Fabian Weißenbacher, BSc

Describing complex intermolecular interactions for structure search

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Technical Physics

submitted to

Graz University of Technology

Supervisor

Oliver T. Hofmann, Assoc.Prof. Dipl.Ing. Dr.techn

Institute of Solid State Physics

Graz, July 2020

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date

Signature

Acknowledgements

This is the place for me to acknowledge all the persons that helped and/or guided me on this long journey towards a finished master's thesis.

First and foremost, I want to thank my supervisor, Oliver T. Hofmann, not only for helping me whenever I ran into problems connected to the act of carrying out and writing a thesis, be they scientific, methodologic or organisational in nature, but also the unbelievable amount of patience he had with me in times when I had lost focus, determination and sometimes even hope to finish this thesis. In these instances he showed me ways to overcome my doubt and fear of failure and helped me to power through this metaphorical last mile to the finish line.

Another source of encouragement and support - both morally and scientifically, were my colleagues of the Advanced Materials Modelling group at the Institute of Solid State Physics. While my sincerest gratitude goes out to every single one of them for giving me advice on scientific methods and presentation skills, I especially want to thank Andi, Lukas and Alex for showing me how to efficiently work with SAMPLE, for challenging my results and conclusions and for answering every question that I had during the work on my thesis.

I also want to retroactively apologize to all my friends for having had too little time for many birthday parties, get-togethers and other activities during the last year and a half that I worked on my thesis. I promise to make it up in the future!

This brings me to my family, which through all of this, at times difficult, section of my academic and personal life has always stood behind me, supporting me both mentally and financially. They always encouraged and never pressured me, for which I am unspeakably grateful.

Last, but definitely not least, I want thank my marvelous girlfriend and better half, Marlene, for putting up with all the bad moods, depressive phases and other unpleaseantries that I was the source of and the long hours that I spend at university and in front of a PC that I could have otherwise spent with her.

I further want to thank her for helping me formulate my thoughts and conclusions as well as for proof-reading my thesis on such short notice. I naturally take responsibility for any errors that might have been missed.

Abstract

Describing complex intermolecular interactions for structure search

Fabian Weißenbacher
Institute of Solid State Physics, Graz University of Technology

The central goal in the field of materials design is to identify novel materials with a certain set of desired properties. One very interesting application is the adsorption of (small) organic molecules on metal surfaces, as these layers can be used to modify the properties of the interfaces they are applied to, for instance to shift work functions of metal electrodes that are contacting organic semiconductors.

These effects depend on the nature of the molecular arrangement, which is more generally known as a *surface polymorph*. A number of structure search algorithms has been proposed/introduced over the years which can be used to find and study these polymorphs, one fairly new addition to this list is the SAMPLE method.

It introduces an efficient surface polymorph generation algorithm which pushes back the configurational explosion by means of a clever coarse-graining algorithm in combination with an effective two-body energy model. At the same time, the need for expensive training data is minimized by employing Bayesian Learning in the form of Gaussian process regression to accomplish the prediction of properties, chiefly adsorption energies.

Since the energy model at the core of SAMPLE does not go beyond two-body interactions, the question arose whether this method is capable of describing complex molecular interactions. Therefore, this thesis aims to test the learning and prediction capabilities of SAMPLE for more complicated test systems than those which have previously been studied, with the focus on intermolecular interactions.

Tests are carried out on four classes of test systems, each tackling a specific aspect of intermolecular interactions, namely molecular symmetries, anisotropic bonds, substituent patterns and the effect of using different functional groups as substituents. The learning characteristics for each test system are presented, analyzed and cross-examined in relation to the other systems in order to judge whether SAMPLE can be seen as a robust structure search algorithm. In addition to that, a demonstration of the principle of transfer learning with SAMPLE is conducted on the subject of pair potentials. In summary, SAMPLE is found to be a robust and reliable structure prediction algorithm, which has a large potential for use and further development.

Kurzfassung

Beschreibung von komplexen intermolekularen Wechselwirkungen im Rahmen der Struktursuche

Fabian Weißenbacher
Institut für Festkörperphysik, Technische Universität Graz

Das Hauptziel von Materialdesign ist es, neue Materialien zu identifizieren, die eine Reihe an bestimmten, gewünschten Eigenschaften aufweisen. Ein sehr interessantes Anwendungsgebiet bildet dabei die Adsorption von (kleinen) organischen Molekülen auf Metalloberflächen, da die dabei gebildeten Schichten verwendet werden können, um die Eigenschaften der jeweiligen Oberflächen zu modifizieren. Auf diese Weise kann beispielsweise die Austrittsarbeit von Metallelektroden, welche in Verbindung mit organischen Halbleitern stehen, verändert werden.

Diese Effekte sind abhängig von der Art der molekularen Anordnung, welche im Allgemeinen als Oberflächenpolymorph bezeichnet wird. Im Verlauf der Jahre wurden einige Algorithmen für die Struktursuche konzipiert, mit welchen solche Polymorphe gefunden und untersucht werden können. Eine relativ neue Ergänzung zu dieser Liste ist die SAMPLE-Methode.

Sie führt einen effizienten Algorithmus zur Erzeugung von Oberflächenpolymorphen ein, welcher die Konfigurationsexplosion mithilfe eines intelligenten Coarse-Graining-Algorithmus in Kombination mit einem effektiven Zweikörper-Energiemodell verlangsamt. Gleichzeitig wird die Abhängigkeit von teuren Trainingsdaten reduziert, indem Bayes'sches Lernen in Form einer Gaußprozess-Regression anwendet wird, um Eigenschaften – insbesondere Adsorptionsenergien – vorherzusagen.

Da das Energiemodell im Kern von SAMPLE nicht über Zweikörper-Wechselwirkungen hinausgeht, stellte sich die Frage, ob diese Methode in der Lage ist, auch komplexe molekulare Wechselwirkungen zu beschreiben. Daher liegt das Ziel der vorliegenen Arbeit darin, zu testen, wie gut SAMPLE lernt und wie gut es Eigenschaften von Testsystemen, die viel komplizierter sind, als jene, die bisher untersucht wurden, vorhersagen kann. Der Fokus liegt dabei auf paarweisen Molekülwechselwirkungen.

Es werden dazu Tests für vier Kategorien von Testsystemen durchgeführt, die sich jeweils mit einem bestimmten Aspekt von intermolekularen Wechselwirkungen befassen. Dazu zählen Molekülsymmetrien, anisotrope Bindungen, unterschiedliche

Substitutionsmuster sowie durch die Verwendung verschiedener funktioneller Gruppen als Substituenten entstehende Effekte. Die Lerncharakteristiken für jedes Testsystem werden präsentiert, analysiert und mit denen der anderen Systeme verglichen, um zu beurteilen, ob SAMPLE als robuster Struktursuchealgorithmus angesehen werden kann. Zudem wird das Prinzip des Transferlernens mit SAMPLE in Bezug auf Paarpotentiale demonstriert. Zusammenfassend kann festgehalten werden, dass SAMPLE ein robuster und verlässlicher Algorithmus zur Vorhersage von Strukturen ist, der ein großes Potenzial für die Verwendung und Weiterentwicklung aufweist.

Contents

Abstract vii

Kurzfassung viii

1. Introduction 1

1.1. Motivation 1

1.2. SAMPLE 3

1.2.1. Selecting local geometries 3

1.2.2. Generating configurations 4

1.2.3. Energy model 5

1.2.4. Identifying similar interactions: introducing the Feature Vector 7

1.2.5. Learning interactions 8

1.3. Determining reference energies 10

2. Methodology 11

2.1. Requirements for structure search algorithms 11

2.2. Quantifying learning performance 12

2.2.1. Prediction uncertainty 12

2.2.2. Evaluation workflow 13

2.2.3. Analysis tools and performance markers 15

2.2.4. Interpreting learning curves 18

2.3. Setting model hyperparameters 21

2.3.1. Defining minimal distance thresholds 22

2.4. Designing test systems 24

2.4.1. Virtual lattice 24

2.4.2. Test molecules 25

2.4.3. Local geometries 25

2.5. DFT calculation settings 26

3. Results and Discussion 27

3.1. Influence of test set selection 27

3.2. Evolution and convergence of pair potentials 30

3.3. Level 1: Impact of molecular symmetries 32

3.3.1. Test molecules 32

3.3.2. Results 33

3.4. Level 2: Adding Hydrogen bonds 37
3.4.1. Test systems and objective 37
3.4.2. Results 38
3.5. Level 3: Influence of substituent patterns 42
3.5.1. Objective and test systems 42
3.5.2. Results and findings 42
3.6. Level 4: Functional groups 46
3.6.1. Objectives 46
3.6.2. Type of functional group 46
3.6.3. Combining multiple functional groups 51
3.7. A look at transfer learning with SAMPLE 54

4. Summary and Conclusion 57

A. Determination of hyperparameters 61
A.1. Calculated minimal distance thresholds 61
A.2. Minimal distance threshold sweeps 62

B. Listings 65
B.1. FHIaims control file for periodic calculations 65
B.2. FHIaims control file for non-periodic calculations 66

List of Figures

1.1. Building configurations of two molecules in a 4x4 supercell 4
1.2. Illustration of the feature vector 8

2.1. Evaluation workflow 13
2.2. Algorithm for computing a learning curve 15
2.3. Illustration of a learning curve and related quantities 16
2.4. Motivation for fit function 20
2.5. How to select minimal distance thresholds 24
2.6. Definition of a test system as a molecular monolayer 25

3.1. Influence of test set on achieved prediction accuracy 29
3.2. Conparison of different test set types for different test systems 29
3.3. Pair potential evolution 31
3.4. Level 1 test molecules 32
3.5. Learning curves for level 1 35
3.6. Barplot of RMSE @ set size 700 for level 1 systems 36
3.7. RMSE at training set size 700 versus number of 2-body fit parameters for level 1 systems 36
3.8. Comparison of electrostatic potentials of test molecules with and without hydrogens 38
3.9. Level 2 test molecules 39
3.10. Level 2 learning curves 39
3.11. RMSE vs number of fit coefficients for level 2 systems 41
3.12. RMSE vs energy spread for level 2 systems 41
3.13. Level 3 test molecules 43
3.14. Comparison of RMSE @ set size 700 for all level 3 systems 44
3.15. Learning curves for NO_2 subset 44
3.16. Learning curves for test molecule *NO_2-ortho* in dependence of feature dimension 45
3.17. Level 4 single group test molecules 47
3.18. Learning curves for single group systems 49
3.19. Range of E_{MLF} of test set configurations for all single-group systems . 49
3.20. Level 4: multi-group test systems 51
3.21. Learning curves for multi group systems 52
3.22. Nitrogeon systems 53

3.23. Test systems for transfer learning demonstration 54
3.24. Learning curves of transfer learning systems 55
3.25. Comparisons between composite and direct NH+OH pair potential . 56

A.1. Distance threshold sweep curves, page 1 62
A.2. Distance threshold sweep curves, page 2 63

1. Introduction

1.1. Motivation

The SAMPLE method is a new type of structure search algorithm that makes use of coarse graining to generate surface polymorphs. It employs a two-body energy model in conjunction with Gaussian process regression [**Rasmussen2006a**] to efficiently predict polymorph properties, with a primary focus on adsorption energies. Recently, it was successfully demonstrated that SAMPLE can be used to predict surface polymorphs and interaction energies for a small number of adsorbate molecules [**Jeindl:masters-thesis**, **Egger:masters-thesis**]. While this shows that the method works in principle, these findings, on their own, are not enough to make the assumption that SAMPLE can achieve similar results under all circumstances. In other words, the question that still needs to be answered is

How flexible is SAMPLE's approach to structure search with regard to arbitrarily complex systems?

In order to answer this question, we need to, among other things, evaluate the ability of SAMPLE's energy model to emulate complex molecular interactions. This evaluation is necessary, particularly due to the fact that the model only includes one two-body term for each pair of molecules.

The aim of this thesis shall therefore be to navigate the limits of SAMPLE in terms of its learning performance. For this purpose, the SAMPLE method is applied to four different classes of test systems, each selected to cover specific aspects of molecular interactions that might pose problems to the way SAMPLE is set up and provides its predictions. The focus of these test classes or *levels* can be summed up as:

1. The impact of molecular symmetries
2. Effects of anisotropic interactions (example: hydrogen bonds)
3. Influence of substituent patterns
4. Functional groups

Based on these benchmarks, it should be possible to discern whether SAMPLE can be seen as a robust structure search algorithm. In order to give the reader a guideline, I will begin by introducing the SAMPLE method as well as connected processes and concepts in the following two subsections.

1.2. SAMPLE

The primary goal of this thesis is to benchmark the performance of the SAMPLE method, which was first explored in the master's theses of Scherbela [**Scherbela:masters-thesis**], Hörmann [**Hoermann:masters-thesis**] and Jeindl [**Jeindl:masters-thesis**] before being properly introduced in a paper by Hörmann et al. [**Hoermann:sample-paper**]. This section is going to describe the key concepts and inner workings of said method and will therefore follow the approaches taken by Hörmann et al. and Jeindl.

First and foremost, **SAMPLE**, which cleverly stands for *Surface Adsorbate Polymorph Prediction with Little Efforts (SAMPLEs)* [**Hoermann:sample-paper**], is a quasi-deterministic algorithm for structure search on surfaces. It makes use of a combination of coarse-graining and Bayesian linear regression to push back the exponential increase in the number of potential polymorphs, also know as the *configurational explosion*. Simultaneously, due to its quasi-deterministic nature, it can give physical insight into molecule-substrate and molecule-molecule interactions.

It was originally designed for the task of finding and describing single-layer periodic arrangements or *surface polymorphs* of (organic) adsorbate molecules on metal substrates, which, in the framework of SAMPLE, are often denoted as *configurations*.

Central to its approach is the assumption of commensurability between surface polymorph and the substrate it sits on, or, in other words: that the polymorph unit cell is a supercell of the substrate (surface) lattice. This assumption implies that the molecule-substrate interactions are dominant compared to the molecule-molecule interactions.

Overall, applying SAMPLE is a multi-step process, whose structure is described in the following subsections.

