University of Nottingham

School of Mathematical Sciences

Determining the number of training points required for machine learning of potential energy surfaces.

M. Pearson.

A thesis submitted to the University of Nottingham for the degree of

MPhil

November 2020

Dedicated to my secondary school teachers who made science my favourite lessons:
Mr Harvey who tolerated my digressions into explosives, and Mrs Norton for
regularly detaining me after school at 'the Mrs Norton appreciation society'.

Abstract

In recent years, there has been an explosion in the use of machine learning, with applications across many fields. One application of interest to the computational chemistry field is the use of a method known as Gaussian processes to accurately derive a system's Potential Energy Surfaces (PES) from ab-initio input-output data. Gaussian processes are a stochastic process, or collection of data, each finite group of which has a multivariate distribution. When modelling the PES of a system with GPs, the cost of computation is proportional to the number of sample points, and in the interests of being economical it becomes imperative to use no more computing time than in necessary.

When examining the $H_2O - H_2S$ system, 10,000 sample points was found to be insufficient to accurately model the PES, raising the question: how many points are needed, and what makes this system so challenging?

The root mean squared error, or RMSE, provides a non-negative measure of the absolute fit of a model to sample data. PESs for a selection of different dimers were modelled using an LHC regime and a GP, and the RMSE tested against a set of test data. An LHC or Latin hypercube is a method of multidimensional distribution used to generate a near random sample of parameter values. From the RMSE data a parametric regression was implemented to find the number of sample points required n_{req} to achieve a benchmark precision of 10^{-5} Hartrees (E_h), and from a collection of these a correlation observed between the relative difficulty of a system and geometric and chemical characteristics of each system.

An exponential correlation was observed between n_{req} and number of Degrees of Freedom (DoF) of a system, making it the principal determinant of difficulty. A strong negative correlation was also observed between the number of permutations in a symmetry group and the difficulty of that system, with a distinction made between the effects of 'flip' and 'interchange' symmetries, which reduce the points required by 50% and 34% respectively. The difficulty of systems also positively correlates with energy well depth, atomic size and atomic size disparity, though these are not so easily unpicked and quantified. With DoF and symmetry in mind, a general equation for estimating n_{req} was formulated, and a 6 DoF system was projected to require upwards of 32,000 sample points to achieve benchmark accuracy.

Since the cost of calculating a PES of a system is proportional to the number of sample points included, and high performance computer time is limited, the ability to estimate n_{req} permits better management of the computational effort. Moving forward, the methodology outlined may be used to appraise further systems of interest before committing processor time.

Acknowledgements

I would like to pay my special regards to my supervisors: Richard Graham and Richard Wheatley, who throughout this project have patiently provided guidance and feedback in abundance. I would also like to thank the EPSRC and ERA for their studentship, and the University of Nottingham for use of their high performance computers.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

BACKGROUND.

Currently the growth in use of Machine Learning has spread to many fields, from commodity price prediction [**Guo. 2012**], to classification of celestial bodies [**Davies. 2015**], to agricultural management [**Liakos. 2018**]. The interest in machine learning often stems from the ability to iteratively optimise a computer's ability to perform a task without the need for explicitly programming the process. As a result, some degree of automation is achievable and human involvement may be minimised, benefiting productivity where expensive or repetitive tasks are involved.

Support Vector Machines (SVMs) are Machine Learning models which make use of algorithms to perform classification [**Suykens. 1999**] and regression analysis [**Smola. 2004**]. To perform these operations the SVM may make use of a range of strategies, amongst which are Gaussian similarity kernels, which in principle are Gaussian Processes. A Gaussian Process is a stochastic process where between each data point there is a multivariate normal distribution, so the calculated data is assumed to be exact at points x_i and y=f(x) is predicted probabilistically for all values of x along with their marginal distribution.

The machine learning in this instance is applied to molecular potential energy surfaces. A potential energy surface (PES) describes the energy of a molecular system with respect to the Cartesian coordinates of a molecule.

Dependent on the number of atoms present in the molecules, this may be a very high dimensional surface, as in the case of the $H_2O - H_2S$ system, which has 6 degrees of rotational freedom. The energy value is highly variable in response to the manipulation of intermolecular distances. There is slight attraction when two molecules sit within one another's energy "well", where various attractive forces cause the molecules to pull one another together. At a larger distance this interaction between the opposing charges rapidly decays, according to Coulomb's law, resulting in an energy close to zero. When intermolecular distances converge to zero the energy can be seen to rise asymptotically due to repulsive forces. The attractive forces in a system may include hydrogen bonding between a weakly positive dipole and a lone pair of electrons. Another permanent source of dipole interaction takes place between a weakly positive and weakly negative atom, which occurs as a more electronegative element pulls electron density away from its neighbouring atom. Without permanent dipoles, interaction is predominantly dictated by van der Waal's forces, which include: the contributions of rotating permanent dipoles with one another, permanent dipoles inducing shifts in neighbouring electron clouds, and the significant contributions of London dispersion forces. London dispersion forces do not rely on permanent dipoles, making them abundant even between non-polar molecules. Mechanistically, temporary positive or negative dipoles within an electron cloud give rise to complimentary charges in the electron fields of the adjacent molecule. The Coulombic repulsive forces are due to negative-negative electrostatic interaction between the electron clouds, and nucleus-nucleus interaction. As the electron fields overlap further, the Pauli Exclusion Principle begins to dominate and a repulsive quantum mechanical effect called the exchange interaction manifests, as electrons may not occupy the same state or location at the same time [**Stone. 2013**]. The contributions of all these many-body interactions make the study of model systems necessarily complex, hence the need for ab-initio simulation rather than decomposition to simple pairwise contributions.

1.1 AB-INITIO CHEMISTRY.

Present methods of determining material behaviours are frequently based upon empirical observations. However, alternatives exist which more accurately model the interactions of molecules by emulating the electron clouds surrounding them, using spatially-dependent functionals. Taking the level of precision a step further, the behaviour of electron clouds can be determined more exactly by finding solutions to the Schrödinger equation, in what is known as ab-initio calculation.

Ab-initio (from the start) computation of the intermolecular PES is effective as it accurately resolves the aforementioned forces by emulating the electron clouds from established quantum principles. The downside is frequently the cost of lengthy calculations, which naturally leads efforts to interpolation of smaller data sets to save computing time while still obtaining an accurate PES [**Cui. 2016**]. The problem arising is that parametric interpolation is a flawed system, making use of a best fit which may not coincide with the true potential at a given point. As a result, machine learning methods such as SVM regression which respect the calculated data are gaining traction for PES interpolation.

Ab-initio molecular simulation is a useful tool for predicting the physical behaviours of chemical species owing to its superior level of accuracy when compared with empirical methods. The cost of this level of accuracy is the computational expense of calculations from first principles. Upon expanding this method to systems with more electrons or degrees of freedom, the computational burden increases by orders of magnitude [**Yazal. 2001**]. As a result it is primarily only used to calculate interactions in small systems, requiring relatively fewer sampled points.

Using this family of methods, two body and three body systems may be examined with a view to resolving the physical properties of a system.

Two body interactions are defined by the expression $\Delta^2 E_{AB} = E_{AB} - E_A - E_B$ where $\Delta^2 E_{AB}$ is the sum of forces such as Coulombic attraction and repulsion, charge transfer polarisation, and Pauli Exchange Interaction, E_{AB} is the pairwise interaction and E_A and E_B are the internal energy of molecule A and B [**Huang. 2015**].

Three body interactions build from this, and are more present in condensed phase systems, due to the increasing particle density compared to a gas. This can be described by the expression $\Delta^3 E_{ABC} = E_{ABC} - \Delta^2 E_{AB} - \Delta^2 E_{BC} - \Delta^2 E_{AC} - E_A - E_B - E_C$ where $\Delta^3 E_{ABC}$ is the non-additive three body interaction [**Huang. 2015**]. These potentials may be obtained and used in a PES to predict such physical properties as boiling point of a system [**Hellmann. 2017**], and for a modest computational cost the additive potentials may be implemented if the system requires it [**Oakley. 2009**]. Ab-initio computational chemistry is an attractive prospect to many, as it offers the opportunity to simulate systems which may be experimentally infeasible due to difficulty, cost, safety or structurally unstable conformations.

Among the family of Ab-Initio methods there is the well known Hartree-Fock method, and post Hartree-Fock methods such as Møller-Plesset perturbation theory and Coupled Cluster methods. Among these, the Coupled Cluster methodology is considered the gold standard to accurately solve the time-independent Schrödinger equation [**Niu. 1997**]. The time independent equation may be derived using the time dependent wavefunction.

The time dependent wavefunction Ψ is defined by

$$i\bar{h}\frac{\partial}{\partial t}\Psi = \hat{H}\Psi, \tag{1}$$

assuming the solution can be factored as follows

$$\Psi(r,t) = \psi(r)\exp\left[\frac{-iEt}{h}\right], \tag{2}$$

where E is the energy value, t is time and h is Planck's constant [**Kuhn. 2009**]. By substituting equation (1) into (2), the time independent Schrödinger equation is derived

$$\hat{H}\Psi = E\Psi, \tag{3}$$

where $\hat{H}$ is the electronic Hamiltonian operator, Ψ is the wave function and E is the total energy associated with the system. The Hamiltonian is a sum of the potential energy $\hat{T}$ of the nuclei T_n and electrons T_e in question, as well as the potential energy $\hat{V}$ between the concerned particles, be they electron-electron, electron-nucleus, or nucleus-nucleus

$$\hat{H} = T_n + T_e + V_{e-e} + V_{e-n} + V_{n-n}, \tag{4}$$

In practice an approximation of the wavefunction is a linear combination of many electron configurations, and each orbital is described using a linear combination of basis functions for post Hartree-Fock methods,

$$\psi_i = \sum_{\alpha} C_{\alpha i}\chi_{\alpha} \tag{5}$$

where the basis functions χ may be one of a number of functions of the interatomic distance, r, including Contracted Gaussian $C_i e^{-\alpha r^2}$ or Slater $e^{-\alpha r}$, with some angular descriptor such as x, y or z for P orbitals. Provided appropriate basis functions are chosen, the larger the ζ number, or number of basis functions used to represent the wave function of the Schrödinger equation, the better the agreement with the actual wavefunction. This improvement in fit is due to the addition of functions resulting in uniform convergence to the electronic wavefunction. Along with this is an increase in computational cost, proportionate to the size of the basis set. The number of basis functions commonly implemented include single, double, or triple ζ, and a Complete Basis Set (CBS) may even be extrapolated using triple and quadruple ζ together for greater accuracy. The basis functions are then added to describe orbitals. In this study the basis set used is the augmented correlation-consistent triple-zeta (aug-cc-pVTZ) basis set.

An established software platform used to perform ab-initio calculations is Molpro, which offers Coupled Cluster Single-Double excitation and Triple perturbation (CCSD(T)) functionality, and the Møller-Plesset (MP2) method used in this study [**Werner et al. 2018**]. Coupled cluster refers to a post Hartree-Fock numerical technique to construct multi-electron wavefunctions using an exponential cluster operator to account for electron correlation. In the case of CCSD(T), the single electron and double electron excitation terms are calculated explicitly, whereas the triple excitation (brought about by the contributions of three electrons) is approximated non-iteratively using perturbation theory. Møller-Plesset methodology is another post Hartree-Fock numerical technique, where electron correlation is approximated using Rayleigh-Schrödinger perturbation theory, with the numeral in MP2 denoting a second order perturbation. Perturbation theory involves taking a known mathematical solution to the electron correlation and adding a perturberance, or small disturbance to the solution. Small changes to the observed physical characteristics may then be characterised using a truncated asymptotic series to find approximations of the remaining terms in electronic correlation series.

1.2 GAUSSIAN PROCESSES.

The practice of regression using a Gaussian method (kriging), is achieved within a type of machine learning called Support Vector Machines. These are supervised methods of machine learning typically used for regression and classification, employing a covariance function (kernel) suitable to the task at hand. This kernel indicates how much two variables x and x' change together, and may be of a linear, polynomial, or radial basis function, also known as the Gaussian similarity function [**Rasmussen. 2006**].

The utility of a Gaussian regression as opposed to a parametric fit is that it constructs a model explicitly composed of the combinations of training data provided, and obliquely interpolates using this data and hyperparameters. This is in contrast to simply placing a trend line based on explicit parameters which may not pass through all points, and in all likelihood would be simply impractical in a high dimensional system. This is accomplished by a kernel function which correlates clusters of points and their inner product in higher dimensional space in what is called the "kernel trick". This ultimately leads to the most probable surface between data points. The kernel itself is commonly of the squared-exponential form.

The squared exponential kernel is

$$\kappa(x, x') = \sigma_0^2 \exp\left[-\frac{1}{2}\left(\frac{x - x'}{\lambda}\right)^2\right], \tag{6}$$

where λ represents the correlation length-scale, and σ_0 is the signal amplitude, comprising the hyperparameters. The parameters x and x' represent two points. From this, the covariance function generalised to N interatomic distances.

This kernel is popular due to both its stationarity, and the fact it has only two hyperparameters per dimension to optimise; the variance σ and length-scale λ. Stationarity refers to the fact the squared exponential is a function of $(x - x')$ making it invariant to translations in the input space, which is to say that there is no spatial trend in the multivariate distribution [**Lapidoth. 2017**]. To best fit the data in multiple dimensions, an Automatic Relevance Determination (ARD) kernel is used, the only difference being that λ is a vector of the same dimensionality as the data being fitted. This is necessary for chemical data as the interaction, for example, between a hydrogen and neon atom differs from the interaction between an oxygen and neon atom due to differences in radius, electron cloud, and potential energy surface.

$$\kappa(x, x') = \sigma_0^2 \prod_{i=1}^{N_D} \exp\left[-\frac{1}{2} \left(\frac{x_i - x_i'}{\lambda_i} \right)^2 \right]. \tag{7}$$

Another component of a Gaussian process is the nugget, a simple noise variance parameter. If Y is a vector of energy potentials in a training set $\{x_i, Y_i\}_{i=1}^{N}$, and there are N values of i, GP regression may be used to find the value of $f(x)$ at point x' by the equation

$$f(x') = K_*^T K^{-1} Y, \tag{8}$$

where K_* is a vector of covariances between x_* and all values of x_i, and K is the positive-definite covariance matrix,

$$K = \begin{bmatrix} k(x_1, x_1) + \sigma_n^2 & k(x_1, x_2) & \cdots & k(x_1, x_N) \\ k(x_2, x_1) & k(x_2, x_2) + \sigma_n^2 & \cdots & k(x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N + \sigma_n^2) \end{bmatrix}, \tag{9}$$

where σ_n^2 is the nugget, accounting for noise in the training set. The nugget is discussed further, later in this section.

With this covariance function, an optimisation algorithm will identify the best kernel hyperparameters to fit the function about the marginal distribution. A key quantity in obtaining the likelihood function, which is the probability of obtaining the training data, given the hyperparameters. Optimising the hyperparameters may include a downhill descent algorithm which identifies the lowest negative log likelihood to find the values of σ and λ that provide the most agreeable fit, whilst also iterating through a number (typically 30 in this application) of random initial hyperparameter values to avoid any local minima providing erroneous results. The benefit of optimising with the natural logarithm of likelihood as opposed to likelihood in numerical methods is its asymptotic nature, which avoids the rapid transitions between zero and large sums which may challenge a numerical method. Respecting symmetry in a system, and making use of the systems permutation group, permits efficient calculation of any equivalent points without wasted computation. As it stands, this kernel does not respect this required symmetry. Ultimately, the GP predictions are sums over the kernel function, hence if the required symmetry can be imposed on the kernel function then the GP predictions will inherit this symmetry. There exists an algorithm to impose the symmetry on the kernel which takes the permutation group of the molecule at hand and adapts the GP to present a symmetric energy surface [**Uteva. 2017**]

$$\kappa_{sym}(x, x') = \sum_{g \epsilon G} \kappa(gx, x'). \tag{10}$$

The symmetric kernel respects permutations in that $\lambda_i = \lambda_j$ when coordinates x_i and x_j permute. This is accomplished by recognising that the interaction potential between, for example, hydrogen and neon will hold true at any other conformation where hydrogen and neon could interact similarly (for example, at an equal interatomic distance). G represents the permutation group, and g represents a permutation within that group.

One further component of the Gaussian Process is the nugget. The nugget is a constant meant to account for noise in the data. It is typically small relative to the total variance observed, and serves effectively to minimise the detriment of small scale variations and rounding in the data set upon the model. The nugget then becomes optimised by the data and settles at a low value for a smooth surface, or a relatively larger one if the data features much short range variation.

1.3 LATIN HYPERCUBES.

Latin hypercube (LHC) sampling is a statistical method of generating an evenly distributed and near random array of multidimensional sample points. The geometry of the hypercube builds from that of the Latin square, in which each row and column of a grid may only be occupied by a single sample point, not unlike the placement of a given number in a sudoku puzzle. When expanded to multidimensional space this allows for a large number of dimensions to be allocated in a fairly homogeneous fashion. Furthermore, there are many possible arrangements that satisfy the LHC requirement for a given LHC size and these can readily be generated via a stochastic algorithm. It is possible to determine the 'best' of these as the LHC with the largest minimum distance between points, with the intention that this indicates the most homogeneously distributed sample array.

The Latin hypercube sampling method was chosen over grid based sampling approaches and sampling methods based upon random distribution such as Monte Carlo, due to perceived disadvantages to these methods. While grid sampling works well in a 2 dimensional or 3 dimensional sampling methodology, as the dimensionality increases in more complex molecular systems, the number of sample points utilised will increase by an order of magnitude each time. Not only is this computationally costly, but it also runs the risk of specifying a sample with many points of limited usefulness. While the pre-processing scripts which select the sampling array with the largest minimum distance between any two points would certainly allow the selection of a homogeneous sample set, the random nature would most likely require far more arrays to be created before finding one of comparable minimum distance to those generated by LHC sampling.

1.4 APPLICATIONS OF GAUSSIAN PROCESSES TO DATA FROM COMPUTATIONAL CHEMISTRY.

Due to the cost of ab-initio calculations only a small number can be performed, so interpolation methods must be used to infer the results between existing calculations. The use of Gaussian machine learning methods has been shown to be a valuable tool when applied to interpolate a discrete data set in many fields, including quantum chemical topology [**Handley. 2009**], where the Popelier research group utilised a Gaussian regression technique to model the polar moments of water molecules. In the field of material chemistry, Gaussian Processes have been utilised in conjunction with calculated energy potentials to reproduce dislocation structures and behaviours in a tungsten crystal [**Szlachta. 2014**]. Most recently, Gaussian Processes have been applied to interpolate intermolecular potentials derived from ab-initio calculations with a high degree of fidelity for a bimolecular system [**Uteva et al. 2017**].

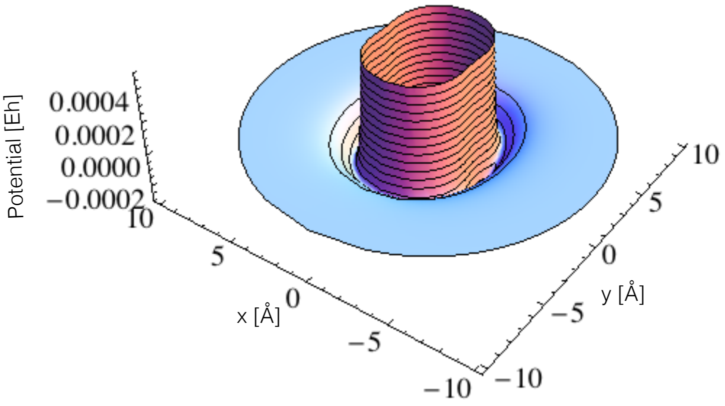

Figure 1.: A plot of the 2D PES for a carbon dioxide - nitrogen system using a GP taken from Uteva et al (2017). The z axis is the interaction potential and the x and y axes denote the position of the Ne atom. The CO2 molecule lies along the y-axis with the carbon atom at the origin.

The study by Uteva made use of a Python library called GPy, which allows the use of Gaussian processes in order to interpolate and extrapolate a data set [**GPy. 2012**]. The training data was constructed spatially using a Latin Hypercube regime, with geometric and energy constraints to focus ab-initio calculations to a region typical of molecular interaction and not waste computing time while still getting a thorough placement about the molecules concerned. This method also made use of inverse centre to centre inter-atomic distances as opposed to centre to centre distance within the LHC construction programme, with a view to placing more sample points proximate to the origin, therefore concentrating more effort on representing close range interactions. These close range interactions are significant as the energy potentials are more pronounced than distant interactions which converge to zero, as observable in figure 1, making them valuable for constructing an accurate PES once a good design strategy had been demonstrated. The study also made use of upgraded ab-initio calculations with complete basis set (CBS) extrapolation of the CCSD(T) interaction energy from the triple ζ and quadruple ζ basis sets, yielding better approximations of the experimentally obtained cross Virial coefficient than the GERG equation of state for a $CO_2 - CO$ system [**Uteva et al. 2017**].

1.5 AIMS OF THIS WORK.

Chapter 3 shows that a direct application of these methods to the $H_2O - H_2S$ dimer fails to give a PES that is as accurate as work done previously on simpler dimers. In this study a series of Potential Energy Surfaces will be constructed for a selection of dimers using en-masse LHC placement, and comparisons made to a large test set composed of MP2 level calculations with triple ζ basis set for the relative accuracy and low computational cost. It is hoped that the data acquired will show that the $H_2O - H_2S$ dimer is hard to resolve to a reasonable precision, given the relative ease with which

$CO_2 - CO$ was modelled in the work of Uteva [**Uteva et al. 2018**], then shed light on possible sources of difficulty. It is hoped that the apparent results will allow conclusions to be drawn about a multitude of parameters inherent to a system, including molecular shape and associated degrees of freedom, numbers of atoms, symmetries and atomic size.

2

METHODOLOGY.

In essence, modelling a dimer requires a series of sufficiently informative arrangements of the two molecules in question. To maximise the usefulness of each conformation, those where the molecules are so far away that the interaction energy tends to zero can be discounted, as can arrangements so close together that the energies involved become thermally inaccessible. A further consideration is precisely what phase space around the molecule is symmetrically distinct. A neon atom for instance, is invariant to an observer from any angle, meaning only distance must be considered when calculating a PES for two interacting neon atoms. A nitrogen molecule by comparison has an end-to-end symmetry which effectively eliminates one hemisphere of indistinct phase space, and due to its invariance when spun about its longest axis, and reflective symmetry about this axis, only 90° in a single plane is distinct. By focusing the placement of dimers with respect to one another into this distinct region, superfluous conformations can be avoided, thereby allowing a more optimal distribution of samples to be taken from the most informative zones of each molecule. It is ensured that points are spaced effectively using a series of random Latin Hypercube based distributions, and the cube with the largest minimum distance between any two sample points is selected.

2.1 MOLECULAR DEGREES OF FREEDOM.

To describe a rigid molecule's placement in space some convenient central point must be designated the datum, and the molecules orientation described from here. A spherical molecule located at the datum point $[0,0,0]$ such as Ar is invariant to rotations about the X, Y and Z axis, a linear molecule is invariant to rotation about its longest axis, and a bent molecule is variant about all three axes. When considering a pair of interacting molecules the distance between them and their orientation must also be considered. The spatial configuration of a rigid molecule relative to another may be described using a center-to-center distance (r) and three angles, α, β, and γ. These parameters are used to describe geometric constraints, during the design of an LHC.

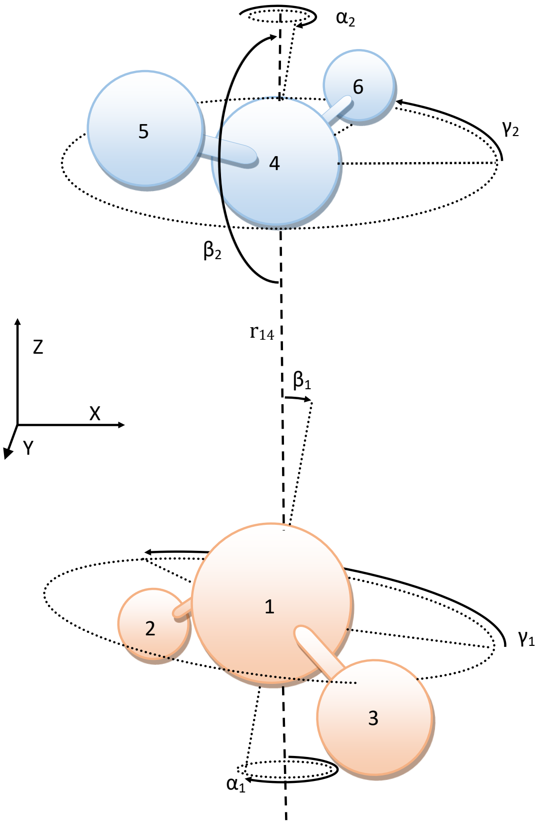

Figure 2.: A diagram of the available rotations about the X, Y and Z axis when observing interaction between two bent molecules.

It may be observed in figure 2 how this combination of angles and distance look with respect to a $H_2O - H_2S$ system. Only a single α angle is required as the two are commutative to one another, β describes a zenith angle, and γ provides an azimuthal angle.

