated by these weaker forces.

Chapter 3

Intermolecular Forces

Research questions: *What are intermolecular forces and what are their origins? How were intermolecular forces discovered? What properties do intermolecular forces influence? What are the different types of intermolecular forces and what are their relative strengths? What are the two ways to compute intermolecular interaction energy?*

Keywords: *Electrostatic, attractive, repulsive, dispersion, inductive, supermolecule, perturbation theory, basis set superposition errors*

3.1 Textbook references

The theoretical backdrop was sourced from: two papers on intermolecular forces,[36],[37] J.S. Rowlinson's 'Cohesion: A Scientific History of Intermolecular Forces',[38] Jacob N. Israelachvili's 'Intermolecular and Surface Forces'[39] and

A.J Stone's 'Theory of Intermolecular Forces'.[40]

3.2 Introduction

All physical processes and structures observed in nature are believed to stem from just four fundamental forces: gravitational, electromagnetic, weak and strong interaction. Gravitational forces are extremely long range, only attractive and may not be transformed or absorbed. Strong and weak interactions operate at a very short range of less than 10^{-5}m). Both these interactions belong to the domain of high-energy and nuclear physics, with weak interactions being responsible for nuclear phenomena such as fusion and strong interactions being the force that binds quarks into protons and neutrons. On the molecular scale, electromagnetic forces dominate and are responsible for the interaction between charged particles. They also underpin the chemical bonds that keep atoms together in molecular systems and the intermolecular forces between atoms and molecules.

The evidence for the existence of intermolecular forces stems from simple observations. The existence of attractive forces can be deduced from the presence of condensed phases of matter. Without the force of attraction, water in a vessel would have no reason to be confined to the container. Without the force of repulsion, there would be no difficulty in compressing water and a liquid would have no definite density. This coexistence of attractive and repulsive forces

was assumed by van der Waals[41] when he derived the Equation of State laws. These laws mathematically captured gases deviating from ideal behaviour. Van der Waals' equation included corrections for the reduced pressure of the gaseous molecules exerting on the container due to the existence of attractive forces between molecules and for the finite volume of space occupied by the molecules due to repulsive forces.

Van der Waals' equation included corrections for attractive forces between molecules (reducing the pressure that gaseous molecules exert on the container), and repulsive forces (by including a finite volume of space occupied by the molecules):

$$(P + \frac{an^2}{V^2})(V - nb) = nRT \tag{3.1}$$

This attractive-repulsive force coexistence also means that the energy of interaction $U(R)$, as a function of the distance, R, between molecules, will take the form of Figure 3.1 where there is a steep repulsive wall (accounting for the low compressibility of condensed phase materials), a region where the force $(-dU/dR)$ is negative and the two molecules experience an attractive force and a minimum energy at a distance R_M, corresponding to the most stable configuration. In line with standard notation, the well depth is denoted by ϵ

and the distance at which the energy of interaction is zero is σ.

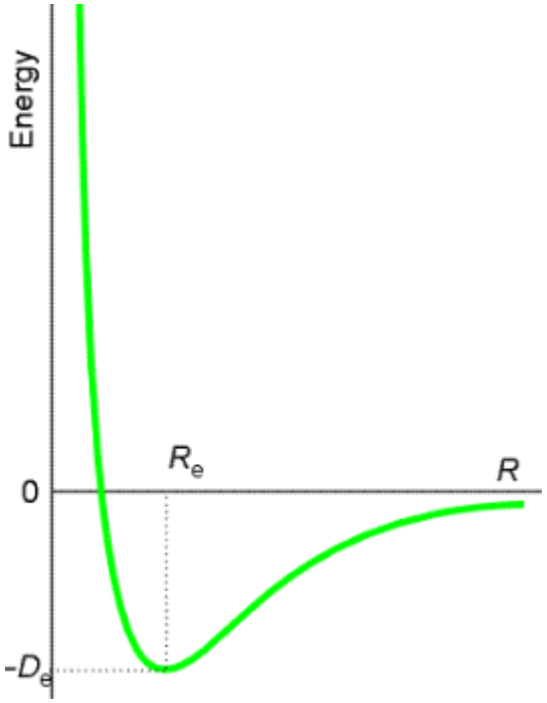


Figure 3.1: Energy of interaction as a function of distance between two molecules. Image sourced from:
http://staff.um.edu.mt/jgri1/teaching/che2372/notes/10_st/04/v_surfaces.html

3.2.1 Origins of intermolecular forces

Intermolecular forces are weak, non-covalent molecular interactions, giving rise to bonding energies of 1-20 kJ/mol. The physical properties that intermolecular forces underpin include the melting and boiling temperatures, the vapour pressure, evaporation, viscosity, surface tension and solubility. They are responsible for the deviation of gases from ideal behaviour and play a role in the properties of condensed phases, the tertiary structure of proteins,[42] the structure of polymorphs and liquid crystals,[43] surface-molecule interactions

and the transportation of molecules and ions across biological membranes.

The development of quantum theory in the 1920s[44] and the elucidation of electronic structure, allowed for the origins of intermolecular forces to be properly examined. The Hellman-Feynman theory[45] allowed the intermolecular forces to be calculated using classic electrostatic theory, once electronic distributions were solved using the Schrödinger equation.

Collectively, all intermolecular forces are known as van der Waals forces, but may also be described in terms of its individual components. The first, broad classification of intermolecular forces are the long- and short-range forces. The short-range forces include electron exchange and are responsible for steric hindrance. They fall off exponentially with the distance between molecules and are repulsive when the interacting monomers are closed-shell. The long-range forces are mainly attractive and include electrostatic, dispersion and induction forces. They fall off with a factor R^{-n} where n can range from 3 to 10, depending on the type of intermolecular force.

The origin of short range repulsive forces can be explained by examining what happens when two electronic clouds come close enough for overlap to occur. Due to the Pauli exclusion principle, electrons cannot occupy the same space. Therefore the electron density (and as a direct result, the mutual shielding of the two nuclei) decreases, leading to a repulsive force.

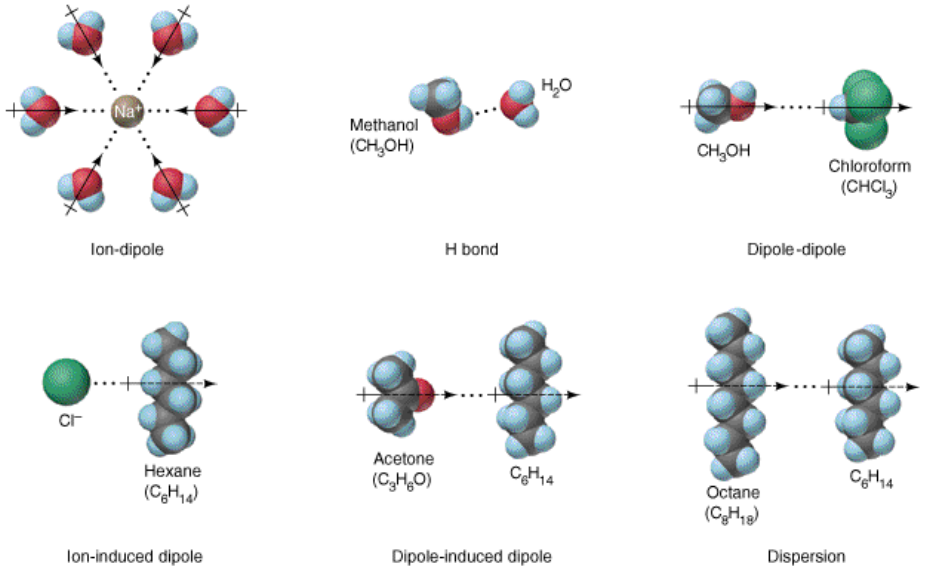


Figure 3.2: Different types of intermolecular forces. Image is sourced from: http://www.science.uwaterloo.ca/ cchieh/cact/c123/intermol.html

When the distance between molecules is considerable larger than the dimensions of the molecule, the interaction is largely determined by electrostatic energy resulting from the interaction of permanent electric moments (charge, dipole, quadrupole, etc) belonging to the isolated molecules. These permanent electric moments are also able to polarise nearby molecules by distorting their electron structure, leading to induction energy, whose value is dependent on the polarisability of the molecules. Neutral atoms and molecules are also interacting through the creation of instantaneous dipoles due to the non-spherical distribution of electrons around the nucleus. The latter case is the only intermolecular force present in all systems.

3.2.1.1 Dipole-Dipole interactions

Dipole-dipole forces, otherwise known as Keesom forces,[46] were discovered by Willem Hendrik Keesom who produced a mathematical description of the interaction in 1921. Dipole-dipole interactions refer to the forces between monomers with permanent multipole moments. They are due to charge distributions within the molecule creating partial negative and positive charges. Energy of interaction between the molecules is orientation-dependent and since no distortion of the electron cloud is needed, it is classed as a first order effect.

3.2.1.2 Dipole-Induced Dipole interactions

Dipole-induced dipole interactions, otherwise known occur when a molecule with a permanent dipole polarises a non-polar molecule and gives it non-permanent dipole moment, which then interacts with the inducing dipole to produce an attractive force between the pair. This type of intermolecular interaction requires a distortion of the electron cloud in the non-polar molecule and is thus classed as a second order effect.

3.2.1.3 Induced Dipole -Induced Dipole interactions

Before the dawn of quantum mechanics, the origins of intermolecular forces in closed-shell atoms were completely enigmatic. Dispersion forces are quantum mechanical in origin and account for a range of physical phenomena including

adhesion, physical adsorption, condensed macromolecular structure, properties of liquids and thin films. The induced dipole induced dipole interactions are otherwise known as London dispersion forces in accordance with their discovery by the scientist Fritz London.[47]

Although a neutral atom may have a time averaged dipole of zero, in one particular instant the electrons may be distributed around the nucleus in such a way that a finite instantaneous dipole is produced. This instantaneous dipole produces an electric field, which polarises a nearby neutral atom and induces an instantaneous dipole in it. The overall result of this is a momentary attractive force between the two atoms, with the force having a finite time-average value.

3.2.2 Computing intermolecular forces

Considering two spherically-symmetrical atoms A and B, at infinite separation the total energy is the sum of the energies of the isolated atoms.

$$E_{tot}(\infty) = E_A + E_B \tag{3.2}$$

For any finite interatomic distance between the two atoms, R, the total energy becomes:

$$E_{tot}(R) = E_A + E_B + U(R) \tag{3.3}$$

where U(R) is the interaction-dependent contribution, the pair potential. The pair potential is the difference between the energy of the system A + B at a specific configuration and isolated particles A and B brought to infinite separation.

$$U(R) = E_{tot}(R) - E_{tot}(\infty) \tag{3.4}$$

For a polyatomic case, the interaction energy becomes dependent on the relative orientation, relative positions and the interatomic coordinates.

In itself, the interaction energy is not an observable but can be calculated either experimentally or theoretically.

Two common approaches to producing ab initio values of interaction energy:

- **Supermolecule approach**, where interaction energy is computed as the difference in energy between the supersystem and the two isolated monomers.
- **Perturbation approach**, which starts with the unperturbed molecules and treats the interaction energy and corresponding wavefunction as small quantities resulting from the pertubation of the molecules via non bonding interactions.

Since all proceeding research will be based around the use of the supermolecule approach, it is the former method that is described in detail.

3.2.3 Supermolecule approach

The supermolecule approach produces values for the interaction energy by subtracting the energy of the individual monomers from the energy of the complete system. The calculations need to be performed at high accuracy as the interaction between constituent monomers is going to be very small compared the energies of the monomers and the dimer. Correlation also needs to be accounted for in the calculation, otherwise the dispersion forces are ignored.

Outside of the importance of performing high precision calculations, two other drawbacks for the supermolecule approach is that:

1. it produces a total value for the interaction energy rather than values for each of the contributions
2. basis set superposition errors must be accounted for

3.2.3.1 Basis set superposition error

Basis set superposition errors (BSSE) can result in the artificial shortening of the intermolecular distances and strengthening of the intermolecular forces. The basis set superposition error comes from the finite nature of the basis set used. In a dimer, as one of the isolated monomers, A, approaches the other monomer, B, the dimer is artificially stabilised as each of the monomers are stabilised by the extra basis functions on the other one.

The smaller the basis set, the more pronounced the effect of the basis set superposition error. BSSE is not strictly speaking an error, but rather an effect of the inconsistent treatment of the monomer's basis sets. Even after eliminating the inconsistency, there is still the error associated with an incomplete basis set. In the limit of a complete basis sets, both the basis set incompleteness and basis set superposition errors are equal to zero.

One way of dealing with the BSSE is the Boys and Bernardi[48] counterpoise correction. A normal supermolecular approach to the calculation of interaction energy of a dimer AB, $\triangle E_{INT}(AB)$, is given by the following equation:

$$\triangle E_{INT}(AB) = E_{AB}^{AB}(AB) - E_{A}^{A}(A) - E_{B}^{B}(B) \tag{3.5}$$

where the superscript denotes the basis sets used, the subscript denotes the geometry and the contents in the bracket denotes the chemical system examined.

The counterpoise correction procedure involves the use of dummy atoms. So to evaluate the energy of monomer A, the basis functions of B are placed on the atomic centres of B and the electrons and nuclear charges of B are ignored.

The equation for the interaction energy then becomes:

$$\triangle E_{INT}(AB) = E_{AB}^{AB}(AB) - E_{AB}^{A}(A) - E_{AB}^{B}(B) \tag{3.6}$$

The nature of dispersion forces needs to be considered to choose the best method for calculating intermolecular potentials of weakly bound systems. HF ignores electron correlation and consequently predicts no van der Waal attractive forces between molecules. Similarly, DFT only accounts for 'local' correlation and does not encompass the non-local electron correlation underlying dispersion forces. Therefore, post-HF methods need to be used when modelling van der Waals systems. Second-order Møller–Plesset perturbation theory is a good choice to give decent accuracy at a relatively decent computational cost. CCSD(T) produces calculations of significantly higher accuracy, but are time-consuming even for small systems.

Chapter 4

Gaussian Processes and Bayesian Learning

Questions: *What are Gaussian Processes? How can they be used to conduct machine learning tasks? What is a covariance and mean function and what role do they have in function prediction? What is Bayesian learning and how does it differ from classical machine learning?*

Keywords: *Gaussian Process, covariance function, mean function, bayesian inference, hyperparameters*

4.1 Textbook references

The theory was sourced from the seminal book on Gaussian Processes by Rasmussen[49] and miscellaneous notes by M. Ebden[50] and Hanna M. Wal-

lach.[51]

4.2 Two approaches to learning

Suppose we have some data from an underlying function that we are attempting to infer (i.e. potential energy as a function of nuclear geometry). One way of generating this underlying function is to restrict the search to one functional class (e.g. linear or quadratic). The corresponding best-fit parameters could then be found through a fitting method that attempts to minimise the error of training data predictions (e.g. least squares method in the case of linear regression). This approach (e.g. force field method) is an example of parametric modelling, whereby the number of parameters remain unchanged for any given model, and once the optimal parameters are selected, the data points are no longer needed to make predictions. The overall result is that adding any more data points to the model would not increase the accuracy of predictions at all. The main pitfall of this method is that our assumptions about which functional class best represents the data may be incorrect. This results in poor fits either because it is too simplistic of a functional form to accurately capture the data trend, or the fixed parameter number leads to over-fitting of the data. The overall effect of this is bad predictive performance outside of regions covered by training data points.

An alternate way of inferring the underlying function is by not placing any re-

strictions on the functional class itself, but by examining all possible functions, making assumptions about which functions are more likely than others, and then ultimately selecting the function that is most likely to have produced the training data set. Although this bypasses the aforementioned drawbacks of fixed parameters and functional class assumptions, it introduces a new major theoretical problem: the mathematical infeasibility of assigning probabilities over an infinite amount of functions. This infeasibility can in principle be bypassed by implementing Gaussian Processes in a Bayesian framework, where the Bayesian framework relates to a field of statistics built around defining probability in terms of a prior belief (a belief before making any observations) that is updated with data.