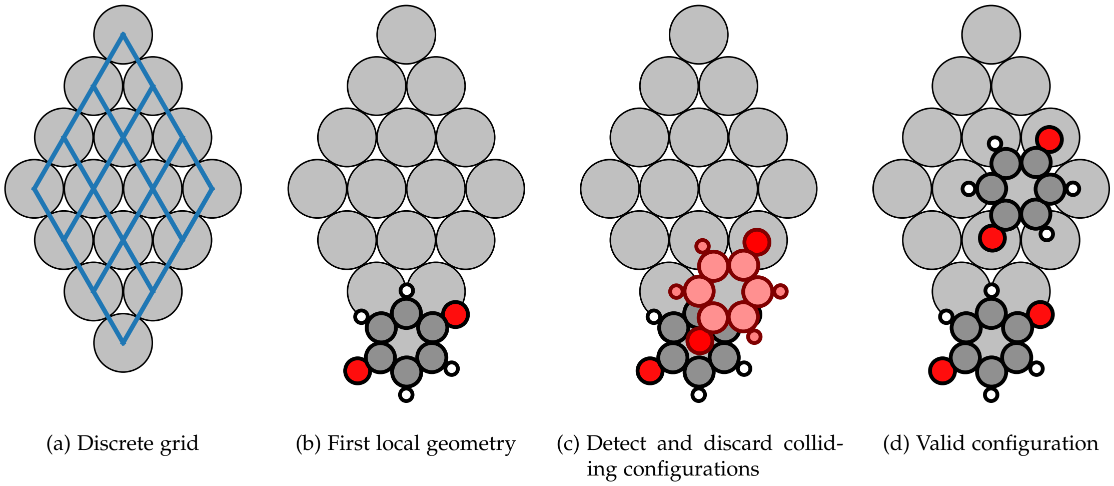
1.2.1. Selecting local geometries

The starting point for each SAMPLE run consists of selecting the *symmetry-inequivalent local geometries (SILGs)*[1]. In principle, they represent local minima in the *potential energy surface (PES)* of an isolated molecule adsorbing to the surface of the substrate and can be found through various types of geometry optimizations.

Each SILG is defined by the specific atomic structure of the molecule plus its orientation and position relative to the primitive substrate unit cell.[2] By applying

[1] called an *original geometry* in SAMPLE's notation

[2] See section 2.4 for details on actual test systems


(a) Discrete grid

(b) First local geometry

(c) Detect and discard colliding configurations

(d) Valid configuration

Figure 1.1.: Building configurations of two molecules in a 4x4 supercell. Starting from the plain substrate supercell and the discrete grid (blue) it defines (a), a single molecule/LG is placed in the origin of said grid (b). Then, a second molecule/LG is added, moved to an unoccupied grid point and the resulting configuration is checked for colliding atoms. If collisions are found, the corresponding configuration is discarded (c). If not, it is added to the list of previously found configurations. This is repeated for all remaining grid points. Adapted from [**Jeindl:masters-thesis**].

point group symmetries of the substrate lattice to the SILGs, the *symmetry-equivalent local geometries*, most often simply called *local geometries (LGs)*, are generated.

They are subsequently used by SAMPLE as immutable building blocks from which it constructs surface polymorphs by means of arranging LGs on a two-dimensional, discrete grid defined by the surface lattice, as illustrated in a. This designation implicitly asserts that, even when molecules are closely packed, the molecule-molecule interaction is not strong enough to alter the geometry of individual molecules or push them out of their respective local minimum positions. This, in turn, implies that the PES needs to be sufficiently corrugated so that local energy minima are pronounced/deep enough.

1.2.2. Generating configurations

The process of generating configurations starts by constructing a list of all substrate or surface *supercells (SCs)* whose cell areas, expressed in multiples of the primitive unit cell area, fall into a specified range. Symmetries of the substrate and certain constraints with regard to the unit cell shape are then applied to discard superfluous cells and thus reduce the total number of SCs.

The next stage of the algorithm is best described as a systematic attempt to fill each supercell with different combinations and numbers of LGs.

It entails a layered, iterative process, that is executed for each supercell separately and is illustrated in figure 1.1. The first layer consists of creating configurations where a single molecule/LG sits at the origin of the cell. Next, it is attempted to add a second molecule/LG to each of the one-molecule configurations. This is done by systematically placing the second LG at some unoccupied grid point, followed by checking the resulting structure for atom collisions.

For this, the distances d_{ij} between atoms i of the first molecule and atoms j of the second molecule are compared against pre-selected *minimal distance thresholds* d_{min}^{AB}, which are defined for each atom species pair (A, B) separately[3].

Configuration candidates that fail the collision check, i.e. that have any $d_{ij} < d_{min}^{AB}$, are discarded, the others are kept. This procedure is repeated with different LGs until all possible two-molecule configurations are found. In a last step, the list of collected configurations is scanned for symmetry-equivalent duplicates, which are eventually filtered out.[4]

Going to configurations with three molecules per cell involves the same steps as above, although this time the starting points are the two-molecule configurations. In effect, this iteration allows, at least in theory, to construct configurations with an arbitrary number of molecules. In practice, due to the exponential growth of the number of configurations, the number of molecules per cell will usually hit an upper limit rather quickly.

1.2.3. Energy model

After completing the generation of configurations, the next task is finding the properties of said configurations/polymorphs. Particular interest is awarded to the polymorphs' energies, as they are needed to determine which polymorph is most likely to form under certain conditions (pressure, temperature, etc.).

As shown by Bernstein [**Bernstein2011**] and Nyman [**Nyman2015**], energy differences between two polymorphs can be as low as 20 meV. Being able to calculate or estimate these polymorph energies with high accuracy is therefore imperative for reliable structure search.

[3] While the term *atom species* is usually equivalent to the chemical element of an atom, it can also be used to distinguish atoms of the same element that are in different chemical environments that influence their interactions.

[4] The methods that are used to detect these duplicates were developed by Lukas Hörmann in his master's thesis, see [**Hoermann:masters-thesis**].

Unfortunately, as of now - and for the foreseeable future - quantum chemical calculations that can deliver this level of accuracy are far too expensive to be run for all of the millions of configurations that SAMPLE typically generates.

Instead, SAMPLE employs a physically motivated *energy model* to estimate energies of configurations. The model is based on a Taylor expansion of the configuration energy in terms of many-body interactions. Since 3-body and higher order terms are expected to play only a small role in the formation of surface polymorphs, the series is broken off after the two-body term. Hence, the energy of a configuration c is described by

$$E[c] = \sum_{LGs\ g} n_g[c] U_g + \sum_{pairs\ p} n_p[c] V_p \,. \tag{1.1}$$

The first sum represents the one-body (i.e. molecule-substrate) interactions and runs over all local geometries g. We can identify U_g with the energy of the interaction between a single molecule in LG g and the substrate. The factor $n_g[c]$ simply denotes how often each LG appears in c, or more accurately, in its unit cell. The second sum runs over all possible pairs of molecules, with V_p being the two-body energy and $n_p[c]$ denoting the number of occurrences of pair p.

Both U_g and V_p represent *model parameters* which are not known a-priori. Thus, they need to be determined by fitting equation (1.1) to measured/calculated configuration energies.

For notational ease, the model parameters are collected in the *interaction vector* $\boldsymbol{\omega} = (U_1, U_2, \ldots, V_1, V_2, \ldots)^T$, whereas the corresponding number of occurrences, n_g and n_p, form the *model vector* $\mathbf{n}$. As a result, equation (1.1) is simplified to

$$E[c] = \mathbf{n} \cdot \boldsymbol{\omega} \,. \tag{1.2}$$

For a set S of configurations, the model vectors can be combined to form the *model matrix* $\boldsymbol{X}$, in which each row represents a single configuration. Equation 1.2 then becomes

$$\boldsymbol{E}[S] = \mathbf{X} \cdot \boldsymbol{\omega} \,. \tag{1.3}$$

In order to better compare configurations with unequal numbers of molecules N_{mol} in their unit cells, it is useful to define configuration energies on a per-molecule basis. Unless stated otherwise, all configuration energies featured in this study conform to this convention.

Even though higher order terms are neglected, equation 1.1, at this point, still contains an infinite number of two-body parameters V_p due to the periodic nature of the described polymorph. The path to making this problem tractable, as explained in the thesis of Jeindl [**Jeindl:masters-thesis**], is footed on two pillars:

Firstly, it can be assumed that at some large enough intermolecular distance the strength of any two-molecule interaction becomes small and eventually falls off to zero. We can therefore define an upper distance limit and omit all pair interactions where the corresponding intermolecular distance is larger. This brings down the number of model parameters V_p from infinity to a manageable, finite number. The distance limit is enforced on an atom-pair basis via suitably chosen *maximum distance thresholds* d^{AB}_{max}, similar to the collision check mentioned before.

The second pillar consists of the assumption that interaction energies V_p of two molecule pairs in similar arrangements will be correlated. This effectively reduces the number of free/uncorrelated model parameters. In order to apply this assumption, we need to define a measure for the similarity of two molecule pairs.

1.2.4. Identifying similar interactions: introducing the Feature Vector

This line of thought brings us to the *feature vector (FV)*, written as $\mathbf{f}$. It represents the arrangement of two molecules as a vector of (scaled) interatomic distances. Using feature vectors, the similarity of two pairs of molecules, p_1 and p_2, can be quantified via their distance in feature space, which is given by

$$d_f(p_1, p_2) = \|\mathbf{f}_1 - \mathbf{f}_2\|_1\,, \tag{1.4}$$

where $\mathbf{f}_1$ and $\mathbf{f}_2$ are the corresponding feature vectors. Relatedly, two pair arrangements are treated as the same *feature*, and thus are mapped to the same interaction energy V_p, if the element-wise differences of their feature vectors are below a (user-defined) *feature threshold* Δf.

Calculating $\mathbf{f}$ for a pair of molecules, g_1 and g_2, starts by determining atom pair distances $d_{ij} = |\boldsymbol{r}_i - \mathbf{r}_j|$, where atom i is part of g_1 and atom j is part of g_2. These distances are then grouped by involved atom species (A, B) and sorted in ascending order. Within each group, the N_{AB} shortest distances are kept while the rest is discarded, as highlighted in figure 1.2. The remaining entries are then normalized by the corresponding minimal distance threshold, d^{AB}_{min} and subsequently exponentiated by n, the (negative) *decay power*. This last operation has the effect that differences in pair arrangement cause a greater separation in feature space when molecules are closer. In the end, each entry in the FV has the form

$$f_\alpha(g1, g2) = \left(\frac{\left\| \mathbf{r}^A_i - \mathbf{r}^B_j \right\|_1}{d^{AB}_{min}} \right)^n \,. \tag{1.5}$$

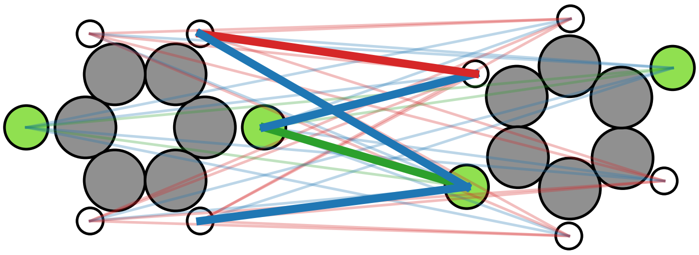


Figure 1.2.: Illustration of the feature vector of a pair of 1,4-diflourobenzenes. The thin lines mark all intermolecular connections between peripheral atoms, in this case, F and H atoms. Drawn in bold are the shortest atom-atom connections which subsequently appear in the feature vector. Graphic based on [**Jeindl:masters-thesis**].

Per convention, the entries in the FV are grouped by atom species pairs, which gives it a more intuitive structure, as is shown in equation 1.6.

$$f(g_1, g_2) = \begin{pmatrix} \left(\frac{\left|\boldsymbol{r}_i^F - \boldsymbol{r}_j^F\right|}{d_{min}^{FF}}\right)^n \\ \vdots \\ \left(\frac{\left|\boldsymbol{r}_i^F - \boldsymbol{r}_j^H\right|}{d_{min}^{FH}}\right)^n \\ \vdots \\ \left(\frac{\left|\boldsymbol{r}_i^H - \boldsymbol{r}_j^H\right|}{d_{min}^{HH}}\right)^n \\ \vdots \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} N_{FF} \text{ entries} \\ \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} N_{FH} \text{ entries} \\ \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} N_{HH} \text{ entries} \end{matrix} \tag{1.6}$$

1.2.5. Learning interactions

Now that the definition of feature vectors is complete, the next step is to train the previously introduced energy model on existing data with the help of Bayesian learning [**Todorovic2019**] and *Gaussian process regression (GPR)* [**Rasmussen2006a**]. The core premise of the former lies in Bayes' theorem, which, in the case of SAMPLE, reads [**Hoermann:sample-paper**]

$$p\left(\boldsymbol{\omega} \mid \boldsymbol{E}_{DFT}\right) = \frac{p\left(\boldsymbol{E}_{DFT} \mid \boldsymbol{\omega}\right) p(\boldsymbol{\omega})}{p\left(\boldsymbol{E}_{DFT}\right)} \tag{1.7}$$

It states that the unknown posterior probability $p\left(\boldsymbol{\omega} \mid \boldsymbol{E}_{DFT}\right)$ of having one- and two-body interactions $\boldsymbol{\omega}$ when having measured energies $\boldsymbol{E}_{DFT}$, can be written as a product of two known probability distributions, namely the *likelihood* $p\left(\boldsymbol{E}_{DFT}|\boldsymbol{\omega}\right)$, which denotes the probability of measuring $\boldsymbol{E}_{DFT}$ given specific model parameters $\boldsymbol{\omega}$, and the *prior* $p\left(\boldsymbol{\omega}\right)$, which describes the a-priori probability of finding specific model parameters $\boldsymbol{\omega}$. The *marginal probability* $p\left(\boldsymbol{E}_{DFT}\right)$ serves as a normalization of the posterior probability, but is assumed to be constant and can therefore be neglected. The functional form of the likelihood,

$$p\left(\boldsymbol{E}_{DFT}|\boldsymbol{\omega}\right) \propto \exp\left(-\frac{1}{2\sigma_{model}^2}\left\|\boldsymbol{E}_{DFT}-\boldsymbol{X}\boldsymbol{\omega}\right\|^2\right), \tag{1.8}$$

is determined by the energy model and the DFT reference energies $\boldsymbol{E}_{DFT}$. Here, the term σ_{model} denotes the *model uncertainty*, which may be seen as the uncertainty of the DFT calculations. The prior can be used to insert physical knowledge about the system into the model, thereby rendering this in principle under-determined problem solvable. Hörmann et al. proposed using a normal distribution with mean $\boldsymbol{\omega}_0$ and covariance matrix $\boldsymbol{C}$,

$$p(\boldsymbol{\omega}) \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\omega}-\boldsymbol{\omega}_0\right)^T \mathbf{C}^{-1}\left(\boldsymbol{\omega}-\boldsymbol{\omega}_0\right)\right), \tag{1.9}$$

as prior, where the covariances C_{ij} are assumed to depend on the distances in feature space via

$$C_{ij} = \sigma_i^*\sigma_j^* \exp\left(-\frac{\left\|\boldsymbol{f}_i-\boldsymbol{f}_j\right\|_1}{\xi}\right). \tag{1.10}$$

The factors σ_i^* and σ_j^* each describe the decay of the corresponding two-body interaction at large intermolecular distances, while ξ defines a characteristic length scale in feature space and is known as the *feature correlation length*.[5]

When both prior and likelihood are Gaussians, equation 1.7 tells us that the posterior probability also takes the form of a Gaussian, namely

$$p\left(\boldsymbol{\omega} \mid \boldsymbol{E}_{DFT}\right) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\omega}-\bar{\boldsymbol{\omega}})^T\boldsymbol{C}_{post}^{-1}(\boldsymbol{\omega}-\bar{\boldsymbol{\omega}})\right). \tag{1.11}$$

In this equation, the posterior covariance $\boldsymbol{C}_{\text{post}}^{-1}$ is equal to

$$\boldsymbol{C}_{\text{post}}^{-1} = \frac{\boldsymbol{X}^\top\boldsymbol{X}}{\sigma_{\text{model}}^2} + \boldsymbol{C}^{-1} \tag{1.12}$$

[5] For a complete list of all hyperparameters of SAMPLE, see table 2.1

and the posterior mean $\bar{\boldsymbol{\omega}}$ is given by

$$\bar{\boldsymbol{\omega}} = \boldsymbol{C}_{\text{post}} \left(\frac{\boldsymbol{X}^T \boldsymbol{E}_{DFT}}{\sigma^2_{\text{model}}} + \boldsymbol{C}^{-1} \boldsymbol{\omega}_0 \right) , \tag{1.13}$$

where σ_{model} is the *model uncertainty*. Training the model and learning the interactions thus means calculating $\bar{\boldsymbol{\omega}}$ and using the expectation values for the one- and two-body interactions that are contained in $\bar{\boldsymbol{\omega}}$ as estimators for the real interaction energies.