The angles shown in figure 2 are used as co-ordinates for the Latin hypercube design. Because of symmetries in the individual molecules and of the overall system the full range of all angles is not required. We define the symmetrically distinct region as the smallest possible range of angles required to cover all possible molecular configurations, when the symmetries of the dimer are accounted for. The symmetrically distinct region comprises of a reduced, but continuous range for each of the angles and depends on the geometry of each molecule in the dimer. Table 2 shows the symmetrically distinct region for a range of dimer types.

Another method for describing molecular orientation is to use all interatomic distances, which is less helpful for a human, due to its raw data form, and the likelihood with which one may over-specify the system. This is shown using the same $H_2O - H_2S$ system in table 1 and 3.

Table 1. Table of vectors

Index	0	1	2	3	4	5	6	7	8
Species	O-S	O-H	O-H	H-S	H-H	H-H	H-S	H-H	H-H
Distance	r_{14}	r_{15}	r_{16}	r_{24}	r_{25}	r_{26}	r_{34}	r_{35}	r_{36}

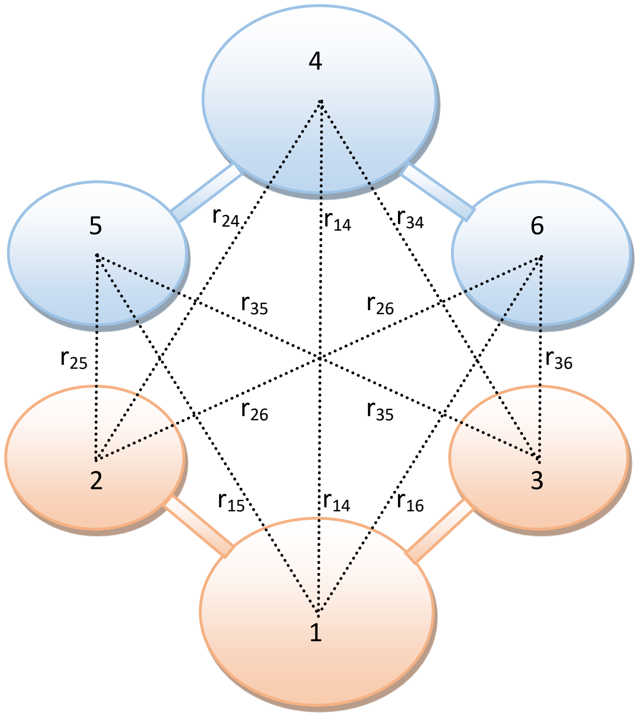

Figure 3.: A diagram of the nine available intermolecular distances when observing interaction between the two bent molecules illustrated in figure 2.

In figure 3 the 9 available intermolecular distances between the atoms of a pair of bent molecules are shown alongside the index and species interaction given in table 1. Given a known bond length within a molecule, this regime functionally over specifies the system but is nevertheless the preferred input format for the GP as a selection of inverse distances each correlate to the energy potential more easily than variables including a single distance and several angles. This strategy also makes specifying permutability within the system straightforward. Since the method is easier for the user to visualise, when designing a Latin Hypercube the extent of the bimolecular conformations are specified using the intermolecular rotations method depicted in figure 2, and these ranges are converted trigonometrically to the intermolecular distances.

2.2 RMSE AS A METRIC OF PRECISION.

To ascertain the accuracy of the Gaussian process, the energy surface derived from the training data must be compared to a much larger independent set. By using training data with various different geometries and an associated energy value,some insight can be obtained as to the effect training set size has on the Root Mean Square Error (RMSE), which is a useful metric by which to judge the quality of interpolated energy values.

$$RMSE = \sqrt{\frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \left[f^{GP}\left(x_i\right) - f^{Molpro}\left(x_i\right) \right]^2}. \tag{11}$$

The symbols f^{GP} and f^{Molpro} represent the potential energy determined by Gaussian process and Molpro respectively, and N_{test} represents the number of points in the test set.

2.3 ACTIVE LEARNING.

Active learning is a further sampling methodology which iteratively selects points to examine potentials at. In this case it makes use of Gaussian processes and a performance metric such as the point with the largest error with respect to a test set, which is then used to select training points sequentially as opposed to en-masse, based on the degree of error between any test point and its counterpart on the evolving potential energy surface.

In practice this process utilises 3 data sets, a normative set for testing against, a reference set which the algorithm selects geometries from, and a working training set composed of the selected points which are used to construct the Potential Energy Surface.

In principle this could offer an accurate insight into a surface with far fewer points than a pre-made sample regime, since the points are each selected after a dynamic analysis of the most recently calculated energy surface [**Uteva et al. 2018**].

2.3.1 *Linear Molecules.*

The nitrogen molecule interacting with neon represents possibly the simplest non-trivial molecule examined, due to the invariant geometry of a linear molecule upon rotation about its longest axis and inversion symmetry. The placement of a neon atom relative to the nitrogen molecule may therefore be described in polar terms, simply using the centre to centre distance r and an angle β from the longitudinal axis of the nitrogen molecule. Since the molecule is symmetric end-to-end, the angle β need only range from $0° - 90°$, thus describing the symmetrically distinct region for nitrogen. The same is true of the symmetrically distinct region of $CO_2 - Ne$, along with any linear molecule possessed of an end-to-end symmetry.

Similarly, carbon monoxide is rotationally invariant about its longest axis, but does not possess the same end-to-end symmetry, and the value of β must range from $0° - 180°$ when interacting with neon to cover the distinct region.

2.3.2 *Bent Molecules.*

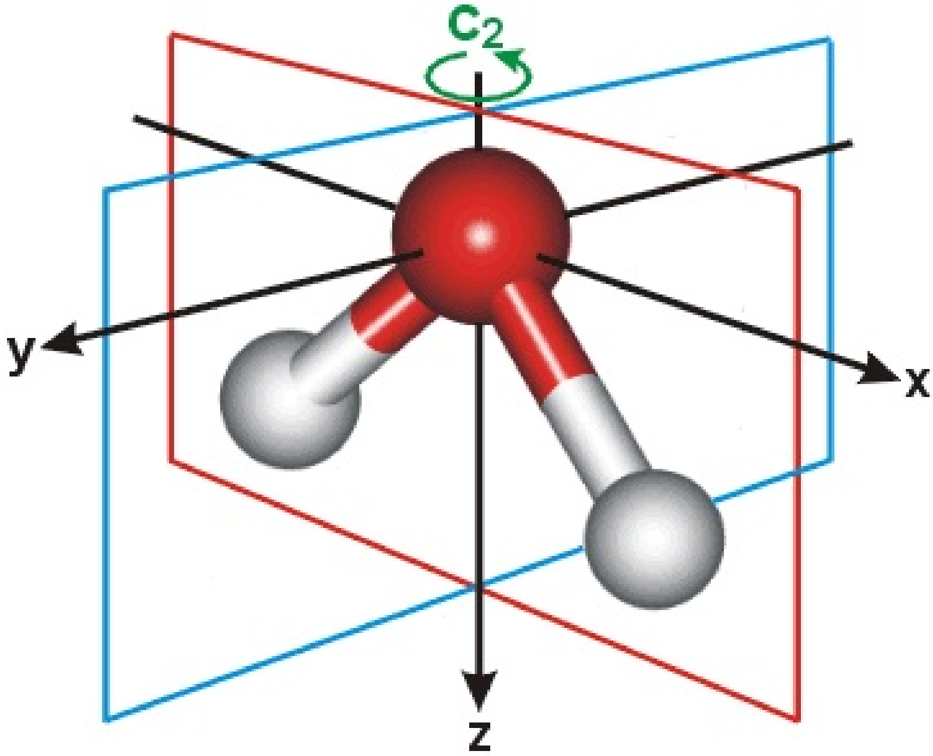

Figure 4.: A diagram of the symmetric planes of a bent molecule, and the axis of rotational symmetry.[**Chaplin. 2019**]

Given the symmetric planes and rotational symmetry illustrated in figure 4, some spatial orientations of a bent molecule are superfluous and the range of angles specified in the construction of the hypercube may be truncated. In figure 2 the rotations α_1 and α_2 can be seen to complement one another and therefore in all the systems examined, only the difference between the two is relevant, so α_1 may be equated to 0. Owing to the two planes of symmetry, $\cos(\beta)$ for bent molecules is specified as -1 to 1, and γ to be between 0 and 180 degrees. The reason for specifying the cosine of β rather than 0 to 180 degrees is to sample more heavily at a relative orientation of 90 degrees, where the angle α is required to sample over a larger phase space.

Using the described angles coupled with knowledge of the systems permutation group permits efficient calculation of any equivalent points without wasted computation [**Cui. 2016**].

2.3.3 *Table of Symmetrically Distinct Phase-Spaces for Dimers of Different Geometries.*

Table 2. Table of symmetrically distinct regions

			molecule 1			
			Spherical	Symmetric Linear	Asymmetric Linear	Symmetric Bent
molecule 2	Spherical	$\cos(\beta_1)$	0	(0,1)	(-1,1)	(-1,1)
		γ_1	0	0	0	(0,90)
		α_2	0	0	0	0
		$\cos(\beta_2)$	0	0	0	0
		γ_2	0	0	0	0
	Symmetric Linear	$\cos(\beta_1)$		(0,1)	-1,1)	(-1,1)
		γ_1		0	0	(0,180)
		α_2		0 to 180	0 to 180	(0,180)
		$\cos(\beta_2)$		(0,1)	(0,1)	(0,1)
		γ_2		0	0	0
	Asymmetric Linear	$\cos(\beta_1)$			(-1,1)	(-1,1)
		γ_1			0	(0,180)
		α_2			(0,180)	(0,180)
		$\cos(\beta_2)$			(-1,1)	(-1,1)
		γ_2			0	0
	Symmetric Bent	$\cos(\beta_1)$				(-1,1)
		γ_1				(0,180)
		α_2				(0,180)
		$\cos(\beta_2)$				(-1,1)
		γ_2				(0,180)

2.4 PROCESSING.

Each dimer system was first selected, respecting the dimensions ($r, \cos\beta_1, \cos\beta_2, \alpha_2, \gamma_1$ and γ_2) in which to orientate the two molecules relative to each other, alongside a range of r between 1.5 and 9.0 Angstroms. Reference data sets were then constructed using Latin Hypercube (LHC) placement strategies, in a series of magnitudes starting with 2 and doubling it each time to give even distribution when plotted against one another on a logarithmic scale. These points were placed with distances ranging from 1.5 to 9.0 Angstroms and allowing free rotation of each molecule within the previously specified bounds. A test set an order of magnitude larger than the largest training set was constructed using the same methodology, serving as a means of comparison in order to obtain a Root Mean Squared Error (RMSE) value of the GP estimates against calculated vales. In active learning the use of an independent test set offers a more accurate comparison against the results achieved using the training data than if a comparison was made against the reference set, as the ever shrinking array of values would ultimately always result in an RMSE converging to zero. To minimise the cost of computing these data sets, the basis set used for calculations is triple ζ with the relatively simple Møller-Plesset methodology discussed in chapter 1.1.

2.4.1 *The Software Algorithm.*

The geometry of the dimer must first be described, then a Latin Hypercube constructed based on these geometric ranges, which follows the algorithm described in chapter 1.3. A geometric constraint is applied to the LHC which eliminates any configuration with an interatomic distance less than 1.5 Angstroms, or no points less than 9 Angstroms. From this LHC data, Molpro performs ab-initio calculations of the interaction energy. Once Molpro has carried out calculations and any failed runs are resubmitted, the data is collated and Gaussian processes are carried out, implementing a high energy cut-off of 0.005 Hartrees to eliminate points of limited interest for molecular simulations, focusing instead on the lower energies, resulting in around 50% of points being eliminated.

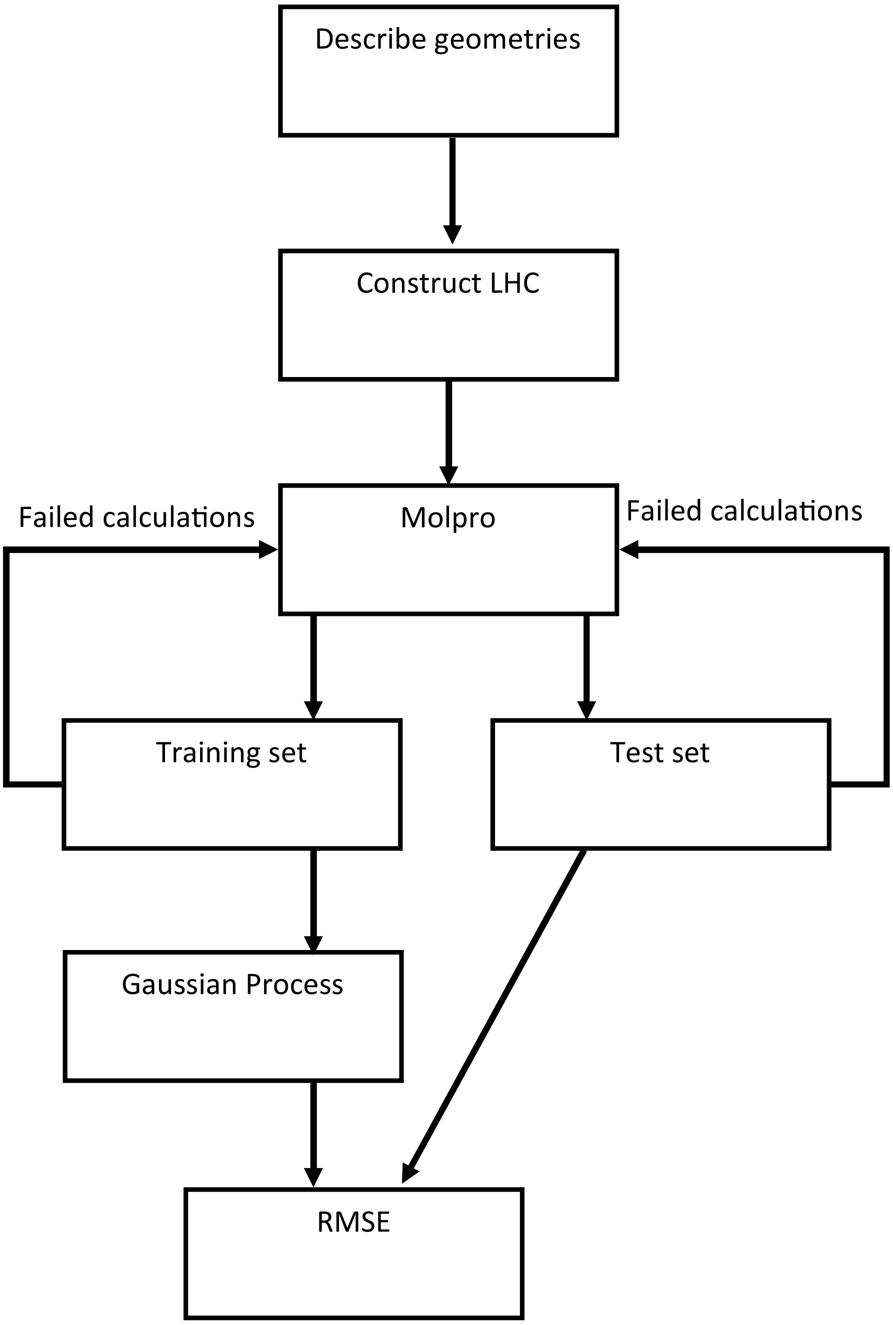

Figure 5.: A diagram of the pre-processing to design the sample regime, the Molpro input and output, and the analysis of data.

The specifications of the LHC to be constructed are then described, making use of the molecular geometry and constraints described in table 2. Other specifications applied include: the number of desired geometries, the number of LHCs the programme should create before deciding on the most homogeneous sample (defined by the LHC with the largest minimum distance between points), and the method of ab-initio calculation desired, which in this instance is the computationally cheap MP2 method with triple ζ basis set. From these specifications, the Molpro input files are constructed. Molpro will then take the input files and perform MP2 calculations, returning the inverse distance vectors and calculated energy of the geometry in an output file. The data from these individual output files is then parsed and concatenated into a single file detailing the inverse distance vectors and energies of every geometry in the LHC, and any calculations which are incomplete are flagged for reprocessing, after which they will also be parsed into the complete data file. A Gaussian process will then be constructed and optimised to fit the training data. Finally the RMSE of each method against the test data may be compared to determine the performance of Gaussian modelling, and some comparison made between sample size and RMSE.

2.5 FITTING A NON-LINEAR CURVE.

In order to best analyse data, it is useful to characterize each learning curve by a few parameters. To this end, this subsection concerns the fitting of an appropriate parametric function.

When comparing the performance (RMSE) of a GP against sample size it becomes apparent as in figure 7 that too few points will elicit no improvement in RMSE, a larger amount will yield an improvement according to a power law trend, and an excessive number of points will saturate the learning curve and yield no further improvement. To accommodate the plateau observed in LHCs with small sample sizes, and the power law trend seen after a certain number of points is satisfied, a cross functional fit is implemented. The basic form of this function is

$$y_{RMSE}(x_{train}) = \frac{R_0}{1 + (\frac{x_{train}}{x_0})^{\alpha}}, \tag{12}$$

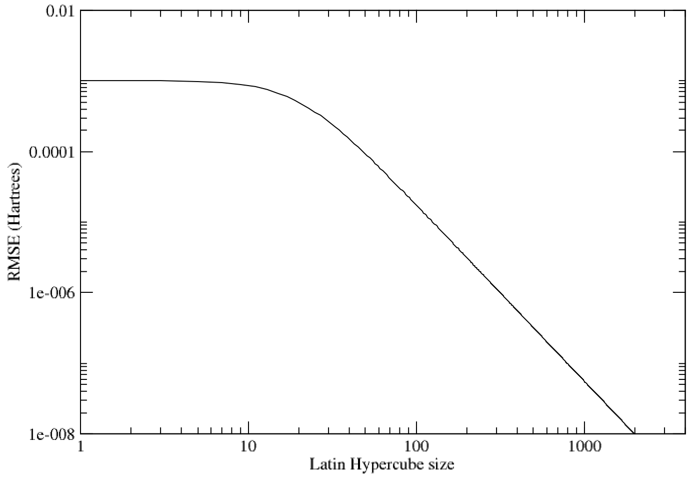

Figure 6.: An example plot of the cross function. This is an empirical plot based on observed data. $y = 1 \times 10^{-3}/(1 + (X/20)^{2.5})$

where R_0 represents the no-model RMSE: the RMSE achieved when the test set is compared to the mean energy value of the training set. The variable x_0 is the number of points required to start seeing an improvement in the RMSE, and α is the power law exponent. Additionally, if a data set is large enough that a sigmoidal shape may be observed due to a minimum value being reached, a modified equation may be implemented:

$$y_{RMSE}(x_{train}) = \frac{R_0 - R_{\infty}}{1 + (\frac{x_{train}}{x_0})^{\alpha}} + R_{\infty}, \tag{13}$$

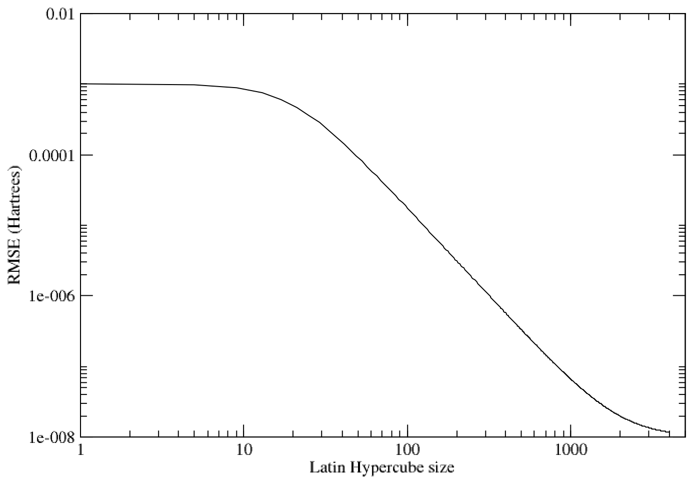

Figure 7.: An example plot of the cross-sigmoidal function. $y = (1 \times 10^{-3} - 1 \times 10^{-8})/(1 + (X/20)^{2.5}) + 1 \times 10^{-8}$

where R_{∞} represents the minimum energy value, most likely to correspond to the root of Molpro's energy convergence threshold at around 2×10^{-8} Hartrees. These curves are fitted with a $1/y$ weighted least squares fitting in order to best capture the contribution of the smallest sample sizes.

3

THE $H_2O - H_2S$ SYSTEM.

The previous year's research concerned the dimer $H_2S - H_2O$ as accurate ab-initio models for this system were not abundant in literature, possibly owing to the magnitude of sample points required to resolve the energy surface in a system with so many degrees of freedom. Active learning was employed in the hope that it may reduce the number of sample points required to accurately resolve the surface, which in turn may reduce the computational expense of modelling such a system.

3.1 PLACEMENT OF SAMPLE POINTS.

The Lennard-Jones potential is considered an archetypical, realistic yet simple model for intermolecular interactions. The key observable characteristics of the Lennard-Jones potential are an asymptotic increase in energy as intermolecular distance converges to zero, a minimum energy at the base of the asymptote, and a convergence to zero at longer ranges. Irrespective of the model used, whether it be Lennard-Jones, or ab-initio, a pairwise molecular interaction should present these same features, just as two real molecules would. As a sanity check for the aforementioned features and therefore some semblance of Lennard-Jones conformity amongst the molecular potentials calculated from first principles, all points in the 500, 1000, 4000 and the two 10000 point sample sets were plotted with their individual oxygen to sulphur distance (r_{14}) versus energy. To conform to the Lennard-

Jones model $V_{LJ} = \epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$, it is expected that at low values of r_{14} the energy will climb asymptotically and there will exist a minimum point in V where $V < 0$ at slightly larger vales of r_{14}. As the separation increases further it is anticipated that the energy will converge to zero. Due to the variance of molecular rotation, there will also be an observable distribution of points about this pattern, with more pronounced variations as the molecules are placed closer together.

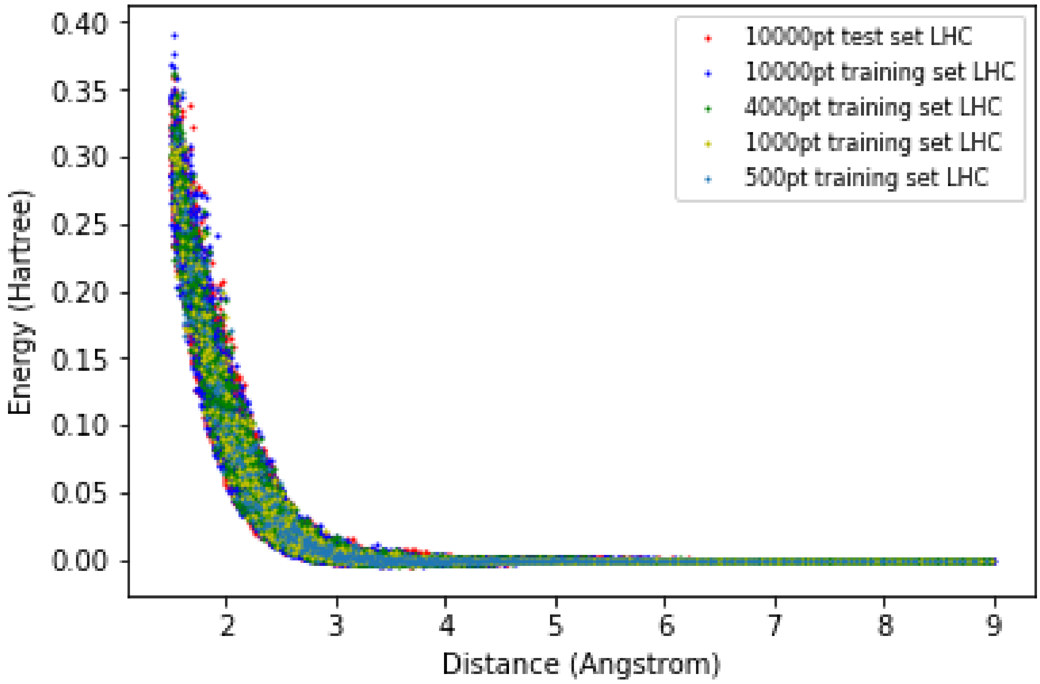

Figure 8.: A plot of the energy and magnitude of vector r_{14}, to observe for Leonard-Jones conformity.

It can be observed from figure 8 that the collective points do indeed feature asymptotic behaviour as distance approaches 0 and energy tends toward ∞ , a minimum point around 3.5 Angstroms, and a convergence to zero beyond 5 Angstroms. A spread can also be observed, particularly as r_{14} tends to 0 due to the different ways the hydrogens of each molecule are oriented about the central sulphur or oxygen. A degree of confidence can be taken in the lack of anomalous points, since this indicates no irregular calculations or anomalies in the data.

3.2 COMPARISON OF TRAINING SET SIZE.

Each of the sets of reference data were used in conjunction with Active Learning to produce a new GP optimised to each successively larger data set, and the RMSE versus dataset magnitude is plotted for each. A point is also plotted to represent the RMSE of the LHC designed GP optimised to each complete training set without any active point selection. This is presented in figure 9 along with the 'no model' RMSE where the training set is equal to zero at all points. Being equal to the square root of the Molpro nugget (1.6×10^{-15}), the random error is 4×10^{-8} Hartrees.