4.3 Bayesian Learning

Broadly speaking, the Bayesian approach to inference can be summed up in the following steps:

1. Define a model, conditioned on unknown parameters, using qualitative knowledge
2. Select a prior probability distribution over the unknown parameters
3. Incorporate the data into the model
4. Compute or approximate a posterior distribution over the parameters, given the data

5. Carry out prediction by averaging over the posterior distribution

The corner stone equation that forms the backbone of Bayesian inference is the Bayes' equation, which was founded $\sim$ 200 years ago by mathematicians Bernoulli, Bayes and Laplace. In mathematical terms, the Bayesian Inference is based on the manipulation of a few simple rules of probability (the conditional and marginal probability laws):

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned} \tag{4.1}$$

where $P(A)$ is the probability of A happening, $P(B)$ is the probability of B happening $P(A \cap B)$ is the probability of and $P(A|B)$ is the conditional probability of A happening given B happens.

In Bayesian regression, one of the events is defined as the model, M, and the other is the data, D. When conducting Gaussian Process Bayesian learning, instead of looking at the probability of a finite set of models, we use a probability distribution over an infinite number of functions before and after data is incorporated. Relating this to the intermolecular potential energy surface, the model will be the potential energy function as a function of geometry, and the data will be a set of known nuclear configurations and corresponding energy

values. The cornerstone equation becomes:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \tag{4.2}$$

where $P(M|D)$ is the posterior distribution of functions (defining the probability of the potential energy functions after considering the data), $P(D|M)$ is the likelihood function (defining the probability of the data arising from the functions) and $P(M)$ is the prior function distribution which reflects the prior knowledge of the function before any data has been introduced.

Since the Bayes' rule is a mathematical relationship that shows how we can update our prior beliefs after being presented with some data, it is inherently a rule that defines learning.

Now there is a method to update our prior probability distribution over functions after the introduction of data into the model, we need to just find a way of defining our prior assumptions about the functions we are likely to see. Gaussian processes provide us a mathematically trivial way of doing just that.

A Gaussian Process is a probability distribution over functions, completely defined by a mean and covariance function. Gaussian Processes are extensions of Gaussian distributions into infinite dimensionality, analogous to joint mul-

tivariate Gaussian distributions being generalisations of univariate Gaussians into finite higher dimensions.

4.4 Gaussian Distributions

4.4.1 Univariate Gaussian Probability Distributions

A Gaussian Probability distribution, (alternatively known as a normal or bell curve distribution) is one of the most common continuous probability distributions in the field of statistics. It has numerous unique mathematical properties that makes it popular in probabilistic machine learning due to tractability and easy manipulation. They key property is that the conditional and marginal distribution of a joint Gaussian distribution is also a Gaussian distribution.

A random variable, X, is said to be normally distributed, with mean, μ and variance, σ^2, if its probability distribution function is equal to:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)}{2\sigma^2}} \tag{4.3}$$

or in shorthand notation:

$$X \sim N(\mu, \sigma^2) \tag{4.4}$$

where $\sim$ denotes 'distributed by' and N denotes a Gaussian distribution.

The result is a bell shaped curve seen in figure 4.1. The variance σ^2 defines the spread of the distribution around the maximum (mean): the larger the variance, the greater the spread.

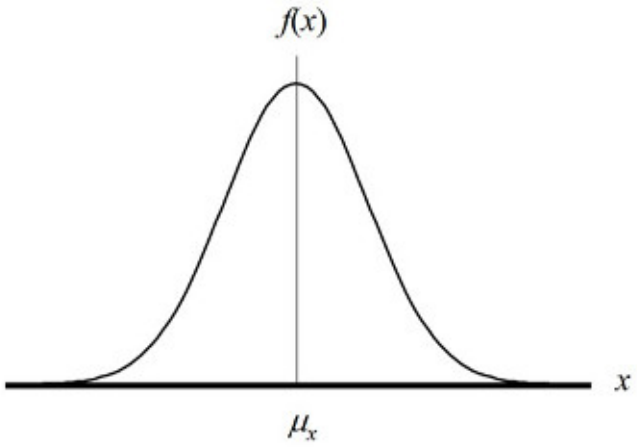


Figure 4.1: A 1D Gaussian distribution. Image sources from http://www.isobudgets.com/probability-distributions-for-measurement-uncertainty/

4.4.2 Multivariate Gaussian Distributions

Extending the univariate distribution into n, dimensions gives a probability density function defined as:

$$p(X;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \tag{4.5}$$

where X is a $n \times 1$ random vector, $|\Sigma|$ is the determinant of the covariance

matrix Σ and Σ^{-1} is the inverse of the covariance matrix.

In the multivariate case, the Gaussian distribution is now fully defined by a mean vector μ and an $(n \times n)$ covariance matrix. The shorthand notation of a multivariate Gaussian distribution is given by:

$$\mathbf{X} \sim n(\mu, \Sigma) \tag{4.6}$$

4.5 Gaussian Process

The leap from Gaussian distributions to Gaussian processes is made when extending the multivariate Gaussian distribution to infinite dimensionality. Just as an n-dimensional distribution is a probability distribution of n-variables, fully expressed by a mean vector of length n and an $(n \times n)$ covariance matrix, an ∞-dimensional Gaussian distribution is a distribution over infinite (∞) variables, fully expressed by a mean vector of length ∞ and an $(\infty \times \infty)$ covariance matrix. Noting that an infinitely long vector may be alternatively viewed as a function, an ∞-dimensional Gaussian distribution can be seen as a probability distribution over functions, fully defined by a mean and covariance function.

The notion of infinitely long vectors may seem mathematically prohibitive when dealing with inference, but the marginalisation property means we only

ever have to deal with finite-dimensional quantities, and calculating the properties of a function at a finite set of data points yields the same result whether you consider only the finite set or take into the account the infinite amount of points. The overall result: we only ever need to define a distribution over the function values at a finite, arbitrary set of points.

4.5.1 Mean function

In the domain of calculus, the mean value is broadly defined as the average value of a function over its domain. In Gaussian processes, each input has a corresponding mean value and the mean function is a function of the mean over the domain. Due to mathematical simplicity, the mean function is usually set to zero. In the case of learning potential energy surfaces, zero mean is most problematic around the high energy region as in the extrapolative region there are no further data points, it would eventually fall back to zero and produce unphysical results. One way to circumvent this is through the use of an asymptotic function and the avoidance of extrapolation in the high energy region.

4.5.2 Covariance Function

In the domain of statistics and probability, the covariance is a measure of how two random variables changes alongside each other. In Gaussian Process

regression (GPR) tasks, the covariance function encodes all the key assumptions about the underlying function we are trying to learn and is the crucial ingredient of GP learning. Loosely, it can be seen as a statistical relationship between two points in input space (x,x') or in broader terms, a similarity metric across input space. Intuitively, points closer to each other in space are expected to be more similar in value to each other than those further apart in space, and this quantitative measure of nearness and similarity is given by the kernel (covariance function).

Covariance functions are parametrised in terms of hyperparameters, and tuning these are a key step in obtaining good predictions using a chosen kernel. The length scale is roughly described as the distance you must move in input space before the function value can change significantly. With short length scales, the error grows rapidly further away from data points and large length scales signify non-important features. Exampes of sample functions generated with different hyperparameters can be seen in Figure 4.2 and Figure 4.3.

4.5.2.1 Squared exponential

The most popular kernel in GPR is the squared exponential function, which produces smooth, universal and infinitely differentiable prior and posterior functions. It is alternatively known as both the radial basis function and the Gaussian kernel, and is given by:

$$K_{SE}(x, x') = \sigma^2 exp(\frac{-(x - x')^2}{2l^2}) \tag{4.7}$$

where l is the lengthscale and σ^2 is the scale factor that defines the average distance of the function from the mean.

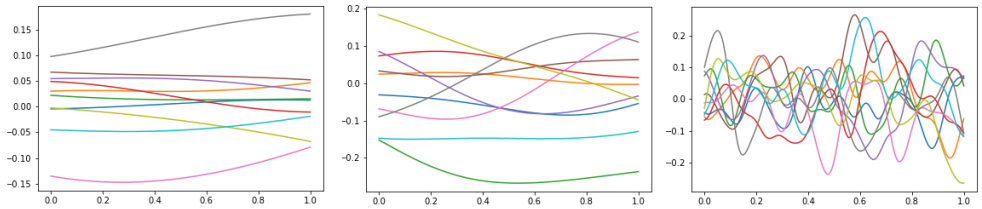


Figure 4.2: Sample prior function of a Squared exponential covariance function with three different lengthscales (1, 0.5 and 0.05 from left to right, with variance set to 1×10^{-2})

4.5.3 Matérn Covariance Function

The Matérn class of covariance functions is given by:

$$K_{Matern}(x, x') = \sigma^2 \frac{2^{1-v}}{\Gamma(v)} (\sqrt{2v}\frac{d}{p})^v K_v(\sqrt{2v}\frac{d}{p}) \tag{4.8}$$

where d is distance, $\Gamma(v)$ is a Gamma function, K_v is the modified Basel function and p and v are positive parameters. As v approaches ∞, the covariance function transforms to a squared exponential function. The Matérn covariance function may be seen as a generalisation of Gaussian Radial Basis functions and are recommended for modelling more realistic functions without infinite differentiability.

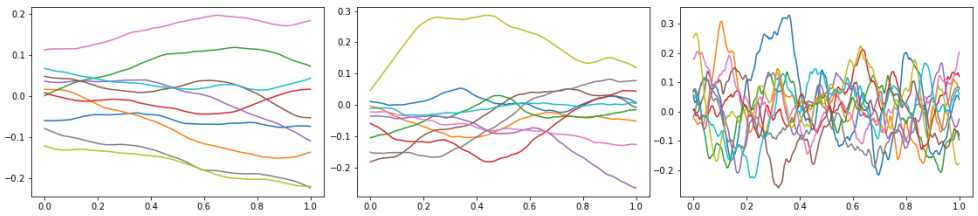


Figure 4.3: Sample prior function of a Matérn exponential covariance function with three different lengthscales (1, 0.5 and 0.05 from left to right, with variance set to 1×10^{-2})

4.5.4 Modifying kernels

The individual kernels mentioned above can be viewed as the basic building blocks to construct even more complex covariance functions through either modification processes (scaling and symmetrising) or combination (summation and multiplication). Modifying kernels is useful when trying to capture multiple features of a function.

4.6 Gaussian Process Learning

Pictorially, Gaussian process learning can be visualised in terms of samples drawn from the prior and posterior distribution over functions. In terms of the modelling data, the prior and posterior samples represent possible functions before and after the introduction of training data. One way of looking at Gaussian process learning is the discarding of functions that don't satisfy the data after training.

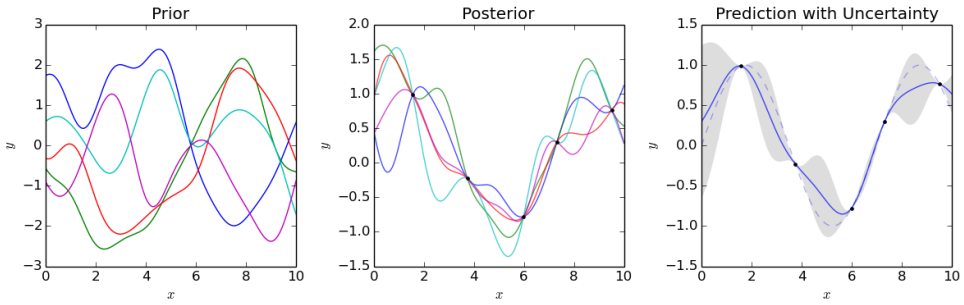


Figure 4.4: Samples drawn from the prior and posterior distribution, and the resulting mean function with uncertainty bounds after training on the data points. Imaged sourced from www.wikipedia.org/wiki/Gaussian_process

4.6.0.1 Prediction with noise free observations

Consider a case where the training data is noise-free. In the general case, a training set is a data set of inputs and corresponding outputs. In the case of a potential energy surface, the training set will have molecular geometries and corresponding energy values, while the test set will only have molecular geometries. Due to the key assumption of Gaussian Process modelling being the ability to represent a dataset as a sample from a multivariate Gaussian, the joint probability distribution of the test points, $y*$ and the training data y, is

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right) \tag{4.9}$$

where T denotes a matrix transposition.

For a data set with n training points and $n*$ test points, K_* is an $n \times n*$ matrix

of all the covariances between the training-test pairs.

To obtain the posterior distribution of functions, the prior distribution is conditioned on the data (given the data, how likely is a certain prediction of the test points).

$$y_*|y \sim \mathcal{N}\Big(K_* K^{-1} y, K_{**} - K_* K^{-1} K_*^T\Big) \tag{4.10}$$

Where $y_*|y$ is the unknown data y_*, conditioned on the known data y.

Prediction of $f*$ from f requires a multivariate conditional density, which is Gaussian distributed as well.

The best predictions for the test points is the mean of the distribution:

$$\mu = K_* K^{-1} y \tag{4.11}$$

The associated uncertainties of the predictions is denoted by the variance:

$$var(y) = K_{**} - K_* K^{-1} K_*^T \tag{4.12}$$

4.6.0.2 Prediction with noisy data

Usually, for the data modelling tasks, the observables are not the true underlying function values, but corrupted by some noise (assumed to be uncorrelated zero mean Gaussian noise of a fixed variance σ^2).

$$y = f(x) + \epsilon \tag{4.13}$$

This noise corruption of data is accounted for by assuming additive noise:

$$\epsilon = N(0, \sigma^2 I) \tag{4.14}$$

where I is the identity matrix.

With the introduction of noise, the joint probability distribution of the test outputs, f, and the training outputs, $f*$ is:

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) \tag{4.15}$$

The best estimate of test points therefore becomes:

$$\mu = K(X_*,X)[K(X,X) + \sigma_n^2 I]^{-1} y \tag{4.16}$$

and the associated variance is calculated by:

$$var(y) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*) \qquad (4.17)$$

4.6.1 Gradient based optimization

The hyper parameters can be found either by sampling methods or optimising the marginal likelihood with respect to the hyperparameters and noise, θ. Gaussian Process inference is conducted within a hierarchal framework where the choice of parameters is obtained directly from the training data, as opposed to other models (i.e. splines and support vector machines) that require a cross validation step.

Hyper parameters are obtained by minimising the negative log marginal likelihood with respect to the hyperparamaters, θ. Mininimisation of $L(\theta)$, the log marginal likelihood, is a non-convex optimization task and can be carried out using standard gradient-based techniques such as conjugate gradient or quasi-Newtown method. Gradient based optimisation refers to finding the optimal choice of parameters by using an iterative algorithm based on the gradient.

$$\begin{aligned} L &= -\log p(y|\theta) \\ &= \frac{1}{2}\log det C(\theta) \\ &= \frac{1}{2} y C^{-1}(\theta) y \\ &= \frac{N}{2}\log 2\pi \end{aligned} \tag{4.18}$$

where

$$C = K_N + \sigma^2 \tag{4.19}$$

To prevent the minimiser settling on a local minimum, random restarts of the optimiser at new starting points may be used.