1.3. Determining reference energies

Training SAMPLE's energy model (see 1.2.3) and evaluating the prediction uncertainty (see section 2.2) requires knowledge of *reference energies E_{ref}* for certain configurations. Under normal circumstances, these energies will be supplied by some type of electronic structure method.

In the present thesis, it was chosen to employ *density functional theory (DFT)*, which, over the last decade(s), has become the de-facto standard method for electronic structure calculations, especially in solid state physics and the material modelling world [**Maurer2019**].

From the start, SAMPLE was conceived to work in tandem with DFT, therefore, DFT will naturally also be used in this study. In addition to that, DFT can be easily applied to both periodic and non-periodic systems, as is required in order to calculate the corresponding *monolayer formation energy (E_{MLF})* via equation 2.2.

While there are many quantum chemistry packages that facilitate DFT calculations, such as Quantum ESPRESSO [**QE-2009**, **QE-2017**], VASP [**Kresse1993**, **Kresse1996**], ORCA [**Neese:ORCA-paper**] or GAUSSIAN [**gaussian16**], to name a few, we chose to use FHIaims [**Blum:aims-paper**], short for *Fritz Haber Institute ab initio molecular simulations*.

FHIaims is an all-electron code that uses atom-centered orbitals as basis functions, whose radial part is defined numerically. With this freedom in the choice of the radial functions we can construct basis functions that take the $1/r$-potential near the nuclei into account and whose degree of localization can be controlled freely in order to reduce the number of costly overlap integrals (cf. [**Blum:aims-paper**]).

For more details and information on the specific calculation settings used in this thesis, the reader is referred to section 2.5.

2. Methodology

2.1. Requirements for structure search algorithms

In order to be able to rate the capabilities of a structure search algorithm, we first need to establish a set of criteria that a "good" algorithm has to meet. While there a certainly many ways to define these, in this thesis we will pose the following three requirements:

1. High prediction accuracy
2. Reasonable computational cost
3. Robustness

Ad) High prediction accuracy First and foremost, the algorithm should be able to generate structures and select those which are best with respect to a desired property or properties. In order to correctly identify the best configurations, the algorithm must be capable of calculating or predicting the value of these properties with high confidence.

Ad) Computational cost In an ideal world, where computational resources are infinite and cost-free, the quality of a structure search algorithm would only be gauged based on whether it is able to find the best possible structure or not. In the real world, however, an algorithm also needs to be efficient, affordable and (optimally) fast to be usable.

Ad) Robustness In order to become a reliable tool in practice, the performance of a structure search algorithm, as determined by the first two requirements, should not vary wildly between different applications. If performance does vary, it should at least be in a predictable way, so that it is possible to estimate the costs of running the algorithm in advance and weigh them against the potential reward in terms of gained insight and knowledge.

In the scope of this thesis, an algorithm (i.e. SAMPLE) will be considered **robust** if for all tested systems we can establish that:

a) Configuration energies can be predicted with an uncertainty of less than 25 meV per molecule.

b) The computational effort needed to reach this level of prediction accuracy can be estimated and overall stays reasonable. In case of SAMPLE, where the compute cost is mostly determined by the number of DFT calculations needed to provide training data, it is decided to set the limit for what is seen as reasonable to 500 calculations.

2.2. Quantifying learning performance

To assess the learning performance of SAMPLE for different systems in accordance with section 2.1, we need to define the "accuracy" of predictions in a quantifiable way. This includes devising a process which assures that this quantity is measured or, more precisely, calculated consistently in all cases. The following subsections 2.2.1 and 2.2.2 are going to discuss these two aspects, while subsection 2.2.3 introduces some additional terms that are helpful in the analysis of learning performances.

2.2.1. Prediction uncertainty

This quantity should represent SAMPLE's ability to predict configuration energies (more concretely: monolayer formation energies $\boldsymbol{E}_{MLF}$) for all possible configurations. Unluckily, due to practical restrictions, i.e. limited compute resources, we are usually going to be unable to calculate the difference between predicted energy and actual (DFT-) energy for all configurations. Therefore, we analyze the configuration-wise prediction errors $\Delta E_{MLF} = E_{prediction} - E_{DFT}$ on a smaller subset of configurations instead, which we will call the *test set* S_{test}. For this subset we can then define the *prediction uncertainty* γ as the *root mean square error (RMSE)* of the ΔE_{MLF} via

$$\gamma[S_{test}] = \mathrm{RMSE}[S_{test}] = \sqrt{\frac{1}{N} \sum_{\text{config c} \in S_{test}} \left[E_{\text{prediction}}(c) - E_{DFT}(c)\right]^2} \,. \tag{2.1}$$

Assuming that S_{test} is a representative sample of all configurations, its γ value can then be used as a measure for the general prediction uncertainty. As such, it becomes an inverse quantifier for the prediction accuracy of SAMPLE.

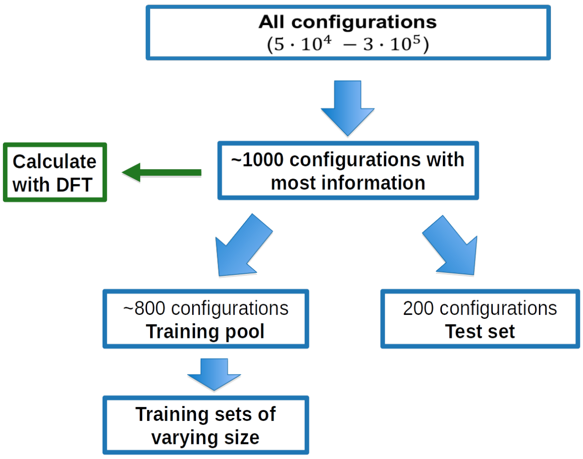

All configurations
$(5 \cdot 10^4 - 3 \cdot 10^5)$
~1000 configurations with most information
Calculate with DFT
~800 configurations Training pool
200 configurations Test set
Training sets of varying size

Figure 2.1.: Graphical representation of steps between generation of configurations and selection of training sets.

2.2.2. Evaluation workflow

Evaluating the prediction uncertainty (see section 2.2.1) follows the same scheme for all test systems. In every case, the starting point is a fully configured SAMPLE setup[1], which is then used to generate a suitably large set of configurations with varying coverages θ. In this thesis, the coverage is varied between 6 and 16 primitive lattice unit cells - or equivalently, substrate atoms - per test molecule. Based on the lattice constant of the selected virtual hexagonal lattice (which is introduced in section 2.4), this corresponds to values of 33.858 Å^2/molecule to 90.289 Å^2/molecule.

Another factor that decides the range of possible configurations is how many test molecules we allow per supercell, since, as explained in section 1.2, adding an extra molecule increases the total number of configurations exponentially. When performing global structure search, such as looking for the structure with lowest energy, we generally want to make the search space as large as possible or feasible.

However, as this thesis just focuses on intermolecular (two-body) interactions, we only require that the generated configurations feature all possible relative orientations of two molecules, given of course certain restrictions on supercell size and a discretization with regard to molecular rotations. With that in mind, there is actually no benefit in generating configurations with more than two molecules per supercell as those configurations do not feature any two-body arrangements that are not found in configurations with one or two molecules per supercell.

For settings as above and test molecules as described in section 2.4, SAMPLE's

[1] internally referred to as *project*; see section 2.4 for details.

generation process[2] yields approximately 5×10^4 to 3×10^5 configurations for each test system. All configurations for a system are aggregated in one configuration set, S_{all}. The actual number of configurations depends on the size and, more importantly, the symmetries of each test molecule.

The subsequent step after the generation process is illustrated in figure 2.1 and consists of selecting approximately 1000 configurations from S_{all} for which single point DFT calculations in FHIaims are executed. These configurations, collectively known as the set S_{DFT}, and the associated calculations are the basis for all subsequent evaluations, their selection should therefore be well motivated.

Firstly, S_{DFT} should represent a diverse sampling of S_{all}. Secondly, in order to give comparisons between different systems a solid footing, the choice of S_{DFT} should be deterministic to avoid the effect of randomness.

Both criteria are met by building S_{DFT} according to the principle of D-optimal design of experiments [**Fedorov1972**]. Doing so makes sure that, as a collective, the selected configurations in S_{DFT} contain the most diverse set of two-molecule arrangements possible.

The actual evaluation starts when the results for the DFT calculations are in. First, the E_{MLF} of each configuration c is determined from the corresponding DFT calculation via

$$E_{MLF}(c) = \frac{1}{N_{molecules}} E_{\text{config}} - E_{SM} \,, \tag{2.2}$$

where $N_{molecules}$ is the number of molecules per supercell, E_{config} is the total (single-point) energy of the configuration and E_{SM} is the energy of a single isolated molecule. At this point, it should be noted that E_{config} stems from a periodic and E_{SM} from a non-periodic FHIaims-calculation[3].

Next, the S_{DFT} set is split into a *test set* S_{test} and the *training pool* S_{pool}. The test set is used to evaluate the learning performance via equation 2.1. For consistency reasons, S_{test} always contains 200 configurations. The size of S_{pool} on the other hand varies depending on the size of S_{all}. As the name suggests, S_{pool} is the pool of configurations from which training configurations are drawn and assigned to *training sets* S_{train}, which are then used to train SAMPLE's energy model via GPR, as discussed in section 1.2.5.

[2] see chapter 1.2

[3] For more computation settings, see section 2.5.

2.2.3. Analysis tools and performance markers

Learning curve The core strategy for gaining insight into how well/badly certain systems can be learned with SAMPLE will be to look at how SAMPLE's prediction uncertainty γ evolves with an increase of the amount of training data inserted into its energy model. This evolution of γ as a function of the training set size x will be called the *learning curve* $\gamma(x)$. The algorithm that is used to compute all learning curves is depicted in figure 2.2.

It consists of a simple `for`-loop, that iterates over a list of targeted training set sizes x_n, which are usually evenly spread between some x_{start} and some larger $x_{end} < x_{pool}$, i.e $x_n = x_{start} + n\Delta x$, with Δx being a suitable increment and x_{pool} the total number of available training configurations in the training pool.

For each x_n, a training set $S_{train}^{x_n}$ of corresponding size is drawn D-optimally from S_{pool} and subsequently used to train the energy model. Based on this - now trained - model, we then predict E_{MLF} for all test set configurations. Comparing these predicted energies to the ones calculated by DFT via equation (2.1) then gives the prediction uncertainty for that training set size, $\gamma(x_n)$.

By repeating these steps for all x_n, we generate the learning curve $\gamma(x)$. It should be noted that it is important to always reset the model to an untrained state upon entering the next iteration to prevent training data from the previous step from influencing the learning performance.

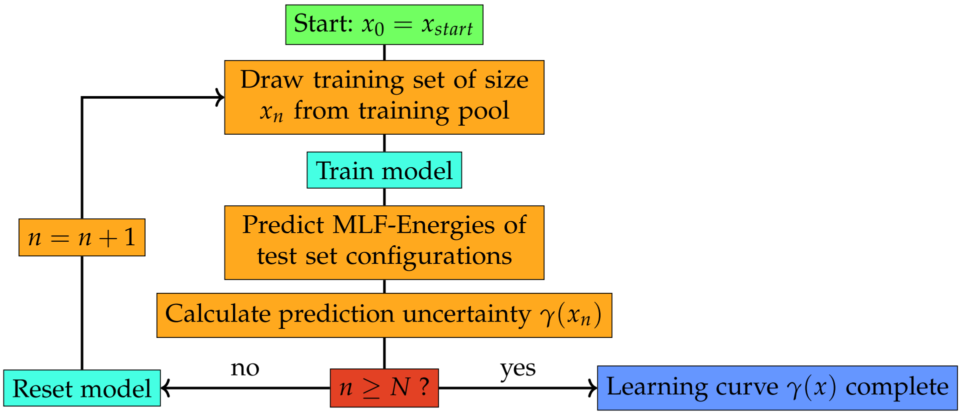

Start: $x_0 = x_{start}$
Draw training set of size x_n from training pool
Train model
Predict MLF-Energies of test set configurations
Calculate prediction uncertainty $\gamma(x_n)$
$n = n + 1$
Reset model
no
$n \geq N$?
yes
Learning curve $\gamma(x)$ complete

Figure 2.2.: Algorithm for computing a learning curve. Training set sizes x_n are usually chosen linearly, i.e $x_n = x_{start} + n\Delta x$, with Δx being a suitable increment. The maximum training set size is naturally given by the size of the training pool S_{pool}.

An example of what a learning curve usually looks like is featured in figure 2.3. The same figure also highlights tools and quantities that serve to analyze and interpret the learning curve.