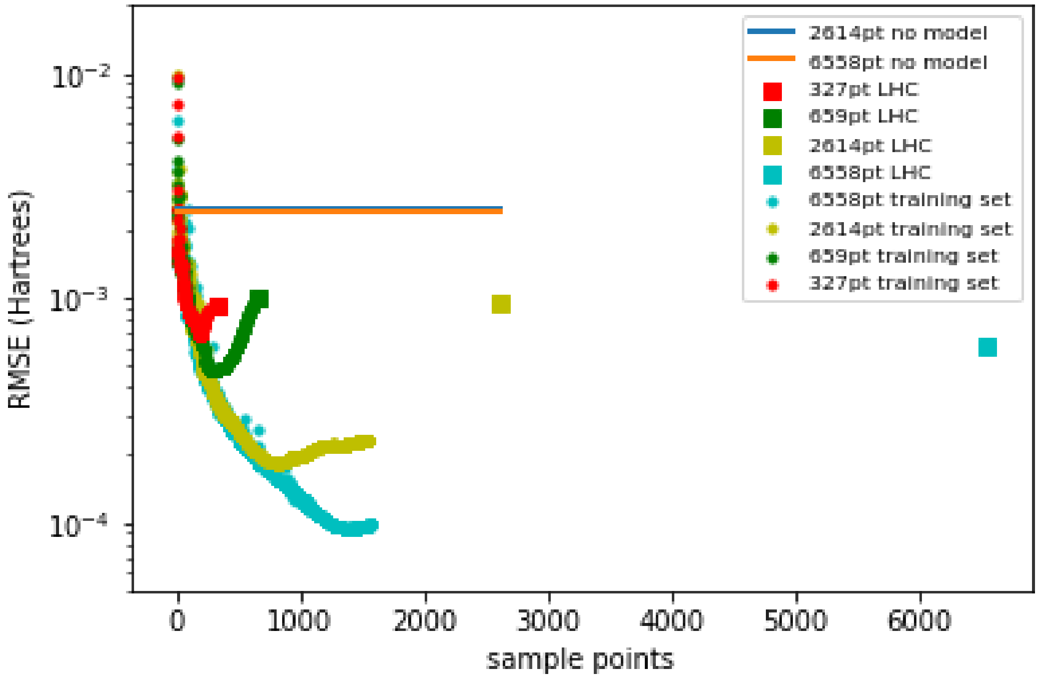

Figure 9.: A graphical illustration of the performance of various size LHCs and associated active learning processes. The independent test set is 20000 points in magnitude.

It can be observed in figure 9 that the LHC derived GPs all have a similar RMSE to one another when the PES is compared to the 6582 point independent test set. Looking at the results for the Active Learning models, the first few points yield a poor model, but as subsequent points are added, the RMSE rapidly outperforms the RMSE of the no model, and within 250 points all training sets outperform the first Latin Hypercube. Using the 2614 points within the energy cut-off in the 4000 point set, the Active Learning can be seen to perform similarly to its LHC designed counterpart with only a quarter of the points, and likewise for the 10000 point Active Learning. It can be observed in the 1000 and 4000 point series that the RMSE worsens after a substantial fraction of available points have been utilised. This impaired precision may be tentatively attributed to the exhaustion of informative points, e.g. those in the short range asymptote and around the minimum. These regions represent significant deviation from the no model, and excessive points outside these areas thereafter may adjust the GP to disproportionately accommodate fairly featureless regions. It should be said at this juncture that none of the methods implemented achieve a comparable RMSE to those observed in simpler systems [**Uteva et al. 2018**].

3.3 EFFECT OF HIGH ENERGY CUT-OFF VALUE ON RMSE.

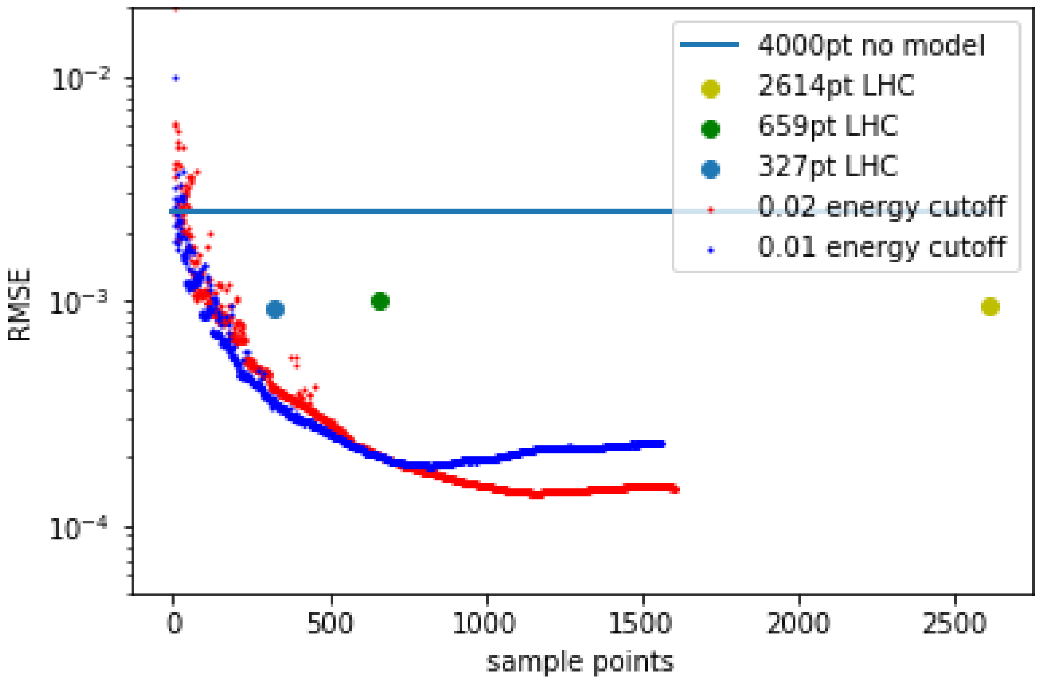

Figure 10.: An illustration of two 4000 point active learning operations making use of the same energy cut-off as the normative data and one twice as high.

Active learning was carried out in the manner described in chapter 3.2, with two 4000 point reference sets. As observed in figure 10, the application of a higher (0.2 as opposed to the default 0.1 Hartree) high energy cut-off promotes a more stable downward trend in the RMSE values obtained by active learning as points are added. This lends credibility to the hypothesis that the short range asymptote is statistically significant, and improves the RMSE of the GP as a result. The trade off however is, the additional cost of calculating the potentials of points falling outside the normal range of interaction distances.

3.4 OBSERVATIONS

Whether using active learning, or LHC driven processing, a good RMSE in this instance would have been 10^{-5} Hartrees, however, even with a relatively large training set this was not achieved. A number of factors may have been responsible for the relative difficulty of this system, namely the number of degrees of freedom, atomic size disparity, and characteristics of the specific atoms involved. To appraise the contributions of these variables to the difficulty of the system, each must be isolated and the RMSE analysed. In the next chapter we present a sequence of dimers of increasing complexity to isolate the effect of each of these factors on the RMSE versus training size behaviour.

4

BEHAVIOURS OF DIMERS.

This chapter concerns a fairly exhaustive selection of molecular pairs, from which comparisons are drawn in the next section. For a summary of findings, refer to the table of fitted functions in appendix A.

In order to observe the effects of symmetry and asymmetry, spherical, linear and bent molecular geometries, as well as atomic sizes and size disparity, a number of representative dimers must be selected. The simplest base case, neon was selected as its small atomic radius is unlikely to result in a large fraction of the sampled energy values being rejected for being too high. For linear molecules, nitrogen was first chosen to represent a longitudinally symmetric diatomic molecule, carbon monoxide was selected to represent a diatomic molecule with longitudinal asymmetry, and carbon dioxide to represent longitudinal symmetry in a tri-atomic linear molecule. In addition, HBr was selected to represent an asymmetric linear molecule with a large disparity between the constituent atomic radii. To observe the relative difficulty of the water and hydrogen sulphide, these were selected to represent two bent molecular geometries. Additionally, to observe the effect of a less pronounced disparity in atomic radii, sulphur dichloride was observed, as were a selection of diatomic halogens. These conformations then made use of the LHC placement strategy so as to give a complete picture of how each interacts with one another.

Once the LHC designs were generated, the interaction for each geometry was calculated in Molpro, a GP was trained to each LHC size and the RMSE data were plotted against LHC size. From the plotted data, a line is fitted; This line is either of the of the form

$$y_{RMSE} = \frac{R_0}{1 + (x_{train}/x_0)^{\alpha}}, \tag{14}$$

or, in the event of a sigmoidal distribution of data

$$y_{RMSE} = \frac{R_0 - R_\infty}{1 + (x_{train}/x_0)^{\alpha}} + R_\infty. \tag{15}$$

The parameter R_0 approximates the initial RMSE value which corresponds to the no-model estimate, x_0 represents the number of values plotted, and x_{train} represents the number of points needed before a downward trend in RMSE is observable. The parameter α describes the functions gradient, and in the sigmoidal function R_∞ represents the value of the observable lower plateau, which is close to the systemic convergence error, and the function will not go below. The fitting of the functions is carried out with a weighting of $1/y^2$ to best respect the small values of RMSE which are of interest. The main purpose of this fitting is to accurately estimate the LHC size required to give an RMSE of 10^{-5} Hartrees, the rationale being that Uteva [**Uteva et al. 2017**] successfully determined the CO2-CO second virial coefficient using a GP PES with an RMSE of $10^{-}5E_h$ so we use this RMSE value as the typical value required to give a PES that is sufficiently accurate to make first principles predictions.

By substituting y_{RMSE} in the two trend equations with a value of 10^{-5} and rearranging to solve for x_{train} a value which corresponds to the desired precision is obtained. This value describes the calculated number of sample points required to achieve an RMSE of 10^{-5} (n_{req}).

$$x_{train} = x_0 \left(\frac{R_0}{y} - 1 \right)^{\frac{1}{\alpha}}, \tag{16}$$

$$x_{train} = x_0 \left(\frac{R_0 - R_\infty}{y - R_\infty} - 1 \right)^{\frac{1}{\alpha}}. \tag{17}$$

With the variety of molecular geometries and corresponding model molecules chosen, there remains the task of allocating some rational combination of each. This is depicted in table 4, with the sub-chapter relating to each dimer specified in its associated cell.

Table 3. Table of Dimers

	Ne	N_2	F_2	Cl_2	Br_2	CO	HBr	CO_2	H_2O	H_2S	SCl_2
Ne	4.2.1	4.3.1	4.3.2	4.3.3	4.3.4	4.3.5	4.3.6	4.3.7	4.4.1	4.4.2	4.4.3
N_2					4.5.1	4.5.2		4.5.3	4.5.4	4.5.5	
CO						4.6.1		4.6.2	4.6.3		
CO_2								4.7.1			
H_2O											
H_2S											

Table 4. The sub-chapters corresponding to molecular pairs are shown at the intersection of each row and column. Grey cells indicate a pairing which is already specified, or not examined within this study.

4.1 METHOD.

For the methods above, training and data sets were generated for each dimer pair using ab-initio calculations and GPs were trained to each training set size. Finally RMSEs were calculated for each training set size. Table 4 provides a directory of subsection numbers for each molecular pair examined.

4.2 two spherical molecules

4.2.1 $Ne - Ne$

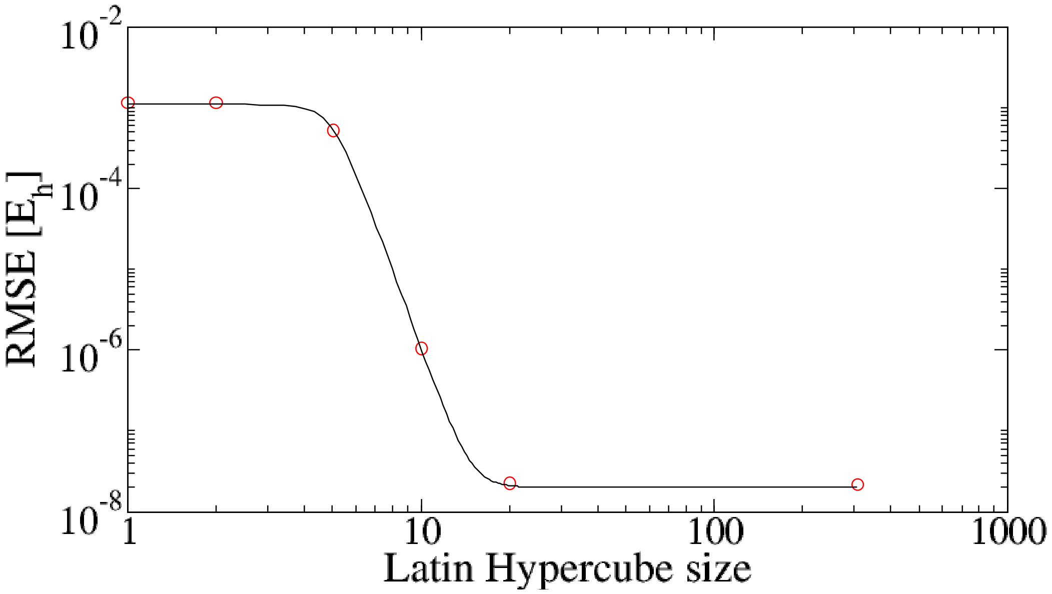

Figure 11.: A plot of RMSE against size of LHC for $Ne - Ne$, showing data points as red circles. The black line shows the fit $y = (1.18 \times 10^{-3} - 2.17 \times 10^{-8})/(1 + (X/4.90)^{9.87}) + 2.17 \times 10^{-8})$

Taking 10^{-5} Hartrees to be indicative of a reasonably precise model, the $Ne - Ne$ system in figure 11 can be observed to achieve the desired standard with a placement strategy using only ten sample points, owing to the single dimensionality of the system. It may also be observed that the logarithmic plot of RMSE versus LHC size is roughly sigmoidal, and is fitted with the associated function and $1/y^2$ weighting. The initial plateau at the no-model value is 0.00118 Hartrees , then a downward power law trend commences as the LHC size exceeds 5, and a final plateau exists at 2.21×10^{-8} Hartrees, roughly the energy convergence threshold of Molpro. Being the simplest base case, the observable pattern should be roughly comparable to further systems tested, with the length of initial plateau increasing and power law gradient becoming less steep as more complexity enters the systems.

4.3 LINEAR MOLECULES WITH A SPHERICAL MOLECULE

4.3.1 $N_2 - Ne$

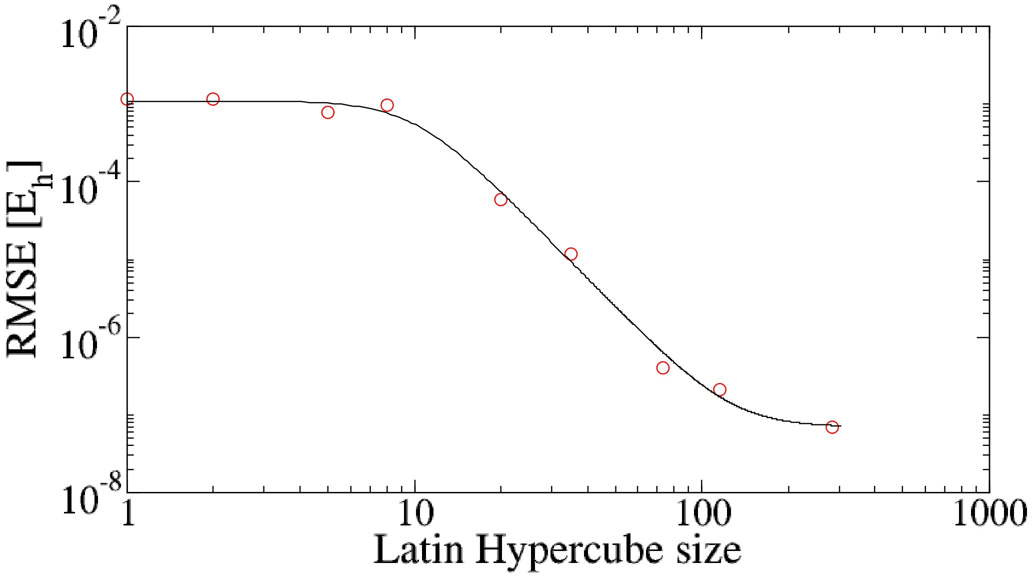

Figure 12.: A plot of RMSE against size of LHC for $N_2 - Ne$, showing data points as red circles. The black line shows the fit $y = (1.05 \times 10^{-3} - 7.22 \times 10^{-8})/(1 + (X/10.2)^{3.97}) + 7.22 \times 10^{-8})$

Much like figure 11, figure 12 shows a sigmoidal trend when plotted logarithmically, with a substantially longer profile. Conformity to the no-model RMSE of 0.00105 Hartrees is apparent up to a LHC size of 10.2, compared to figure 11's 4.9. The RMSE then commences a downward power law, with a less pronounced gradient than seen previously and shows evidence of a second plateau at 7.22×10^{-8} as the number of sample points approaches 300.

4.3.2 $F_2 - Ne$

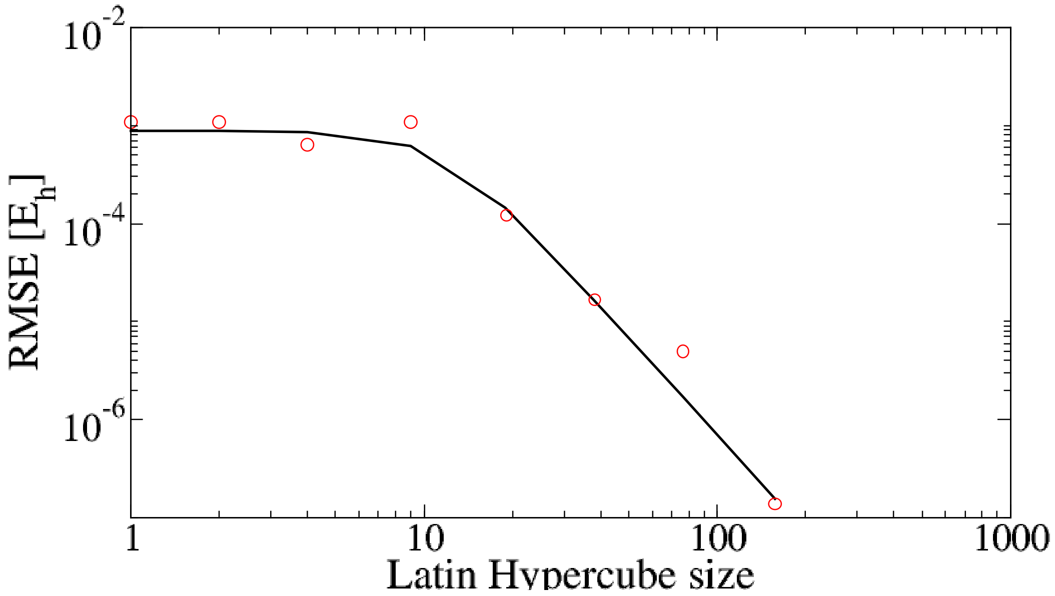

Figure 13.: A plot of RMSE against size of LHC for $F_2 - Ne$, showing data points as red circles. The black line shows the fit $y = 8.88 \times 10^{-4}/(1 + (X/11.5)^{3.31})$

The $F_2 - Ne$ system depicted in figure 13 performs similarly to $N_2 - Ne$, requiring 45 points to reach the benchmark value. This shows a good fit with a reasonably sized LHC as expected, since the two exist in a linear form and possess similar atomic radii.

4.3.3 $Cl_2 - Ne$

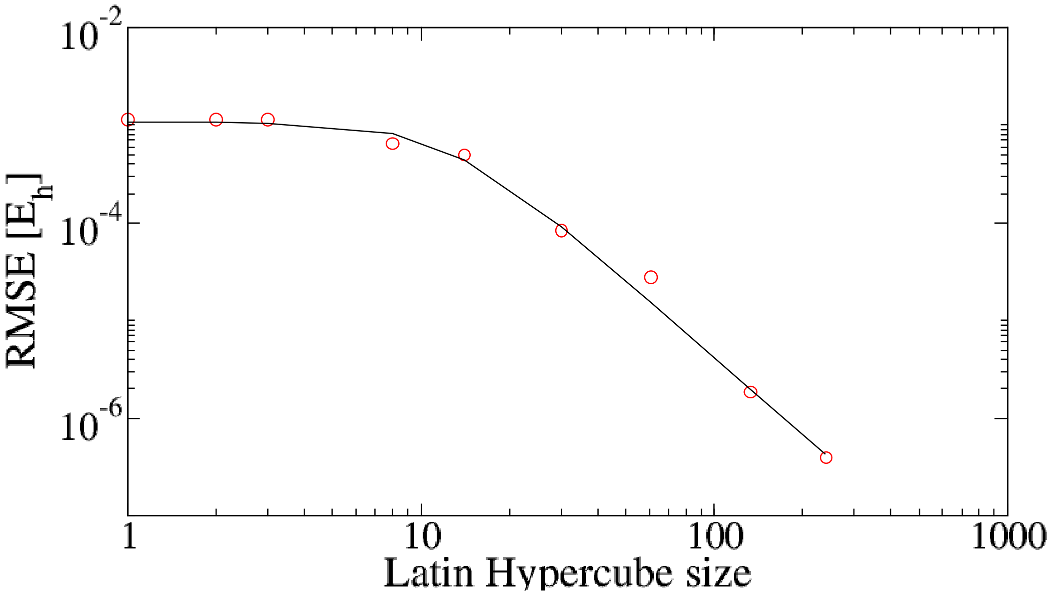

Figure 14.: A plot of RMSE against size of LHC for $Cl_2 - Ne$, showing data points as red circles. The black line shows the fit $y = 1.08 \times 10^{-3}/(1 + (X/12.5)^{2.65})$

The $Cl_2 - Ne$ system depicted in figure 14 performs slightly worse than $F_2 - Ne$, requiring 73 points to achieve the benchmark.

4.3.4 $Br_2 - Ne$

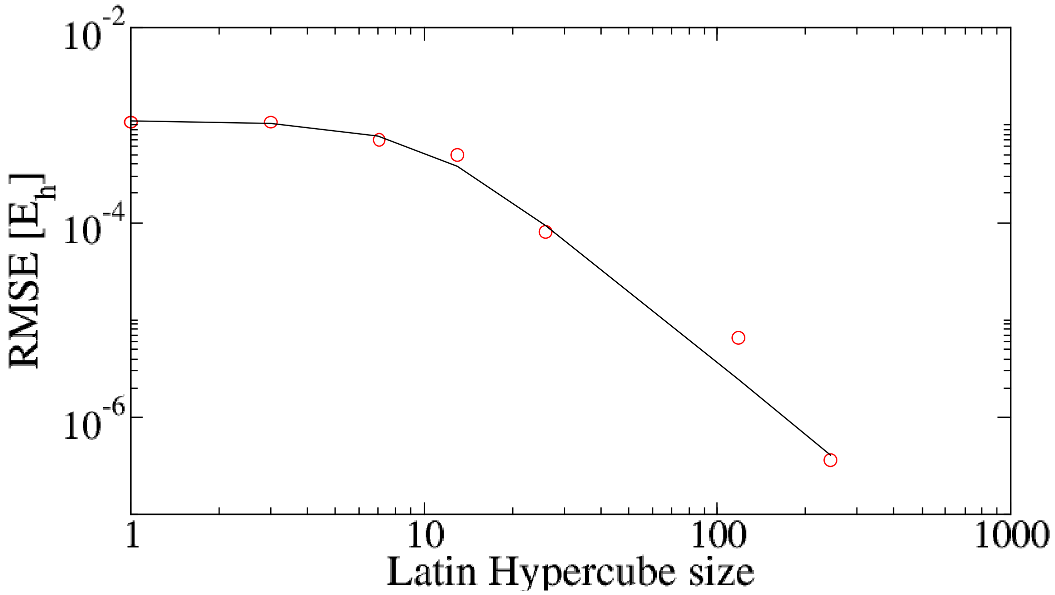

Figure 15.: A plot of RMSE against size of LHC for $Br_2 - Ne$, showing data points as red circles. The black line shows the fit $y = 1.10 \times 10^{-3}/(1 + (X/9.98)^{2.47})$

The $Br_2 - Ne$ system depicted in figure 15 also requires more points to achieve the benchmark, at 67 points.