Chapter 5

Gaussian Process Interpolation of Potential Energy Surfaces

Research questions: *Is Gaussian Process Regression an effective method for interpolating potential energy surfaces for Van der Waals complexes? What's the best choice of covariance function? How should symmetry be taken into account? Does incorporation of symmetry into the model improve its performance? What is the best way to assess model accuracy?*

Keywords: *Gaussian Process Regression, carbon capture storage, bimoleculer complexes, Van der Waals interactions, symmetric covariance function, squared exponential kernel, latin hypercube*

5.1 Introduction

A central problem in molecular physics is the quantitative prediction of macroscopic material properties from first-principles molecular physics. Since intermolecular potential energies can often be calculated accurately enough to be quantitatively useful by solving the Schrödinger equation approximately, the problem above can in principle be solved by firstly computing molecular interactions from ab initio quantum mechanics, and then successively using statistical mechanics to generate macroscopic properties of the system under study. In practice however, the computational cost of directly combining these two techniques is hopelessly expensive in all but exceptional cases of small systems and computationally less demanding levels of theory such as DFT. The cost of evaluating the energy of a single geometric point is usually so significant (often minutes or hours of time), that for computational feasiblity, the production of a global mesh of points usually requires the use of data fitting or machine learning methods in conjunction with a grossly reduced set of data points.

One such important application of ab initio study of macroscopic properties using interatomic potentials, is the modelling of carbon dioxide in the presence of the impurities found in the post-sequestration process of Carbon Capture Storage (CCS). CCS refers to the technology related to the entrapment of carbon dioxide from industrial process emissions due to environmental considerations. Due to the limited amount of experimental thermophysical data

available on different CO_2 mixtures, and the time it takes to produce just one data point on a relevant phase diagram, using machine learning techniques to model potential energy surfaces of carbon dioxide in the presence of the impurities found in the separated flue gas, presents itself as a way of generating phase behaviour information about the molecular systems relevant to the CCS process, ensuring safety and efficiency of the transport and storage stages by helping produce homogeneous one phase pipe flow through manipulation of the pipeline temperature and pressure.[52][53]

Aside from the specific application to CCS bridging technology, the study of Van der Waals bimolecular complexes in general, is the stepping stone to understanding more about molecular systems dominated by dispersion forces, and gaining a richer understanding about the various chemophysical properties that stem directly from the interaction[54][55][56] . Simply defined, Van der Waals molecules are weakly bound complexes formed from chemically stable neutral molecules. Gaining an in-depth understanding about their structural and dynamic properties[57] can both give insight into anisotropic potential energies and act as a model prototype for large-amplitude motions found in reactive systems.

In the majority, Van der Waals potentials have been generated using a combination of extensive grid data consisting of thousands of points spread across input space, and either an analytical functional form,[58] or a localised fitting

method like interpolative least moving squares (ILMS).[59] The motivation of this research therefore is the development of a Gaussian process based method of ab initio data interpolation to produce potential energy surface of bimolecular complexes at a fraction of the usual data size needed.

5.2 Methodology

5.2.1 Input Coordinates

Properties of the GP model, including differentiability, continuity and stationarity, are inherited directly from the covariance function used in the model. The intermolecular energy is a non-stationary function of distance, varying rapidly at small interatomic separations, but more gently at larger separation. Although it is possible to specify a non-stationary covariance function, in practice it can be challenging to find a flexible form with the correct non-stationary behaviour. It is simpler to transform either the inputs or outputs to achieve approximate stationarity, which is addressed here by using the inverse interatomic distances as covariates in the GP. Thus the GP coordinates are $\mathbf{x} = (1/r_1, ..., 1/r_{N_D})$ where r_i is the interatomic distance, running over all pairs of nuclei on different molecules shown in Figure 5.4. This results in an over-specified system, for example with $N_D = 6$ dimensions for CO_2-H_2. It is shown later that this change in variables leads to a dramatic improvement in performance.

The training data should ideally cover evenly a single symmetry-distinct subregion of $\mathbf{x}$ space, and respect the geometric constraint. The general strategy is to generate many candidate data sets (coordinates only, not energies), exclude points outside the symmetric and geometric constraints, and select the candidate data set with the best distribution of points.

5.2.2 Data generation

The approach involves two sets of data. A set of training data (between 20 and 1000 points) is used to train the model, and a larger set of grid data is used to test the model's predictive performance. Both datasets are described below. No knowledge of the test data is used during training.

Data sets of the intermolecular interaction energy of the bimolecular complexes CO_2-Ne, CO_2-H_2, CO_2-CO, HF-HF, and CH_4-N_2 are calculated as a function of their configurational geometry. These complexes are chosen to cover a range of intermolecular bond strengths and symmetries and to be small enough that an extensive set of data can be produced for testing the interpolations. For the same reason, all molecules are approximated as fixed-bond length, linear rigid rotors in their vibrational ground state, although the interpolation method can be extended straightforwardly to non-rigid molecules. Energy calculations are carried out in Molpro using second-order Møller-Plesset (MP2) perturbation theory and augmented correlation-consistent triple-zeta (aug-cc-pVTZ) basis sets. Basis set superposition error is corrected using the full counterpoise cor-

rection procedure.

Jacobi coordinates are used to describe the multi-dimensional potential energy hypersurfaces. In all cases r is the distance between the molecular centres in the complexes. For CO_2-Ne, θ is the angle between r and the CO_2 axis.

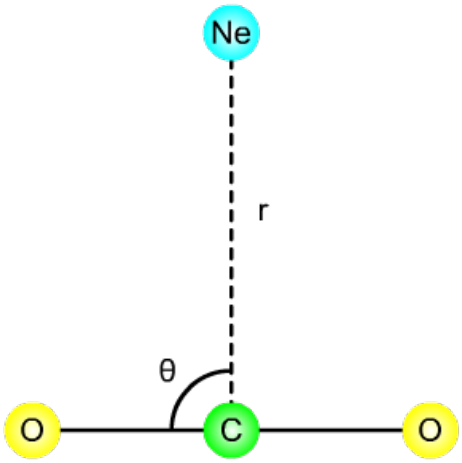


Figure 5.1: The Jacobi coordinate system for CO_2 - Ne. Figures 5.1 to 5.4 were created by Liam Goulding

For CO_2-H_2, θ_1 is the angle between r and the CO_2 axis, θ_2 is the angle between r and the H_2 axis, and ϕ is the torsional angle of the H_2 axis as shown in Figure 5.3

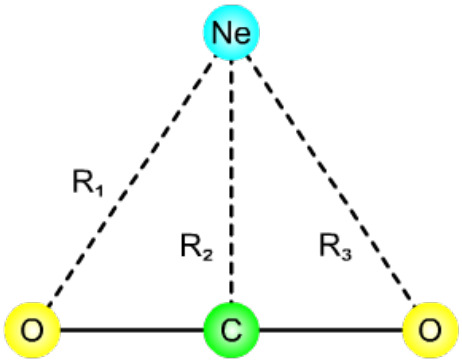


Figure 5.2: Interatomic distance coordinate system for CO_2 - Ne. The three interatomic distances are represented by dashed lines.

5.2.2.1 Latin Hypercube Sampling

A latin hypercube (LHC) is used to sample a space-filling combination of points in multidimensional space. LHC sampling aims to spread the sample points more evenly across all possible values than random sampling. In 2D, a latin square is defined as a set of samples with exactly one sample in each row and column. For a given number of rows and columns there are many permutations that satisfy this. A LHC generalises this to an arbitrary number of dimensions by requiring that each hyperplane normal to one of the co-ordinates contains exactly one sample.

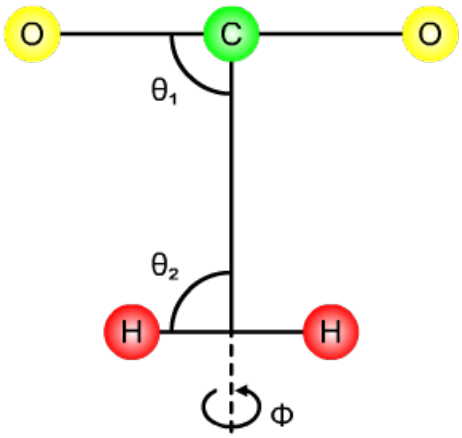


Figure 5.3: The jacobi coordinate system of CO_2 - H_2.

A dataset of model evaluations, $\{\mathbf{x}_i, f(\mathbf{x}_i)\}_{i=1}^{N}$, is needed to train the Gaussian process, where $\mathbf{x}_i$ represents N distinct molecular geometries. Each element of $\mathbf{x}_i$ is the inverse distance between two atoms, one from each of the molecules under consideration. The design only needs to contain points in a symmetry-distinct subspace. For example, in CO_2-Ne the O nuclei are denoted O_1 and O_2, and the symmetry-distinct subspace is defined such that Ne is always nearer to O_1 than to O_2. Space filling designs are held to be good choices for Gaussian process models, and so we will use a maxi-min criterion to evaluate candidate designs, which aimes to maximise the minimum distance between any two design points. Latin hypercube (LHC) designs are used as candidate

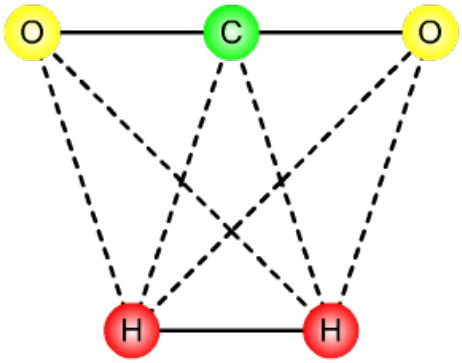


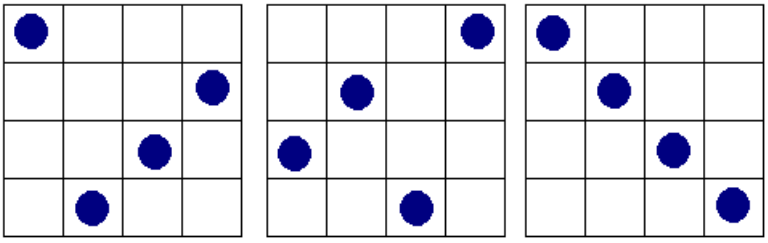
Figure 5.4: Interatomic distance coordinate system for CO_2 - H_2. The six distances are represented by dashed lines.

designs, as they naturally fill space to some extent, and then a prefered design is chosen from a large number of candidates. The effective distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the design is defined to be:

$$|\mathbf{x}|_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_i - \mathbf{x}_j) \tag{5.1}$$

A training design is generated using the following algorithm:

- Generate a LHC in $1/r$ and rigid-body rotation angles. (For non-rigid molecules, intramolecular coordinates would also be used.)
- Convert the LHC data to atomic positions and compute all interatomic



A trio of 2D examples of latin hypercubes consisting of 4 points

distances for pairs of atoms on separate molecules.

- Reject the geometries that do not obey the geometric constraint or lie outside the symmetry-distinct region of coordinate space.
- Reject the entire LHC if it does not contain at least the target number of geometries (usually the mean number of remaining points after the geometric constraint is applied).
- Find the minimum $|\mathbf{x}|_{ij}^2$ within the current LHC.
- Repeat for N_{it} new LHCs and return the LHC with the largest minimum $|\mathbf{x}|_{ij}^2$.

5.2.2.2 The Nugget Term

The absence of a nugget term in a Gaussian process regression task results in a non-interpolative task of the data, as the data is surrounded by a non-zero uncertainty bound. While this may seem counter-intuitive for a computer emulation task that deals with deterministic software outputs, previous research has demonstrated that the presence of a nugget term gives both better numerical stability to the matrix inversion step and offers better statistical coverage

for higher dimensional space. The enhanced numerical stability relates to both a more stable matrix inversion step during training and prediction, and the correlation between the smoothness of the spatial function and the size of the deviations to have a substantial effect on functional predictions. Since the most common covariance function used in regression tasks is the infinitely-differentiable squared exponential function, the use of a a nugget term helps negate any possible consequences of small deviations in the data set.

5.2.3 Model Construction

Results are obtained using the GPy package, modified to include symmetric covariance functions. Zero-mean Gaussian observation error is assumed on the function outputs (referred to as *nugget* in geostatistics), with standard deviation σ_n. Thus the model's hyperparameters, σ_f, σ_n and $\{l_i\}$, are estimated by optimising the log-likelihood over ≈ 30 random restarts, which typically finds the optimal values multiple times.

The choice of inverse internuclear distances to transform to stationarity is important. To illustrate this. a 'basic model' GP is created, which uses internuclear distances r as co-ordinates rather than $1/r$, but is otherwise identical to the non-symmetric GP above, having the same test and training data.

5.2.4 Symmetric Kernels

A major bonus of Gaussian Process modelling of potential energy surfaces is that the symmetry can be incorporated directly into the covariance function itself.

5.2.4.1 A single symmetry

Suppose we want to model f where f is invariant under the single permutation σ, where $\sigma^2 = e$. If we assume

$$f(x) = g(x) + g(\sigma x)$$

for some arbitrary function g, then f has the required symmetry. Modelling $g(\cdot) \sim GP(0, k(\cdot, \cdot))$, the covariance function for f is

$$k_f = k(f(x), f(x')) = k(x, x') + k(\sigma x, x') + k(x, \sigma x') + k(\sigma x, \sigma x')$$

5.2.4.2 Simplification

Given that k is an isotropic kernel, where isotropy is only required for each pair of vertices that swap in σ, then $k(x, x') = k(\sigma x, \sigma x')$, $k(x, \sigma x') = k(\sigma x, x')$ and $k_f(x, x') = k(x, x') + k(\sigma x, x')$, and consequently the computational time is reduced by half.

5.2.4.3 Example system

The motivating problem in this chapter for the symmetrising of the kernel is modelling the H_2 - CO_2 system, which is parameterised by 6 interatomic distances:

- $r_1 = \mathrm{H}^1 \rightarrow \mathrm{C}$
- $r_2 = \mathrm{H}^2 \rightarrow \mathrm{C}$
- $r_3 = \mathrm{H}^1 \rightarrow \mathrm{O}^1$
- $r_4 = \mathrm{H}^2 \rightarrow \mathrm{O}^1$
- $r_5 = \mathrm{H}^1 \rightarrow \mathrm{O}^2$
- $r_6 = \mathrm{H}^2 \rightarrow \mathrm{O}^2$

The potential function f between the two molecules obeys the following symmetry relations

$$f(123456) = f(214365) = f(125634) = f(216543)$$

where $f(123456)$ denotes $f(r_1, r_2, r_3, r_4, r_5, r_6)$.

5.2.4.4 Invariance under permutations in K_4

For functions that are invariant to permutations in K_4, where $f(x) = g(x) + g(\sigma_1 x) + g(\sigma_2 x) + g(\sigma_3 x)$ and $g(\cdot) \sim GP(0, k(\cdot, \cdot))$, the symmetrised kernel

defined as:

$$k_f(x,x') = k(x,x')+k(\sigma_1 x,x')+k(\sigma_2 x,x')+k(\sigma_3 x,x')+k(x,\sigma_1 x')+k(\sigma_1 x,\sigma_1 x')+\ldots k(\sigma_3 x,\sigma_3 x') \tag{5.2}$$

If k is isotropic, then

$$k(x,\sigma_i x') = k(\sigma_i x,x')$$

Therefore, for $i \neq j \neq k$

$$k(x,x') = k(\sigma_i x,\sigma_i x')$$

$$k(x,\sigma_i x') = k(\sigma_i x,x')$$

$$k(\sigma_i x,\sigma_j x') = k(\sigma_k x,x')$$

and the symmetrised covariance function can be defined as

$$k_f(x,x') = k(x,x') + k(\sigma_1 x,x') + k(\sigma_2 x,x') + k(\sigma_3 x,x')$$

Using this covariance function for the function f instead of Equation (5.2) results in the reduction of computations required by 75%.

For the simplification, k does not have to be completely isotropic. The only requirement is that the covariance function is isotropic for any pair of inputs that swap in any of the permutations (e.g. for the H2$-CO_2$ system example, we thus require the length-scales to be the same for inputs 1 and 2, and the

same for inputs 3, 4, 5 and 6.)

5.3 Results and Discussion

Predictive performance is measured using the root mean square error (RMSE) of the GP predictions of the test data. Since the test data extensively covers the potential energy surface, the RMSE is a reasonable guide to the expected 'accuracy' of the interpolation. The methodology used here represents a demanding test of accuracy, since the GP has no advanced knowledge of the test data, relying only on the far more limited training data instead. Equal weighting is also used for all test data, including positive interaction energies up to the potential energy cut-off.