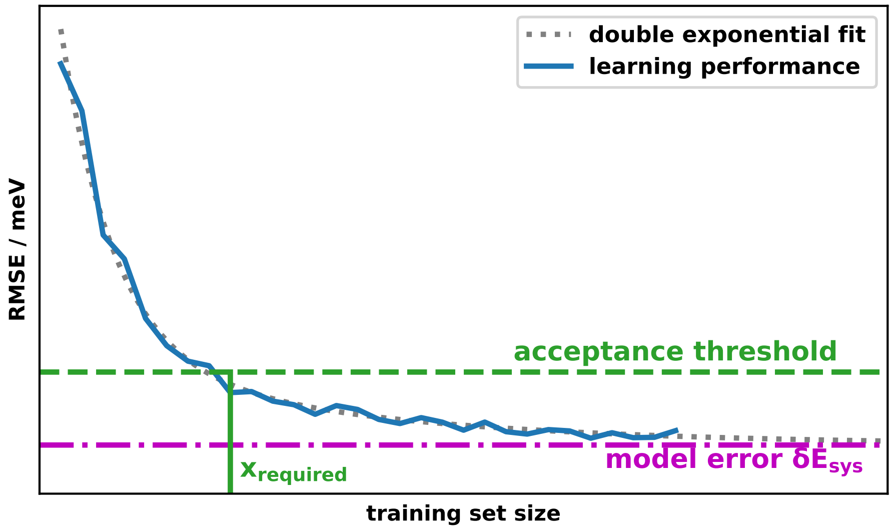


Figure 2.3.: Illustration of a learning curve and related quantities. Drawn in blue with a thick stroke is an exemplary learning curve. The dotted gray line in the back is a double exponential fit to said learning curve. Notice how for large training set sizes, the fit function approaches the model error, marked by a dash-dotted magenta line. Also depicted are the acceptance threshold (green, dashed line) and the required training set size ($x_{required}$, denoted by a vertical green line).

Acceptance threshold Defines the level of prediction uncertainty that a SAMPLE run needs to fall below of in order to be considered a success in terms of prediction accuracy. For this thesis, it is decided to set an acceptance threshold of 25 meV, which is close to 1 $k_B T$ for systems at room temperature. In the exemplary plot in figure 2.3, the acceptance threshold is depicted as a green dashed line.

Required training set size Denoted as $x_{required}$, it marks the (approximate) number of training configurations that is needed to achieve a RMSE that is equal to or lower than the acceptance threshold. In this thesis, we will identify $x_{required}$ with the smallest training set size x for which the measured $\gamma(x)$ is below the acceptance threshold.

Pair potentials Generally speaking, a pair potential is a way of visualizing the interaction between two (identical) molecules of a function of their (relative) positions and orientations. We can formally write the corresponding *pair interaction energy* Φ_{pair} as

$$\Phi_{pair} = \Phi_{pair}(\mathbf{r}_A, \mathbf{r}_B, \hat{\varphi}_A, \hat{\varphi}_B)\,, \tag{2.3}$$

where $\mathbf{r}_A$ and $\mathbf{r}_B$ are the positions of the (symmetry-)centers of A and B, whereas $\hat{\varphi}_A$ and $\hat{\varphi}_B$ describe the orientation of each molecule. Since this thesis focuses on free-standing monolayers, we will assume that both molecules are centered in the xy-plane, which we interpret as the plane of the monolayer.

Let $\tilde{\mathbf{x}}_i^{(A)}$ be the position of i-th atom of molecule A relative to the center of that molecule. We can then define a matrix $\tilde{X}_A$, whose rows represent the molecule-centered positions of all atoms in A. Subsequently, its absolute atomic coordinates $\mathbf{x}_i^{(A)}$ can be represented as a matrix

$$X_A = \mathbf{r}_A I + \tilde{X}_A \,, \tag{2.4}$$

with I being a 3x3 identity matrix. The same can be done for molecule B.

Next we express the position and orientation of B relative to that of A. For this, we let Λ_{AB} be the unitary transformation which moves molecule A into the orientation of molecule B by mapping the molecule-centered coordinates of A onto those of B:

$$\tilde{X}_B = \Lambda_{AB} \tilde{X}_A \,. \tag{2.5}$$

Together with $\mathbf{r}_{AB} = \mathbf{r}_B - \mathbf{r}_A$, the above leads to

$$X_B = \mathbf{r}_B I + \tilde{X}_B = \mathbf{r}_{AB} I + \Lambda_{AB} \tilde{X}_A + \mathbf{r}_A I \,. \tag{2.6}$$

Lastly, by fixing molecule A to the origin of our reference frame, we achieve $\mathbf{r}_A = 0$ and can define Φ_{pair} as

$$\Phi_{pair}^{AB} = \Phi_{pair}(\mathbf{r}_{AB}, \Lambda_{AB}) \,, \tag{2.7}$$

where $\mathbf{r}_{AB} = (x_{AB}, y_{AB}, 0)^T$ is effectively two-dimensional.

Mapping out this pair potential generally consists of the following steps:

1. put one molecule into a defined orientation and place it at the origin of a two-dimensional reference frame
2. orient molecule B relative to molecule A and place it at some position $\mathbf{r} = (x, y)$, so that the molecules do not collide and determine the energy of this arrangement
3. keeping its orientation fixed, systematically move molecule B around and record the energy as a function of $\mathbf{r}$
4. Repeat steps (1) and (2) for different orientations of molecule B (=different Λ_{AB})

The procedure above is the basis of how pair potential maps are constructed in the framework of SAMPLE. Based on the definition of SAMPLE's energy model in equation 1.1, it can be seen that the energy of a pair arrangement $(\mathbf{r}, \Lambda)$ is equivalent to one of the two-body energy parameters V_p and can therefore, in theory, be easily extracted from the trained model.

In practice, the number and type of pair arrangements for which Φ_{pair} can be determined as described above, are controlled by the number of local geometries and the underlying virtual lattice (as introduced in section 1.2). The possible values for ($\mathbf{r}$, Λ) also depend on the chosen minimum and maximum distance thresholds, d_{min} and d_{max} respectively, as discussed in section 2.3.

2.2.4. Interpreting learning curves

In order to estimate the systematic model error δE_{sys}, an attempt was made to quantify at which training set size the prediction accuracy flattens out and becomes stationary or even starts to worsen because of overfitting. At first it was tried to define such a potentially stationary regime, dubbed a *learning plateau*, as the interval where the slope of the learning curve stays below a certain threshold. As the prediction accuracy is only known at discrete set sizes x_n and because it does not decrease strictly monotonically, two alternative definitions for the learning plateau were explored. The first,

$$\left| \frac{\gamma(x_{n+1}) - \gamma(x_n)}{x_{n+1} - x_n} \right| = \left| \frac{\Delta\gamma}{\Delta x} \right| < \epsilon \,, \tag{2.8}$$

is based on the absolute slope of γ while the second,

$$\left| \frac{1}{\gamma(x_{n+1})} \frac{\gamma(x_{n+1}) - \gamma(x_n)}{x_{n+1} - x_n} \right| < \epsilon \,, \tag{2.9}$$

is based on the relative change of γ. Unfortunately, both definitions turned out to be not very reliable: the extent of the calculated learning plateaus was highly dependent on the value of ϵ, rendering the result somewhat arbitrary and susceptible to noise in the data.

To better control the influence of apparently noisy data, the thresholding approaches via equations 2.8 and 2.9 were discarded in favor of fitting the learning curves with a suitable model. With the assumption that the prediction accuracy is bounded by the systematic error δE_{sys}, the general form of this model function is

$$\gamma^{fit}(x) = \delta E_{sys} + f(x) \,, \tag{2.10}$$

where $f(x)$ is monotonically decreasing and is restricted to $[0, \infty)$. The actual choice of $f(x)$ was motivated by a visual analysis of several learning curves. As demonstrated exemplary in figure 2.4, it was found that if a suitable estimate for the model error δE_{sys} is subtracted from the learning curve, the resulting curve appears to be made up of two approximately linear sections when plotted semilogarithmically. This behavior is similar to that of a linear combination of two exponential functions e^{-x/λ_1} and e^{-x/λ_2} given that λ_1 is significantly larger than λ_2. Due to this similarity, we define f as

$$f(x) = K_1 e^{-\frac{x}{\lambda_1}} + K_2 e^{-\frac{x}{\lambda_2}} \, . \tag{2.11}$$

Combined with (2.10), the full learning curve fit model then reads:

$$\gamma^{fit} = \delta E_{sys} + K_1 e^{-\frac{x}{\lambda_1}} + K_2 e^{-\frac{x}{\lambda_2}} \, , \tag{2.12}$$

with $\lambda_1 > 0$, $\lambda_2 > 0$, $K_1 > 0$ and $K_2 > 0$. To avoid ambiguity between the λ- and K-parameters, we further require that $K_1 > K_2$ and $\lambda_1 < \lambda_2$. By doing so, we effectively link K_1 and λ_1 to the fast decrease of prediction errors at the start of each learning curve, while assigning K_2 and λ_2 to the slowly decaying part for large training set sizes.

The task of finding the best parameters of (2.12) for a specific learning curve is solved via non-linear least-squares optimization, whose basic premise is minimizing a weighted sum of residuals $r_i = y^{fit} - y^{data}$:

$$\chi^2 = \sum_i^N w_i r_i^2 = \sum_i^N w_i \left(\gamma^{fit} - \gamma^{data}\right)^2 \, . \tag{2.13}$$

The weights w_i control the impact of individual residuals r_i. In the most generic case, uniform weights ($\forall w_i = 1$) are used and all data points are treated equally. For a more sophisticated approach, the weights can be set to $w_i = 1/\sigma_i^2$ to reflect the uncertainty of each data point. For the case at hand, we unfortunately do not know the variances σ_i, but we are going to assume that they are proportional to the data value:

$$\sigma_i \propto \gamma_i \, . \tag{2.14}$$

By doing so, we effectively assign more weight to the tail of each learning curve. This is intentional as these data points are assumed to contribute more information with regard to the model error δE_{sys} than data points for small training set sizes. This choice of weights also corrects the large mismatch in magnitudes between data points from the beginning and those from the end of the learning curve.

Last but not least, it should be noted that different weighting schemes were used to fit artificial learning curves[4], where it was observed that the the data-dependent approach mentioned above yielded the most consistent fits.

[4] data was based on equation 2.12 plus random noise

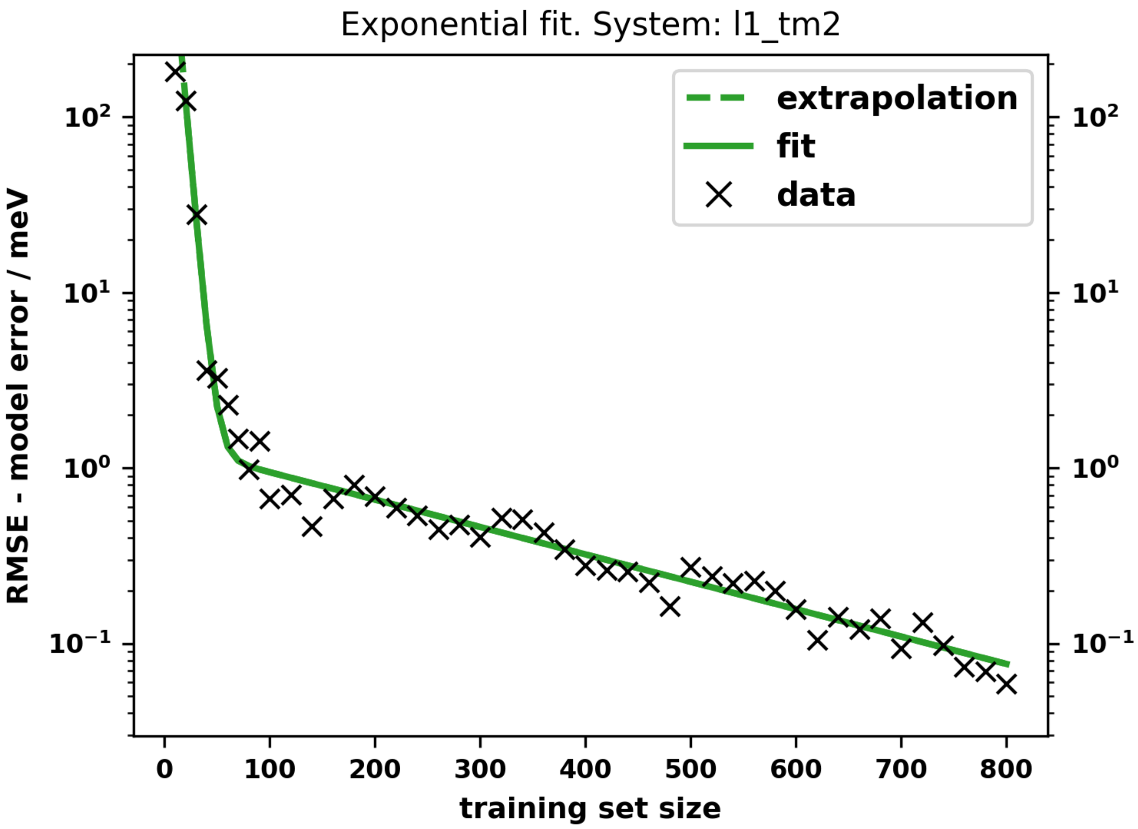


Figure 2.4.: Visual analysis of a learning curve for test system l1-tm2 (C3h). Drawn with black 'x's is the curve that results from subtracting an estimate for the model error δE_{sys} from the measured learning curve. In this semilogarithmic depiction, the curve appears to consist of two (approximately) linear sections with significantly differing slopes, with the transition around a training set size of ≈ 80. This observation motivates choosing $f(x)$ in equation 2.10 as the sum of two exponential functions. The corresponding (optimized) fit function $f^{fit}(x)$ is drawn as a solid green curve, with green dashes representing the extrapolation to points beyond the fitted data range.

2.3. Setting model hyperparameters

Besides the one- and two-body interaction energies, which directly enter into SAMPLE's model and can be determined from training data, there are several other parameters which need to be set a-priori. These *hyperparameters (HPs)* control the behavior of the model in general. Table 2.1 lists different types of HPs and describes their significance.