4.3.5 *CO* − *Ne*

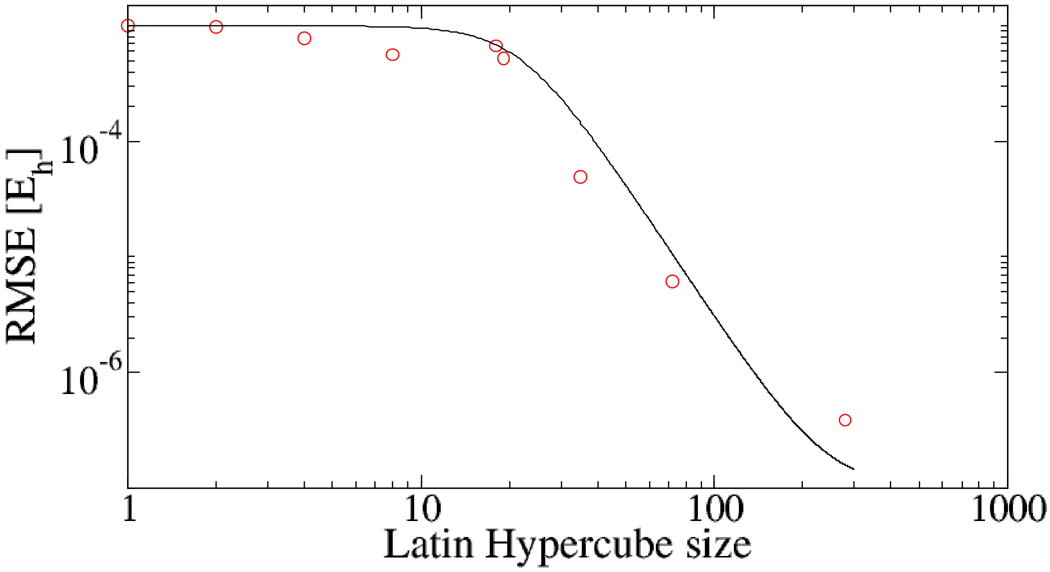

Figure 16.: A plot of RMSE against size of LHC for $CO - Ne$, showing data points as red circles. The black line shows the fit $y = (7.53 \times 10{-}4 - 2.47 \times 10{-}7)/(1 + (X/18.0)^{3.70}) + 2.47 \times 10{-}7)$

The carbon monoxide-neon system differs from $N_2 - Ne$ in that CO lacks the end-to-end symmetry of N_2. As discussed in chapter 2.3 ,this results in a relatively larger symmetrically distinct region of phase-space, requiring more points to accurately capture the PES. The plot of RMSE vs LHC size shown in figure 16 shows some conformity to the no-model value of 0.000753 Hartrees for sample sizes up to 18, nearly twice as many as observed with $N_2 - Ne$. Similarly, the gradient of the power law appears around half that of figure 12. With the sample sizes used, the lowest RMSE observed is 3.91×10^{-7} Hartrees, an order of magnitude above Molpro's systemic error, and well below the desired benchmark.

4.3.6 $HBr - Ne$

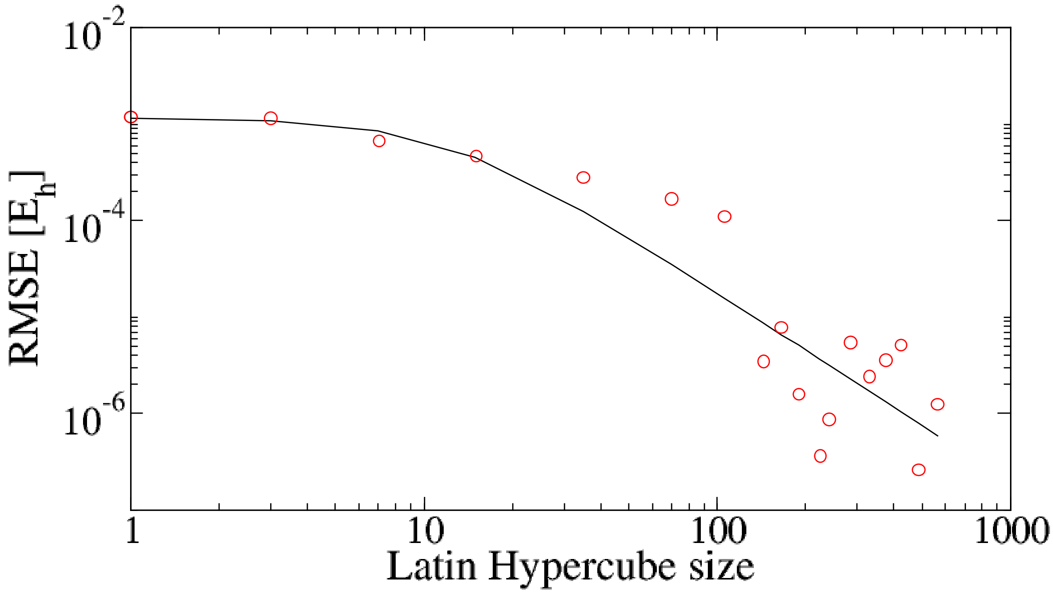

Figure 17.: A plot of RMSE against size of LHC for $HBr - Ne$, showing data points as red circles. The black line shows the fit $y = 1.16 \times 10^{-3}/(1 + (X/11.9)^{1.97})$

The HBr molecule possesses the same end-to-end asymmetry as CO but the constituent atoms have a greater disparity in radius. Repeated attempts to model the system revealed some substantial noise beyond 100 points. To best fit this a weighting of $1/y$ was utilised rather than the usual $1/y^2$.

4.3.7 $CO_2 - Ne$

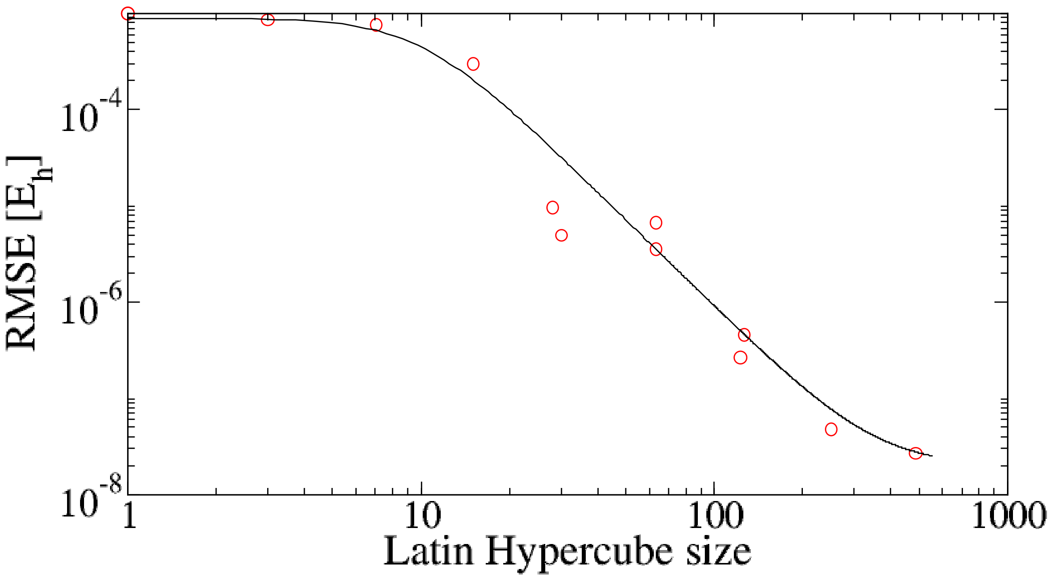

Figure 18.: A plot of RMSE against size of LHC for $CO_2 - Ne$, showing data points as red circles. The black line shows the fit $y = (9.83 \times 10^{-4} - 1.95 \times 10^{-8})/(1 + (X/9.51)^{3.01}) + 1.95 \times 10^{-8})$

The $CO_2 - Ne$ system resembles the $N_2 - Ne$ system insofar as it possesses end-to-end symmetry, while the constituent atoms more closely resemble $CO - Ne$, with an additional oxygen. With respect to the observable trends in figure 18, the data diverges from the no-model RMSE of 0.000983 Hartrees at around 9.5 sample points, and trends downward to a minimum of 2.73×10^{-8}. The fitted function passes the 10^{-5} Hartrees benchmark with 45 sample points, unsurprisingly resembling the $N_2 - Ne$ systems 33; the end-to-end asymmetric $CO - Ne$ and $HBr - Ne$ systems for comparison required 58 and 84 points respectively. It may also be conjectured based on $CO_2 - Ne$ requiring more samples than $N_2 - Ne$ to resolve to within 10^{-5}, that the increased number of atoms in CO_2 over-specifies the data for the Gaussian process involved, and more computational effort is required to resolve the PES.

4.4 BENT MOLECULES WITH A SPHERICAL MOLECULE

4.4.1 $H_2O - Ne$

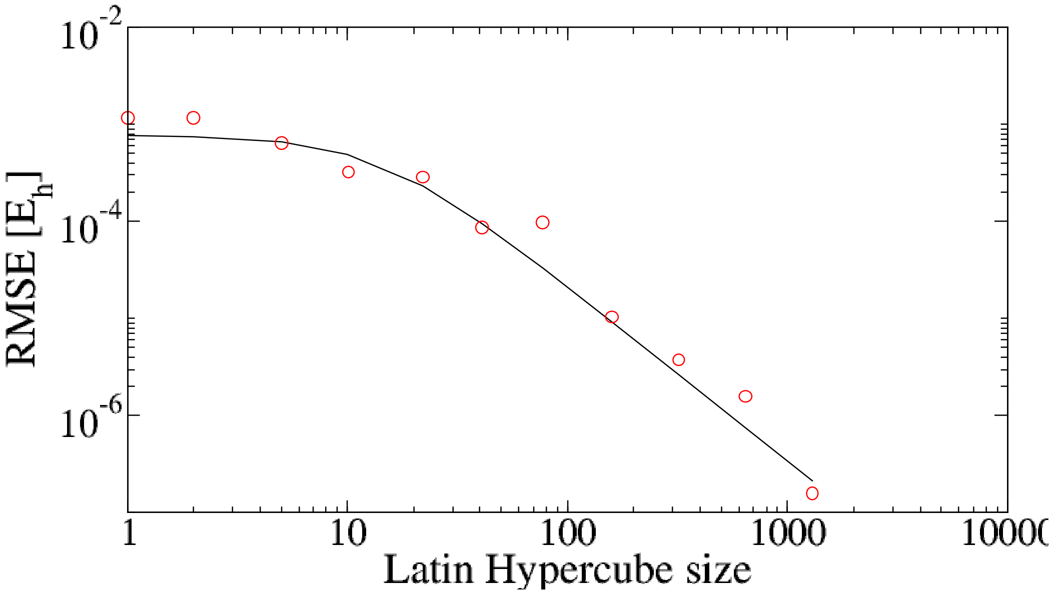

Figure 19.: A plot of RMSE against size of LHC for $H_2O - Ne$, showing data points as red circles. The black line shows the fit $y = 7.82 \times 10^{-4}/(1 + (X/13.5)^{1.80})$

From figure 19 it can be observed that either the triatomic bent geometry or intramolecular size difference is a significant confounding factor compared to a triatomic linear molecule. In the case of the $H2O - Ne$ the calculated PES rapidly overtakes the no-model RMSE of 0.000782 Hartrees, and continues to improve, surpassing the 10^{-5} mark with a sample size of 152. For comparison, the comparable system $CO_2 - Ne$ system requires only 44, as shown in figure 18.

4.4.2 $H_2S - Ne$

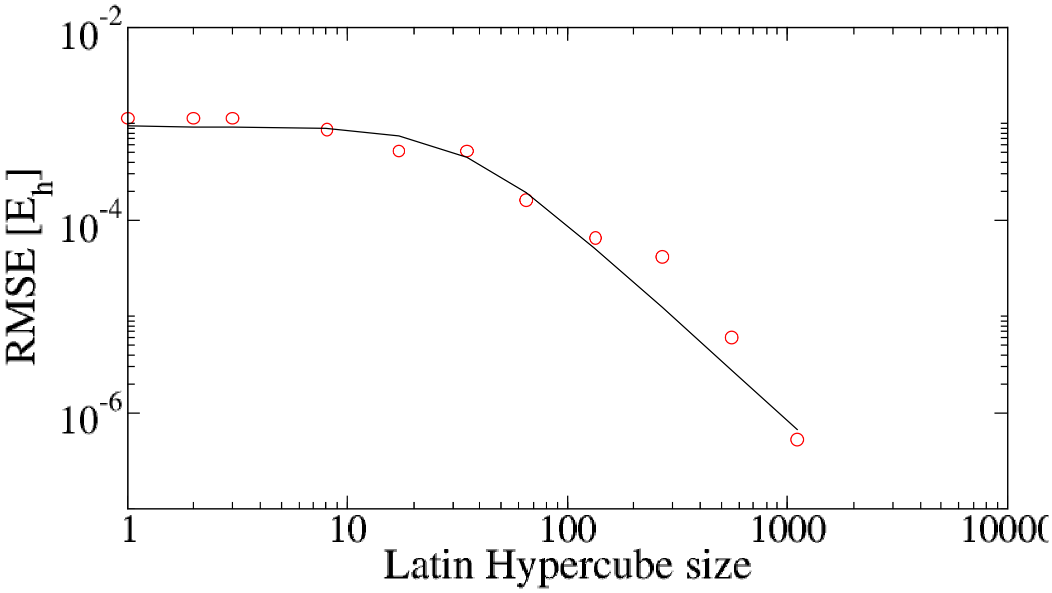

Figure 20.: A plot of RMSE against size of LHC for $H_2S - Ne$, showing data points as red circles. The black line shows the fit $y = 9.28 \times 10^{-4}/(1 + (X/34.4)^{2.09})$

Figure 20 shows a superficially similar pattern for the $H2S - Ne$ system as observed for $H2O - Ne$ in figure 19, however achieving an RMSE of 10^{-5} Hartrees requires far more points (300) as opposed to 152. This suggests some complicating factor inherent to H_2S more so than H_2O.

4.4.3 $SCl_2 - Ne$

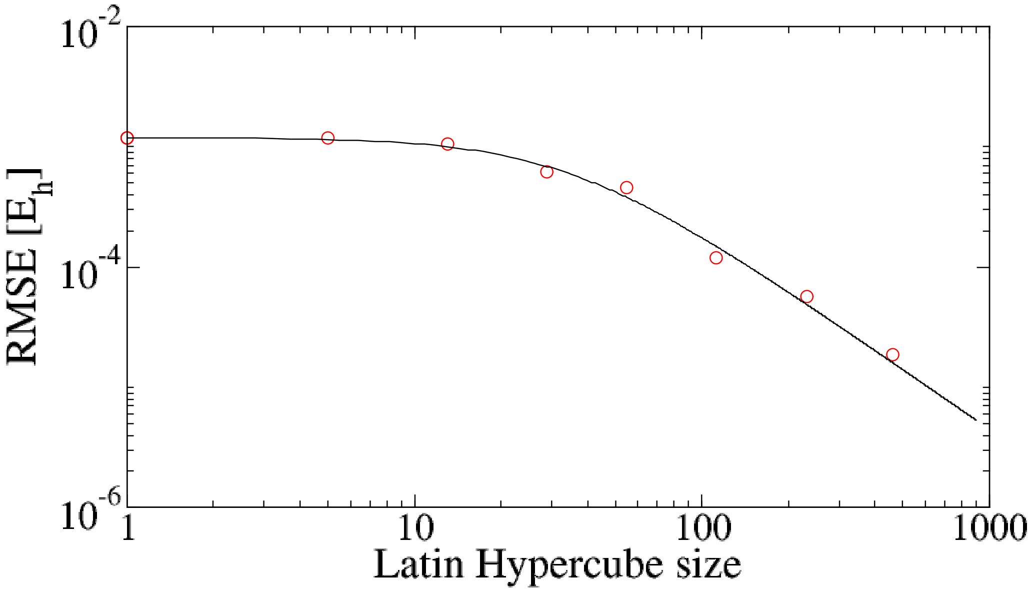

Figure 21.: A plot of RMSE against size of LHC for $SCl_2 - Ne$, showing data points as red circles. The black line shows the fit $y = 1.22 \times 10^{-3}/(1 + (X/30.8)^{1.55})$

A $SCl_2 - Ne$ system was selected due to the relatively similar atomic radii within the SCl_2 molecule. From figure 21 the trend can be seen to resemble that of $H_2S - Ne$ in figure 20, but requiring relatively more sample points (686) to achieve the desired RMSE.

4.5 INTERACTIONS WITH A NITROGEN MOLECULE

4.5.1 $N_2 - N_2$

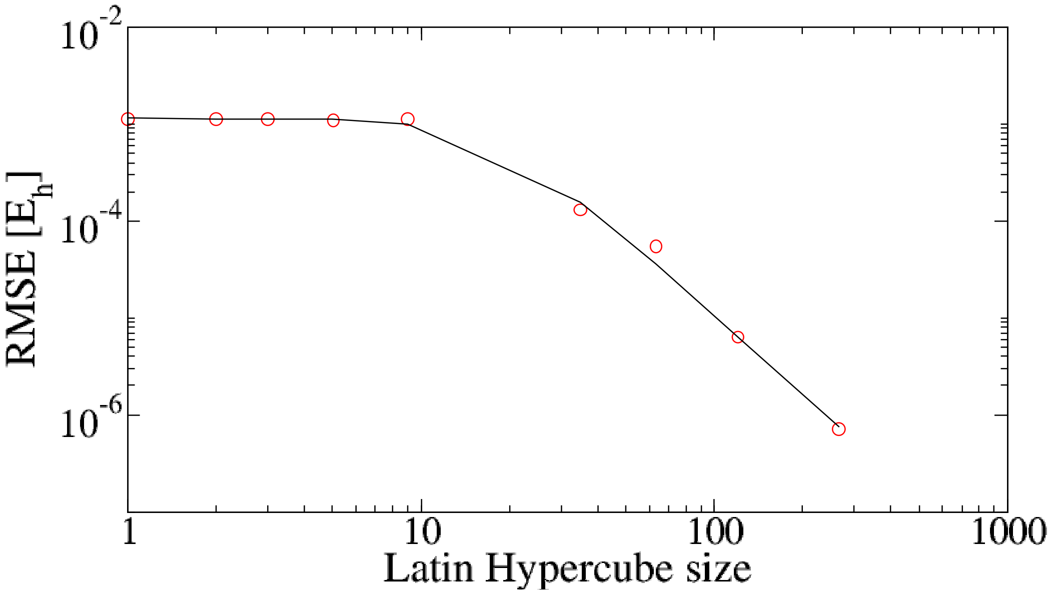

Figure 22.: A plot of RMSE against size of LHC for $N_2 - N_2$, showing data points as red circles. The black line shows the fit $y = 1.14 \times 10^{-3}/(1 + (X/17.7)^{2.70})$

Figure 22 shows the RMSE of the nitrogen dimer, which remains around the no-model RMSE of 0.00114 Hartrees for LHCs of up to 18 points, which is not uncommon among the systems examined. The trend then ventures downward, crossing the 10^{-5} Hartrees mark with 102 points. Compared with the $N_2 - Ne$ dimer in figure 12 the initial plateau is of a similar magnitude but persists for almost twice as many sample points, and the downward gradient after this is less pronounced, and the desired RMSE requires around 3 times as many sample points.

4.5.2 $CO - N_2$

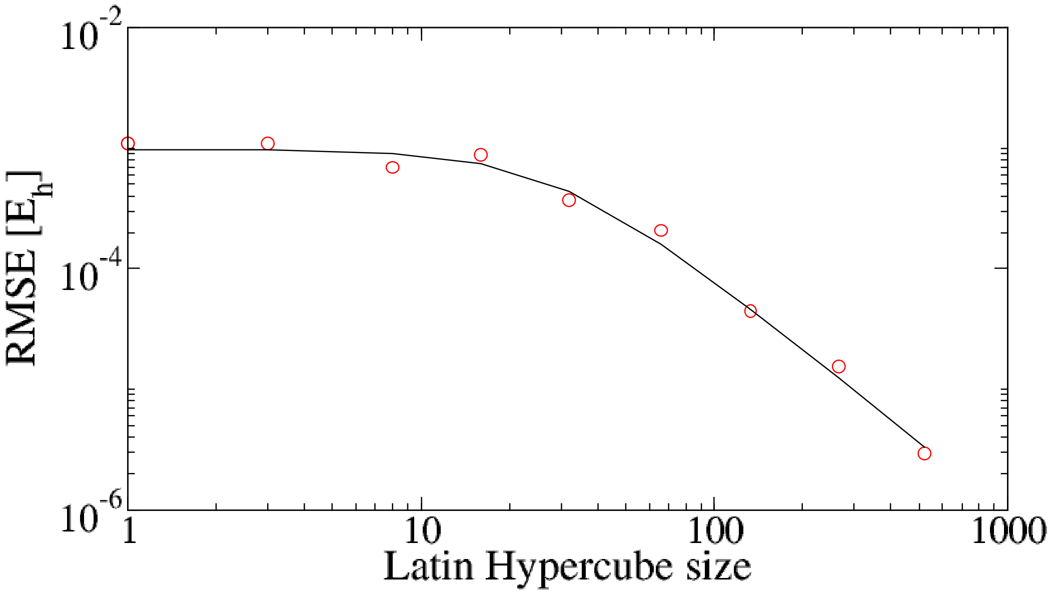

Figure 23.: A plot of RMSE against size of LHC for $CO - N_2$, showing data points as red circles. The black line shows the fit $y = 9.70 \times 10{-}4/(1 + (X/29.2)^{1.97})$

Figure 23 shows the performance of the $CO - N_2$ dimer with increasing sample sizes. The No model RMSE is observed for LHCs up to 30 points, before trending down and crossing 10^{-5} Hartrees at around 295, approximately 3 times higher than the comparable $N_2 - N_2$ dimer, and 5 times higher than the $CO - Ne$ dimer.

4.5.3 $CO_2 - N_2$

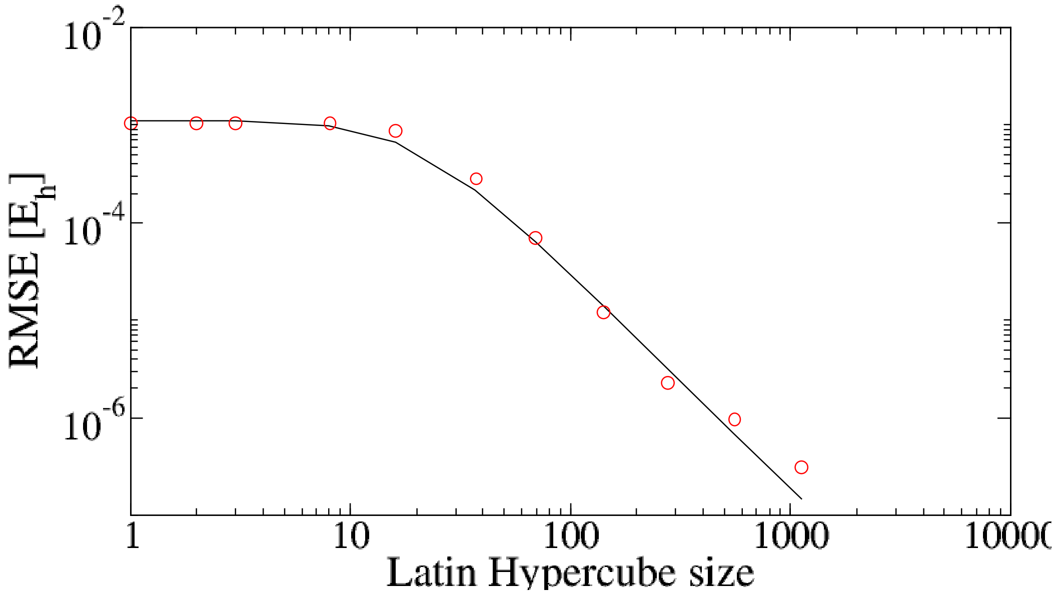

Figure 24.: A plot of RMSE against size of LHC for $CO_2 - N_2$, showing data points as red circles. The black line shows the fit $y = 1.11 \times 10^{-3}/(1 + (X/19.4)^{2.199})$

As may be expected, the $CO_2 - N_2$ dimer requires larger sample sizes than $CO_2 - Ne$ owing to the increased degrees of freedom. Figure 24 shows conformity to the no-model RMSE of 0.0011 Hartrees for LHCs of up to 19 points, followed by a downward trend to 10^{-5} at 165 sample points. This represents a sample around 60% larger than the comparable $N_2 - N_2$ system, and 4 times the points required by $CO_2 - Ne$. Despite these relative difficulties in resolving the PES, the sample size required still appears to be half that of the $CO - N_2$ dimer, lending credence to the assertion that a symmetrically distinct region twice the size will require twice the points in order to resolve to the same fidelity.

4.5.4 $H_2O - N_2$

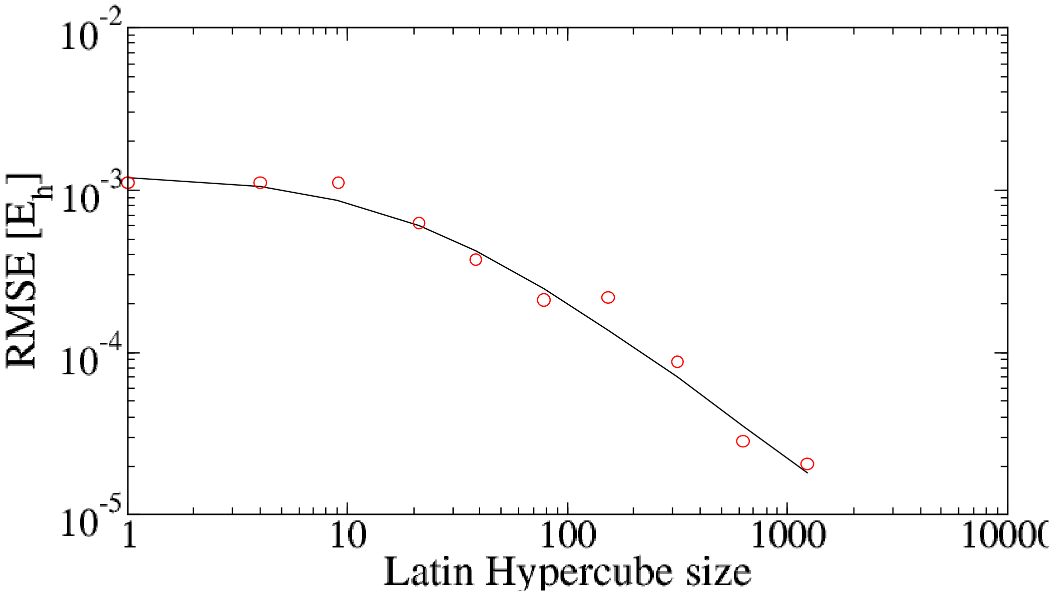

Figure 25.: A plot of RMSE against size of LHC for $H_2O - N_2$, showing data points as red circles. The black line shows the fit $y = 1.25 \times 10^{-3}/(1 + (X/19.9)^{1.02})$

The $H_2O - N_2$ results depicted in figure 25 perform as may be expected based on the relative difficulty observed in the $H_2O - Ne$ system, with a plateau for LHCs up to 20 samples large, then trending downward with a less pronounced gradient than observed in figure 19 for $H_2O - Ne$. With a test set of 8192 points before energy cut-offs are applied and a largest training set of 4096 before cut-offs, this plot never reaches an RMSE of 10^{-5} Hartrees. Extrapolating from the fitted function, the LHC size required to achieve the desired accuracy is 2214.