The results for CO_2-Ne are shown in Figure 5.5. Models based on inverse intermolecular distances dramatically outperform the basic model, being typically 2-3 orders of magnitude more accurate for a given LHC size. Furthermore, even though CO_2-Ne has only one symmetry, the symmetric model is typically twice as accurate as the non-symmetric model. Figure 5.6 shows similar results for CO_2-H_2. Here, the inverse distance models again strongly outperform the basic model, achieving RMSEs $< 10^{-6}$ E_h for a reasonable number of training points. The symmetric kernel typically gives a factor of 2-10 improvement, with the greater improvement compared to CO_2-Ne probably resulting from the greater number of symmetries.

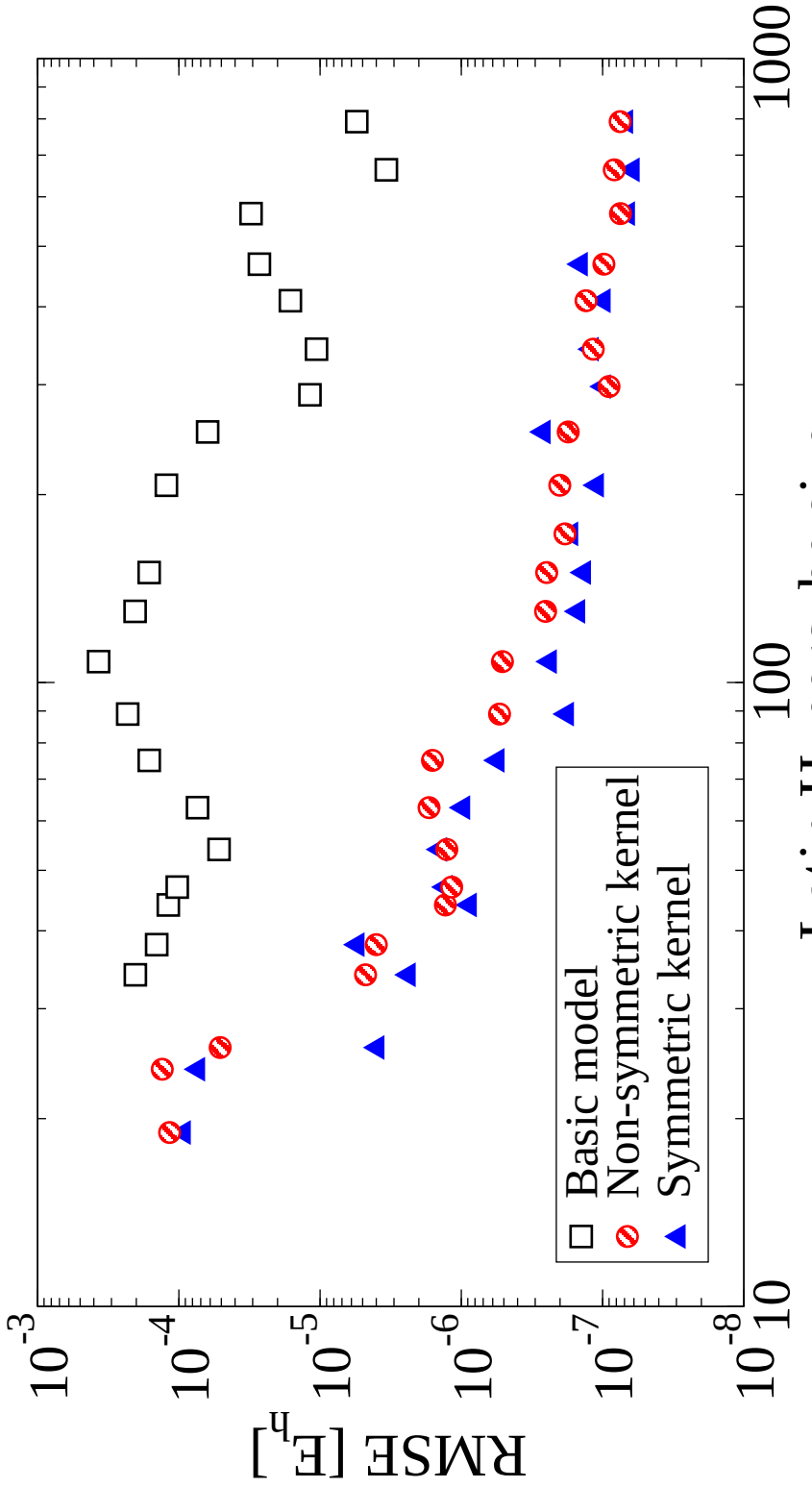


Figure 5.5: RMSE against LHC size for CO_2-Ne. The lowest energy in the grid data is -2.90×10^{-4} E_h.

The results for HF-HF are shown in Figure 5.7. For HF-HF, the minimum energy in the test data is -6.17×10^{-3} E_h, which is about an order of magnitude larger than for the other interactions. As a consequence of this, it is necessary to include training points up to at least 10^{-2} E_h, otherwise the prediction of points on the repulsive wall is poor. Using a cutoff of 2×10^{-2} E_h gives an RMSE of 1.6×10^{-4} E_h for a symmetric GP with 59 training points, and the RMSE generally decreases with increasing numbers of training points, to 1.8×10^{-5} E_h for 327 training points. The RMSE in the negative-energy region is about 5×10^{-6} E_h for the latter GP; one or two high-energy points dominate the overall RMSE. The inclusion of symmetry in the GP has little effect on the RMSE for this interaction.

The results for CH_4-N_2 are shown in Figure 5.8. For CH_4-N_2, all 48 symmetry elements are included in the GP. Consideration of symmetry is important for this interaction, even though all the training and test data are confined within a single symmetry-distinct region of space. The minimum energy in the test data is -6.98×10^{-4} E_h. With a training set of 106 points, the RMSE is 51×10^{-6} E_h for the nonsymmetric GP and 6.8×10^{-6} E_h for the symmetric GP. Using 327 training points reduces these to 17×10^{-6} E_h and 1.3×10^{-6} E_h respectively.

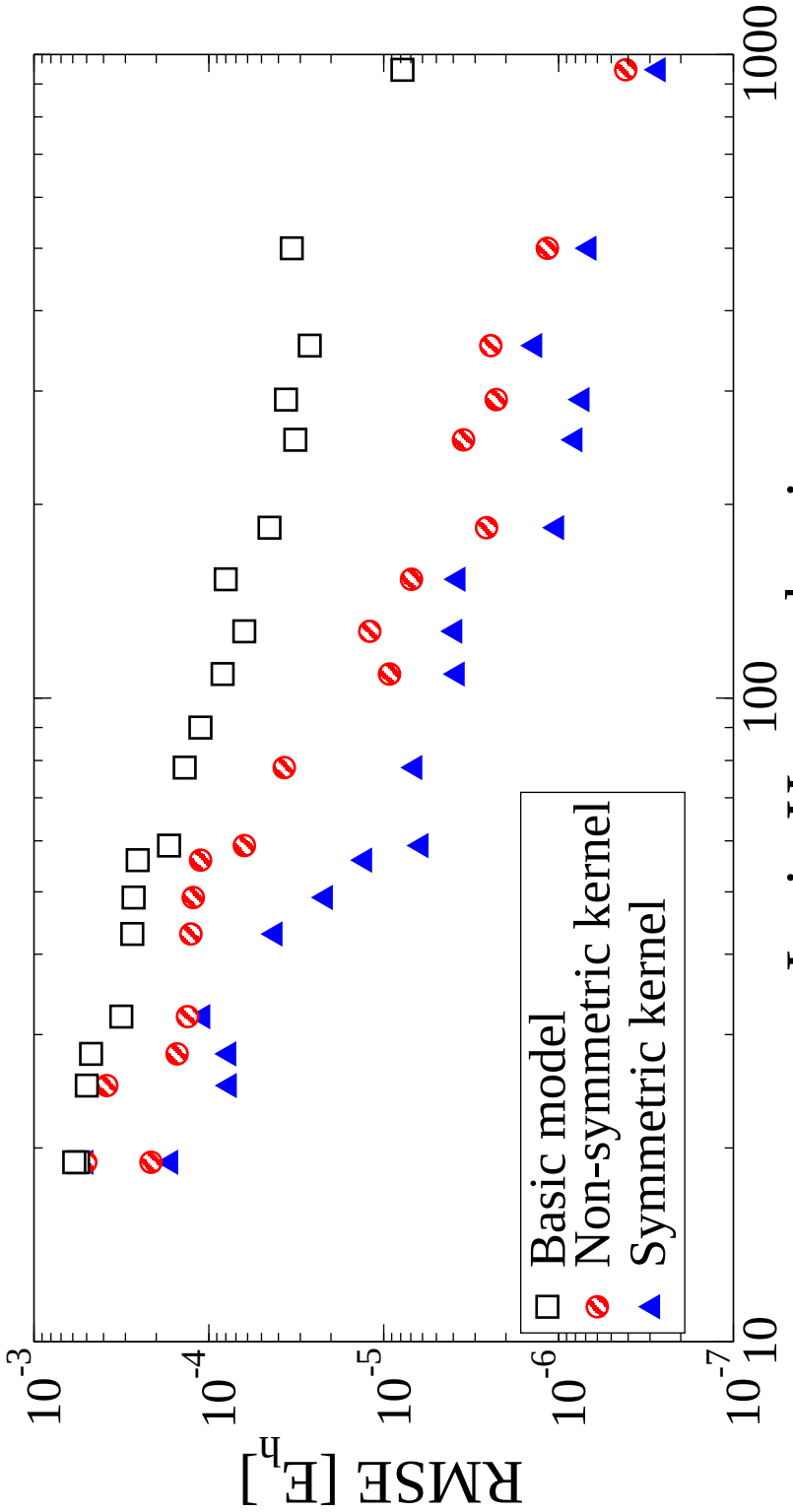


Figure 5.6: RMSE against LHC size for CO_2-H_2. The lowest energy in the grid data is -8.25×10^{-4} E_h.

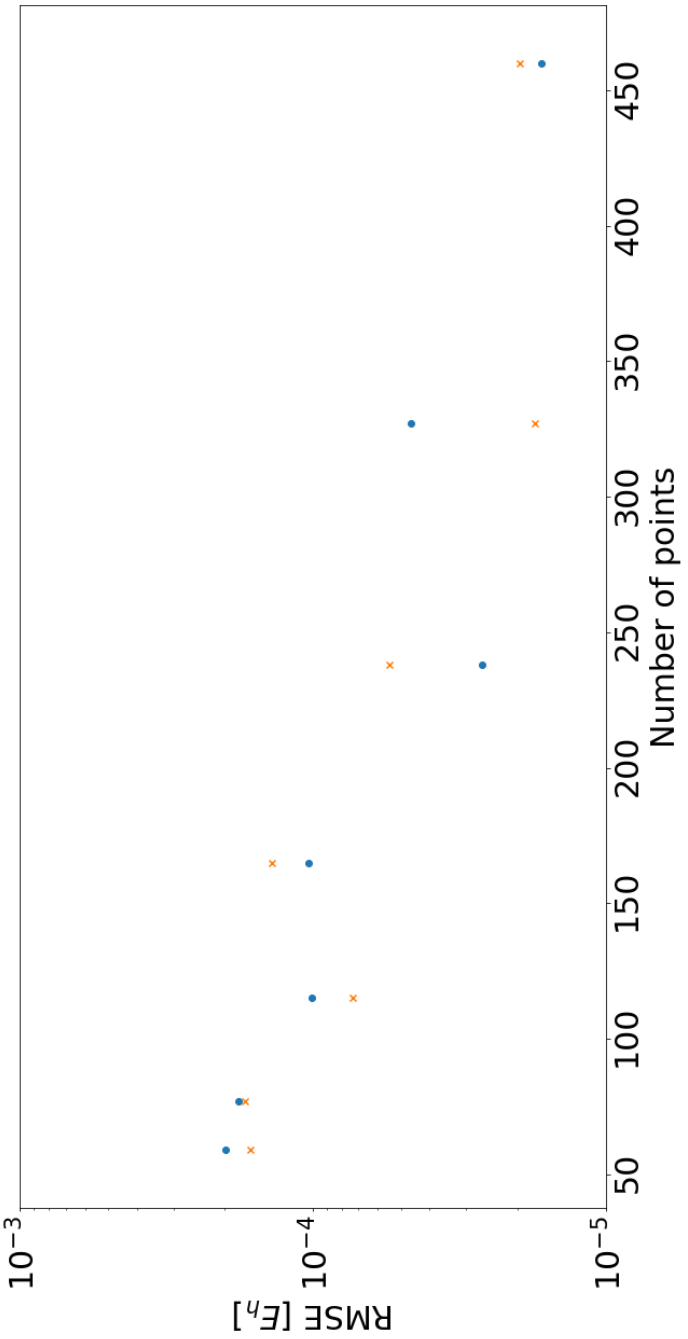


Figure 5.7: RMSE against LHC size for HF-HF. The lowest energy in the grid data is -6.17×10^{-3} E_h.

The results for CO_2-CO are shown in Figure 5.3. For CO_2-CO, with the symmetric GP, the RMSE is 6.7×10^{-5} E_h for 132 training points, reducing to 1.1×10^{-5} E_h for 345 training points. The RMSE resulting from the non-symmetric GP is generally a factor of ~ 2 larger. Generating CO_2-CO training data by replacing LHC sampling of $1/r$ with $1/r^2$ improves performance, leading to RMSEs of 2.4×10^{-5} and 3.0×10^{-6} for 135 and 366 training points respectively, with the RMSEs computed using the same test set as the above results.