Table 2.1.: Hyperparameters of SAMPLE

Parameter name	Symbol	Description
one-body interaction uncertainty	σ_{1-body}	estimated uncertainty of one-body energies U_g
two-body interaction uncertainty	σ_{2-body}	estimated uncertainty of two-body interaction energies V_p
DFT uncertainty	σ_{DFT}	estimated uncertainty of DFT calculations
feature threshold	Δf	sets the maximum separation in feature space which two FV can have and still be considered equal
feature correlation length	ξ	characteristic length scale in feature space, see equation 1.10
decay power	n	controls decrease of feature vector entries as a function of the inter-atomic distance
real space decay lengths	τ^{AB}	control the distance above which the two-body interactions converge to their prior mean
maximum distance cutoff	d_{max}^{AB}	maximum interatomic distance up to which an atom pair of species A and B is considered in the FV
minimal distance thresholds	d_{min}^{AB}	minimum allowed distance for atoms of species A and B; structures with smaller interatomic distances are deemed unstable due to high Pauli repulsion

While four of the seven HPs are just scalar values, the rest, namely the real space decay lengths τ^{AB} and the maximum/minimum distance thresholds, d_{max}^{AB} and d_{min}^{AB}, can actually be defined separately for each pair of atom species (A, B). Hence there can be much more than just 7 individual HPs. For the test systems featured in this thesis, it was chosen to only define the minimal distance thresholds on a per-atom-species basis, setting scalar values for all other HPs.

Table 2.2.: Optimized hyperparameters used for all test systems

Parameter	Value
σ_{1-body}	100 meV
σ_{2-body}	100 meV
σ_{DFT}	5 meV
Δf	0.005
ξ	3.0
n	-3
τ	3 Å
d_{max}	16 Å

With the exception of minimum/maximum distance thresholds, which have to be known prior to generating configurations, selecting an appropriate value for an HP can be done more easily by taking a mid-sized set of configurations (≈ 100 structures) backed by DFT calculations and trying to minimize the prediction error by varying said HP. This approach was used to set all single value HP, except for the maximum distance threshold d_{max}, which was manually set to 16 Å.

After it was found that most of the first batch of test systems (level-1 and level-2 systems, see sections 3.3 and 3.4), shared the same set of optimized hyperparameters, it was decided to forgo hyperparameter optimizations for all later systems based on the assumption that this set of hyperparameters would be suitable for all test systems due to their overall structural similarity. Table 2.2 below lists the values of these hyperparameters.

The attentive reader will have noticed that the minimal distance thresholds d_{min}^{AB} are missing in table 2.2. They are omitted on purpose, for once, due to a lack of space, but also to highlight that these hyperparameters are determined through a different process, which is elaborated in the next section.

2.3.1. Defining minimal distance thresholds

As briefly discussed in section 2.3, the minimal distance thresholds d_{min}^{AB} set the lower limit of how close two atoms of species A and B can be in a SAMPLE-conforming polymorph. These limits were introduced to avoid generating configurations that can be expected to be unstable due to the appearance of highly repulsive interactions based on Pauli exclusion. An example of this behavior is depicted in figure 2.5.

As the repulsive contribution rises continuously with decreasing distance, there is no obvious point where the threshold should be put and thus this choice is left to the user. Jeindl [**Jeindl:masters-thesis**], for instance, positioned two molecules[5] in such a way that a specific pair of atoms were most likely to interact and then varied the intermolecular distance and calculated interaction energies at each point. This resulted in curves similar to those depicted in figure 2.5, one for each type of atom species pair (A, B). Based those he determined the atom-atom distance where the repulsive interaction energy (per molecule) would cancel out the attractive adsorption energy.

It was decided to replicate this approach for the thesis at hand, albeit with two adaptations. Firstly, since all test systems represent free-standing monolayers, there is no interaction with the substrate and hence the adsorption energy is not defined. Instead it was decided to set each d_{min}^{AB} in such a way that the maximum repulsive interaction energy is no more than 500 meV per molecule and per interacting atom pair.

The second adaptation is based on the desire to have test systems with multiple, varied functional groups. Upon realizing that finding reasonable minimal distance thresholds through running distance sweeps, like the one presented in figure 2.5, for all appearing atom species pairs would be time-consuming, it was decided to estimate the d_{min}^{AB}s instead based on the sum of the van-der-Waals radii of the involved atoms:

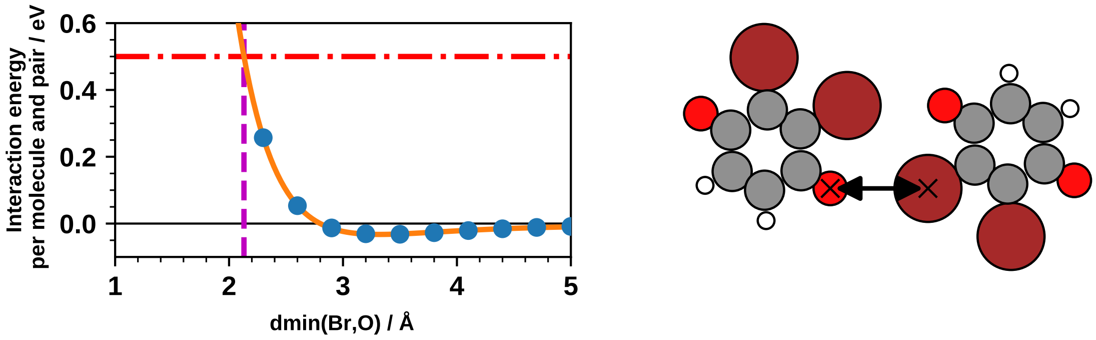
$$d_{min}^{AB} = \beta \left(r_{vdW}^{A} + r_{vdW}^{B} \right) \tag{2.15}$$

Here, r_{vdW}^{A} and r_{vdW}^{B} are the van-der-Waals radii of atom A and atom B, while β is an empirical factor which is chosen so that equation (2.15) fits the results of distance sweeps for several different atom species pairs.

When it was tried to fit β to the results of dmin-sweeps for different atom pair combinations, with calculated curves depicted in figures A.1 and A.2, it was noticed that no single value for β resulted in reasonably good agreement for all tested cases. Instead, the observation was made that the tested combinations could be divided into two distinct groups. The first and also larger group consisted of atom pair combinations that, when compared to the rest, were more repulsive. This encompasses pairs such as O-Br, O-F and F-Cl, as well as all homogeneous combinations like O-O, F-F, Cl-Cl etc. The d_{min} for these *repulsive* atom pairs could be well estimated by inserting a value of $\beta = \beta_{rep} = 0.65$ into equation 2.15. The remaining atom pairs were found to allow for smaller d_{min} and could be well characterized with a value of $\beta = \beta_{atr} = 0.55$. Examples for these *attractive* atom pair combinations were O-H, Br-H and F-H.

There was, however, one outlier in this group, namely the combination O-H. While the corresponding sweep suggests a d_{min}^{OH} of 1.35 Å, the empirical formula returns

[5] 1,4-Benzoquinone in his case



(a) Calculated sweep curve

(b) Test geometry

Figure 2.5.: How to select minimal distance thresholds

a relatively high value of 1.50 Å, which would effectively cut off potentially valid interactions. Contrasting this with the data for F-H and Br-H pairs, where the difference between calculated and estimated d_{min} is much smaller, it seems that the strong attractive component of the O-H interaction allows the atoms to be closer than would be expected based on the respective van-der-Waals radii.

Based on the tested atom combinations, the difference in d_{min} between attractive and repulsive atom pairs seems to come down to the fact that the tested attractive pairs combine an electron donor atom (always H) with an electron acceptor atom with high electronegativity. This allows the formation of hydrogen bridges, which partly negates the repulsive Pauli exclusion interaction.

2.4. Designing test systems

The path towards answering the main research question of this thesis starts with defining feasible test systems. As discussed in section 1.2, in the framework of SAMPLE, any target system is made of two parts: an adsorbate molecule and some kind of substrate on whose surface the former is placed. Both parts need to be chosen in a way that allows the four testing scenarios introduced in chapter 1.1 to be carried out.

2.4.1. Virtual lattice

Since the focus of this thesis lies solely on molecule-molecule interactions, there is actually no need to incorporate a real substrate into our test systems. And because every theoretical physicist should strive to be efficient in their use of computational resources, we will thus avoid a lot of unnecessary compute effort by using a purely virtual lattice instead. All test systems therefore represent free-standing molecular monolayers, as depicted in figure 2.6. The virtual lattice for this thesis is based on a Cu(111) surface. With copper inhabiting a *face-centered cubic (fcc)* crystal structure, this makes the virtual lattice a two-dimensional hexagonal lattice. Its lattice constant a is set to 2.553 Å.

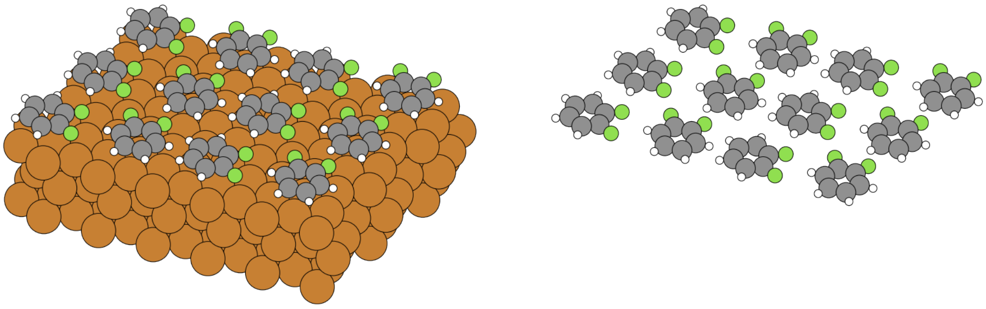


Figure 2.6.: Definition of a test system as a molecular monolayer. The graphic on the left illustrates a monolayer on a Cu(111) substrate, which serves as the base for the actual virtual lattice used in SAMPLE. The graphic on the right shows only the monolayer.

2.4.2. Test molecules

Creating the test molecules was preceded by setting several design goals. These were:

- All test molecules should be derived from a common, simple base structure
- Small molecules are preferred to reduce compute costs
- Molecules should be (mostly) flat in order to focus on in-plan interactions
- Test molecules should be chemically sound, although they do not need to exist in the real world
- It should be possible to prepare test molecules for specific test scenarios by adding functional groups to the base structure

In the end, it was decided that the goals above are best met by using a planar ring of carbon atoms as the base unit for all test molecules.

2.4.3. Local geometries

As explained in section 1.2, in order to apply SAMPLE, we need to assign one or more SILGs. These define the positions and rotations relative to the (virtual) lattice, that a molecule can inhabit. Therefore, they control the level of detail with which molecular arrangements can be represented in SAMPLE.

For this work, it is decided to place molecules onto three distinguishable positions in the unit cell. Written in relative lattice coordinates, these positions, i.e. *adsorption sites* in usual SAMPLE terminology, are: *top* at $(0, 0)$, *hollow-1* at $(0, 1/2)$ and *hollow-2* at $(1/2, 1/2)$. Through this definition, we effectively achieve a spacing of half the lattice constant.

With regard to the second factor, rotations, the aim is to define enough local geometries to assure that all distinguishable orientations of the test molecule with an angle increment of 30° can be realized. The number of LGs for each system is thus based on the (rotational) symmetry of each test molecule, with low-symmetry molecules requiring more LGs than high-symmetry ones.

2.5. DFT calculation settings

All calculations for this thesis, for periodic as well as for non-periodic systems, are run with *Fritz Haber Institute ab initio molecular simulations (FHIaims)* using the *Perdew-Burke-Ernzerhof (PBE)* exchange-correlation functional[**PBE1996**], with multipole correction and van-der-Waals corrections (TS^{surf}[**Tkatchenko:TSsurf**]) enabled. FHIaims' "tight" default species settings[6] are used as basis set definitions.

The two-dimensional nature of the targeted molecular monolayers is handled via a repeated-slab approach, i.e. by setting the height of the unit cell in the (non-periodic) z-direction large enough, so that there is no more interaction between the monolayer and its periodic images. In the present thesis, a unit cell heigh of 100 Å is used.

Even more impactful than the height of the unit cell is the choice of the reciprocal lattice points $\mathbf{k}$ (k-points) for which the KS-equations (see 1.3) are solved. The calculations at hand employ uniform, Γ-point centered k-point grids spaced along the reciprocal lattice vectors $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{v}_z$.

To conform with the repeated-slab approach, only a single k-point is chosen in the z-direction, i.e. perpendicular to the surface. The number of k-points in $\mathbf{v}_1$- and $\mathbf{v}_2$-direction, n_1 and n_2 respectively, are set for each calculation individually and in dependence of the shape of the corresponding supercell in order to yield a k-point density of (approximately) 36 k-points per primitive substrate unit cell.

Additionally, a dipole correction via the introduction of a virtual dipole layer is enabled for all periodic calculations to account for any net polarization of the monolayer. For a complete listing of all computation settings, please see the exemplary control files for both periodic and non-periodic calculations in appendix B.

[6] These basis set settings are included in a FHIaims installation. They can be found under `fhi-aims/species_defaults/tight`.

3. Results and Discussion

This chapter is going to discuss the setup and reasoning behind each of the four test levels, followed by a presentation and an interpretation of the achieved results.

3.1. Influence of test set selection

When using equation 2.1 to measure the prediction accuracy, we have to be aware of the fact that the choice of the *test set* S_{test} will inadvertently influence the result. Naturally, this prompts the question of how to best allot configurations to S_{test}. In order to show the range of influence that this choice can have, we will take a look at three edge cases:

Option A: Draw test set D-optimally
Option B: Draw training pool D-optimally
Option C: Draw test set randomly. Repeat multiple times and average

For each of these three options, we determine the learning curves (see 3.3) by drawing training sets of increasing sizes from the respective S_{pool}, training our energy model and then using S_{test} to get a value for the prediction uncertainty. In the case of option C, we measure multiple learning curves, each based on a different randomly drawn test set. The obtained RMSE values are then averaged point-by-point to yield a single learning curve. The variance of this average curve is estimated by the respective root-mean-square errors for each training set size.

Figure 3.1 compares the measured learning performance in these three cases for a specific molecule (*l4-NH*). We see that choosing the test set via option A results in the highest RMSE, whereas option B produces the lowest value. Except for very small training sets (50 configurations), the "averaged" prediction uncertainty (option C) falls within the values for options A and B. It should be noted that the curve for option C is the average of just five separate learning curves. This rather small sample size naturally reduces the significance of associated averages and standard deviations.

The better performance of A with respect to B can be explained by the amount of information that is contained in the configurations of their respective training pools, S^A_{pool} and S^B_{pool}. In the context of SAMPLE, the information content of a configuration set is linked to the number of distinguishable features which can be found in the configurations that constitute the set. In more general terms, an information-rich configuration set contains a wide array of interaction pairs with a large variation in pair orientations and distances relative to its size. Thus, such a "diverse" configuration set allows sampling the overall PES better compared to a less information-rich set of comparable size.