4.5.5 $H_2S - N_2$

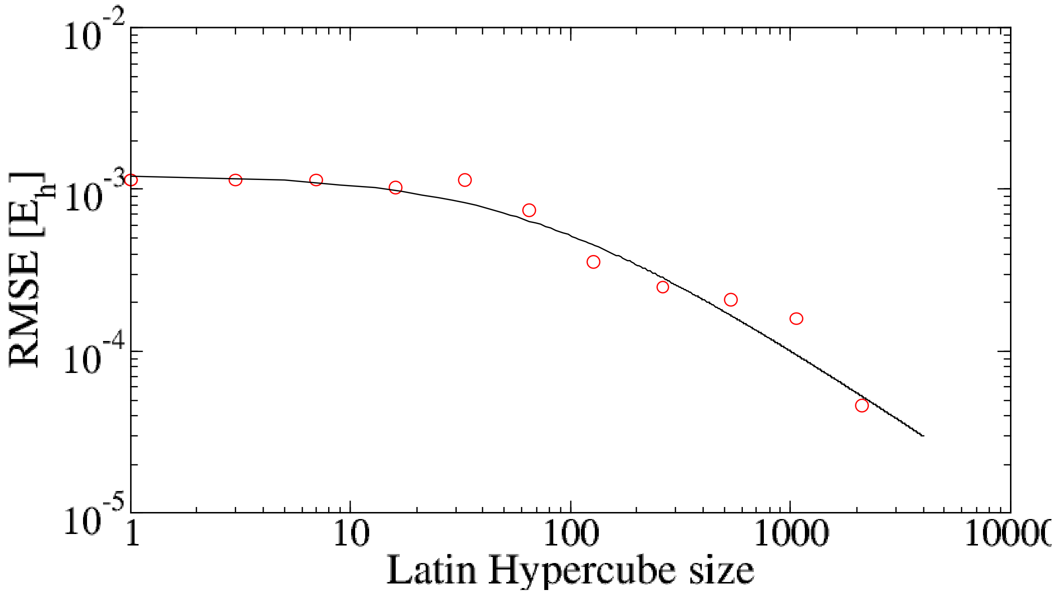

Figure 26.: A plot of RMSE against size of LHC for $H_2S - N_2$, showing data points as red circles. The black line shows the fit $y = 1.24 \times 10^{-3}/(1 + (X/69.7)^{0.913})$

The $H_2S - N_2$ results depicted in figure 26 show a substantial under-performance relative to the comparable system $H_2O - N_2$ in figure 25. This may be due to the size disparity between constituent atoms being larger in H_2S or something intrinsic to the chemistry of the system.

4.6 INTERACTIONS WITH A CARBON MONOXIDE MOLECULE

4.6.1 $CO - CO$

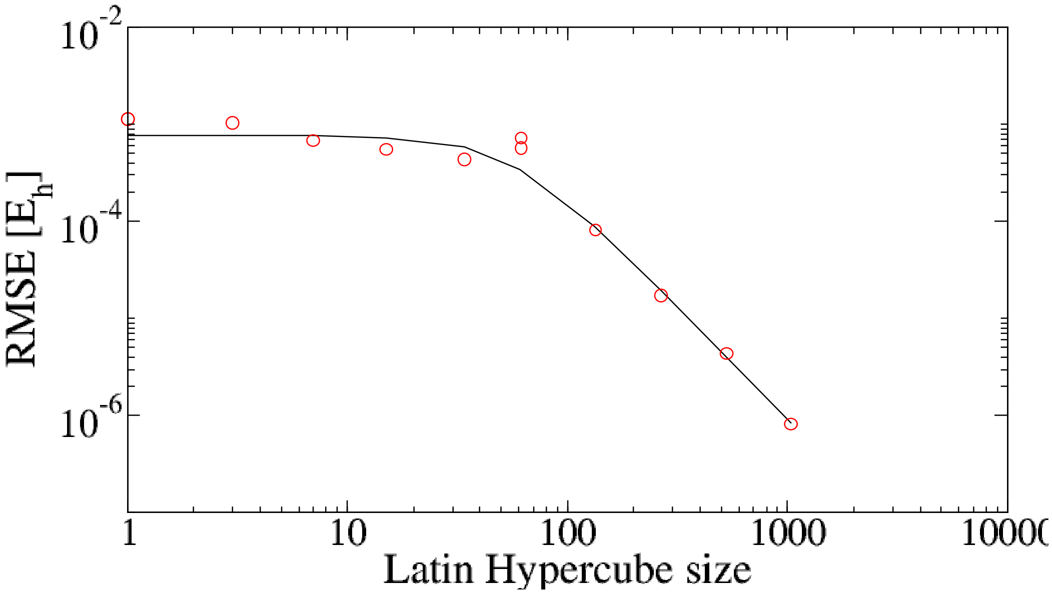

Figure 27.: A plot of RMSE against size of LHC for $CO - CO$, showing data points as red circles. The black line shows the fit $y = 7.61 \times 10^{-4}/(1 + (X/56.3)^{2.34})$

The plot of $CO - CO$ in figure 27 performs closely to that of $CO - N_2$ in figure 23, both passing 10^{-5} Hartrees at around 356 and 295 respectively. This is unsurprising given the symmetrically distinct regions will cover the same amount of phase space, owing to the two CO molecules being identical.

4.6.2 $CO_2 - CO$

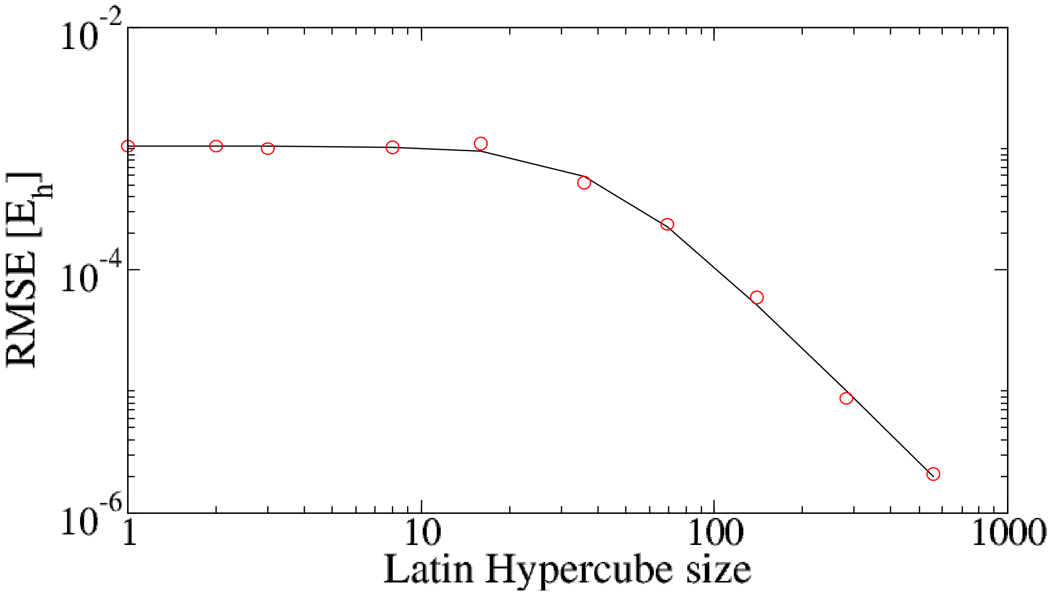

Figure 28.: A plot of RMSE against size of LHC for $CO_2 - CO$, showing data points as red circles. The black line shows the fit $y = 1.06 \times 10^{-3}/(1 + (X/40.0)^{2.38})$

Much like previous plots have shown, substituting N_2 with CO_2 frequently gives a similar result, the plot of $CO_2 - CO$ in figure 28 being no exception. Once again, by having the same phase space encapsulated within the symmetrically distinct region, 10^{-5} Hartrees is achieved with 283 points.

4.6.3 $H_2O - CO$

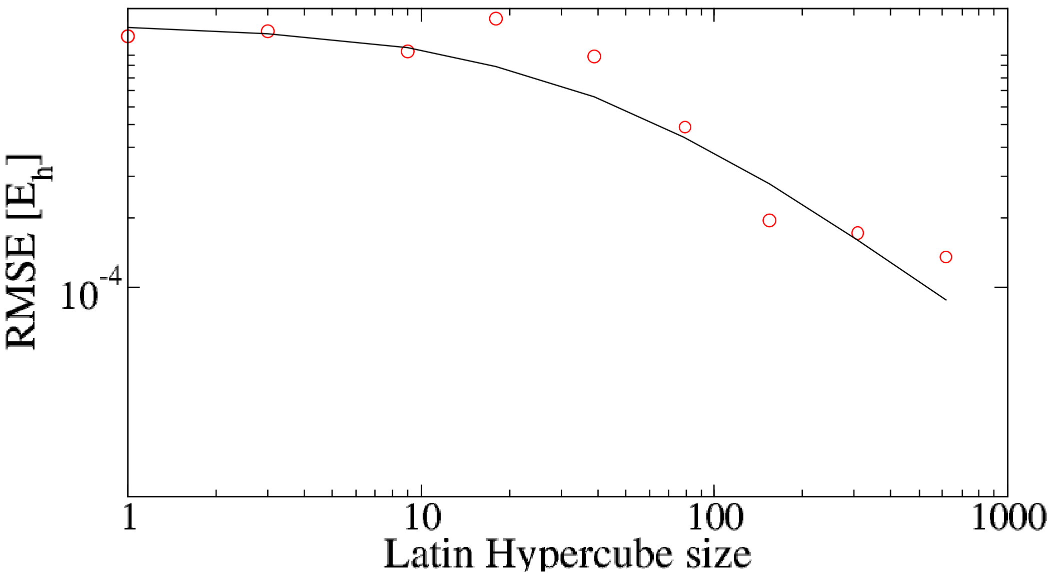

Figure 29.: A plot of RMSE against size of LHC for $H_2O - CO$, showing data points as red circles. The black line shows the fit $y = 1.35 \times 10^{-3}/(1 + (X/36.5)^{0.942})$

The data in figure 29 featuring a $H_2O - CO$ system gives some metric of the difficulty of resolving a system with a bent molecule, and the resulting added degree of rotational freedom compared to the more linear $CO_2 - CO$ system that precedes it. The system precedes from the no-model estimate of 0.00135 Hartrees, and as indicated by the fitted function, trends slowly downward after around 36 points. After plotting 614 points, the system still only approaches an RMSE of 10^{-4} Hartrees, so the number of samples required to obtain the benchmark RMSE must be extrapolated from the trend to be around 6655.

4.7 INTERACTIONS WITH A CARBON DIOXIDE MOLECULE

4.7.1 $CO_2 - CO_2$

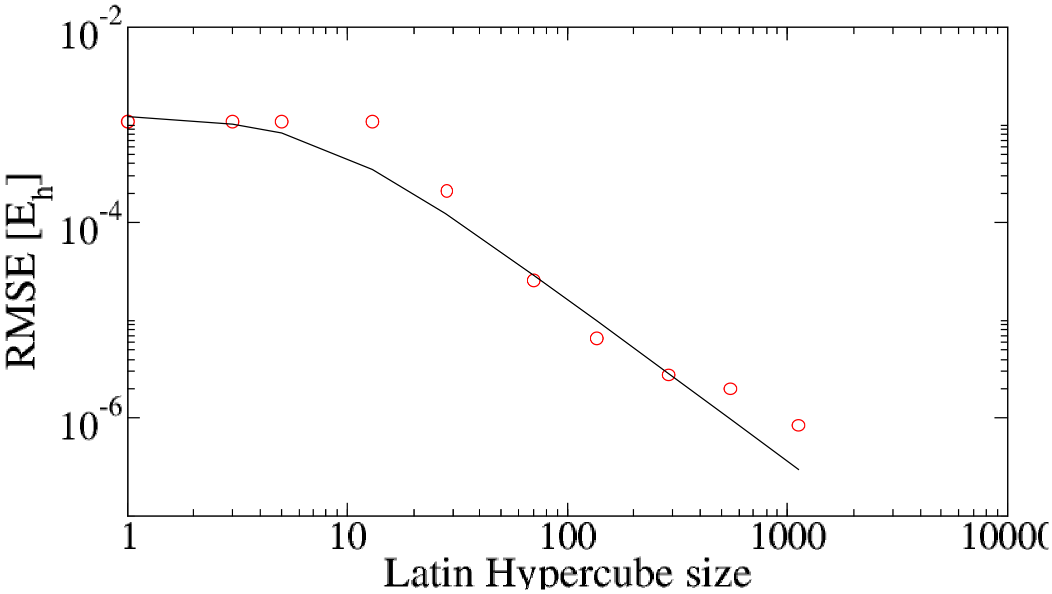

Figure 30.: A plot of RMSE against size of LHC for $CO_2 - CO_2$, showing data points as red circles. The black line shows the fit $y = 1.26 \times 10^{-3}/(1 + (X/6.82)^{1.62})$

Like most of the systems examined, $CO_2 - CO_2$ pictured in figure 30 begins to move from its no-model RMSE of 0.00126 Hartrees with an x_0 value of 6.8, trending downward with a steep gradient to exceed the desired precision with 136 points. It has been observed previously that the symmetrically distinct region of CO_2 has the same phase space as N_2, and conjectured that by having more atoms from which to resolve vectors it may require more computational effort than a nitrogen molecule in its place; in this instance, by referring back to figure 22 it can be seen that $CO_2 - CO_2$ and $N_2 - N_2$ achieve an RMSE of 10^{-5} Hartrees with almost the same number of points. The performance relative to $CO_2 - N_2$ in figure 24 also illustrates how having two identical molecules reduces the phase space to be covered and consequently reduces the number of samples required.

5

COMPARISON OF DIFFERENT GEOMETRIES AND HOW EACH AFFECTS PRECISION.

5.1 OBSERVATIONS ON MOLECULAR SHAPE.

		Ne	F_2	Cl_2	Br_2	N_2	CO	HBr	CO_2	H_2O	H_2S	SCl_2
Ne	DoF	1	2	2	2	2	2	2	2	3	3	3
	n_{req}	8	45	73	67	33	58	134	44	152	300	686
N_2	DoF					4	4		4	5	5	
	n_{req}					102	295		165	2214	13546	
CO	DoF						4		4	5		
	n_{req}						356		283	6655		
CO_2	DoF								4			
	n_{req}								136			

Table 5. n_{req} compared with Degrees of Freedom. n_{req} represents the interpolated number of sample points required to achieve the benchmark of RMSE 10^{-5} Hartrees. DoF describes the number of Degrees of Freedom in a system.

From table 5, the performance of systems may be directly compared with respect to their number of degrees of freedom (DoF). It is immediately apparent that a general correlation exists between more degrees of freedom, and a higher number of points required (n_{req}) to achieve the benchmark ac-

curacy of 10^{-5} E_h . Also apparent is the performance of different geometries when interacting with Ne; from this table it is easily deductible that asymmetry and bent shape are significant confounding factors.
By plotting the quantitative data above, the correlation between DoF and n_{req} may be examined further.

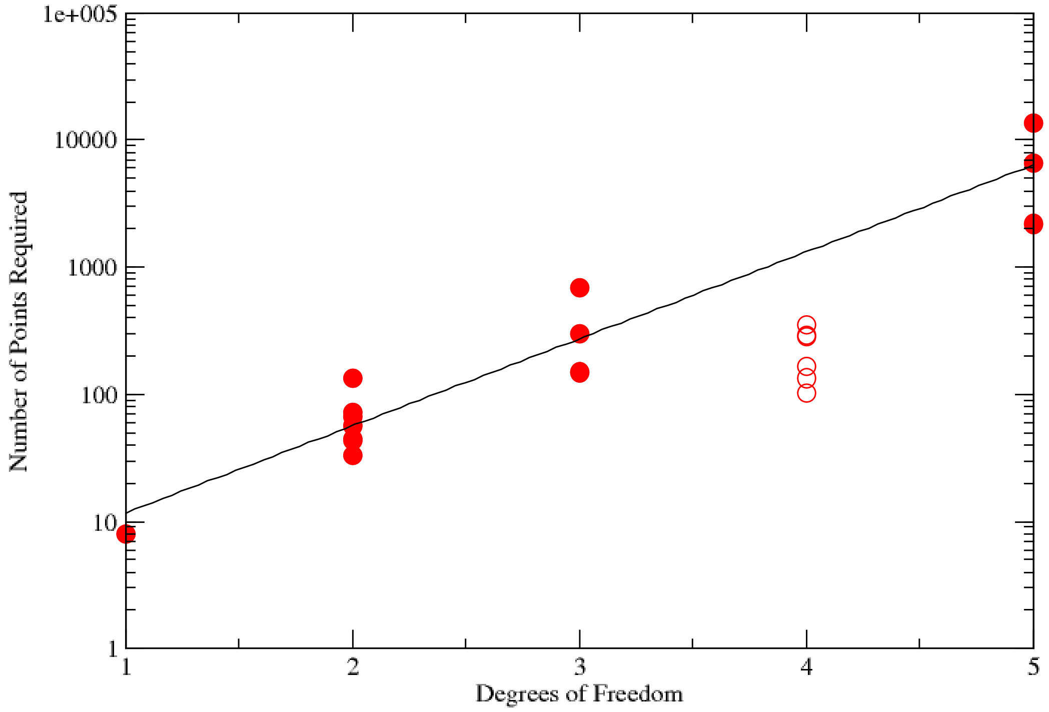

Figure 31.: An exponentially fitted plot of DoF and the points required to achieve an RMSE of 10^{-5} Hartrees.

Figure 31 shows an exponential correlation of $y = 2.4201e^{1.5758x}$ between the 1, 2, 3, and 5 dimensional systems, with the number of sample points required for the most challenging 5 DoF system exceeding 10^4. This explains at least partially why the $H_2S - H_2O$ dimer may have proven so computationally expensive to model with little progress made even at 10,000 points, especially given the two non-identical molecules will lack interchange symmetry and require large rotations through phase space. Extrapolating this trend it would appear that a similar 6 DoF system would require around 31,000 points to achieve the benchmark precision. This trend omits the systems with 4 degrees of freedom, as all deviate from the trend due to the lack of a challenging asymmetric system with this number of DoF.

Table 6. Dimers and their associated x_0 and n_{req} results.

		Ne	F_2	Cl_2	Br_2	N_2	CO	HBr	CO_2	H_2O	H_2S	SCl_2
Ne	x_0	4.9	11.5	12.5	10.0	10.2	18	11.9	9.5	13.5	34.4	30.8
	n_{req}	8	45	73	67	33	58	134	44	152	300	686
N_2	x_0					17.7	29.2		19.4	19.9	69.7	
	n_{req}					102	295		165	2214	13546	
CO	x_0						56.3		40	36.5		
	n_{req}						356		283	6655		
CO_2	x_0								6.8			
	n_{req}								136			
H_2O	x_0											
	n_{req}											
H_2S	x_0											
	n_{req}											

Table 6 shows the estimated number of points required (calculated with equations [14] and [15]) to see some improvement from the no-model, x_0, and the number of samples required to achieve an RMSE of 10^{-5} Hartrees as determined by the non-linear regression, n_{req}. The molecules depicted were chosen to represent the spherical, symmetric linear, asymmetric linear and bent geometry. From the available results it is apparent that the length of the shoulder, or conformity to the no-model RMSE remains relatively low even in systems with the most degrees of freedom.

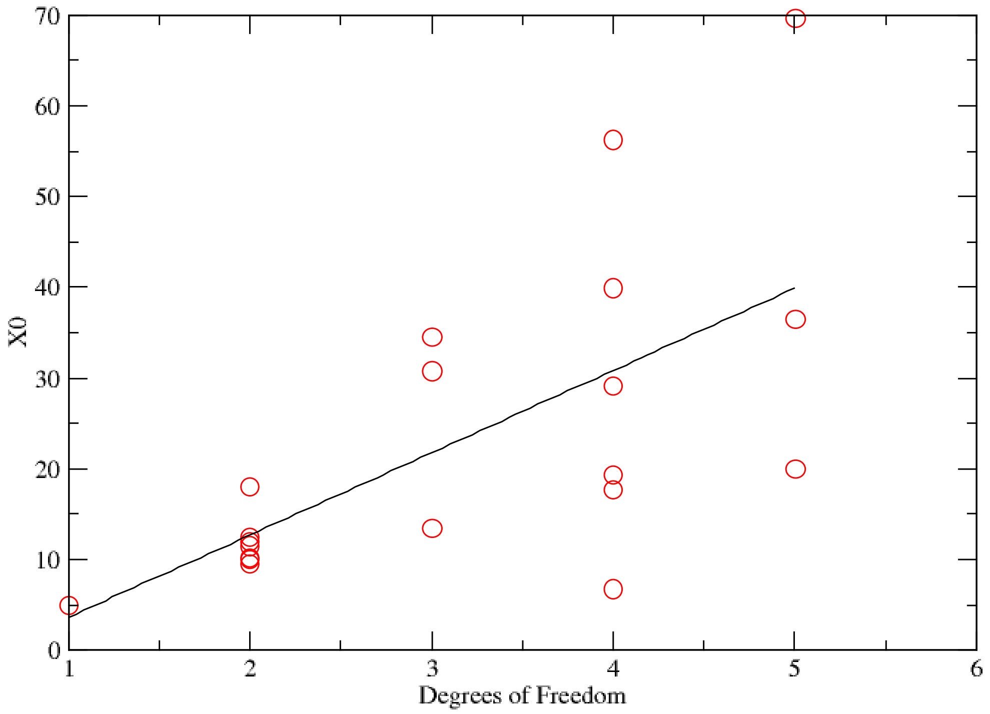

Figure 32.: A fitted plot of DoF and the value of x_0. The trend fitted is: $y = 9.0646x - 5.4187$.

Figure 32 shows the plotted points appear scattered, as in figure 31, only this time there is only weak support for a linear relationship, so x_0 is not responsible for the predominant exponential behaviour observed.

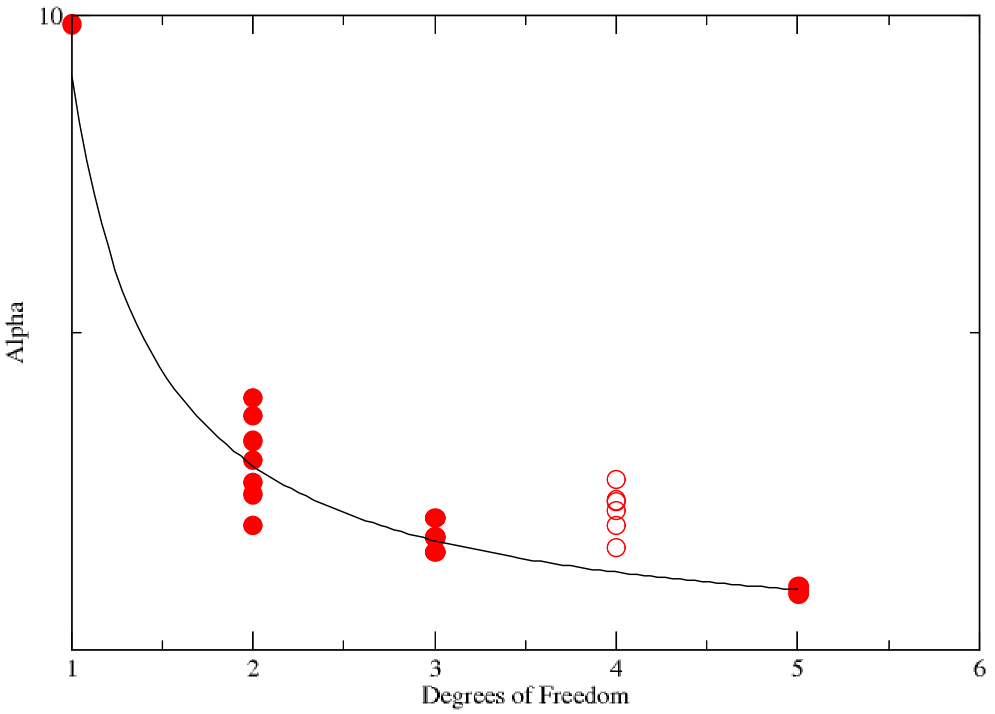

Figure 33.: A fitted plot of DoF and the value of α. The trend fitted is: $y = 1/(0.233x - 0.122)$.