The performance of the GP outside the training region ($E > 0.005E_h$ and $r > 8.5$Å) is shown for CO_2-Ne in Figure 5.9; results for other geometries and interactions are qualitatively similar. The extrapolation errors for short-range geometries satisfying the geometric constraint, but with $E > E_{\mathrm{cut}}$, are mostly a few percent or less. Even though the maximum energy in the training set is limited to 0.005 E_{h}, the predicted repulsive wall continues up to about 0.1 E_{h}. For small values of r the GP returns to its mean value of zero. This unphysical behaviour can be corrected by crossing over to a strongly repulsive function outside the geometric constraint. One example of the many possible choices

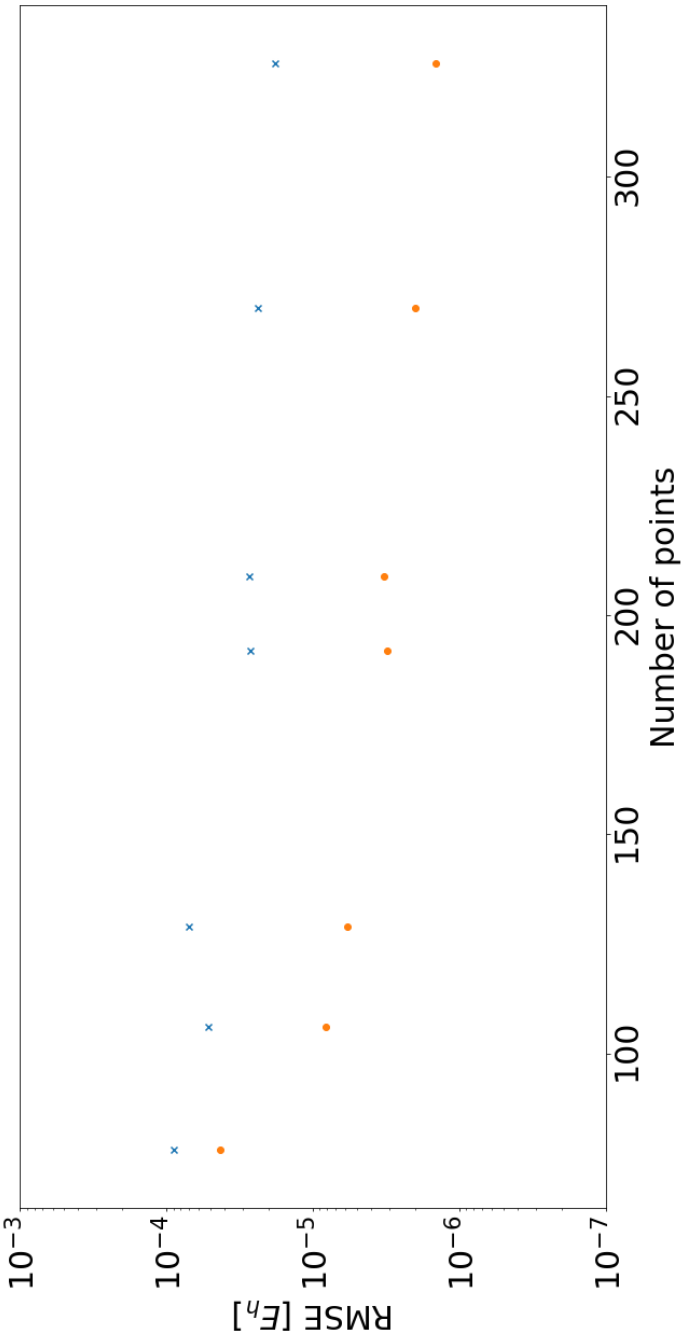
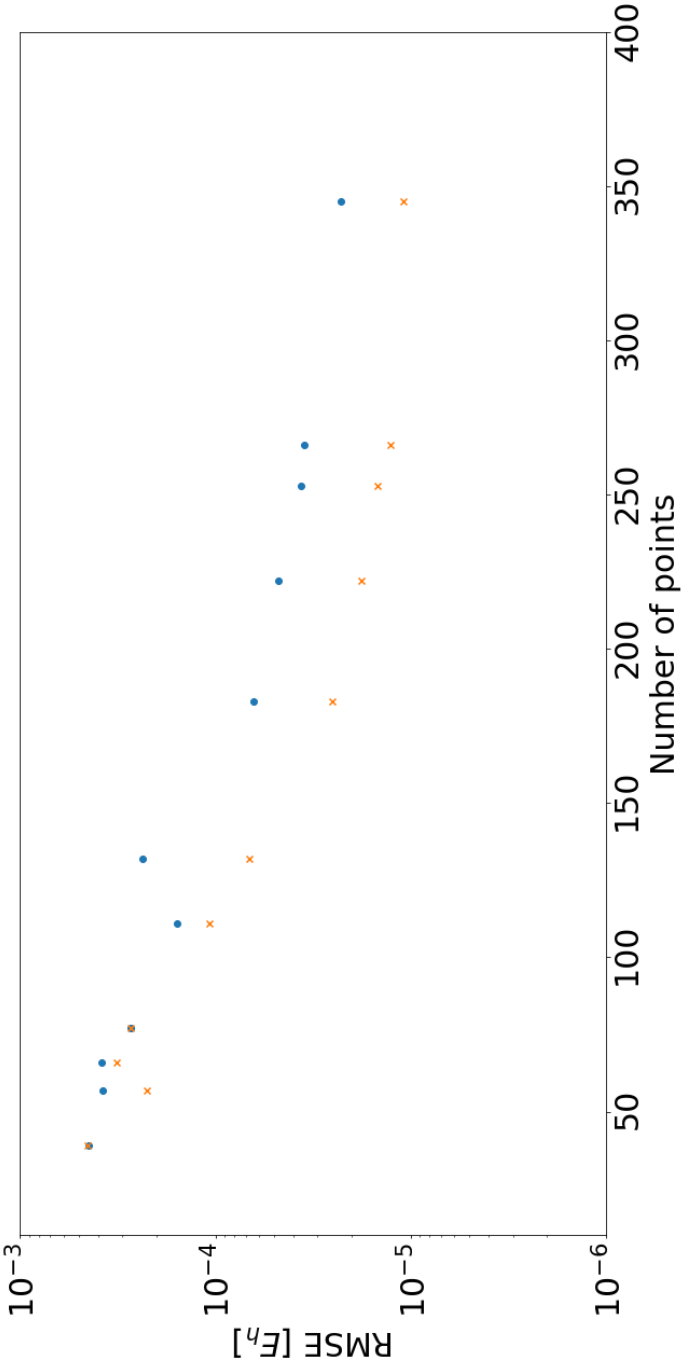


Figure 5.8: RMSE against LHC size for HF-HF. The lowest energy in the grid data is -6.98×10^{-4} E_h.


RMSE [E_h]
10^{-3}
10^{-4}
10^{-5}
10^{-6}
50
100
150
200
250
300
350
400
Number of points

is plotted in Figure 5.9(a), namely

$$E = E_{\max} \frac{1}{N} \sum_{i=1}^{N} (x_i / x_{\max})^{12}, \tag{5.3}$$

where $E_{\max}$ is an estimate of the typical energy at the small-r edge of the geometric constraint and $x_{\max}$ is the maximum inverse distance allowed by the geometric constraint (0.67 Å^{-1} in this case). $E_{\max}$ can be obtained from the maximum energy, before applying the energy cut-off, over either the test data or training data. For large separations the GP tends to a small, but non-zero constant. This can be corrected for points beyond the geometric constraint, by crossing over to the long-range asymptotic expansion obtained from a truncated multipole expansion of the interaction energy (to second order) from intermolecular perturbation theory. Figure 5.3 shows that smooth interpolation between the GP and this function will be straightforward.

5.4 Future work

There are numerous extensions that follow from this approach. Application to many other chemical systems is straightforward. Furthermore, the model's performance for small training set sizes could be optimised by sequentially adding training points through active learning methods. This could be achieved either with or without *a priori* knowledge of the test data, depending on the potential energy data to be modelled. Another promising application is the interpolation of non-additive potentials, which are known to be difficult

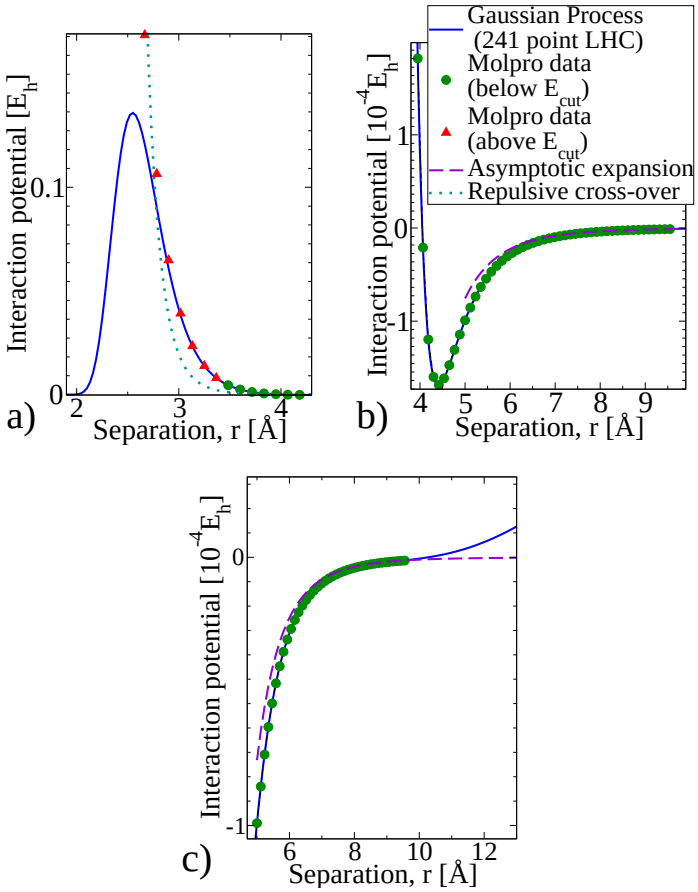


Figure 5.9: CO_2-Ne Molpro calculations and the GP model, at $\cos\theta = 1$ (linear geometry), in the repulsive (a), attractive (b) and long-range (c) regions. The long-range asymptotic expansion is $E = -(0.570 + 0.182\cos^2\theta)r^{-6} - (1.704 + 7.266\cos^2\theta + 1.785\cos^4\theta)r^{-8}$.

to fit. Such data usually have high-dimensionality, vary strongly and rather unpredictably with geometry, and can have many symmetries. Finally, existing high-precision calculations could be used as training and test data for interpolation by the algorithm. Here, a sparse GP could select a subset of the pre-existing data to train the model, leading to numerically cheap, yet highly accurate, potential energy surfaces.

5.5 Conclusions

The procedure described here produces intermolecular potential energy surfaces efficiently from relatively few input data points. The algorithm is straightforward to generalise to new molecular pairs. It uses a symmetric Gaussian process, with the inverse interatomic distances as input variables. The GP is trained using data on a Latin hypercube design, with a maximin criterion for the inverse internuclear distances.

The general utility and robustness of the approach have been demonstrated by testing against three systems involving CO_2, a system with a deep energy minimum HF-HF, and a system with 48 symmetries CH_4-N_2. In all cases the approach accurately predicts an extensive test data set, with no *a priori* knowledge, and gives RMSE values that are similar to, or better than, the best fits in the literature, which were generally based on thousands of training points. Furthermore, the interpolation method can be readily and directly applied to any pairwise interaction, at least for simple molecules, with no bespoke work beyond identifying the relevant symmetries. The approach contains two key features: the use of inverse interatomic distances as the GP input variables, and a strategy for positioning training data on a latin hypercube design with a maximin criterion on the inverse intermolecular distances.

Chapter 6

Sequential Design of Potential Energy Surfaces

Research questions: *Can sequential design be used to produce better data sets than space filling designs? What is the best criterion for new point selection? Where in space are the new points placed? Does the accuracy of sequential design vary with the number of dimensions?*

Keywords: *Active learning, sequential design, space filling design, optimal experimental design, modified highest variance search, expected improvement, two set design.*

6.1 Introduction

For any given model, the choice of training set dictates both the overall predictive accuracy of the model and the total number of points required to reach a certain error threshold. With the computational cost of calculating potential energies increasing with both the size of the basis set and the complexity of the system, generating an optimal choice of points that yields the lowest overall error using the smallest data set possible becomes more and more imperative, the longer each single point calculation takes to run.

One possible way to achieve an optimal training set is through the use of active learning - a family of machine learning methods with a strong anchoring in information theory that has been applied to domains such as chemoinformatics,[60] financial crime detection[61] and speech recognition problems.[62] The overall aim of active learning is to make machine learning more economical and CPU time-efficient by allowing the machine to make queries as to where to add more training data rather than remaining passive to the data acquisition process. In the context of regression-based machine learning models, active learning (alternatively known as sequential design, online fitting or adaptive fitting), is the process of using previously learned information obtained during the training phase, to guide the process of new point placement through iterative methods that add points either one-by-one or in batches.

Active learning (AL), in the domain of PES generation, is an alternative to the current status quo of relying on either grid-based sampling[63] (where points are sampled in intervals between a range of values of each geometric variable) or space filling designs[15] (where points are spread out across design space to get approximate uniform coverage) to produce training data. This work applies active learning to Gaussian processes, a machine learning method based around a Bayesian framework, capable of capturing flexible functions using relatively small data sets. Gaussian processes have been successfully applied to model the PES of several systems, including the generation of a reactive PES for the lowest triplet state of SH_2,[16] a ground state global PES for N_4,[15] and numerous bimolecular van der Waals complexes including CO_2-Ne, CO_2-H_2, HF-HF, CH_4-N_2, CO_2-CO[64] and CO_2-N_2.[65] In all these instances, however, training data were generated using a space filling design known as latin hypercube sampling, which despite surpassing grid-based sampling methods, still does not achieve an optimal accuracy for a given computational cost.

Relative to batch designs (where batch design is a broad term encompassing experimental designs where the required number of data point are produced in one go, rather than sequentially choosing the points one by one after retraining the model), there are numerous potential benefits of active learning that make them an attractive prospect in the pursuit of optimal training sets:

1. A design might not be able to achieve the required level of accuracy for

a given number of data points, so additional points may be needed to reduce the error of predictions. Since some information has already been gained, it follows naturally that active learning of subsequent points will lead to the best improvement in predictive accuracy relative to a non-adaptive data point selection.

2. AL offers the potential to factor in functional non-stationarity into the experimental design, by including more points in regions of space where there is a greater functional variation.

3. AL offers a way of selecting the best subset of points when presented with a large pre-existing data set.

4. AL allows for the training data to be generated directly in the input space of the Gaussian process model when using over-specified variables that are dependent on each other.

For general non-GP active learning problems, there have been many heuristics to guide point selection, including selection of points where there is a lack of data,[66] low confidence,[67] the biggest expected change in the model[68] and where there has been previous selection of data that has resulted in learning.[69]

Being a probabilistic machine learning method, Gaussian process regression inherently produces values of uncertainty for all predictions, making it an ideal candidate for sequential design methodology. The use of variance in Gaussian

process sequential design tasks was first introduced by McKay,[70] who proposed a method that aimed to maximize the expected information gain about the parameter values of the model by choosing data with the highest predictive variance. This was then later expanded on by Cohn,[68] who proposed a method which aimed to maximize the expected information gain about the model's parameter values by selecting points that led to the biggest reduction in average variance instead of just finding the highest single-point predictive variance.

Although there has been some prior research dedicated to active learning in the context of both Gaussian process regression and general computer emulation tasks, there has been relatively little work done on the use of active learning techniques in the generation of intermolecular potentials for different molecular systems. Active learning techniques in the field of PES generation have so far been most prominent in instances of on-the-fly molecular dynamics, where the data points are generated over the course of the sampling. Rupp[71] used an adaptive learning scheme alongside a hybrid Quantum Mechanics and Machine Learning method based on Gaussian process regression to run molecular dynamics simulations of the complex natural product, Archazolid A, and found that using the model's inherent predictive variance to decide whether to carry out additional electronic structure calculations led to a reduction of the amount of calculations carried out by 40%. Research was also done by Podryabinkin *et al.*[72] on linearly parametrized interatomic potentials. An

active learning scheme based on a D-optimality criterion was applied to moment tensor potentials, querying whether the machine learning value at the sampled configuration was expected to be a good enough estimate or whether new *ab initio* data needed to be generated instead. The results were then compared to the "classical" on-the-fly method proposed by De Vita *et al.*[73] The classical method has a general framework of starting with an initial short ab initio molecular dynamics (AIMD) trajectory (a few picoseconds long) and then adding on new ab initio points after a fixed number of steps have been taken. It was found that while the classical method increased the failure time from 15 ps to 150 ps following the addition of 1500 ab intio points (with failure time is defined as the simulation time after which half the trajectories have been terminated), the active learning method increased simulation time to 0.5 μs following the addition of only 50 new points.

Although these methods are promising in the domain of molecular dynamics, there has been little work on active learning of global intermolecular potentials. Since one of the main factors underpinning the predictive power of the model is training point location, finding a way to select an optimal selection of data points is a critical step in generating chemically useful potential energy surfaces at a fraction of the computational cost of extensive ab initio calculations.

Recently, Zhang *et al.*[26] used predictive variances (weighted to bias lower energy point selection) to generate training data for H_3 and two prototypical

reactive systems. Although this lays out the first framework for sequential design of global potentials, only methods involving predictive variance were considered, and the active learning scheme was not compared to space filling design methods of equal size. The following research aims to compare latin hypercube design in PES generation against three separate methods: a prototypical highest variance search, an active learning scheme that adds points in the place of highest absolute error and a third scheme based on biggest discrepancy in predictions between two models. The latter method is a special case of producing a jack-knife estimate of variance, where uncertainty in the parameters is factored by looking at the sensitivity of results to different parameter values.