Building a configuration set by drawing configurations based on a D-optimality criterion **[Fedorov1972]** assures that the most important configurations are allotted to the set, thereby providing it with the highest possible information density. In the case of option B, the training pool is such a D-optimally drawn set. Since test set and training pool are disjunct by design, i.e. do not share any configurations, the test set S^B_{test} will be comprised of configurations with a lesser degree of variability than S^B_{pool}. Thus, its information density will be low, making S^B_{test} an "easy" test set. Consequently, a model that is trained on subsets of S^B_{pool} should receive enough information to predict energies of the test set configurations fairly well.

The opposite is true for option A. Here, the test set S^A_{test} is drawn D-optimally, which makes it a "hard" test set. What is more, this choice leaves only the least important configurations for S^A_{pool}. Because of this, S^A_{pool} will provide less information for the training sets/the model than S^B_{pool}. In combination with the higher information density of S^A_{test} compared to S^B_{test}, this explains why the prediction accuracy in case B turns out better than in case A. By the same logic, we can interpret case C as an intermediate between options A and B. For each of the multiple evaluations, the information content of the test set S^C_{test} (and training pool S^C_{pool}) is determined by chance, but it will never be lower than that of S^B_{test} (S^A_{pool}) or higher than that of S^A_{test} (S^B_{pool}). It is therefore not surprising that the average prediction uncertainty for case C, γ^{avg}, lies between the values for A and B. We can hence identify option A as the worst-case scenario whereas yields an upper bound for the prediction uncertainty, γ^{worst}, while option B represents the best-case scenario, with γ^{best} as the lower limit.

To get a more complete picture, the three test set selection modes were benchmarked on an array of 16 test systems, which are formally introduced in section 3.6. Figure 3.2 shows a comparison between measured prediction uncertainties γ^A, γ^B and γ^C of the different systems, each for a training set size of 700. We observe that for most of the discussed systems, the type of test set influences the measured γ similarly to what is seen for *l4-tm1*, i.e. that "type A" test sets yield nominally better γ than "type B" ones. Still, we can find a system where this order is actually flipped as well as three systems, *l4-CN*, *l4-CO* and *l4-NH2*, where there is only a very small difference between γ^A and γ^B.

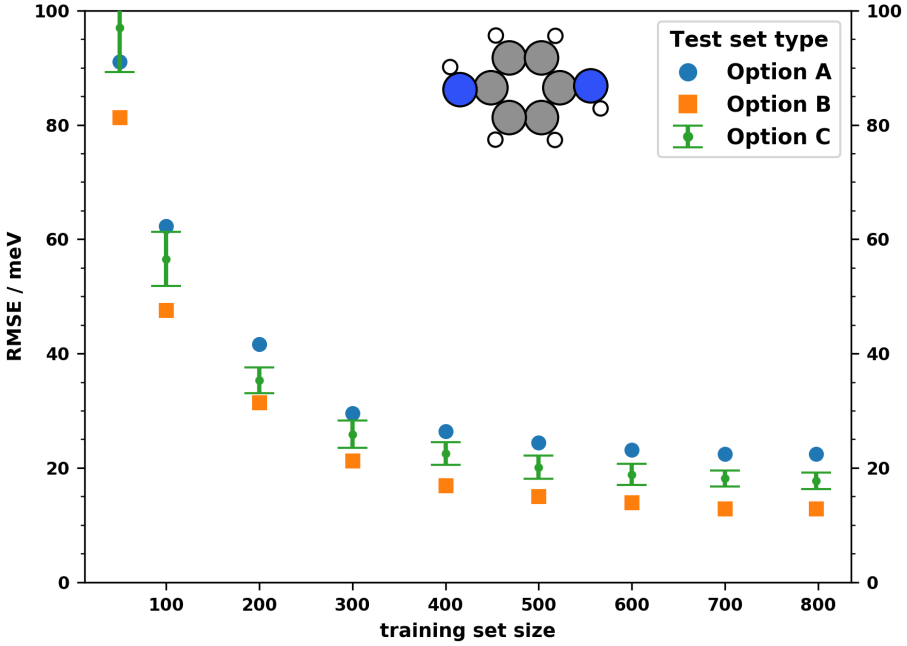


Figure 3.1.: Influence of test set on achieved prediction accuracy. Compares learning curves for molecule *l4-NH* where the test set was chosen in three different ways (A, B and C). The difference between the selection methods is explained in the main text.

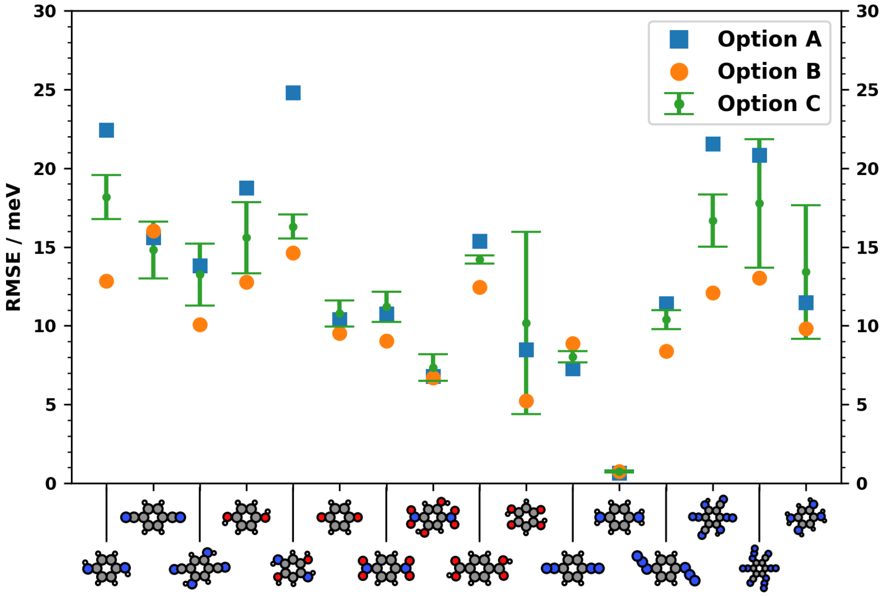


Figure 3.2.: Comparison between different types of test sets for all level 4 molecules (see 3.6). Drawn are the achieved prediction accuracies γ^A(blue squares), γ^B(orange dots) and γ^C(green dots with errorbars) for training sets with 700 configurations.

A detailed look at the distribution of prediction errors ΔE_{MLF} for type A and type B test sets reveals that when there is a significant difference between γ^A and γ^B, then the prediction error distribution of the worse performing test set most often contains strong outliers (more than 2σ deviation). When these outliers are excluded, the RMSEs decrease, sometimes even drastically, and become closer in value.

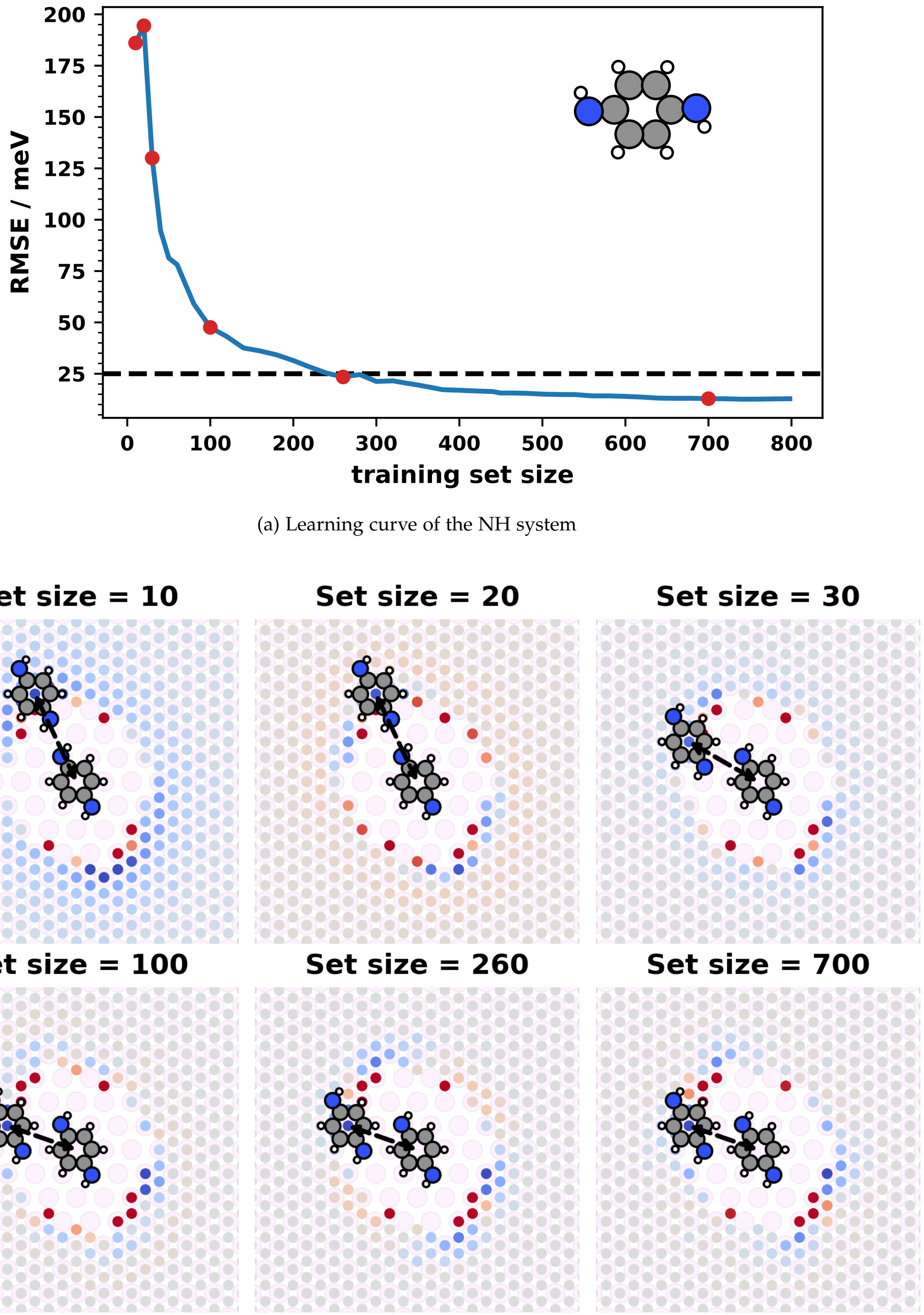
Now that the influence of the test set selection is understood, the only thing left is to decide which type of test set(s) should be used going forward. From a statistical standpoint it would probably be best to go ahead with option C since it allows specifying the average learning performance and the corresponding confidence interval. It comes, however, with the disadvantage of a relatively high computational demand when compared to option A and B. As this whole thesis is focused on testing as many systems as possible, we want to keep compute effort low and therefore we are going to dismiss option C.

The choice between the two remaining options, A and B, comes down to the question, whether we want to determine the worst-case (A) or the best-case (B) learning performance. Although selection mode A might be a safe choice, as it will generally give a conservative estimate of the prowess of SAMPLE, we are still going to choose the test sets via method B. This decision is motivated by the fact that approach B is the one most similar to how we would use SAMPLE in an actual production scenario, as, in practice, we would train on the most important configurations.

3.2. Evolution and convergence of pair potentials

As discussed in section 2.2.3, for the most part, we analyze the learning performance of our model by means of discussing learning curves like the one displayed in figure 2.3. While this approach is sound from a mathematical and statistical point of view, it is somewhat hard to conceptualize intuitively how the model improves as more training data is added. One alternative to learning curves, that presents the learning/training process in a more accessible and visual fashion, is to look at the evolution of *intermolecular* or *pair potentials*, as defined in section 2.2.3.

Figure 3.3 shows how the pair potential of two test molecules (both *l4-NH*) develops as the training set size is increased and relates this evolution to the corresponding learning curve. We observe that Φ_{pair} oscillates wildly for small training sets, most notably changing from overall attractive to repulsive and back again in the progression of 10, 20 and 30 training configurations. From a set size of around 100 configurations onward, the variations of Φ_{pair} quickly die down. Upwards of a set size of approximately 260, the potential appears to have more or less converged, at least qualitatively, as seen from the comparison with the pair potential map based on 700 training configurations.



(a) Learning curve of the NH system

(b) Pair potential evolution

Figure 3.3.: Pair potential evolution. Panel (b) depicts the evolution of the intermolecular potential of a pair of test molecules (*l4-NH*) as a function of the training set size. Panel (a) shows the corresponding learning curve. Position and orientation of the central molecule are fixed and the second molecule is drawn in the most favorable relative position. Points in the learning curve that correspond to the featured potentials are marked with red dots in panel (a)

3.3. Level 1: Impact of molecular symmetries

Due to the nature of the feature vector (see section 1.2.4), the number of fit coefficients (=interaction energies) in SAMPLE's energy model (see section 1.2) is related to the number of unique two-body interaction pairs and thus depends on the symmetry properties of the molecule in question.

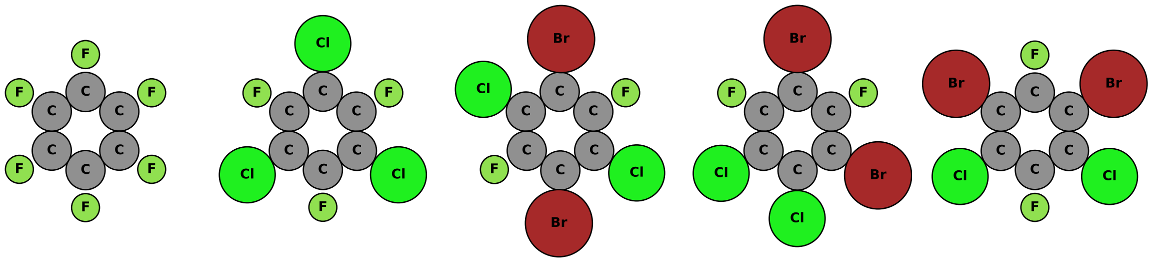
In general, the symmetry class or group of a molecule is defined by the set of symmetry operations (rotations, reflections or inversions) that, when applied to the molecule, map all atoms back onto themselves. The degree of symmetry of a molecule is therefore linked to the number of symmetry operations it fulfills. In this particular case, where we inspect only flat molecules, it should suffice to only investigate symmetries with respect to the surface plane.