As in figure 31, the 4 dimensional systems are discounted and a trend established in figure 33, in this instance with a downward inverse function.

It is clear from figures 31, 32 and 33 that the number of degrees of freedom predominates, but is not the sole determining factor of the difficulty of the system; based on the variability in n_{req} for a given number of DoF, other factors clearly have a significant impact on ones ability to resolve a system. To dissect the involvements of other factors, similar molecule pairs will be compared, and further observations made.

5.2 SYSTEMS WITH 2 DEGREES OF FREEDOM.

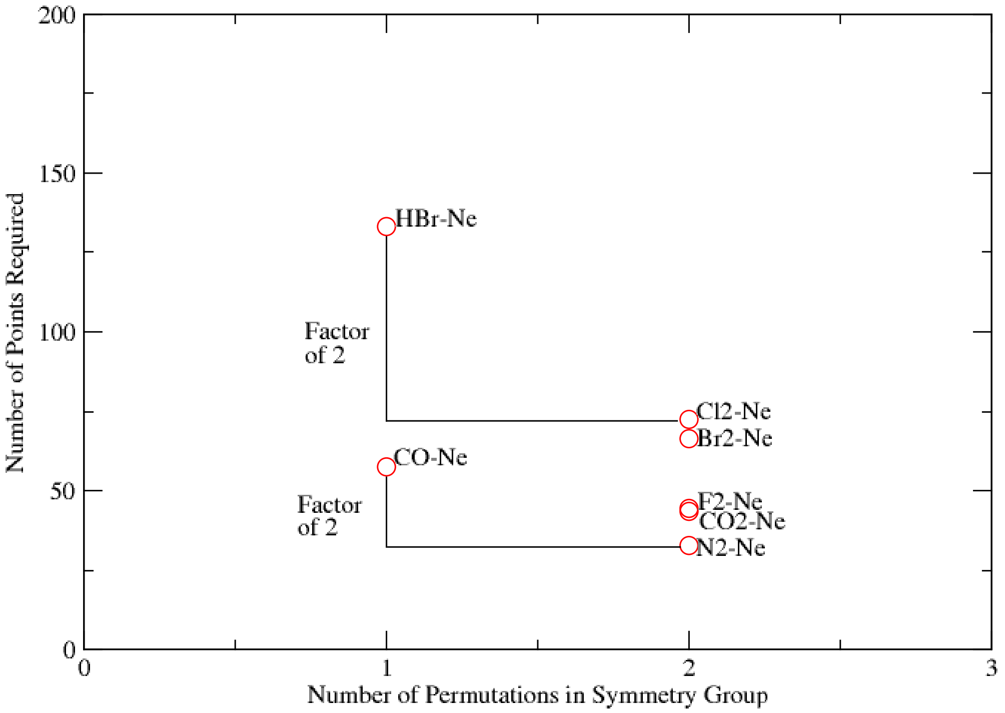

Figure 34.: A plot of points required versus number of permutations in the symmetry group for systems with 2 degrees of freedom.

Among the systems with two degrees of freedom shown in figure 34, those with two symmetric permutations generally require fewer sample points than those with no additional permutability. A trend-line is not fitted as there only exists 2 discrete numbers of symmetries, but the mean of the single permutation group is 95.3 and the two permutation group's mean is 52.0, lending credibility to the assumption that doubling the permutations halves the number of points required.

Table 7. Atomic size disparities of HBr and CO.

	CO	HBr
atomic size disparity (%)	15.3	35.5
n_{req}	58	134

The two systems with a single degree of symmetry: HBr and CO, require markedly different numbers of points which may be explained by the size disparity between hydrogen and bromine. Carbon and oxygen by comparison have more similarly sized atomic radii as shown in table 7.

Atomic size disparity is defined as

$$Disparity = \sqrt{\left(\frac{(r_a - r_b)}{r_a}\right)^2} \times 100, \qquad (18)$$

where r_a and r_b are the atomic radii of atom A and atom B.

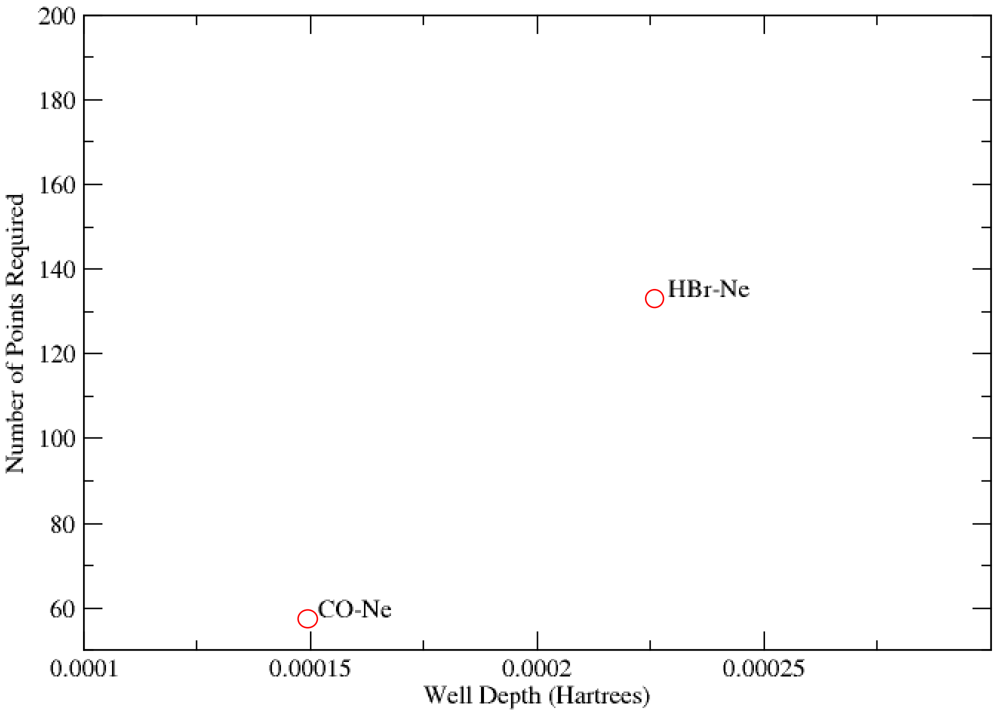

Figure 35.: A plot of well depth for HBr and CO interacting with Ne. Each has 2 DoF and one element in the permutation group.

Figure 35 shows a correlation between increasing well depth (defined in this instance as the lowest observed energy, and probable binding energy) and n_{req} in the comparable $HBr - Ne$ and $CO - Ne$ systems. Since the atomic size disparity is also increasing it is impossible from this plot to unpick which of the two properties may be responsible, or if both, or neither are contributing.

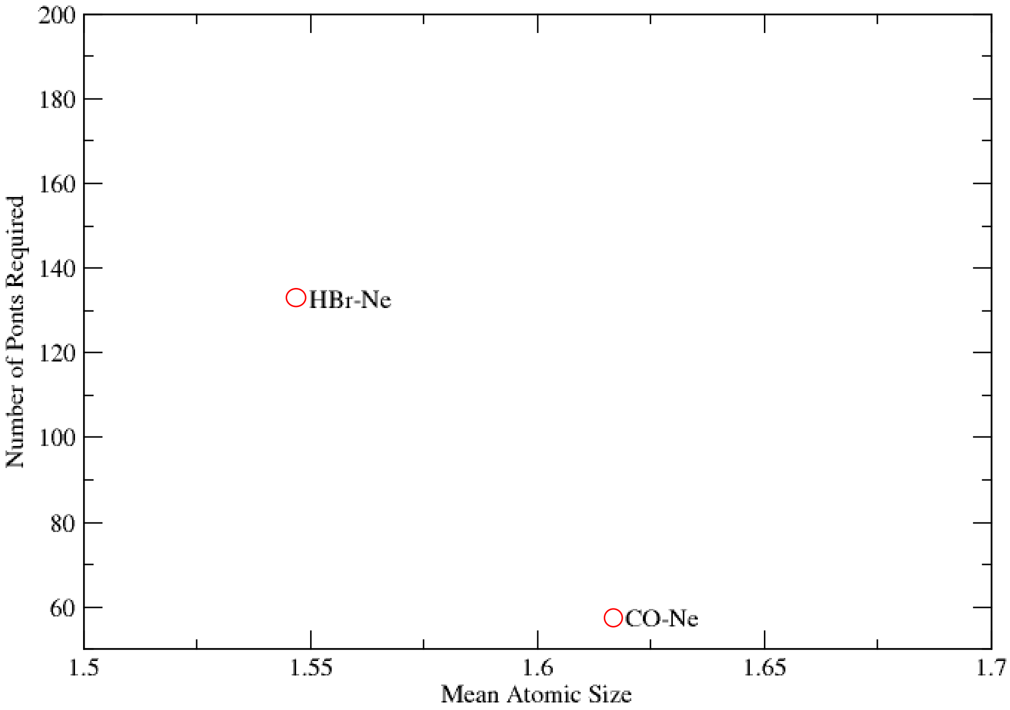

Figure 36.: A plot of mean atomic size against n_{req} for HBr and CO interacting with Ne.

The mean atomic size of all molecules in the $HBr - Ne$ and $CO - Ne$ dimer is presented against n_{req} in figure 36, with the larger atomic size corresponding to a smaller number of required points. This is at odds with the conjecture that larger atoms make the system more difficult, based on the observable difficulty of $SCl_2 - Ne$, seen below.

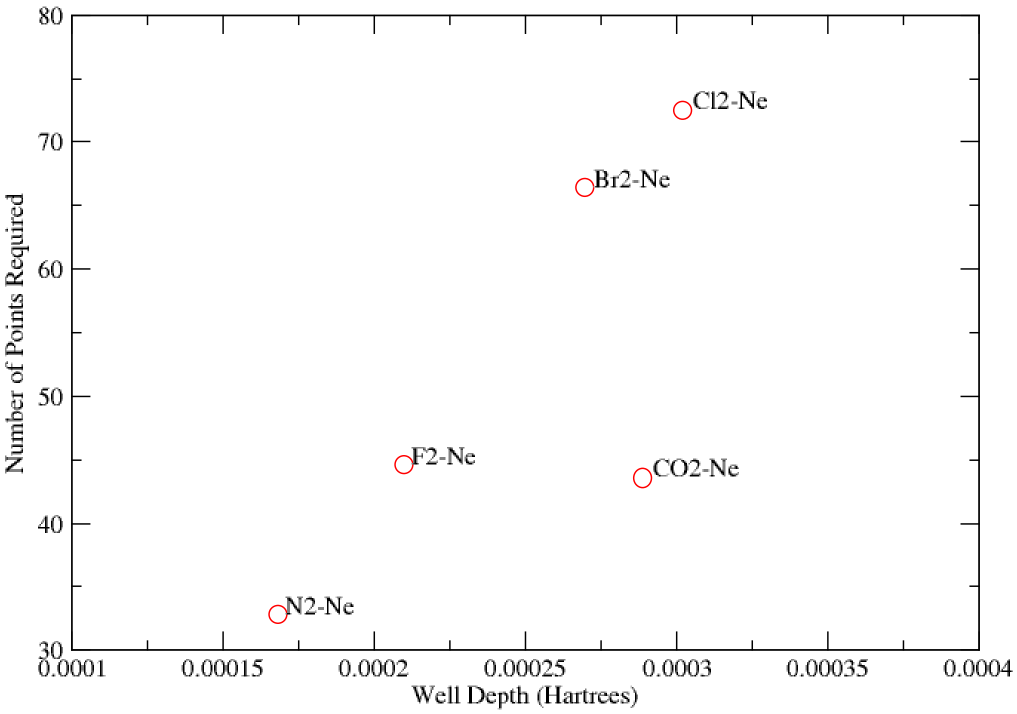

Figure 37.: A plot of well depth for N_2, F_2, CO_2, Br_2 and Cl_2 interacting with Ne.

The systems with two symmetries depicted in figure 37 exhibit a strong correlation with well depth, the exception being $CO_2 - Ne$ which may be somewhat easier due to the triatomic structure of CO_2. Well depth is obtained by examining the molecular simulation output values, from which the minimum energy value is take to be representative of the energy well.

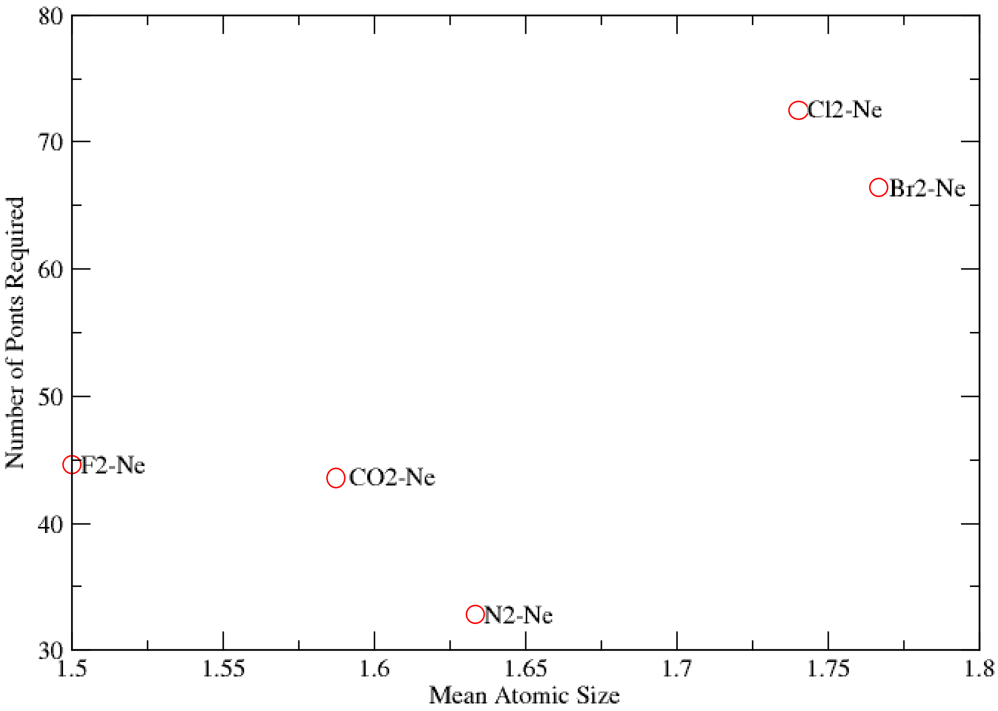

Figure 38.: A plot of mean atomic size against n_{req} for the centrosymmetric linear N_2, F_2, CO_2, Br_2 and Cl_2 molecules interacting with Ne.

Figure 38 shows a noisy plot of mean atomic size against n_{req} with an overall upward trend, at odds with figure 36 but in agreement with the conjecture that larger atomic sizes increase the difficulty of the system. It may also be observed that the halogens vary more widely in size than the systems composed of carbon, nitrogen and oxygen. This may be part of the reason the $CO_2 - Ne$ system's n_{req} does not vary from $N_2 - Ne$ as much as we may assume from the well depth.

5.3 SYSTEMS WITH 3 DEGREES OF FREEDOM.

All the systems with 3 degrees of freedom sampled have two permutations to their symmetry group so no useful correlation may be struck between this and the number of points required ; The system does however show an expansive variability in in number of points required due to other factors.

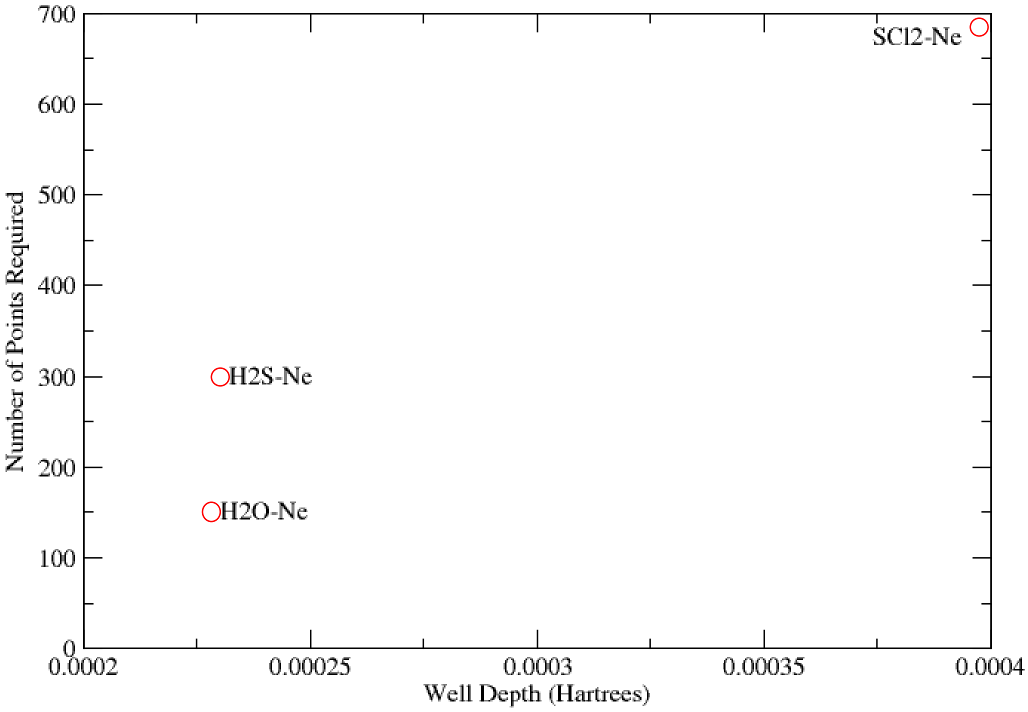

Figure 39.: A plot of points required versus well depth for systems with 3 degrees of freedom.

From the observable correlation in figure 39 it seems reasonable to assume that deeper energy wells require more sampling to be accurately resolved, with $SCl_2 - Ne$ having the deepest well at 0.0004 E_h and requiring 686 points. $H_2S - Ne$ and $H_2O - Ne$ by comparison have well depths of only 0.000230 and 0.000228 E_h respectively and each require 300 and 152 sample points, suggesting that well depth may be the reason that SCl_2 is more difficult to resolve.

Table 8. Atomic size disparities of H_2O and H_2S.

	H_2O	H_2S
atomic size disparity (%)	20	36.5
n_{req}	152	300

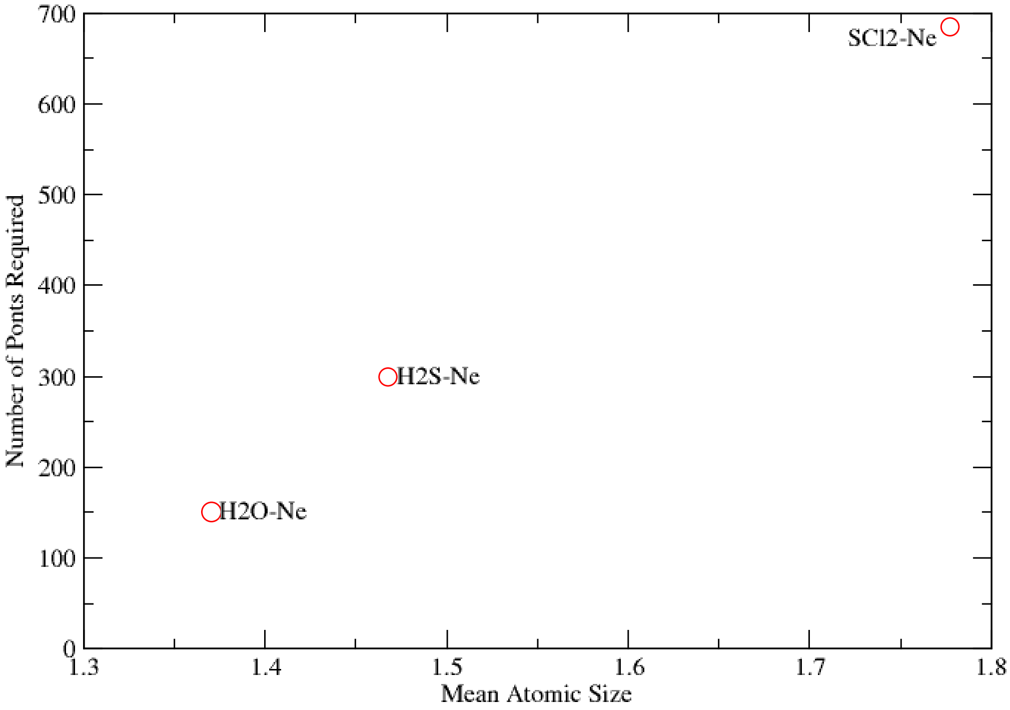

Figure 40.: A plot of mean atomic size against n_{req} for H_2O, H_2S, and SCl_2 interacting with Ne.

Figure 40 Shows a linear correlation between atomic size and n_{req}, suggesting this may indeed be at least a contributing factor to the difficulty of $SCl_2 - Ne$ compared to similar systems.

Table 8 shows the atomic size disparity of the two remaining systems: $H_2O - Ne$ and $H_2S - Ne$. Both have similar well depths, but the system involving H_2S is almost twice as difficult, and it is clear from the table that this orders correctly with the size disparity which is also almost double.

5.4 SYSTEMS WITH 4 DEGREES OF FREEDOM.

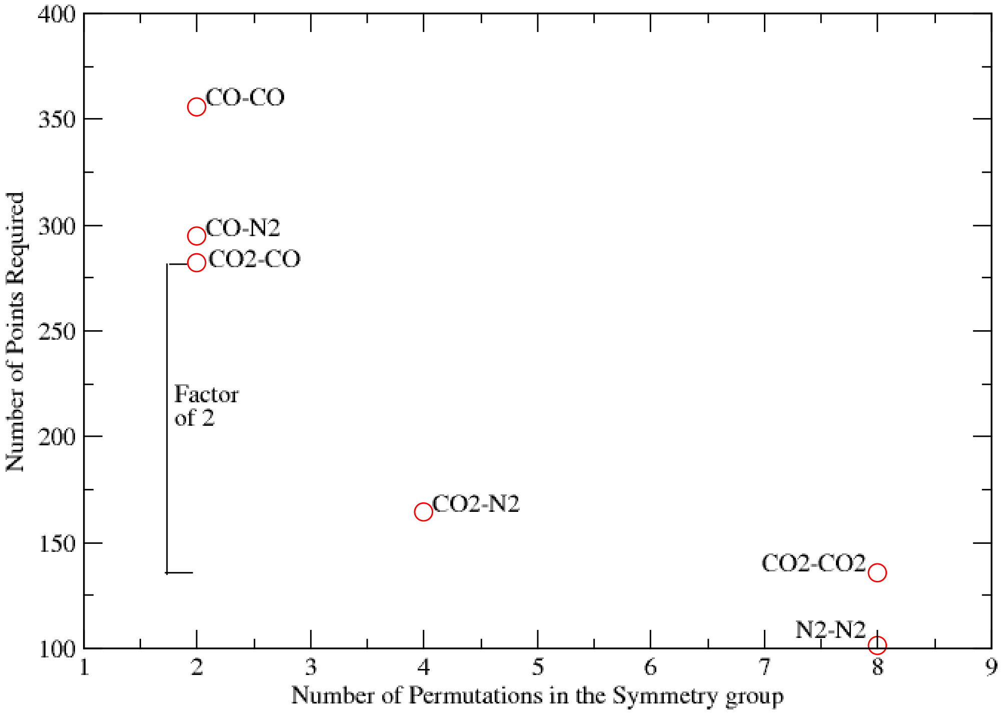

Figure 41.: A plot of points required vs number of permutations in the symmetry group for systems with 4 degrees of freedom.

Figure 41 shows a downward trend in the number of samples required as the number of permutations increases from 2 to 4 to 8. The $CO - CO$ system requires the most points, and has a single interchange symmetry between molecules. Faring slightly better, and resembling one another closely are $CO - N_2$ and $CO_2 - CO$, which each have a single flip symmetry within

CO_2 and N_2 respectively. The $CO_2 - N_2$ system requires approximately half the points of the systems preceding it at 165, and features two flip symmetries, suggesting that an additional flip symmetry has a halving effect on the difficulty of the system. The final two systems, $CO_2 - CO_2$ and $N_2 - N_2$ each feature two flip symmetries and one interchange symmetry, and require about two thirds of the points that $CO_2 - N_2$ does. This suggests that the benefit of an interchange symmetry is less than that of a flip symmetry, reducing difficulty by a third.