6.1.1 Molecular systems

Three molecular systems are investigated (CO_2-Ne, CO_2-H_2 and Ar_3). The two CO_2 systems are chosen because they produced impressive benchmark results in previous work[64] using exclusively a batch design strategy, namely latin hypercube sampling. Thus any further improvements in the RMSE would be a significant result. The Ar_3 system was chosen in order to test the algorithm on a non-additive potential, which is generally harder to fit due to greater functional variation.

The intermolecular interaction energies of these complexes are calculated as a function of their configurational geometry. All molecules are approximated as

linear rigid rotors in their vibrational ground state, with fixed bond lengths, although the interpolation method can be extended straightforwardly to non-rigid molecules. Energy calculations are carried out in Molpro using second-order Møller-Plesset (MP2) perturbation theory for the two CO_2 systems and CCSD(T) perturbation theory for Ar_3. In all cases the augmented correlation-consistent triple-zeta (aug-cc-pVTZ) basis set is used and basis set superposition errors are accounted for by the counterpoise correction technique.

6.1.2 Data sets

The data sets contain a number of cluster geometries, along with their corresponding intermolecular interaction energies. The method considers two types of data set: training sets and reference sets. The training set is used to train the GP. The reference set serves two purposes: to provide candidate data points to be added to the training set during the active learning, and to compute the RMSE of the GP.

6.1.3 Co-ordinates

All reference data are generated using the distance and angular coordinates presented in Table 6.1. In all binary cases, r is the distance between the molecular centre of masses. For CO_2-Ne, θ is the angle between r and the

Table 6.1: Co-ordinates for the reference data (grid or LHC) for each system

System	Coordinate	Range	Spacing	Reference points
CO_2-Ne	r	1.5-10 Å	0.116 Å	1122
	$\cos\theta$	0-1	0.05	
CO_2-H_2	r	1.5-10 Å	0.5 Å	12844
	$\cos\theta_1$	0-1	0.111	
	$\cos\theta_2$	0-1	0.111	
	ϕ	0-180°	20°	
Ar_3	r_{12}, r_{13}, r_{23}	2.88-9.0Å	LHC	5476

CO_2 axis. For CO_2-H_2, θ_1 is the angle between the vector $\mathbf{R}$ and the CO_2 axis, θ_2 is the angle between r and the H_2 axis, and ϕ is the torsional angle of the H_2 axis. For Ar_3, r_{ij} is the distance between Ar atoms i and j. For CO_2-Ne and CO_2-H_2, the reference data are positioned on an evenly spaced grid, as generated for previous work[64] and with the co-ordinates detailed in table 6.1. The Ar_3 reference data are newly generated from a LHC design strategy.[64] For the dimer systems energy, points above the high energy cut off (0.005 E_h) are excluded from the reference set. For the non-additive system, Ar_3, points are excluded if the binary interaction potential between any of the 3 pairs of atoms exceeds the high energy cut off. The most straightforward way to impose this high energy cut off is to pre-calculate the reference data set and exclude points above the cut off, as is done here. However, this pre-calculation can be avoided, in principle, if a reliable way of accounting for the high energy barrier can produced. Such a method is straightforward for non-additive interactions if a PES is available for all dimers in the system.

6.1.4 LHC generation for Ar_3

The trimer system is described by three distances (r_{12}, r_{13}, r_{23}) as seen in Figure 6.1. These distances must obey the triangle constraint $r_{12}+r_{13}+r_{23} \geq 2\max(r_{12}, r_{13}, r_{23})$. Also, as the atoms are identical, the following symmetry constraint is used to restrict the size of the space: $r_{12} \leq r_{13} \leq r_{23}$. LHC data are generated via a 3-dimensional unit LHC and scaling this to an LHC on ($1/r_{12}$, $1/r_{13}$, $1/r_{23}$) using the ranges in table 6.1. These are converted to (r_{12}, r_{13}, r_{23}) and geometries that disobey the triangle or symmetry constraints are rejected. Taking 2.88 Å as the minimum value for the r_{ij} ensure that no pairwise interaction exceeds the high energy cut off. If an LHC rejects more than the mean number of rejected geometries then the entire LHC is rejected. The minimum distance within an LHC is calculated as in our previous work and a long series of candidate LHCs is generated by this algorithm.

The LHC with the largest minimum distance is returned as the best LHC. The test LHC data for Ar_3 are generated on $1/r$ as above. However, the training LHCs are generated using $1/r^2$ as this scaling has been seen to give a slightly improved RMSE for the CO_2-CO dimer.

6.1.5 Active learning and GP training

During active learning, points are removed sequentially from the reference set and added to the training set. Each method starts with a small initial set of

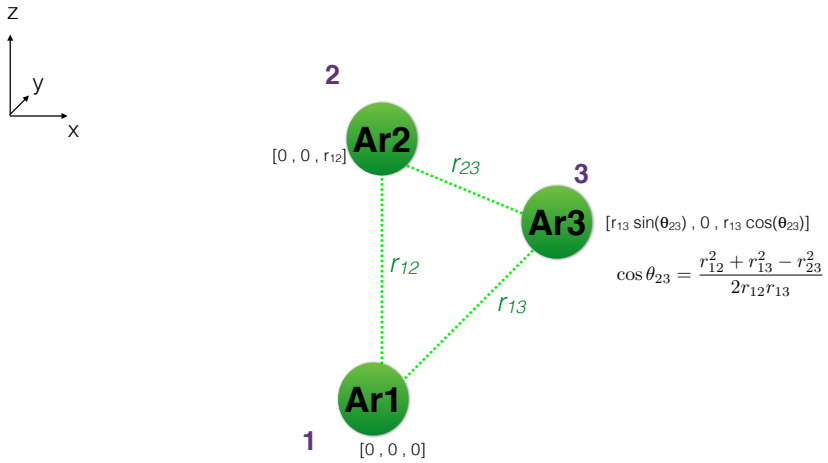


Figure 6.1: Co-ordinates used to describe the Ar_3 system

data points and adds one point at a time, updating the model between each addition. The following three distinct learning strategies for selecting new training points are used for all systems:

- Highest variance search, which adds points on the basis of the highest predictive variance over a pool of input data, as detailed in section 6.1.5.1. **(Method A)**.
- Absolute highest error search, which uses the reference data and adds points with the highest absolute error iteratively, as detailed in section 6.1.5.1. **(Method B)**
- Two-set search, which uses two different GPs from different training sets and places points into both sets on the basis of the biggest discrepancy between the two different models, as detailed in section 6.1.5.2. **(Method C)**

Learning was carried out using the GPy package, modified to include symmetric covariance functions. The co-ordinates in table 6.1 are converted to inverse interatomic distances and are used as over-specified covariates in the GP. The covariance function, a symmetrised squared exponential function with a nugget term, is used alongside a zero mean function, and zero Gaussian observation error is assumed. All of these approaches are identical to previous work.[64] After each new training point the RMSE is computed against the reference data. Points that have been moved from the reference to the training set no longer contribute to the RMSE.

6.1.5.1 Highest variance and absolute highest energy search : METHOD A and B

One data point is chosen at random from the reference set, and used as the initial training set. Hyperparameters are optimised using 20 restarts. Predictions are made over the reference set. One point is added to the training set on the basis of the highest predictive variance (Method A) or the highest absolute error of the predictions (Method B). The point is removed from the reference set and the hyperparameters are re-optimised. The process is iterated until the desired number of points is reached.

6.1.5.2 Two-set search : METHOD C

The two-set search method utilises two training data sets. These two training sets each begin with a single point, chosen randomly from the reference set. The starting point is different for each set and both points are discarded from the reference set. To add a new point, each training set is used to train a GP. The predictions for both GPs are then made over the reference data and the discrepancy between the predictions of the two GPs is calculated at each point. The point with the biggest absolute discrepancy in energy value is then removed from the reference data and is inserted into both training sets. Further points are added by repeating this process until the desired number of training points has been reached. Note that, in this algorithm, as in Method A, it is not essential to pre-compute the reference data. Instead, computations of the intermolecular interaction energy can be performed sequentially as new points are added. Thus, these potentially expensive calculations need only be performed for the much smaller training set.

6.1.6 Comparison to a batch space-filling design

The efficacy of the active learning is then assessed by comparing the resulting RMSE against a non-active approach.[64] These non-active results are obtained by training GPs to independently generated maxi-min latin hypercube data of increasing size, based on the co-ordinates in table 6.1. For the two CO_2

systems the series of LHC data sets generated previously[64] is used. For Ar_3, a new series of LHC training sets is generated using an LHC algorithm.[64]

6.2 Results

For CO_2-H_2 and Ar_3, the predictive performance is measured using the root mean square error (RMSE) of the GP predictions for the reference set. Equal weighting is used for all data points, including positive interaction energies up to the potential energy cut off. Since the reference set is the data set from which new points are selected during the active learning algorithms, any points that have been added to the actively learned set are discarded from the reference set prior to RMSE calculation. For fair comparison with the batch design method based on latin hypercubes, the same reduced test set is consequently used to calculate the error of the LHC of equal size to the actively learned set. In the CO_2-Ne case, the RMSE is calculated using a new 2095-point latin hypercube set with all points above the high energy cutoff discarded. Since this test set is completely independent of the set from which points are added, only points above the energy cut off are removed during error calculation.

The results for CO_2-Ne are shown in figure 6.1. The graphs show the RMSE of all three active learning methods and the error produced by the non-active latin hypercube method. The data sets generated using active learning out-

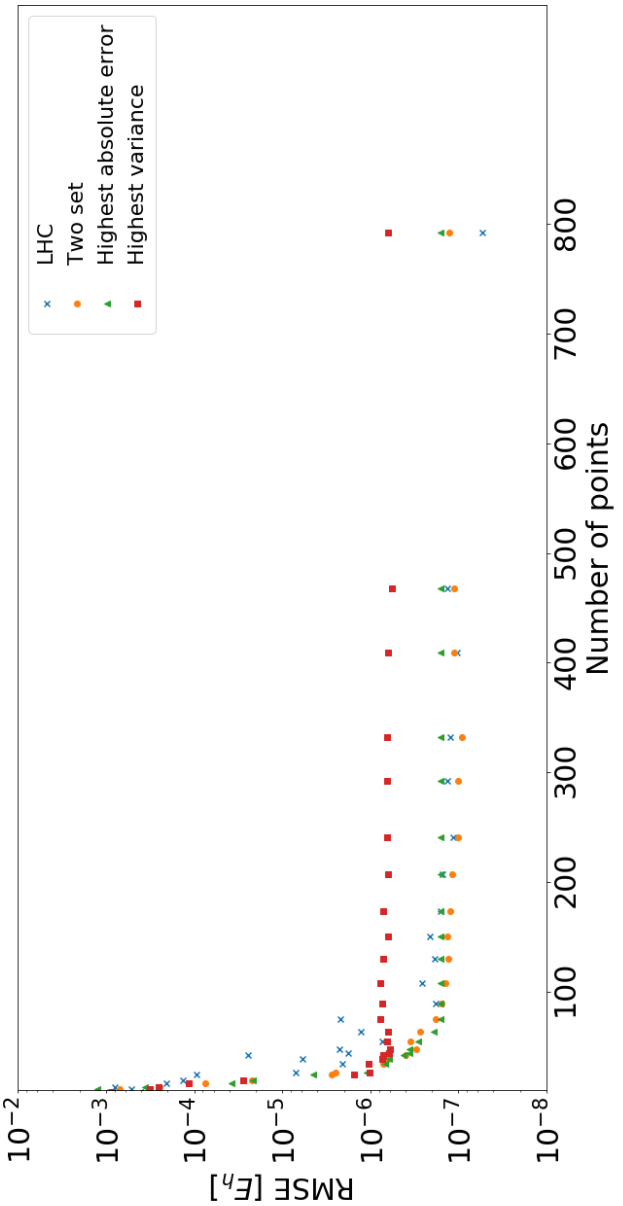


Figure 6.2: RMSE against training set size for CO_2-Ne. The lowest energy in the reference data is -2.90×10^{-4} E_h.

perform the latin hypercube sets until around the 90 point mark, after which the LHC produces RMSE similar to both the highest actual error and the two set methodology. Although the highest variance method performs very well at low points, producing the best RMSE in three cases, it plateaus at $\sim 10^{-6}$ E_h despite the addition of many new data points. This may be due to an accumulation of different factors. As mean variance approaches the value of the nugget term, the algorithm may act more erratically in its point addition process. The highest variance algorithm also tends to bundle points around the border, which have higher uncertainty than central points. This bundling of points produces a high density coverage of a region where such point density is not needed, hence not producing the improvement in accuracy expected for a large addition of points, and causing problems for the optimiser in its search for the best hyperparameter values as the bundling of points close in space to each other can result in matrix irregularities.

Figure 6.2-6.4 shows the location of the first 75 points added using the three individual methods for the CO_2-Ne. Observing the placement of points by the algorithm may offer insight into why certain methods perform better than others and gives important comparative information between the methods. Looking at the RMSE values versus the number of points, it can be seen that at 75 points, the methods placed sequentially from the least to most erroneous are the highest absolute error method, the two-set method, the highest variance search and lastly the LHC design. The highest error method

in figure 6.4 bundles a lot of points around the high energy barrier, which is intuitively where the highest absolute errors are likely to be. The highest variance method points are spread out but contain few high energy points, which seems to be a key component of achieving low predictive error. Due to the way the LHC has been generated in 1/r, points are more densely packed around the high energy barrier, mimicking the general pattern of the highest absolute error method.

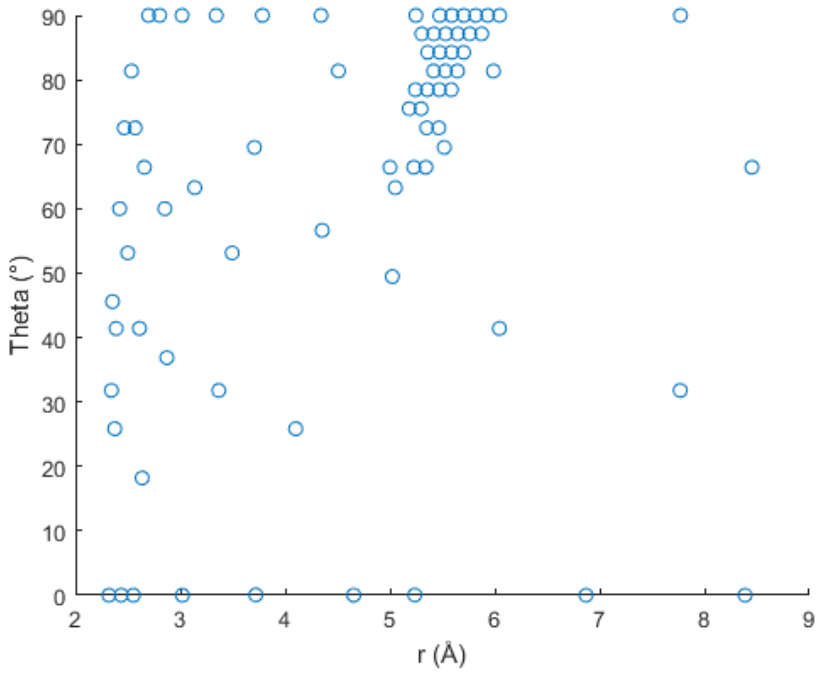


Figure 6.3: The placement of the first 75 points in the two-set method

Figure 6.6 shows results for CO_2-H_2. When the number of points is very low, there is no clear advantage of the active learning over the LHC in terms of predictive accuracy. However, after the addition of about 30 points, all three of the active learning methods outperform the LHC sets. The error of the highest variance method again plateaus around 10^{-6} E_h, similar to the error of the