For a highly symmetric molecule, many of the possible interaction pairs are equivalent and therefore the number of fit coefficients/model parameters should be lower than for a low symmetry molecule of comparable size and composition.

Since, as a general rule, the amount of required training data increases with the number of fit parameters, it seems likely that the molecular symmetry will have an impact on learning performance. The focus of this suite of test systems lies in testing whether this hypothesis holds and if so, in trying to quantify the effect.

3.3.1. Test molecules

This level consists of 4 test molecules, each belonging to a different symmetry class. Each test molecule is a benzene derivative and is constructed by adding halogen atoms (F, Cl, Br) as substituents to a central C6-ring. Figure 3.4 depicts all 4 test systems, labeled by their symmetry classes.



(a) D6h (b) C3h (c) C2h (d) C2v (e) C1h

Figure 3.4.: Geometries of test molecules from level 1, labeled by the corresponding symmetry class

3.3.2. Results

Figure 3.5 shows the learning curves[1] of each test system for training set sizes of 10 to 800 configurations. Drawn as thin lines are fits to the individual learning curves. These double exponential fits were carried out in order to estimate the systematic model error δE_{sys} for each test system and are based on equation 2.11. A detailed discussion of the fitting procedure can be found in section 2.2.4. The resulting optimized fit parameters are listed in table 3.2.

Upon taking a look at the curves in figure 3.5, we notice that almost no training is needed for the test molecule with D6h symmetry. Even for the smallest training set (10 configurations), the prediction accuracy is already far better than the acceptance threshold (γ_{accept}) of 25 meV. Upwards of around 40 configurations, the curve essentially flattens out and becomes stationary at a prediction uncertainty of just $0.63(2)$ meV, far better than expected beforehand.

Going on to the next test molecule, where the molecular symmetry is decreased from *D6h* to *C3h*, it can be seen that now around 40 training configurations are needed to bring the RMSE below γ_{accept}. Still, above of a training set size of ≈ 100, the learning curve of the *C3h* system also levels out and apparently hits the lower limit in terms of prediction uncertainty, very much similar to the learning curve for the *D6h* system. Of the tested systems, only *D6h* and *C3h* show such a rapid decrease of the prediction uncertainty, which equates to a remarkable learning speed. Furthermore, by correlating the training set sizes at which these two learning curves turn stationary with the symmetry properties of the corresponding test molecules in table 3.1, we observe that the former are close in value to the number of fit coefficients included in the energy model.

The results for the next two molecules, i.e. the ones with C2h and C2v symmetry, show a further degradation in learning performance. This is visible from the corresponding learning curves in figure 3.5: they show a much slower decrease with increasing training set size. Overall, SAMPLE performs better for fest molecule *C2v* than for test system *C2h*, particularly in the mid range of training set sizes (100 to 350 configurations), where the learning curve for the former is lower by ≈ 10 meV. Likewise, the prediction uncertainty of *C2v* goes below the acceptance threshold at a training set size of ≈ 180, in comparison to the ≈ 260 configurations needed for *C2h*. However, as training sets become considerably larger, the learning curves for C2v and C2h steadily get closer and even switch places at around 500 training configurations. In the end, at least according to the δE_{sys}s estimated via the fit functions, the two curves remain separated by around 1.5 meV ($\delta E_{sys}^{C2h} = 6.4(3)$ meV vs. $\delta E_{sys}^{C2v} = 5(1)$ meV).

[1] see section 2.2.3

Table 3.1.: Symmetry properties of level 1 test molecules

Property	Test systems				
system name/symmetry class	D6h	C3h	C2h	C2v	C1h
number of in-plane symmetry transformations (including identity)	12	4	2	2	1
number of fit coefficients	47	156	1457	1171	3759

The test system for which SAMPLE performs worst is the one with the non-symmetric *C1h* molecule. Initial prediction errors are much higher than for the other systems and, as shown in figure 3.5, learning progress is noticeably slower. This finding is also reflected in the optimized fit parameters in table 3.2. Even though none of the individual parameters for system *C1h* stand out against those of the other systems, as a collective, they signify the worst learning performance since all of them are close to the top of their respective value ranges.[2]

Relatively slow learning progress aside, SAMPLE is nonetheless capable of reaching γ_{accept} for the *C1h* system, albeit only with a training set consisting of ≈450 configurations. This is almost 1.5 times the required test set size of the next-worst system (*C2h*, $x_{required} \approx 260$ configurations).

Another possibility to gauge learning performance lies in comparing the prediction accuracies for a fixed training set size. Such a comparison is depicted in figure 3.6. Shown are the RMSEs of all level 1 test systems for training sets with 700 configurations. We see an overall trend that a decrease in molecular symmetry goes hand in hand with an increase in RMSE. The increase is biggest in the step from C2h (7.0 meV) to C1h (12.2 meV), but, remarkably, there is almost no difference between the learning performance of C2v (7.0 meV) and C2h (6.7 meV).

This can be explained as follows: the C2h and C2v test molecules fall into different symmetry classes, and thus fulfill different symmetry transformations. But since SAMPLE only applies in-plane rotations and mirror transformations (based on symmetries of the substrate) in order to generate possible local geometries, it is only relevant how many of the molecular symmetry transformations match those of the substrate. As both the C2h and the C2v molecule share two transformations with the substrate - identity and one mirror symmetry in the case of C2v and identity and a 180 degree rotation for C2h - it does make sense that they can be learned equally well or badly.

[2] With the exception of λ_2 for the *D6h* system. But because the double exponential fit function is not well equipped to describe an almost completely flat learning curve, fit values for the parameters K_2 and λ_2 are not be well defined for this system.

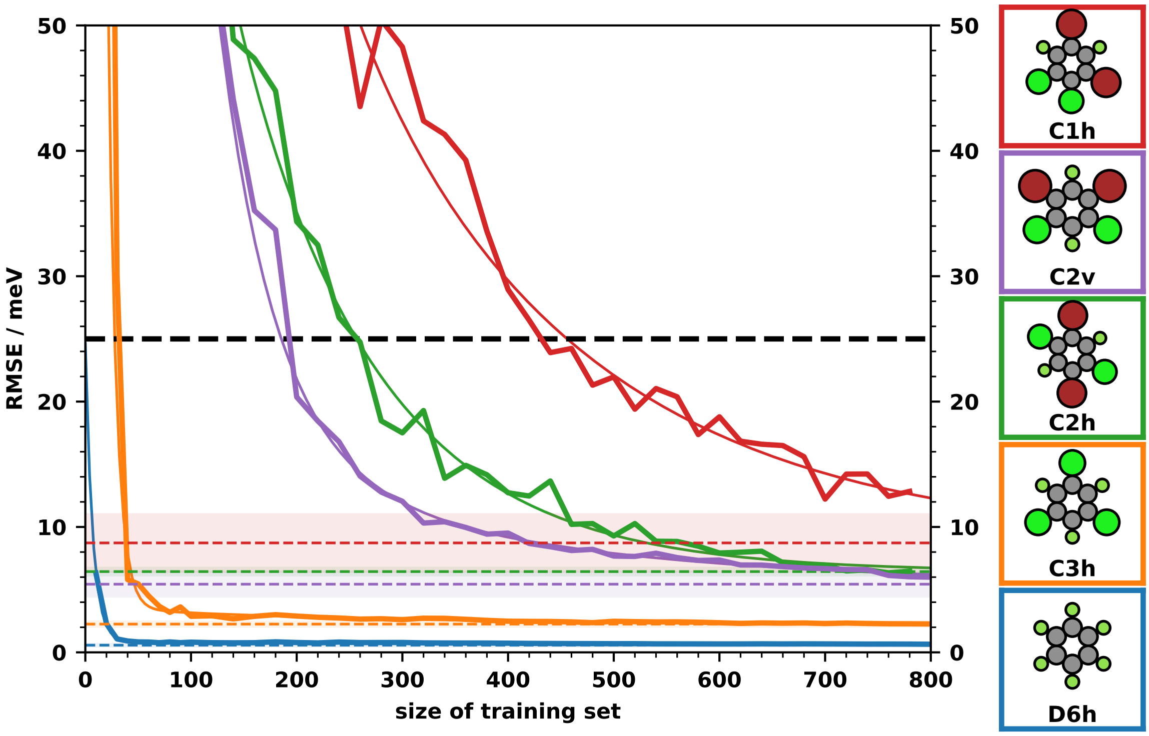


Figure 3.5.: Comparison of learning performance for level 1 test systems. For each test system, the measured learning curve is drawn with a thick solid line and the fitted curve (based on equation 2.12) is represented by a thin solid line of the same color. The colored dashed lines mark the estimated model errors, whose confidence intervals are visualized as semitransparent bars. The dashed black horizontal line marks the acceptance threshold of 25 meV. Test system geometries are depicted on the right hand side.

Table 3.2.: Optimized parameters of fit model 2.12, based on level 1 learning curves. Listed are the average value for each parameter (mean) and its standard error (std).

Molecular symmetry		D6h	C3h	C2h	C2v	C1h
δE_{sys} / meV	mean	0.57	2.3	6.4	5	9
	std	0.08	0.2	0.3	1	2
K_1 / meV	mean	24	760	550	380	330
	std	2	120	80	30	30
λ_1	mean	6.9	7.8	22	53	56
	std	0.2	0.4	3	4	10
K_2 / meV	mean	0.27	1.4	131	16	120
	std	0.07	0.3	13	8	30
λ_2	mean	700	240	131	260	230
	std	400	140	7	130	50

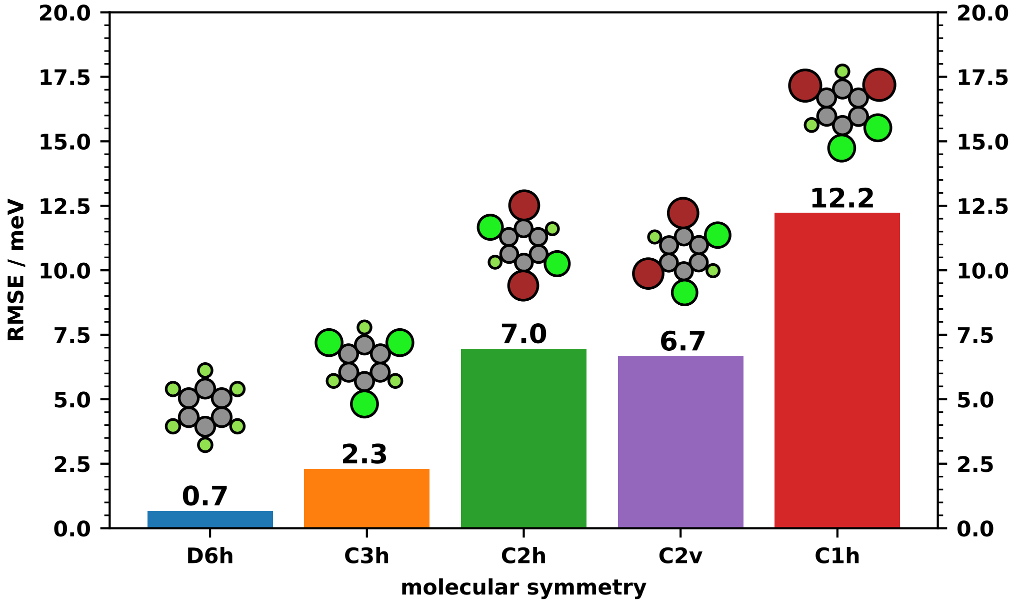


Figure 3.6.: RMSE for training set of size 700 and test systems with different molecular symmetries

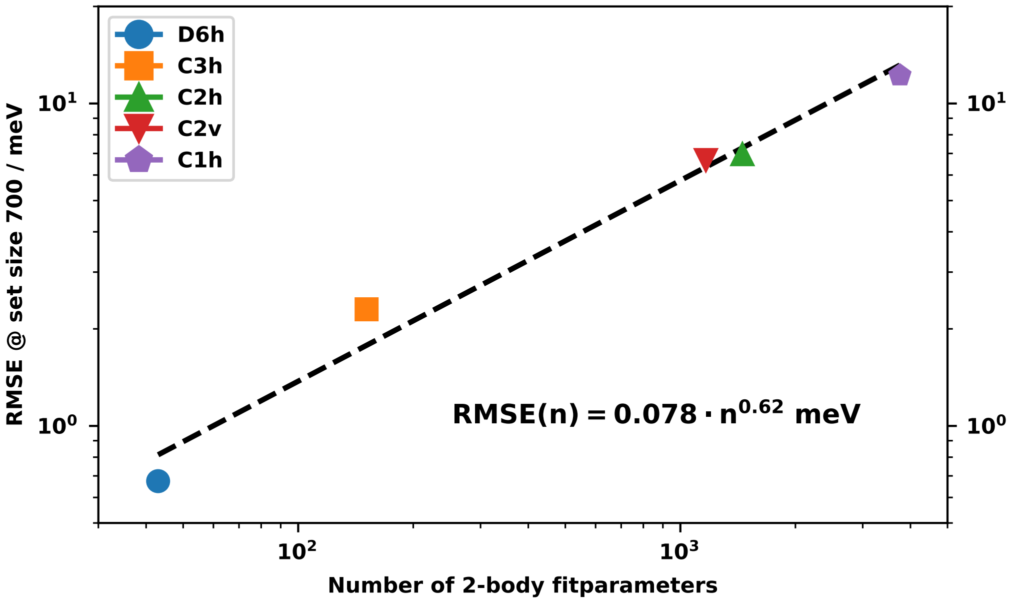


Figure 3.7.: RMSE at training set size 700 versus number of 2-body fit parameters for level 1 systems. The dashed line symbolizes the trend that the prediction errors grow proportionally to the number of fit parameters. Information on the corresponding trend function is given by the insert near the bottom right corner. It should be noted that the scaling behavior is specific to this case and not a general finding.

3.4. Level 2: Adding Hydrogen bonds

3.4.1. Test systems and objective

Due to its nature, SAMPLE's energy model treats all interactions as if they were non-directional (isotropic). In reality, there inter-molecular interactions exist that are anisotropic, where the interaction strength, i.e.the connected potential is direction-dependent. The prime example for this kind of interactions are hydrogen bonds.