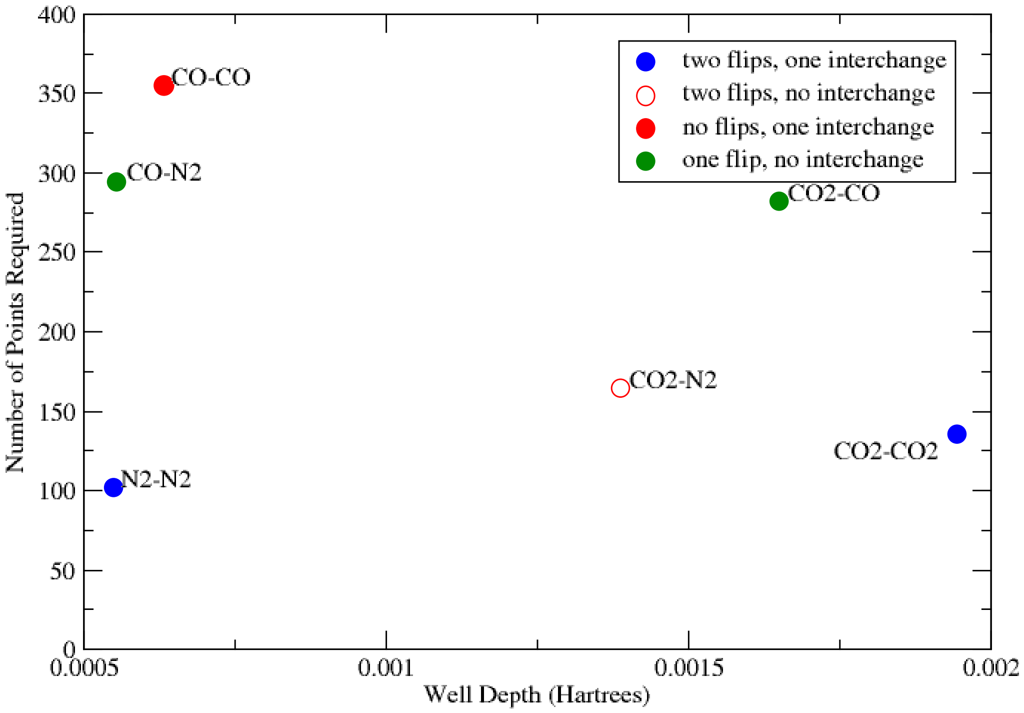

Figure 42.: A plot of points required vs energy well depth for systems with 4 degrees of freedom.

It is apparent from figure 42 that there is much noise and no correlation between well depth and the number of points required when comparing dimers with the same symmetry group, eliminating this as the cause of the

Table 9. Atomic size disparities of $N_2 - N_2$ and $CO_2 - CO_2$.

	Pair A		Pair B	
	$N_2 - N_2$	$CO_2 - CO_2$	$CO_2 - CO$	$CO - N_2$
atomic size disparity (%)	0 & 0	15.3 & 15.3	15.3 & 15.3	15.3 & 0
n_{req}	101	136	283	295

trend observed in figure 41. When comparing systems in the plot with the same symmetry group, n_{req} is almost constant despite changes in well depth. It is also worth observing that the atomic size disparity between carbon and oxygen is a relatively modest 15.3%, and among this data the only samples which have the same combination of flip and interchange symmetries are $N_2 - N_2$ and $CO_2 - CO_2$, and $CO_2 - CO$ and $CO - N_2$. These systems are compared in table 9 and a 36% difference in difficulty is observed in the more disparate $CO_2 - CO_2$ system; while this contribution is significant, the magnitude of differences imposed by the symmetries clearly predominates. In the second pair compared, the more disparate system actually requires slightly fewer points.

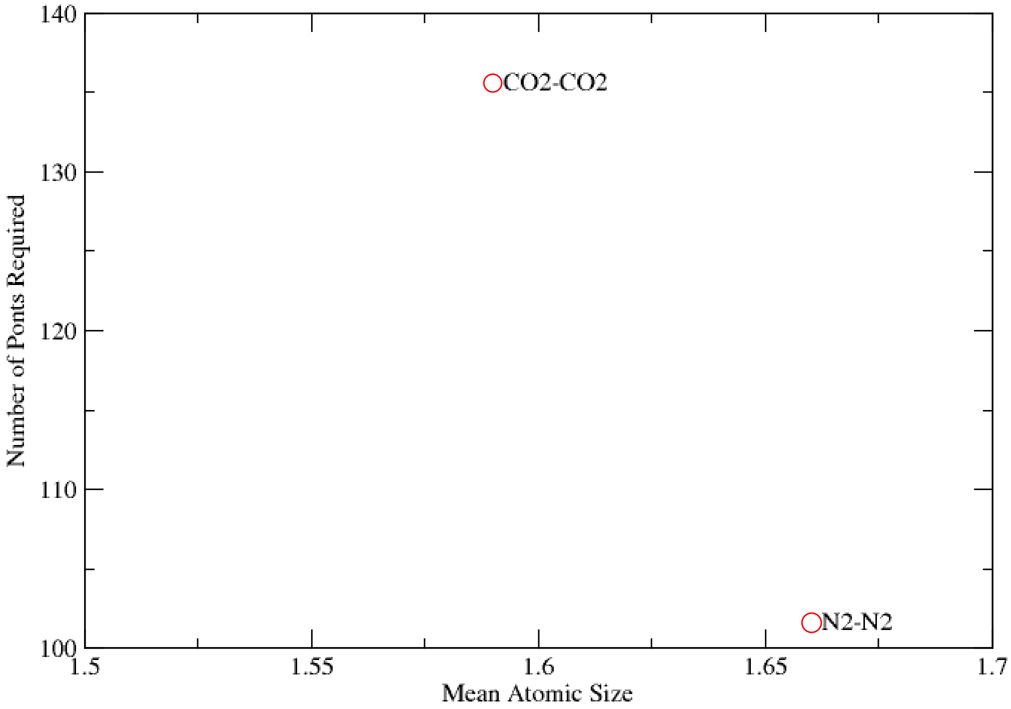

Figure 43.: A plot of mean atomic size against n_{req} for $N_2 - N_2$, and $CO_2 - CO_2$.

Dimers of CO_2 and N_2 may be compared together since they posses similar symmetries. Conversely to the earlier conjecture that increasing mean atomic size correlates with difficulty, figure 43 shows an inverse correlation. This observation is confounded however by the two systems being composed of diatomic vs triatomic molecules, so no clear relationship may be stated.

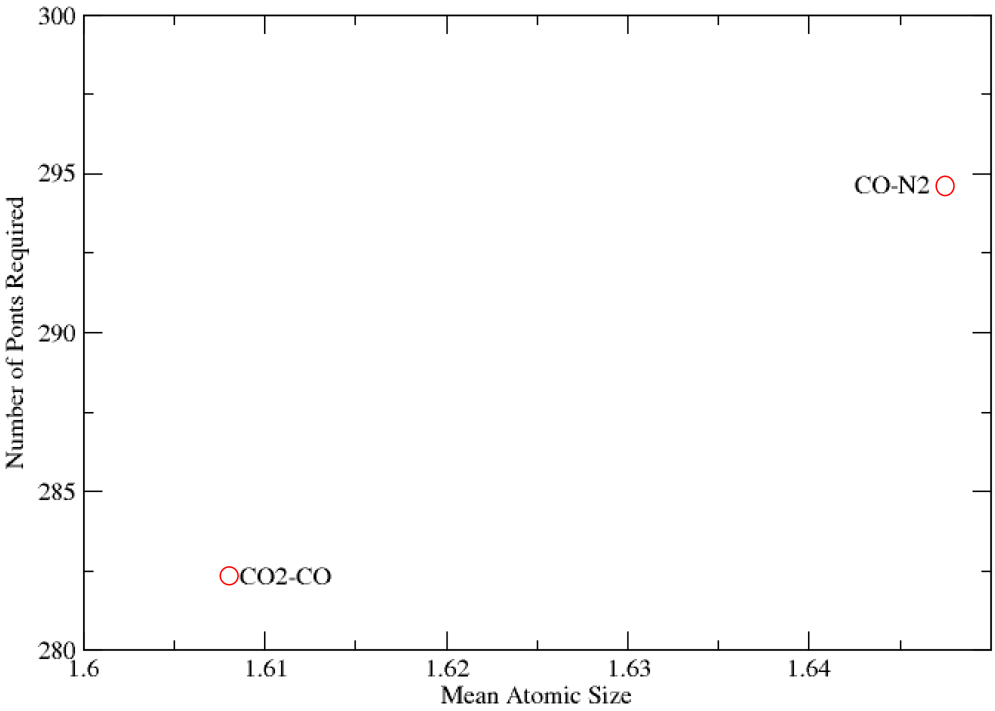

Figure 44.: A plot of mean atomic size against n_{req} for $CO_2 - CO$, and $CO - N_2$.

Since CO_2 and N_2 are similarly symmetric, $CO_2 - CO$ and $CO - N_2$ are grouped together for comparison. The observed trend agrees with the conjecture that increasing atomic size correlates with an increasing number of required points. It must also be emphasized that comparing diatomic and triatomic N_2 and CO_2 is not a like for like comparison, and other factors may contribute to the observed trend.

5.5 SYSTEMS WITH 5 DEGREES OF FREEDOM.

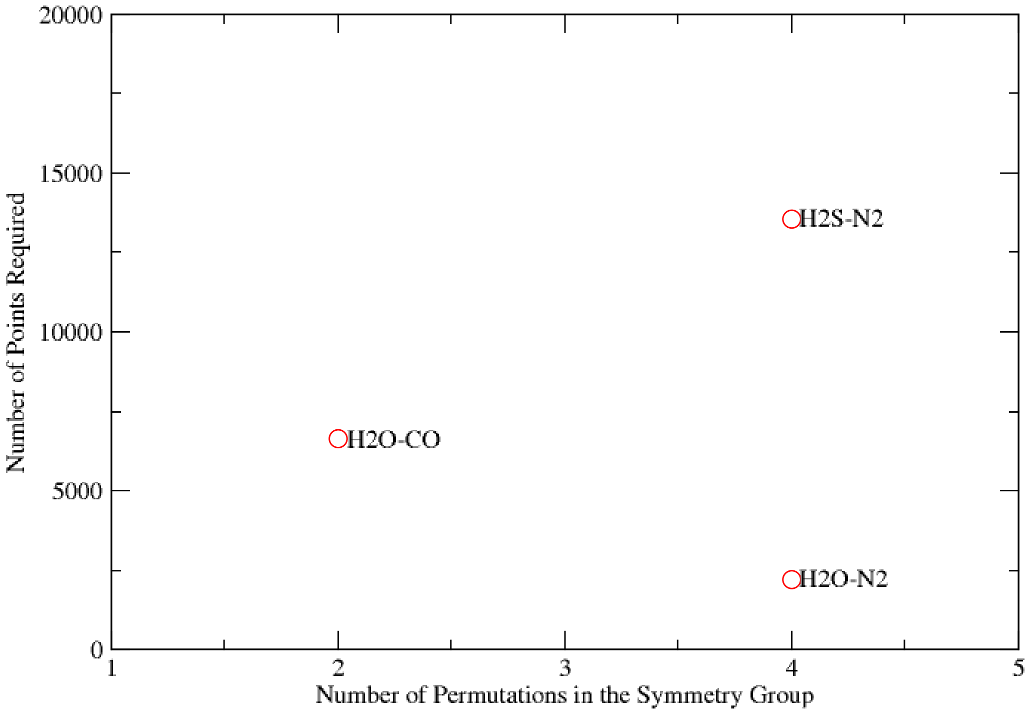

Figure 45.: A plot of points required vs number of permutations in the symmetry group for systems with 5 degrees of freedom.

In figure Figure 45 the downward correlation is not so easily observed, due to the $H_2S - N_2$ system requiring vastly more points to resolve than those composed of $H_2O - N_2$ and $H_2O - CO$, an effect which has been previously observed with systems containing H_2S. We have previously ascribed this to the disparity in size between hydrogen and sulphur. It is worth noting that since the $H_2S - N_2$ system required extrapolation to obtain a value for n_{req}, it has a higher level of uncertainty than the other plots, but the extrapolation otherwise causes no visible features. Putting aside the hydrogen sulphide system, the $H_2O - N_2$ system requires 2214 sample points and $H_2O - CO$

requires around 3 times this many at 6655, exceeding the halving effect one may assume the extra symmetry may provide.

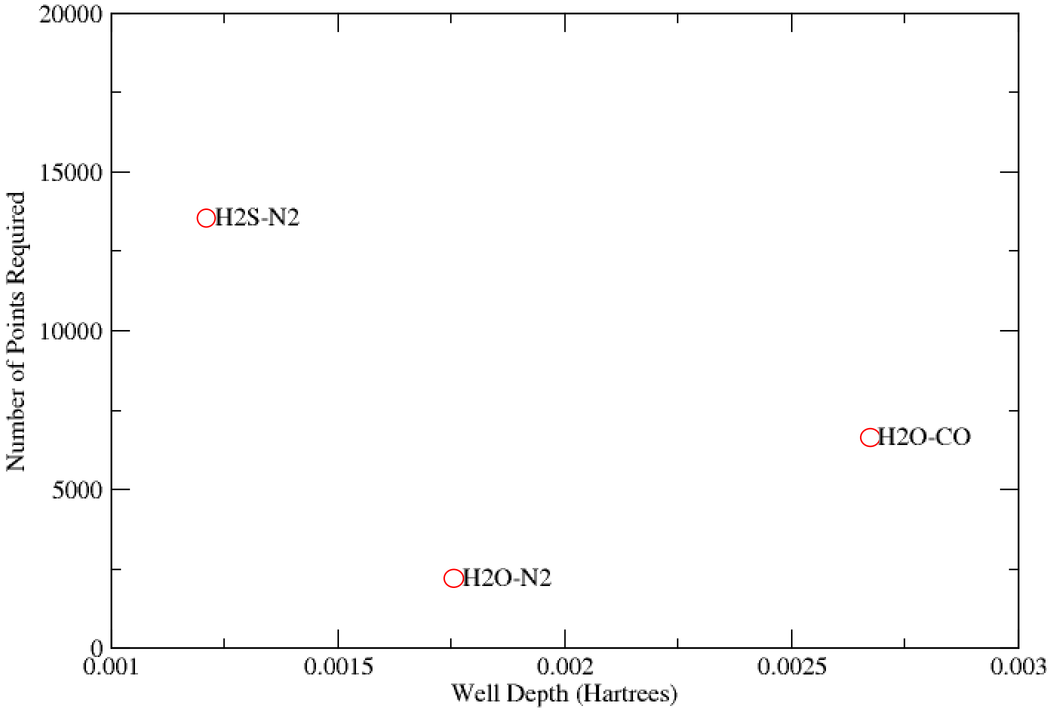

Figure 46.: A plot of points required vs minimum observable energy for systems with 5 degrees of freedom.

Figure 46 resembles figure 45, in that the extrapolated value for the $H_2S - N_2$ system stands alone, indicating that well depth offers no compelling explanation for the difficulty for this system relative to those containing H_2O, and another factor such as atomic size disparity predominates. Nevertheless, it can be observed that $H_2O - CO$ has a well depth of 0.00267 E_h and $H_2O - N_2$ has a well depth of 0.00175E_h; referring back to figure 45 it was discussed how the number of permutations logically should halve the number of points required to 3327.5, but the plot indicates a more substantial reduction. By normalising the two well depths to one another a coefficient of 1.52 is obtained, the reciprocal of which conveniently scales 3327.5 to 2183, closely approximating the observed value of 2214 points for $H_2O - N_2$ and suggesting a linear relationship between well depth and difficulty.

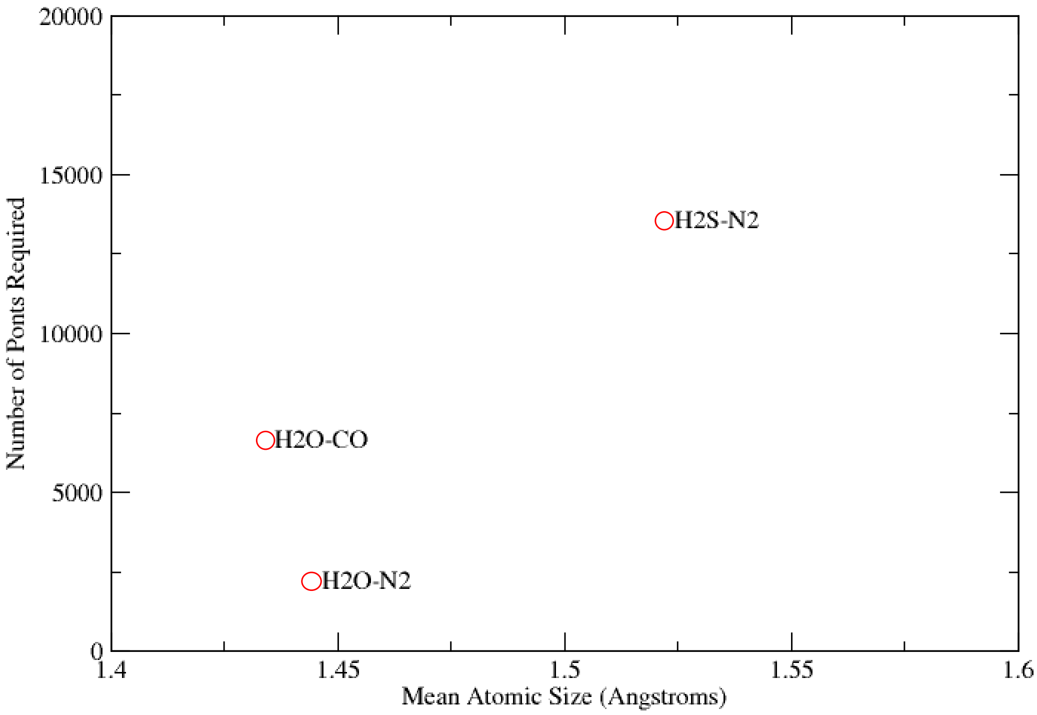

Figure 47.: A plot of mean atomic size against n_{req} for $H_2O - N_2$, $H_2S - N_2$, and $H_2O - CO$.

Once again, in figure 47 an upward trend in difficulty may be observed in response to increasing mean atomic size. While $H_2O - CO$ does not appear to conform to this trend, it may be discounted due to the reduced permutability of CO compared to N_2, and observations made from $H_2O - N_2$ and $H_2S - N_2$ alone.

5.6 EXAMINING WELL DEPTH, MEAN ATOMIC SIZE, AND ATOMIC SIZE DISPARITY.

To examine the contributions of well depth, mean atomic size and mean atomic size disparity in isolation, the effects of degrees of freedom and symmetry must be eliminated so far as is possible. Since the relationship between n_{req} and DoF is shown in figure 31 to be $n_{req} = 2.42 \exp 1.5758 dof$, individual n_{req} values may be divided by the value of n_{req} corresponding to their number of DoF. It is also shown in figure 41 that a flip symmetry halves the expected value of n_{req}, and an interchange symmetry reduces it by a third. From these relationships an equation may be formulated to normalise the difficulty with respect to DoF and symmetry:

$$normalised\ difficulty = \frac{n_{req}\left(2^{n_{flips}}\right)}{2.4201 e^{1.5758 dof}\left(0.66^{n_{interchanges}}\right)}, \tag{19}$$

from which comparisons may be made against well depth, mean atomic size and mean atomic size disparity.

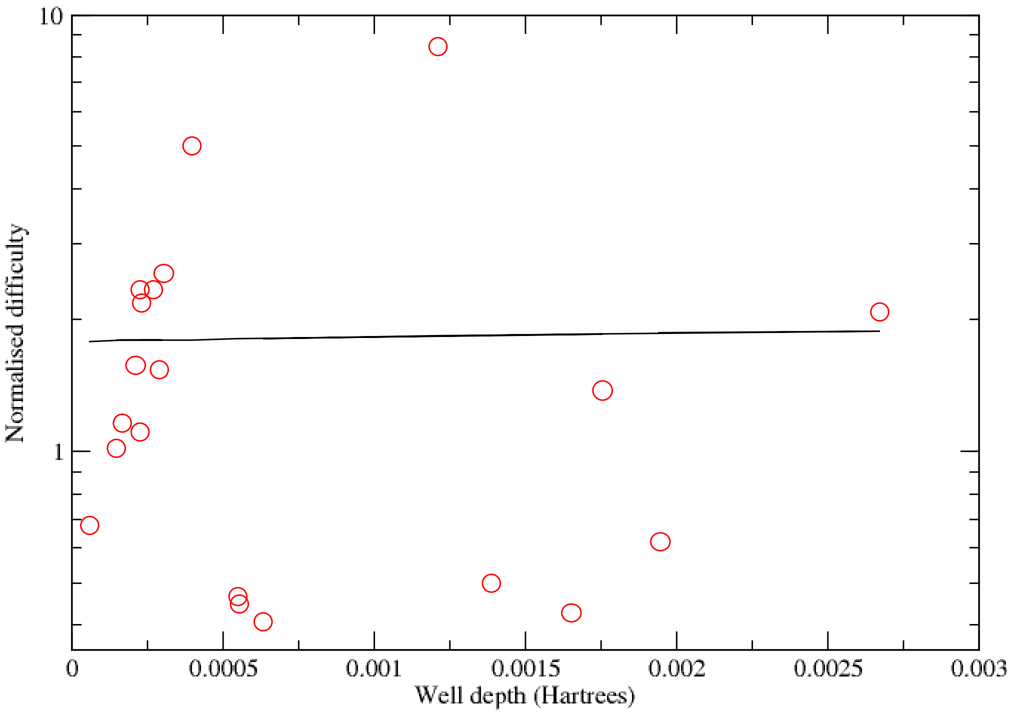

Figure 48.: A plot of the normalised difficulty of systems vs well depth.

Figure 48 displays considerable scatter, and shows such a weak correlation with so much scatter that no clear relationship can be drawn between the difficulty and well depth, which is in agreement with effects observed previously.

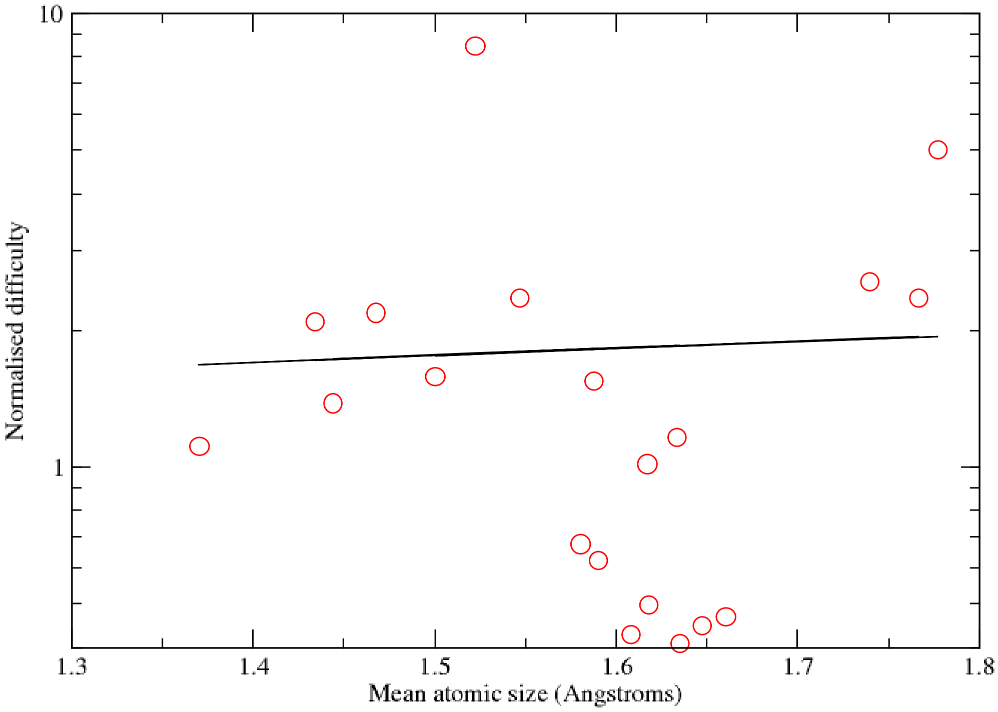

Figure 49.: A plot of the normalised difficulty of systems vs mean atomic size.

The plot of difficulty against atomic size in figure 49 shows such a weak positive correlation with such considerable scatter that no real relationship may be observed.

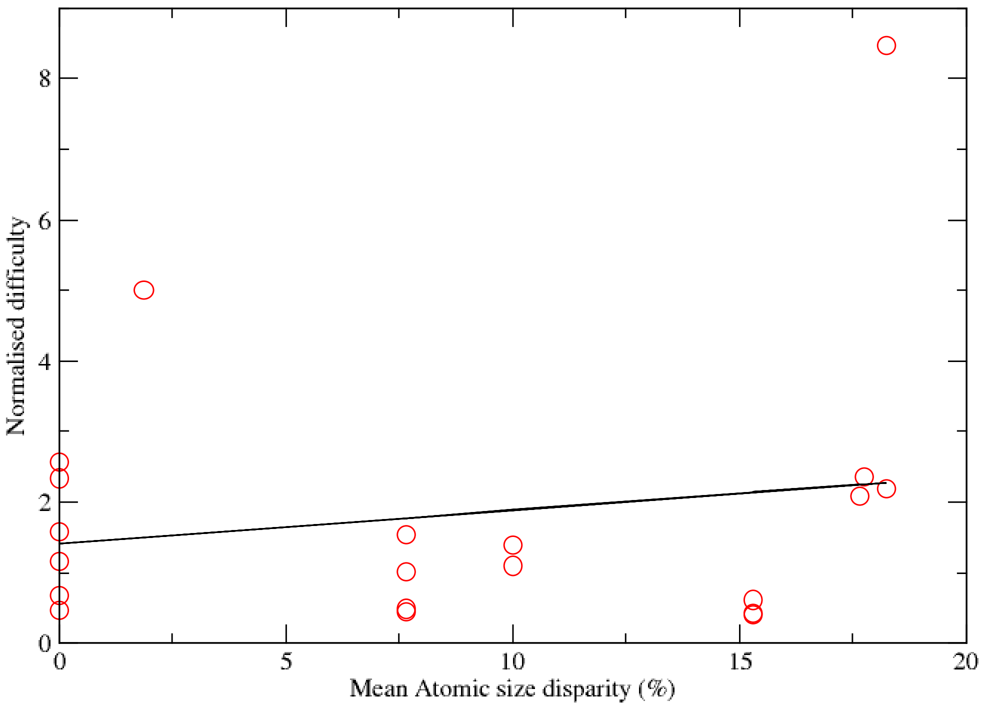

Figure 50.: A plot of the normalised difficulty of systems vs mean atomic size disparity.