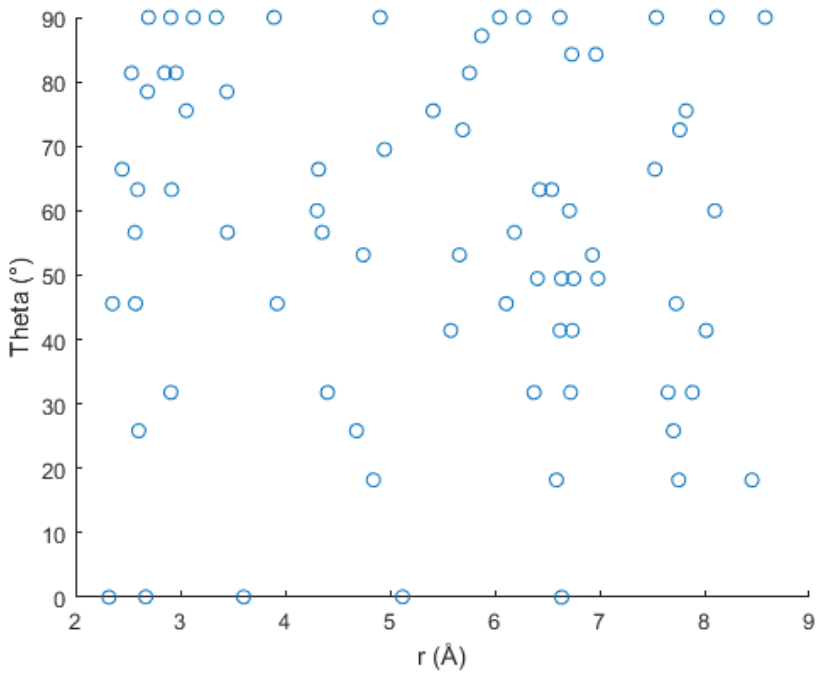


Figure 6.4: The placement of the first 75 points in the highest variance method

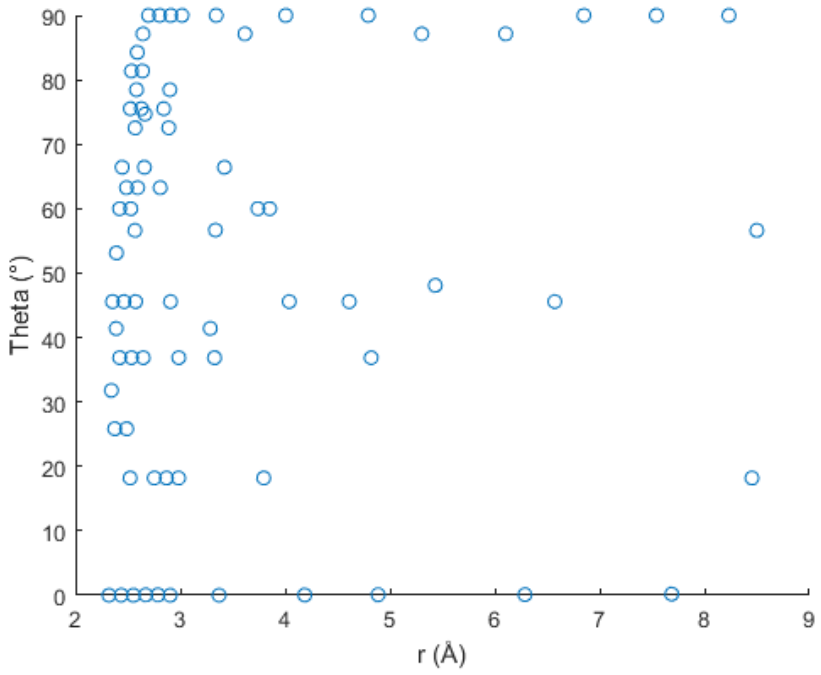


Figure 6.5: The placement of the first 75 points in the highest error method

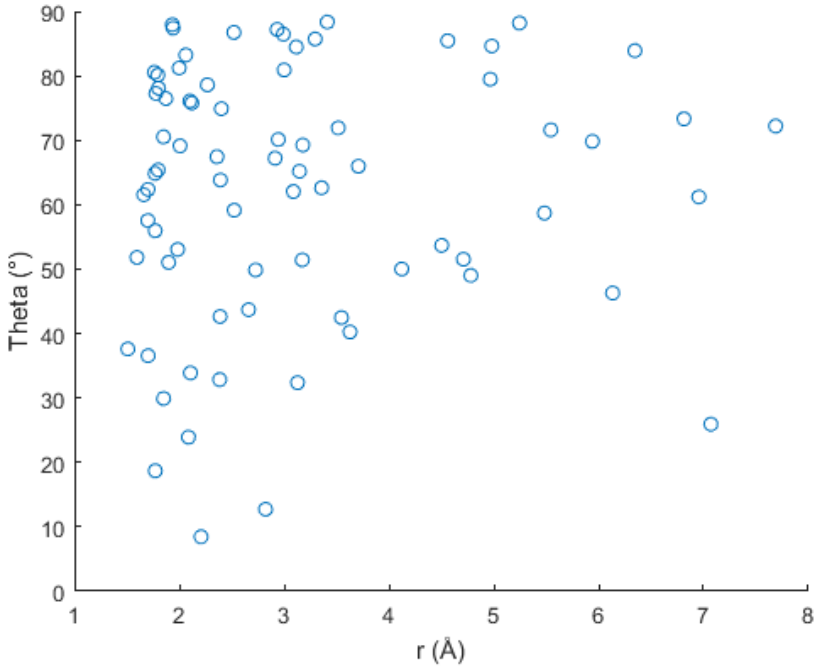


Figure 6.6: LHC point location

large LHC sets, whereas the RMSE of the two set and the highest absolute error method keep reducing towards $\sim 10^{-7}$ E_h and significantly outperforms the LHC design. Since the two set method adds points based on mean discrepancy and does not require energy data for the entire reference set, the fact that it performs the best is a significant result. Note that the LHC has been transformed to account for the fact that a greater density of points at short range will improve the result, so the actively learned data is being compared to an already somewhat intelligent' design that allocates more points in subregions of space that have more variation.

Results for the three-body non-additive potential of argon are presented in

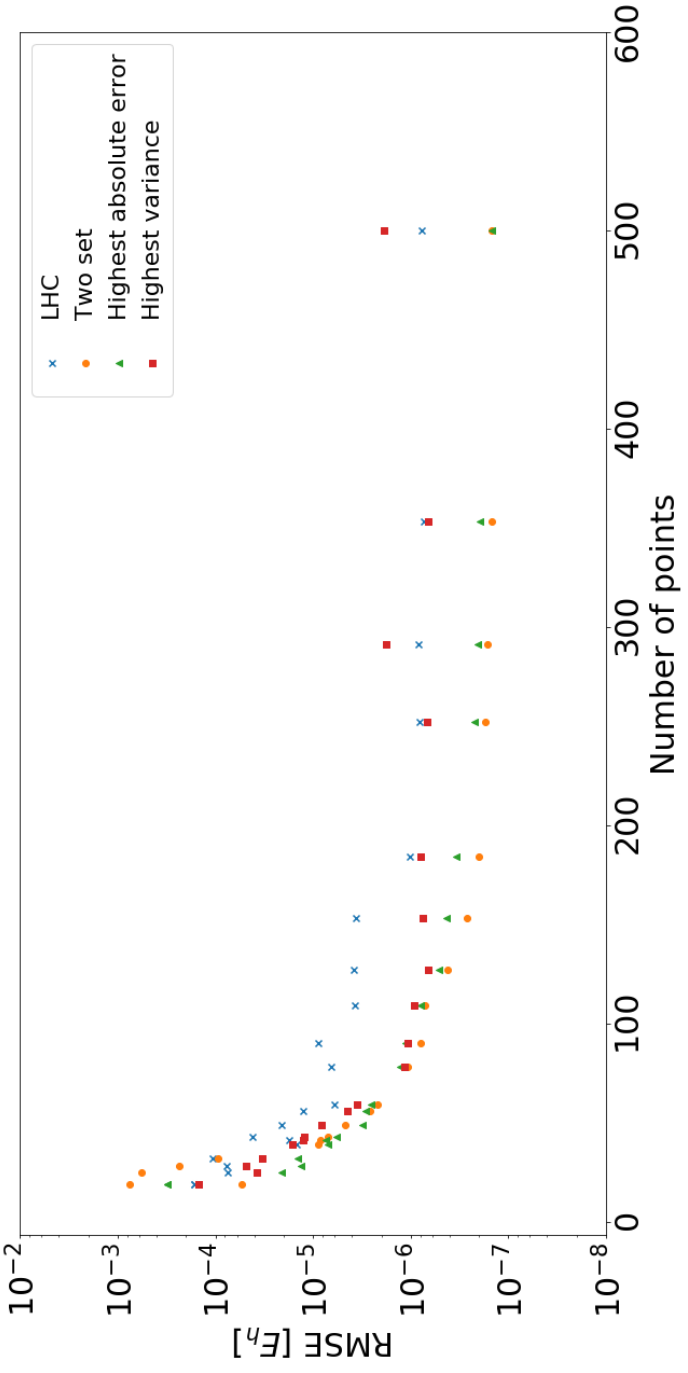


Figure 6.7: RMSE against training set size for CO_2-H_2. The lowest energy in the reference data is -8.25×10^{-4} E_h.

figure 6.7. The two set and the highest absolute error method outperform the space filling design, especially towards the high number of points where the AL methods produce errors an order of magnitude lower. The highest variance method performs erratically as there is no consistent trend with the number of points and is the worst out of all the models.

Highest absolute error relies on having a complete set of calculated energy points prior to conducting active learning, so it is limited exclusively to producing an optimal smaller subset of points from a large pre-existing data set. The other two methods, the highest variance and the two set learning, can both be expanded to generating data sets *de novo* if a way of accounting for the high energy region can be introduced to prevent the active learning algorithm adding points that are then excluded by the high energy cut-off. One possible way of doing this is to use an initial smaller latin hypercube to train a GP and in the active learning, only add points predicted to be below the high energy threshold. Another possible way of excluding the high energy points is by using a classifier that becomes more accurate as the number of points increases, allowing for a complete 'de novo' generation of data and a more refined high energy boundary. An additional consideration when modifying the algorithm to produce training sets *de novo* is to automate the energy point generation and the insertion of the point back into the GPy model for re-training. This may create computational inefficiency, as new data points need to be calculated one at a time.

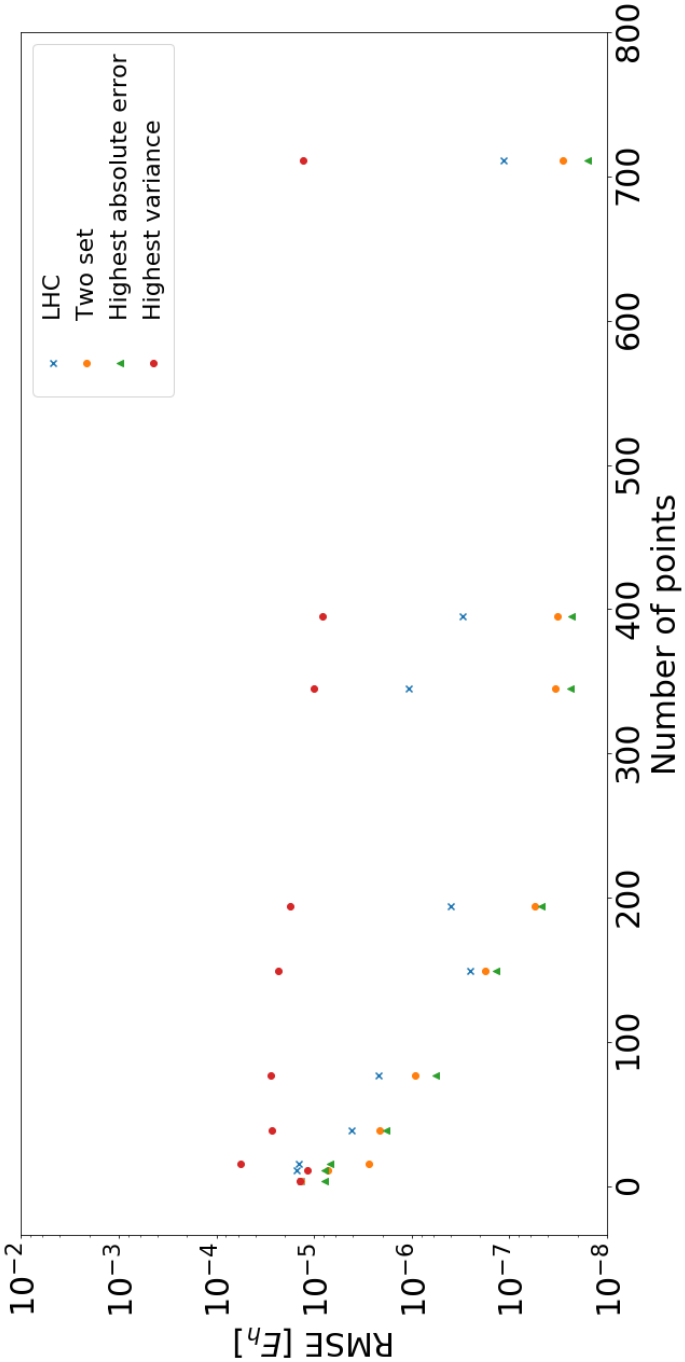


Figure 6.8: RMSE against training size for Ar_3. The lowest and highest energies in the reference data are -2.10×10^{-4} E_h and 5.31×10^{-5} E_h and the root mean square over the reference data is 1.28×10^{-5} E_h.

6.3 Future work

There is a large scope for work that builds on these results. An obvious extension is the application of the algorithms to larger chemical systems, where the need for smaller data sets becomes more imperative due to the computational unfeasibility of carrying out extensive calculations at different geometries. Increasing computational cost per point also motivates work producing data sets completely de novo, using iterative design alongside a classifier (where a classifier is a machine learning algorithm that predicts the class of a set of data points) that aims to predict high energy points and exclude them from the search over the course of the active learning algorithm. It would also be interesting to compare the use of active learning to select the best subset of points from a large data set against the use of sparse GPs. Sparse GPs are used to handle big data sets by using inducing points instead of all the data points. Inducing points are a reduced effective number of data points which do not have to be part of the training data set but could be found by (unconstrained) gradient-based optimisation of the marginal likelihood. The location of the inducing points could be compared to the actively learnt data. The algorithms that we have used rely on a random point selection for the initial training point, and one possible way of improving the robustness of the algorithm is by devising a more systematic way to choose the location of the starting point (eg. the biggest distance between points, the most central points, the highest and lowest energy points in the set etc.) Another possible

extension of the method is combining active learning with space partitioning where multiple active learning loops are run at the same time over different subsections of space. This might hold several advantages over the current full-space method used in the iterative design. It might potentially accelerate the active learning procedure as each individual model optimisation will deal with a smaller subset of points. It also has the potential advantage of making sure all different areas are sufficiently well covered and each small subsection reaches the desired accuracy threshold by using different error loop cut-offs for all individual areas.

6.4 Conclusions

Active learning of training data in the field of potential energy surface interpolation is a promising development, as it allows for the production of more time-efficient global potential energy surfaces that can be systematically improved until the desired level of accuracy has been achieved.

For the CO_2 - Ne system, the biggest relative accuracy of the two set and highest error method, compared to the space filling design, was between the 13 and 89 point range, where the methods performed 58-fold better than the space filling design and RMSE was two orders of magnitude lower.

The two set, highest error and the LHC methods reached convergence at $\sim$ 100

points. Past 100 points, the highest variance fluctuated around an error of 6.5×10^{-7} Eh which is higher than the other methods and maybe due to the added not contributing to improving the accuracy of the high energy region.

For CO_2 – H_2, the two set and highest error methods reach convergence at a much lower energy than the LHC around 200 points. This effect is even more pronounced for Ar_3 where the RMSE reaches a rough plateau (where plateau relates to only a slight reduction of error being reached upon the addition of an extensive amount of points) at around 350 points. The plot of error to training date size starts tailing of around 2.4×10^{-8} Eh for the two set method and 3.2×10^{-8} Eh for the highest error method, whereas the RMSE associated with the highest variance error fluctuates around 1.5×10^{-5} Eh and the lowest error achieved by the LHC is 1.00×10^{-5} Eh.