Figure 50 shows a weak positive correlation between the difficulty and atomic size disparity. Once again there is considerable scatter so no definite relationship may be observed, as is the case in figure 48 and 49.

Given the extent of scatter alongside weak positive correlation in Figures 48, 49 and 50, it seems difficult to unpick the contributions of each. Despite the scatter in these plots, it is clear that the normalised difficulty falls within a fairly narrow band, supporting the idea that DoF and number of symmetries are the key determinants of difficulty.

5.7 QUANTIFYING AN EMPIRICAL FORMULA.

An estimate of n_{req} may be made using the data in figure 31 along with the observations that flip symmetries half n_{req} and interchange symmetries reduce n_{req} by 34%. Each value of n_{req} may be divided based on the contributions of its symmetry elements as shown:

$$n_{asymm} = \frac{n_{req}\left(2^{n_{flips}}\right)}{\left(0.66^{n_{interchanges}}\right)}, \tag{20}$$

yielding the plot shown below.

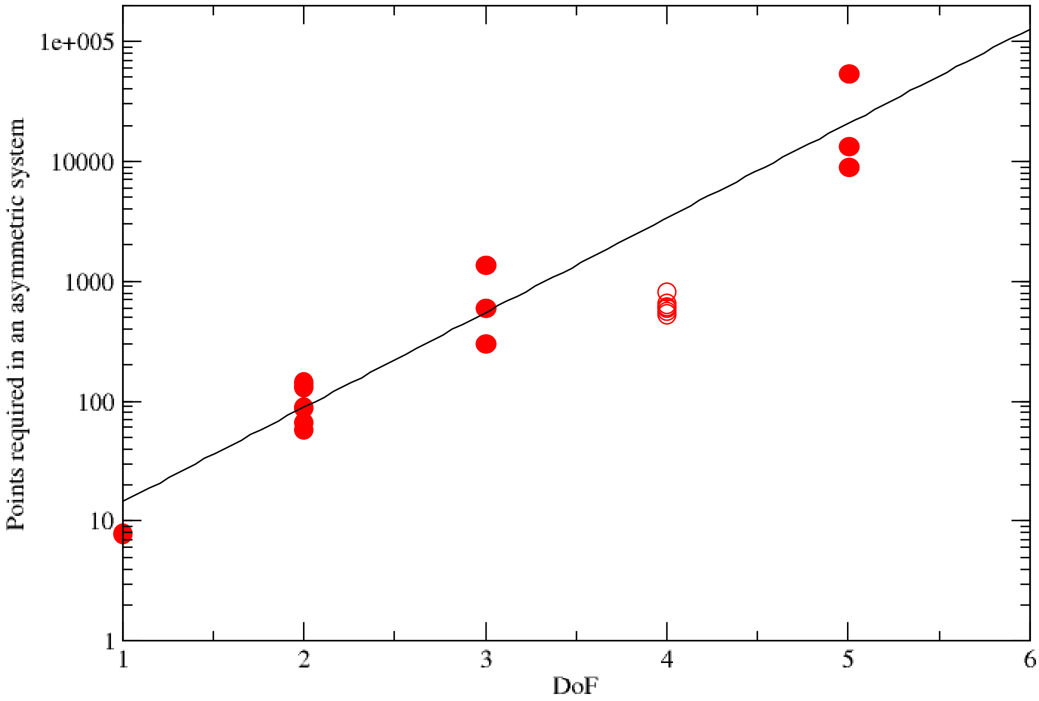

Figure 51.: A plot of n_{asymm} vs DoF. The trend fitted is: $y = 2.3514e^{1.8172x}$

Once again a line is fitted omitting the 4 DoF systems, which deviate downward from the general trend. The equation of this line may now be used to substitute n_{asymm} in equation (20), and a value of n_{req} determined

Table 10. Comparison of 5 and 6 DoF systems.

	H2O-N2	H2O-CO	H2S-N2	H2S-H2O
DoF	5	5	5	6
well depth (Hartrees)	0.00175	0.00267	0.00121	0.00356
size disparity (%)	20 & 0	20 & 15.3	36.5 & 0	36.5 & 20
flip symmetries	2	1	2	2

for a system with any combination of flip and interchange symmetries using the appropriate scaling factors:

$$n_{req} = \frac{\left(0.66^{n_{interchanges}}\right) 2.3514 e^{1.8172x}}{\left(2^{n_{flips}}\right)}, \tag{21}$$

in this manner it is possible to estimate n_{req} for a system of interest with 6 DoF.

5.8 EXTRAPOLATIONS TO SYSTEMS WITH 6 DEGREES OF FREEDOM.

Extrapolation to six degrees of freedom is possible using equation (21), which places n_{req} at 127,800. The $H_2S - H_2O$ system previously examined features two flip symmetries and no interchanges, bringing the estimate down to 31,950, showing good agreement with the 31,000 points projected by figure 31, and still far greater than the largest training set used in previous attempts to model the system.

In addition to two flip symmetries and 6 degrees of freedom, the $H_2S - H_2O$ features an observable well depth of 0.0035612 E_h and reasonably large atomic size disparities at 20% for H_2O and 36.5% for H_2S.

Since $H_2S - H_2O$ possesses a significantly greater well depth than the 5 DoF systems interrogated, as seen in table 10, as well as two large size disparities while possessing a comparable number of flip symmetries, it would be reasonable to assume that the system's additional complexity would warrant slightly more than 32,000 sample points. Given the increase in difficulty associated with well depth in figure 46, and the relative performance of $H_2S - Ne$ versus $H_2O - Ne$ in figure 39 the number of points required may, in a worst case scenario, be twice as high (64,000). During the previous attempt to model the system only 10,000 sample points were used in the largest training set, and 20,000 in the test set, explaining why the benchmark accuracy was not achieved.

6

SUMMARY AND CONCLUSIONS.

When modelling a potential energy surface with a Gaussian Process the cost of computation is proportionate to the number of sample points. The number of sample points also determines how many calculations need to be done when upgrading to intensive calculations such as CCSD(T) thus it becomes essential to calculate only the minimum number of sample points required (n_{req}). It was demonstrated with a $H_2S - H_2O$ system that Active Learning (AL) can be used to lower the number of points required to achieve a benchmark precision, subject to the Latin Hypercube (LHC) used for the initial training set being sufficiently large. Following the study on $H_2S - H_2O$, a selection of pairwise interactions were scrutinised, with a variety of geometries, symmetries, and valences in order to interrogate the contributions of each. These dimers were modelled from first principles using Møller-Plesset perturbation theory, and a Latin hypercube design strategy to allocate a representative sample of geometries. By using a number of different sample sizes compared to a large test set, some measure of the sample size required to make a satisfactorily precise model was observed. The systems were namely: Ne, F_2, Cl_2, Br_2, N_2, CO, HBr, CO_2, H_2O, H_2S, and SCl_2 interacting with Ne; N_2, CO, CO_2, H_2O, and H_2S interacting with N_2; CO, CO_2, and H_2O interacting with CO and finally CO_2 interacting with CO_2. By obtaining RMSE values for a number of different sized LHCs with each of these systems, a trend was fitted to describe the relationship between LHC size and RMSE.

By rearranging the fitted function with a known RMSE benchmark of 10^{-5} E_h and solving for n of LHC points, a value of n_{req} was obtained. The benchmark RMSE of 10^{-5} E_h was chosen because this error level was shown to be suitable to make first principles prediction of physical properties [**Uteva et al. 2018**].
It was shown that for many systems involving spherical, linear and bent molecules 10^{-5} Hartrees was an achievable RMSE using a LHC design regime, but as the number of degrees of freedom increase, so does the number of samples required. By estimating and plotting the number of points required to achieve the required precision, it was shown that the relative difficulty of a system increases exponentially with the addition of degrees of freedom. These results suggest that the number of atoms in a molecule does not appear to impact the precision of a model in any meaningful way, as long as the shape of the molecule does not change, nor the angular descriptors of the symmetrically distinct region. It was also demonstrated that there is an approximately linear relationship between the reciprocal of the number of symmetric permutations as well as a linear relationship to the energy well depth, and the number of points needed to resolve a system, for some systems.
The principal determining factor of n_{req} was shown to be the number of degrees of freedom (DoF) in the system, exhibiting an exponential correlation. Next in order of contribution is the permutability of the system, with an important distinction to be made between 'flip' symmetries which half the value of n_{req}, and 'interchange' symmetries which reduce n_{req} by around a third. Thirdly, the energy well depth of an interaction shows an approximately linear correlation with n_{req}, as does the mean atomic size,and size disparity between a molecules atoms. It was also demonstrated that although Active learning techniques offer improved determination of Gaussian processes, in one case yielding one tenth the RMSE of a Latin Hypercube design strategy, neither the AL nor LHC method was able to model the $H_2O - H_2S$ system's potential energy surface in a satisfactory fashion,

with 10,000 training points.
By quantifying these effects an approximation of the number of points required to model $H_2S - H_2O$ may be placed at 32,000 and possibly as high as 64,000. Since only 10,000 sample points were included in the largest training set previously studied it is abundantly clear that the sheer volume of points required to model a system with 6 degrees of freedom was the reason for an unsatisfactory RMSE. The pay off of this investigation is that an estimate of n_{req} permits accurate cost appraisal for calculating a system. As the complexity of a system grows this will become increasingly important, as high performance computer time is limited and large scale computing projects require reasonable estimates of the required compute effort during the planning stage.

6.1 FUTURE WORK.

Because of the exponential increase in difficulty, efforts must be made to formulate ways to more easily model systems with many degrees of freedom, along with systems possessed of great size disparity.

A measure which could be implemented to save on computation is to have different lengthscales at short range but to equate lengthscales at longer range, where the interaction of molecules resembles the interaction of a pair of atoms. This might be achievable using a new non stationary covariance function in the GP. The important decision to be made when doing this is to accurately decide what point to transition to a single lengthscale without wasting computation or adversely affecting the model. This could probably be achieved when the GP hyperparameters are optimised. Adaptations to improve the modelling of molecules with significant size disparity may also include the use of non-stationary covariance functions, so as to account for the different energy characteristics of atoms with different radii.

Based on the significant effect that an energy well may have, it is also recommended that in future, the high energy cut off be a fixed multiple of this value so as to avoid wasting computation on thermodynamically inaccessible regions. With this change it would also be reasonable to make the target RMSE proportional to the well depth, which may reduce the influence of the well depth.

Since active learning was demonstrated to reduce the size of training set required to model a potential energy surface, future works should examine whether this methodology is similarly effective in systems with different numbers of DoF. If this method proves effective in systems with many DoF it may help lessen the associated exponential increase in difficulty. To verify whether this is the case, a selection of the dimers examined herein could be modelled with active learning, and the relative reduction in sample points required to achieve the benchmark RMSE compared across 1 to 5 DoF systems.

It may be that high dimensional dimers intrinsically require many training points. In this case applications will need to adapt around the constraint of dramatically more expensive PES than are currently typical. Here parallelized calculations of the GP would assist in spreading the computational load over many processors. GP evaluation appears to be a natural candidate for parallelization as they require the summation of many exponential terms. Such parallelized GPs could be implemented in molecular simulations.

BIBLIOGRAPHY

[1] X. Guo, D. Li, and A. Zhang. *Improved Support Vector Machine Oil Price Forecast Model Based on Genetic Algorithm Optimization Parameters* 2012: AASRI Procedia, 1, (1), 525-530.

[2] A. Stone. *The theory of intermolecular forces.*2013: Oxford. Oxford University Press, second edition, 1-7.

[3] G. R. Davies, V. Silva Aguirre, T. R. Bedding, R. Handberg, M. N. Lund, W. J. Chaplin, D. Huber, T. R. White, O. Benomar, S. Hekker, S. Basu, T. L. Campante, J. Christensen-Dalsgaard, Y. Elsworth, C. Karoff, H. Kjeldsen, M. S. Lundkvist, T. S. Metcalfe, D. Stello. *Oscillation frequencies for 35 Kepler solar-type planet-hosting stars using Bayesian techniques and machine learning* 2015: Monthly Notices of the Royal Astronomical Society, 456, (2), 2183-2195.

[4] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis. *Machine Learning in Agriculture: A Review* 2018: Sensors, 18, (9), 2674.

[5] J. Cui, and R. V. Krems. *Efficient non-parametric fitting of potential energy surfaces for polyatomic molecules with Gaussian processes.*2016: Journal of Physics B: Atomic, Molecular and Optical Physics, 49, (22).

[6] J.A.K. Suykens, and J. Vandewalle. *Least Squares Support Vector Machine Classifiers*1999: Neural Processing Letters, 9, (3), 293-300.

[7] A. J. Smola, and B. Schölkopf. *A tutorial on support vector regression* 2004: Statistics and Computing, 14, (3), 199-222.

[8] J. El Yazal, and Y.-P. Pang. *Comparison of DFT, Møller–Plesset, and coupled cluster calculations of the proton dissociation energies of imidazole and N-*

methylacetamide in the presence of zinc(II) 2001:Journal of Molecular Structure: THEOCHEM. 545, (3), 271-274.

[9] H. Kuhn, H.-D. Forsterling, and D. H. Waldeck. *Principles of Physical Chemistry.*2009: Hoboken, USA. Wiley, Second edition, 67-71.

[10] E. Uteva, R. S. Graham, R. D. Wilkinson, and R. J. Wheatley. *Interpolation of intermolecular potentials using Gaussian processes* 2017: Journal of Chemical Physics, 147, (16).

[11] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning* 2006: Massachusetts Institute of Technology.

[12] E. Uteva, R. S. Graham, R. D. Wilkinson, and R. J. Wheatley. *Active Learning in Gaussian Process Interpolation of Potential Energy Surfaces* 2018: Journal of Chemical Physics, 149, (17).

[13] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz et al. 2018: www.Molpro.net.

[14] GPy: A Gaussian process framework in python. 2012: [ONLINE] Available at: http://github.com/SheffieldML/GPy Accessed 21 August 2019.

[15] R. Hellmann. *Nonadditive three-body potential and third to eighth virial coefficients of carbon dioxide* 2017: Journal of Chemical Physics,146, (5), 054-302.

[16] M. T. Oakley, and R. J. Wheatley. *Additive and nonadditive models of vapour-liquid equilibrium in CO_2 from first principles* 2009: Journal of Chemical Physics, 130,(3), 34-110.

[17] S. Niu, and M. B. Hall. *Comparison of Hartree-Fock, Density Functional, Møller-Plesset Perturbation, Coupled Cluster, and Configuration Interaction Methods for the Migratory Insertion of Nitric Oxide into a Cobalt-Carbon Bond.* 1997: Journal of Physical Chemistry A, 101, (7), 1360-1365.

[18] C. M. Handley, G. I. Hawe, D. B. Kell, and P. L. A. Popelier. *Optimal construction of a fast and accurate polarisable water potential based on multipole moments trained by machine learning* 2009: Physical Chemistry Chemical Physics, 11, (30), 6365-6376.

[19] W. J. Szlachta, A. P. Bartok, and G. Csanyi. *Accuracy and transferability of Gaussian approximation potential models for tungsten* 2014: Physical Review B, 90, (10), 104-108.

[20] Y. Huang, and G. J. O. Beran. *Reliable prediction of three-body intermolecular interactions using dispersion-corrected second-order Møller-Plesset perturbation theory.* 2015: Journal of Chemical Physics, 143, (4).

[21] M. Chaplin. *Water Symmetry* [ONLINE] Available at: `www1.lsbu.ac.uk/water/h2o_orbitals.html` Accessed 21 August 2019.

[22] A. Lapidoth. *A Foundation in Digital Communication.*2017: Cambridge. Cambridge University Press, Second edition, 566-567.

[23] WebElements. 2020: [ONLINE] Available at: http://www.webelements.com Accessed 17 November 2020.

A

TABLE OF FITTED FUNCTIONS

The table of fitted functions is an abstract from a spreadsheet used to collate fit data for analysis, and provides a summary of the cross and cross-sigmoidal fits of RMSE vs number of sample points, as presented in chapter 4. The table also has a record of the projected values of n_{req}, well depth, degrees of freedom and number of symmetries, which are discussed in chapter 5.

molecule 1	molecule 2	fit function	R_0	X_0	α	R_∞	n_{req}	well depth	DOF	n_{sym}
Ne	Ne	sigmoidal	0.00118	4.89591	9.86588	2.17E-08	7.933228	-5.84E-05	1	1
N_2	Ne	sigmoidal	0.001047	10.2098	3.97213	7.22E-08	32.84443	-0.00017	2	2
F_2	Ne	cross	0.000888	11.5469	3.31175	-	44.59516	-0.00021	2	2
Cl_2	Ne	cross	0.001076	12.4685	2.65285	-	72.47422	-0.0003	2	2
Br_2	Ne	cross	0.001095	9.98412	2.47354	-	66.41094	-0.00027	2	2
CO	Ne	sigmoidal	0.000753	17.9541	3.70065	2.47E-07	57.50663	-0.00015	2	1
HBr	Ne	cross	0.001157	11.9234	1.96599	-	133.026	-0.00023	2	1
CO_2	Ne	sigmoidal	0.000983	9.5126	3.0057	1.95E-08	43.63181	-0.00029	2	2
H_2O	Ne	cross	0.000782	13.4728	1.79672	-	151.4005	-0.00023	3	2
H_2S	Ne	cross	0.000928	34.4482	2.08863	-	299.818	-0.00023	3	2
SCl_2	Ne	cross	0.001221	30.8315	1.54685	-	685.154	-0.0004	3	2
N_2	N_2	cross	0.001144	17.6749	2.70442	-	101.6415	-0.00055	4	8
CO	N_2	cross	0.00097	29.1527	1.97328	-	294.6439	-0.00055	4	2
CO_2	N_2	cross	0.001113	19.355	2.19869	-	164.34	-0.00139	4	4
H_2O	N_2	cross	0.001246	19.9289	1.02274	-	2213.641	-0.00175	5	4
H_2S	N_2	cross	0.001313	50.8139	0.734707	-	38411.01	-0.00121	5	4
CO	CO	cross	0.000761	56.3139	2.34409	-	355.5346	-0.00063	4	2
CO_2	CO	cross	0.001061	39.9933	2.38194	-	282.3389	-0.00165	4	2
H_2O	CO	cross	0.001354	36.5272	0.941546	-	6655.024	-0.00267	5	2
CO_2	CO_2	cross	0.001263	6.81787	1.61557	-	135.5891	0.001944	4	8

B

TABLE OF PERMUTATIONS

The table of permutations details the possible interchanges of atoms according to the numbering convention used in this study. These permutations are presented to the algorithm to allow it to populate its entire phase space with corresponding energies selected from only its symmetrically distinct region.

system	quantity	permutations
$Ne - Ne$	1	[1]
$N_2 - Ne$	2	[1,2], [2,1]
$F_2 - Ne$		
$Cl_2 - Ne$		
$Br_2 - Ne$		
$HBr - Ne$	1	[1,2]
$CO - Ne$		
$CO_2 - Ne$	2	[1,2,3], [1,3,2]
$H_2O - Ne$		
$H_2S - Ne$		
$N_2 - N_2$	8	[1,2,3,4], [1,3,2,4], [2,1,4,3], [2,4,1,3] [3,4,1,2], [3,1,4,2], [4,3,2,1], [4,2,3,1]
$CO - N_2$	2	[1,2,3,4], [2,1,4,3]
$CO_2 - N_2$	4	[1,2,3,4,5,6], [2,1,4,3,6,5], [1,2,5,6,3,4], [2,1,6,5,4,3]
$H_2O - N_2$		
$H_2S - N_2$		
$CO - CO$	2	[1,2,3,4], [1,3,2,4]
$CO_2 - CO$	2	[1,2,3,4,5,6]
$H_2O - CO$		[1,2,3,4,5,6]
$CO_2 - CO_2$	8	[1,2,3,4,5,6,7,8,9], [3,2,1,6,5,4,9,8,7], [7,8,9,4,5,6,1,2,3], [9,8,7,6,5,4,3,2,1], [1,4,7,2,5,8,3,6,9], [3,6,9,2,5,8,1,4,7], [7,4,1,8,5,2,9,6,3], [9,6,3,8,5,2,7,4,1]

C

TABLE OF ATOMIC SIZE DISPARITY

Atomic size disparity is defined as

$$Disparity = \sqrt{\left(\frac{(r_a - r_b)}{r_a}\right)^2} \times 100, \tag{22}$$

where r_a and r_b represent the radius of atom A and atom B. These values are obtained from the table of van der Waals radii, and the equation provides the values of atomic size disparity as found in the table of atomic size disparities. The atomic size disparity values are discussed in chapter 5 as a possible factor in the determination of how many sample points are required to accurately resolve a system.

	H	C	N	O	F	Ne	S	Cl	Br
van der Waals Radius (Angstroms)	1.2	1.77	1.66	1.5	1.46	1.58	1.89	1.82	1.86

Table 11. The values of the van der Waals radius for a selection of relevant atoms, which may be used to calculate atomic size disparity.

	Ne	F2	Cl2	Br2	N2	CO	HBr	CO2	H2O	H2S	SCl2
Atomic size disparity (%)	-	0	0	0	0	15.3	35.5	15.3	20.0	36.5	3.70

Table 12. The calculated values of atomic size disparity for a selection of relevant molecules, calculated using their constituent atoms respective van der Waals radii.

D

A VERBOSE PROCESS DIAGRAM

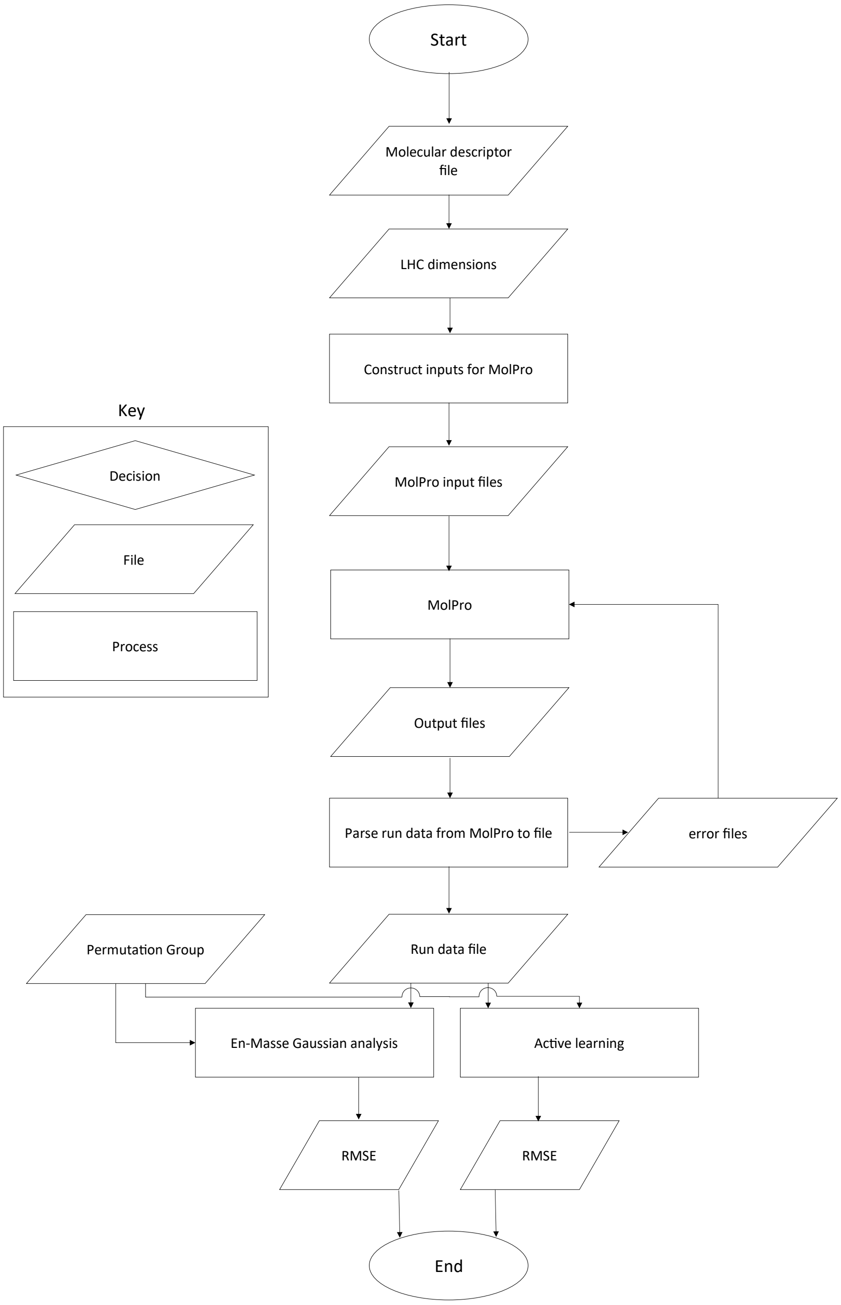

Figure 52.: A diagram of the pre-processing to design the sample regime, the Molpro input and output, and the analysis of data, including files.