Since the main appeal of potential energy surfaces is their consequent application in different techniques to extract macroscopic and atomistic properties of different molecules, subregions of space where the model fared badly due to point sparsity may cause problems in the methodology of Monte Carlo and molecular dynamics simulations. This can be rectified by the use of active learning, where the goal is to place points in the most problematic areas, and hence can be used to identify these subregions of space and fill them with additional data prior to application. Since the ultimate aim of machine learning methods in potential energy surface generation is to produce high-accuracy

potentials at the lowest possible computational cost, active learning based methods of experimental design are a promising tool in achieving optimality in smaller training sets.

Active learning seems particularly well suited for the problem of interpolating a potential energy surface where there is function non-stationarity and one region has a significantly greater impact on the total predictive error.

This seems to underpin the remarkable performance of the two set and highest error methods relative to both the space filling design and highest variance as they are not distance-based learning tactics that place points based on sparsity, but can be viewed as deviation-based approached that prioritise the functional mean (i.e. the potential energy).

Chapter 7

Final Discussion

The results have shown that Gaussian Processes are a very powerful interpolative method for small molecular systems, capable of reproducing global potential energy surfaces to a very high degree of accuracy using only a relatively small pool of data points. When comparing to previous analytical and neural network fits, which achieve the same level of error using thousands of data points, Gaussian Process regression shows a lot of promise in becoming one of the default methods used in the generation of potentials of smaller systems to high quantitative accuracy. The use of inverse interatomic distances accounting for non-stationarity, and a symmetrised covariance function accounting for permutational invariance has shown to improve the efficacy of the method significantly, and are recommended techniques when creating interatomic distance based inputs for GP regression potentials. All in all, there are several avenues for future work, building up on both the research done so

far and expanding its applicative reach. One continuation is just the chemical application of the potential energy surfaces to asses the accuracy of the results obtained when using the machine learning potentials rather than solely reporting error of the generated potential. One other major expansion of this work is automation of the whole procedure, making it even easier and faster to apply routinely to different systems.

The chapter on sequential design presents the relatively unexplored field of iterative experimental design of intermolecular potentials and shows that using sequential point selection based on some information criterion can lead to big improvements in the RMSE, requiring fewer overall points to achieve exceptional levels of predictive accuracy and offers a novel way of selecting points in input space for PES-learning tasks. One possible next step for our research into sequential design is its application in a blind manner to experimental design, where blind manner is used to denote the scenario where there are no pre-existing data to chose from and the high-energy cut-off needs be determined using a classification task over the course of the algorithm. Sequential design can also be investigated as a 'topper' method for when a space-filling or grid-based sample hasn't achieved a required level of accuracy and additional points are needed to boost predictive power, in which case the sequential learning algorithm would be used alongside a stopping criterion based on error. Another option is to tackle sequential design of non-additive potentials, which are much harder to model due to greater functional variation.

The two-set design method can also form the basis for more investigative research. The jack knife approach of examining the uncertainty in hyperparameters can be expanded into the highest variance search approach, by using multiple models to try and estimate the true variance. Additionally, the effect of the number of sets used to calculate the discrepancy between predicted means on the final RMSE can be quantitatively measured as it would be interesting to see whether the three- or four-set method outperforms the two-set version. On top of the number of data sets, the other variables that might underpin the predictive accuracy of the two set method (including the number of initial starting points) form interesting points of study.

One other promising avenue of work is exploring high-dimensional molecules. Since the bottleneck of Gaussian Process learning tasks is high dimensionality, a way of space-dividing large molecules into smaller subsections and learning each localised section separately is an important stepping stone into eventually applying analogous GP algorithms to more complex molecular systems. One way of doing this is using the nearest-neighbour method of dividing up the points on the basis of the n-nearest atoms and then applying a smoothing function to smooth the boundaries between different subsections. The space subdivision method can also be used in conjuction with active learning, with points added to different subsections of space until a desired level of predictive accuracy is achieved. One benefit of this approach is that it will add more

points in subregions that are more difficult to learn and fill the easier-to-learn regions more sparsely. Additionally, since the active learning process is possible to run in parallel, it may be quicker than trying to apply the active learning algorithm over the entire compound space.

To summarise the benefits of Gaussian Process in Potential Energy Surface emulation, it is worthwhile to present a brief overview of the properties of an ideal potential emulator, (where an emulator refers to a statistical model that aims to predict the results of a computer program by acting as a 'model of a model'):

1. The emulator should produce accurate results and should ideally be developed using high level ab initio data with minimal errors.
2. The emulator should be easy to use and modify to other systems, with no bias towards one specific system or bonding type.
3. There should be a way to incorporate functional information into it directly, e.g. symmetry and differentiability.
4. The predictions should be cheap to evaluate when reference to electronic structure calculations.
5. The emulator should offer a systematic way of improving predictive performance if the desired level of accuracy hasn't been reached or if there are areas where the model performs badly.

6. The emulator should be able to produce reactive potentials that allow for bond breaking and forming.

7. There should be as much automisation as possible to reduce human labour expenditure and chance of error.

The Gaussian Process algorithm presented manages to satisfy these criteria. Making modifications to apply the algorithm to novel systems is trivial as the only major input that is needed to be entered differently when constructing the model is the number of interatomic distances and their relevant symmetries. Symmetry and functional features can easily be incorporated by altering the covariance function. Although bond reactivity was not directly addressed, this can easily be studied within the frames of the algorithm. High accuracy results have been obtained for different molecular systems and the sequential design schemes have shown to offer a way of systematically improving the predictive accuracy of the model. All in all, Gaussian processes show huge promise in being established as the routine method for developing high accuracy potential energy surfaces based on ab initio data for small to medium systems, and by tackling the method's limitations, its reach and applicability can be boosted further.

Bibliography

[1] E. G. Lewars, *Computational Chemistry. Introduction to the Theory and Applications of Molecular and Quantum Mechanics*, Springer Netherlands, 2nd edn., 2011.

[2] D. C. Young, *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*, John Wiley Sons, Inc., 2001.

[3] J. Behler, in *Chemical Modelling: Applications and Theory Volume 7*, The Royal Society of Chemistry, 2010, vol. 7, pp. 1–41.

[4] M. Born and R. Oppenheimer, *Annalen der Physik*, 1927, **389**, 457–484.

[5] J. Long, *Ph.D. thesis*, Chemistry Department, Engineering and Nature sciences, University of Iceland, 2010.

[6] D. Truhlar, *Potential Energy Surfaces and Dynamics Calculations: for Chemical Reactions and Molecular Energy Transfer*, Springer US, 1981.

[7] H. B. Schlegel, *Journal of Computational Chemistry*, **24**, 1514–1527.

[8] J. Breidung and W. Thiel, in *Prediction of Vibrational Spectra from Ab Initio Theory*, American Cancer Society, 2011.

[9] T. L. Nguyen, L. Vereecken, X. J. Hou, M. T. Nguyen and J. Peeters, *The Journal of Physical Chemistry A*, 2005, **109**, 7489–7499.

[10] A. Owens, S. N. Yurchenko, A. Yachmenev and W. Thiel, *The Journal of Chemical Physics*, 2015, **143**, 244317.

[11] R. Matthias, *International Journal of Quantum Chemistry*, 2015, **115**, 1058–1073.

[12] C. M. Handley and P. L. A. Popelier, *The Journal of Physical Chemistry A*, 2010, **114**, 3371–3383.

[13] N. Sathyamurthy, G. E. Kellerhals and L. M. Raff, *The Journal of Chemical Physics*, 1976, **64**, 2259–2261.

[14] D. Hu, Y. Xie, X. Li, L. Li and Z. Lan, *The Journal of Physical Chemistry Letters*, 2018, **9**, 2725–2732.

[15] J. Cui and R. V. Krems, *Journal of Physics B: Atomic, Molecular and Optical Physics*, 2016, **49**, 224001.

[16] B. Kolb, P. Marshall, B. Zhao, B. Jiang and H. Guo, *The Journal of Physical Chemistry A*, 2017, **121**, 2552–2557.

[17] C. M. Handley, G. I. Hawe, D. B. Kell and P. L. A. Popelier, *Phys. Chem. Chem. Phys.*, 2009, **11**, 6365–6376.

[18] F. Zielinski, P. I. Maxwell, T. L. Fletcher, S. J. Davie, N. D. Pasquale, S. Cardamone, M. J. L. Mills and P. L. A. Popelier, Scientific Reports, 2017.

[19] N. Di Pasquale, S. J. Davie and P. L. A. Popelier, *The Journal of Chemical Physics*, 2018, **148**, 241724.

[20] P. Maxwell, N. di Pasquale, S. Cardamone and P. L. A. Popelier, *Theoretical Chemistry Accounts*, 2016, **135**, 195.

[21] T. L. Fletcher and P. L. A. Popelier, *Journal of Computational Chemistry*, 2017, **38**, 1005–1014.

[22] W. J. Szlachta, A. P. Bartok and G. Csanyi, *Phys. Rev. B*, 2014, **90**, 104108.

[23] A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B*, 2013, **87**, 184115.

[24] S. De, A. P. Bartok, G. Csanyi and M. Ceriotti, *Phys. Chem. Chem. Phys.*, 2016, **18**, 13754–13769.

[25] Z. Li, J. Kermode and A. De Vita, 2015, **114**, year.

[26] Y. Guan, S. Yang and D. H. Zhang, *Molecular Physics*, 2018, **116**, 823–834.

[27] F. Jensen, *Introduction to Computational Chemistry*, Wiley, 2nd edn., 2006.

[28] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, John Wiley and Sons, 2nd edn., 2004.

[29] D. A. Gilbert, *Introduction to Computational Chemistry Theory*, 2011.

[30] C. D. Sherill, *An Introductiont to Hartree Fock Molecular Orbital Theory*, 2000.

[31] P. Cassam-Chenaï, *Chemical Physics Letters*, 2006, **420**, 354 – 357.

[32] A. S. Ichimura, *Introduction to Computational Chemistry*, 2004.

[33] C. C. J. Roothaan, *Rev. Mod. Phys.*, 1951, **23**, 69–89.

[34] G. G. Hall and J. E. Lennard-Jones, *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 1951, **205**, 541–552.

[35] W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, 1133–1138.

[36] G. Alagona and C. Ghio, *Ab Initio Theoretical Methods for Studying Intermolecular Forces*, 2003.

[37] A. D. Buckingham and B. D. Utting, *Annual Review of Physical Chemistry*, 1970, **21**, 287–316.

[38] J. Rowlinson, *Cohesion: A Scientific History of Intermolecular Forces*, Cambridge University Press, 1st edn., 2002.

[39] J. N. Israelachvili, *Intermolecular and Surface Forces*, Academic Press, 3rd edn., 2011.

[40] A. Stone, *The Theory of Intermolecular Forces*, Oxford, 2nd edn., 2013.

[41] J. van der Waals, *Over de continuïteit van den gas- en vloeistoftoestand*, Sijthoff, 1st edn., 1873.

[42] N. Winter, in *Forces Maintaining the Stability of Tertiary Structure*, ed. E. Bell, Springer New York, 2021, pp. 1–3.

[43] *Understanding Intermolecular Interactions in the Solid State*, ed. D. Chopra, The Royal Society of Chemistry, 2019, pp. 1–339.

[44] H. R. Jagdish Mehra, *The Historical Development of Quantum Theory Volume 1*, Springer-Verlag New York, 1st edn., 2001.

[45] R. P. Feynman, *Phys. Rev.*, 1939, **56**, 340–343.

[46] *Physik. Z*, 1921, **22**, 129–141.

[47] R. Eisenschitz and F. London, *Zeitschrift für Physik*, 1930, **60**, 491–527.

[48] S. Boys and F. Bernardi, *Molecular Physics*, 1970, **19**, 553–566.

[49] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2005.

[50] M. Ebden, *Gaussian Processes for Regression: A Quick Introduction*, 2008.

[51] H. M. Wallach, *Introduction to Gaussian Process Regression*, 2005.

[52] B. Wetenhall, H. Aghajani, H. Chalmers, S. Benson, M.-C. Ferrari, J. Li, J. Race, P. Singh and J. Davison, *Energy Procedia*, 2014, **63**, 2764 – 2778.

[53] Y. Yongchao, T. Hongsheng, Y. Jinghua, S. Qinghua and C. Shuangge, *MATEC Web Conf.*, 2017, **100**, 02004.

[54] T. Morawietz, A. Singraber, C. Dellago and J. Behler, *Proceedings of the National Academy of Sciences*, 2016, **113**, 8368–8373.

[55] D. L. Chen, W. Saidi and J. Karl Johnson, 2012, **24**, 424211.

[56] M. del Cueto, A. S. Muzas, G. Füchsel, F. Gatti, F. Martín and C. Díaz, *Phys. Rev. B*, 2016, **93**, 060301.

[57] J. M. Hutson, *Advances in Molecular Vibrations and Collision Dynamics*, 1991, **1**, 1–45.

[58] R. Prosmiti, C. Cunha, P. Villarreal and G. Delgado-Barrio, *The Journal of Chemical Physics*, 2002, **117**, 7017–7023.

[59] M. Solimannejad, H. Jouypazadeh and H. Farrokhpour, *Molecular Physics*, 2015, **113**, 3303–3311.

[60] M. Karthikeyan and R. Vyas, in *Machine Learning Methods in Chemoinformatics for Drug Discovery*, Springer India, New Delhi, 2014, pp. 133–194.

[61] X. Deng, V. R. Joseph, A. Sudjianto and C. F. J. Wu, *Journal of the American Statistical Association*, 2009, **104**, 969–981.

[62] G. Riccardi and D. Hakkani-Tur, *IEEE Transactions on Speech and Audio Processing*, 2005, **13**, 504–511.

[63] K. Toyoura, D. Hirano, A. Seko, M. Shiga, A. Kuwabara, M. Karasuyama, K. Shitara and I. Takeuchi, *Phys. Rev. B*, 2016, **93**, 054112.

[64] E. Uteva, R. S. Graham, R. D. Wilkinson and R. J. Wheatley, *The Journal of Chemical Physics*, 2017, **147**, 161706.

[65] A. J. Cresswell, R. J. Wheatley, R. D. Wilkinson and R. S. Graham, *Faraday Discuss.*, 2016, **192**, 415–436.

[66] S. D. Whitehead and D. H. Ballard, *Machine Learning*, 1991, **7**, 45–83.

[67] S. Thrun, in *Handbook for Intelligent Control: Neural, Fuzzy and Adaptive Approaches*, Van Nostrand Reinhold, Florence, Kentucky, 1992.

[68] L. Atlas, D. Cohn, R. Ladner, M. A. El-Sharkawi and R. J. Marks, II, in *L. Atlas, D. Cohn, R. Ladner, M. A. El-Sharkawi and R. J. Marks, II*, ed. D. S. Touretzky, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990, ch. Training Connectionist Networks with Queries and Selective Sampling, pp. 566–573.

[69] J. Schmidhuber, J. Storck and S. Hochreiter, *Proc ICANN'95 (Paris)*, 1995, **2**, 159–164.

[70] D. J. C. MacKay, *Neural Computation*, 1992, **4**, 590–604.

[71] M. Rupp, M. R. Bauer, R. Wilcken, A. Lange, M. Reutlinger, F. M. Boeckler and G. Schneider, *PLOS Computational Biology*, 2014, **10**, 1–8.

[72] E. V. Podryabinkin and A. V. Shapeev, *Computational Materials Science*, 2017, **140**, 171 – 180.

[73] Z. Li, J. R. Kermode and A. De Vita, *Phys. Rev. Lett.*, 2015, **114**, 096405.