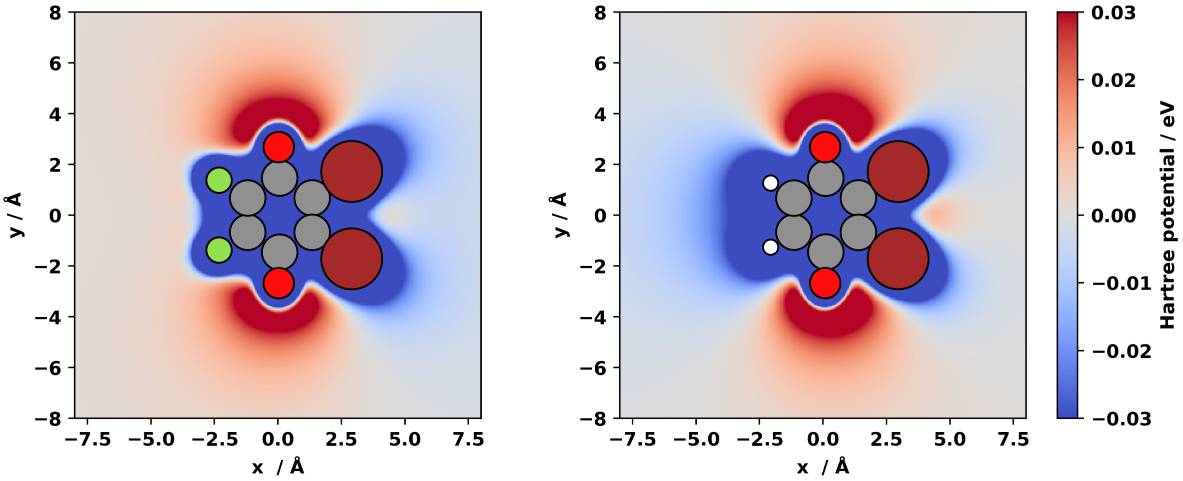
Up to this point (see section 3.3), we treated test molecules, which, due to the choice of substituents, will mostly form repulsive interactions.[3] Under the (reasonable) assumption that all substituents repel each other (more or less) equally, these interactions are in essence isotropic. Once hydrogen atoms are allowed as possible substituents, this picture changes as the electrostatic potential of the molecule gets more heterogeneous. We can, for instance, compare the Hartree potentials of test molecules *l2-tm1* and *l2-tm3*, as is shown in figure 3.8.

If we look at the mid to long-range behavior of the respective potentials, we observe that the PES of the hydrogen-free molecule (left panel) is more homogeneous than its hydrogen-carrying counterpart. The difference is most noticeable in the sectors on the left side of the respective molecules. The suspected reason for this is the difference in the electronegativities of fluorine and hydrogen. Both interact with the neighboring oxygen atoms, which causes a rearrangement of electron density which in turn creates a gradient in the electrostatic potential. Since the difference in electronegativity χ between O ($\chi_O = 3.5$) and F ($\chi_F = 4.1$) is smaller than between O and H ($\chi_H = 2.2$), the slope of the electrostatic potential is less pronounced for the *tm1* system compared to that of the *l2-tm3* system.

Therefore, in order to test whether interaction anisotropies actually affect SAMPLE's performance or not, a second set of test systems was designed, which features molecules with various hydrogen bonding configurations as well as two hydrogen-less molecules that serve as a control group. To avoid spurious effects that arise due to differing molecular symmetries [4], it was made sure that all test molecules adhere to the C2v symmetry class. An overview of the structure of all 7 molecules used can be found in figure 3.9. The keen-eyed reader will, of course, have noticed that test molecule *l2-tm6* is equal to molecule *C2v* from level 1. Its original identifier will not be used in this section to minimize confusion.

[3] The structures of the mentioned test molecules is shown in figure 3.4.

[4] again, see section 3.3 for details



(a) hydrogen-free (*l2-tm1*)

(b) with hydrogens (*l2-tm3*)

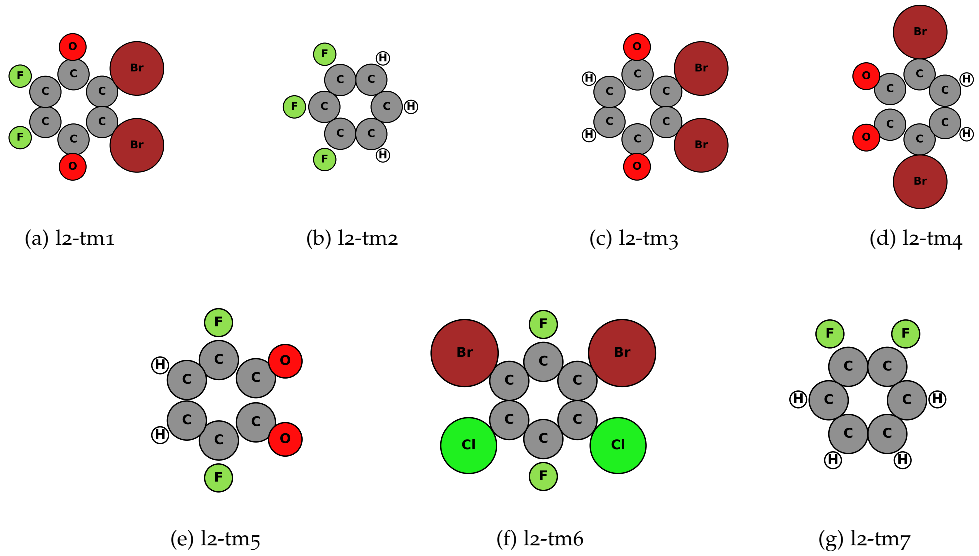
Figure 3.8.: Comparison of electrostatic potential of test molecule (a) with and (b) without hydrogens

3.4.2. Results

Same as for the previous level, we start with an overview of the learning performances of all systems. This is plotted in figure 3.10, accompanied by double exponential fits. The corresponding fit parameters are listed in table 3.3.

A lot of information can be extracted from figure 3.10, starting with the observation that for all level 2 systems, SAMPLE was able to push RMSEs below the acceptance threshold of 25 meV with less than 350 training configurations. We also clearly notice that fluorine/hydrogen test molecules (*l2-tm2* and *l3-tm1*) were by far the easiest to train, with less than 100 training configurations needed to pass γ_{accept}. We also see, that above a set size of 400, the curves for *l2-tm4* and *l2-tm3* are basically equivalent, which is interesting, since they share the same substituents (O, Br, H), but in a different order. The double exponential fits appear to suggest that for even higher training set sizes *l2-tm3* eventually beats *l2-tm4* and ends up at a lower δE_{sys}, but as the corresponding confidence intervals slightly overlap, this finding does not stand on solid ground.

While we could continue to interpret the results for level to based on figure 3.10, due to the high information density, it might actually be more instructive to analyze the prediction uncertainty at a fixed training set size in order to see trends in the data. In figure 3.11, RMSEs at a set size of 700 are plotted against the number of fit parameters (in SAMPLE's energy model). If we then only look at the systems where hydrogen bonds can form, which, in figure 3.11 are highlighted with a dashed box, we see a linear increase of the model error with the number of fit parameters. In


(a) l2-tm1

(b) l2-tm2

(c) l2-tm3

(d) l2-tm4

(e) l2-tm5

(f) l2-tm6

(g) l2-tm7

Figure 3.9.: Geometries of test molecules from level 2, labeled by their identifier tag. Notice: Test molecule *l2-tm6* is also featured in level 1 (*C2v*) and test molecule *l2-tm7* is equivalent to molecule *F-ortho* from level 3.

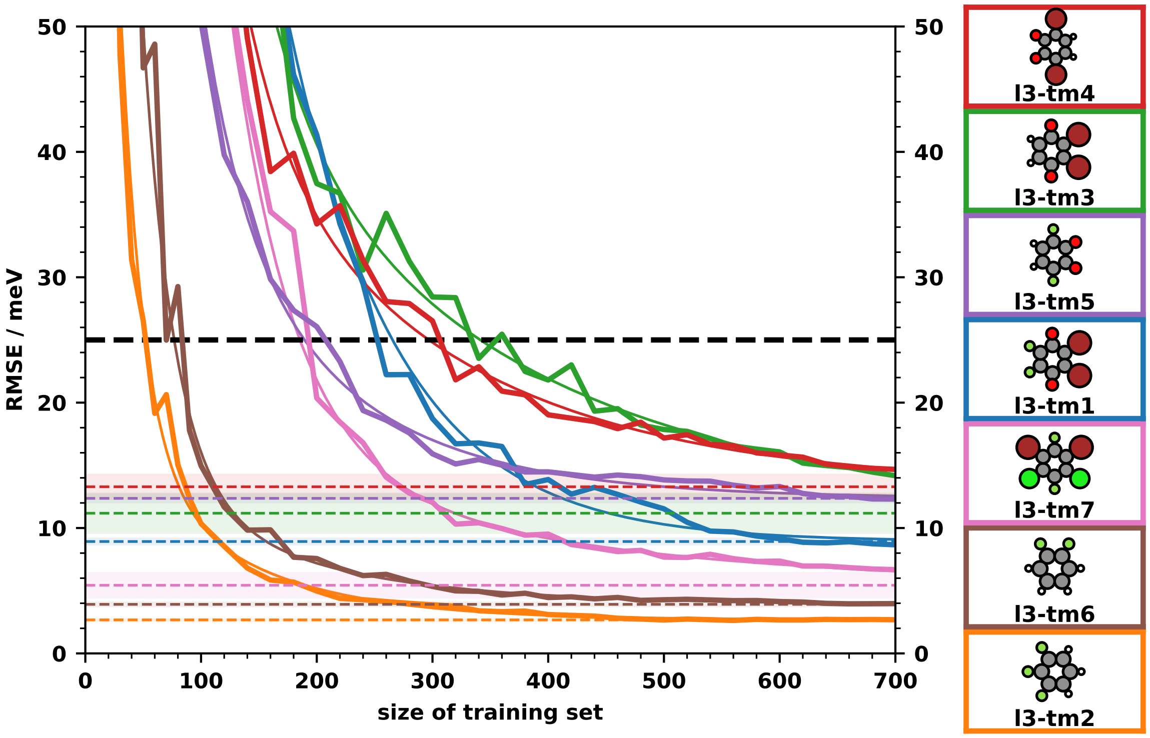


Figure 3.10.: Summary of all level 2 learning curves. Double exponential fits are denoted by thin lines of the corresponding color. Estimated systematic model errors δE_{sys} are marked by dashed lines.

Table 3.3.: Optimized parameters of fit model 2.12, based on level 2 learning curves. Listed are the average value for each parameter (mean) and its standard error (std).

		l3-tm1	l3-tm2	l3-tm3	l3-tm4	l3-tm5	l3-tm6	l3-tm7
δE_{sys} / meV	mean	8.9	2.67	11.2	13.3	12.4	3.91	5
	std	0.2	0.06	1.5	0.9	0.3	0.12	1
K_1 / meV	mean	400	170	390	455	250	240	380
	std	100	30	30	30	20	20	30
λ_1	mean	14	20	46	39	38	27	53
	std	3	2	4	3	4	2	4
K_2 / meV	mean	260	17	56	54	50	14	16
	std	20	3	12	12	16	3	8
λ_2	mean	95	104	240	190	126	135	260
	std	3	10	50	40	20	20	130

general, such an increase is to be expected when using a linear regression model as is the case with SAMPLE. However, if the number of fit coefficients were the only determining factor, we would expect that the non-hydrogen systems follow the same trend, but as can be seen in figure 3.11, the two systems in question yield a significantly (10 meV to 15 meV) lower model error than suggested by the linear trend. This finding suggests that the inclusion of hydrogen atoms, respectively the appearance of hydrogen bonds, makes learning the affected systems measurably harder, although not to an extent that achieving good prediction results becomes impossible. Nonetheless, before we can accept this interpretation, we first need to look at other possible explanations for the observed behavior.

For instance, it could be that the non-hydrogen molecules fare better (in relation to the hydrogen systems), because the energy spread of the configurations in their test sets is smaller. Talking in more detail, the *energy spread* of a configuration set is defined as difference in E_{MLF} between the configurations with lowest and highest E_{MLF}. As such, it is an indicator for the diversity and range of interactions that appear in the configuration set. A large energy spread signifies that the explored part of the PES is relatively corrugated due to the inclusion of highly repulsive interactions and also often attractive interactions as well.

Looking at the results in figure 3.12, reveals that there is no categorical difference in energy spread between hydrogen and non-hydrogen systems. It can therefore be argued that the trend-breaking behavior of the non-hydrogen systems is not caused by differences in the energy spread. We thus can accept our initial interpretation that it is the addition of hydrogen atoms that makes learning more complex.

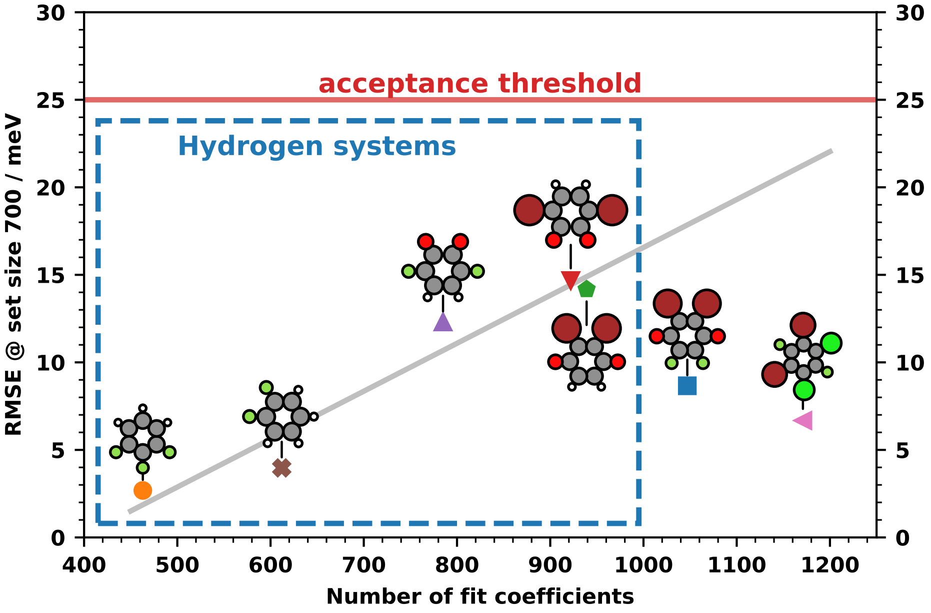


Figure 3.11.: RMSE versus number of fit coefficients for all level 2 systems and a training set size of 700. The grey line illustrates the linear correlation between RMSE and the number of fit coefficients that is observed for the test molecules with hydrogens. The prediction uncertainties for the non-hydrogen test molecules, *l3-tm1* and *l3-tm7*, clearly do not follow the same trend as they are noticeably lower than would be expected based on the corresponding number of fit coefficients for both of the two systems.

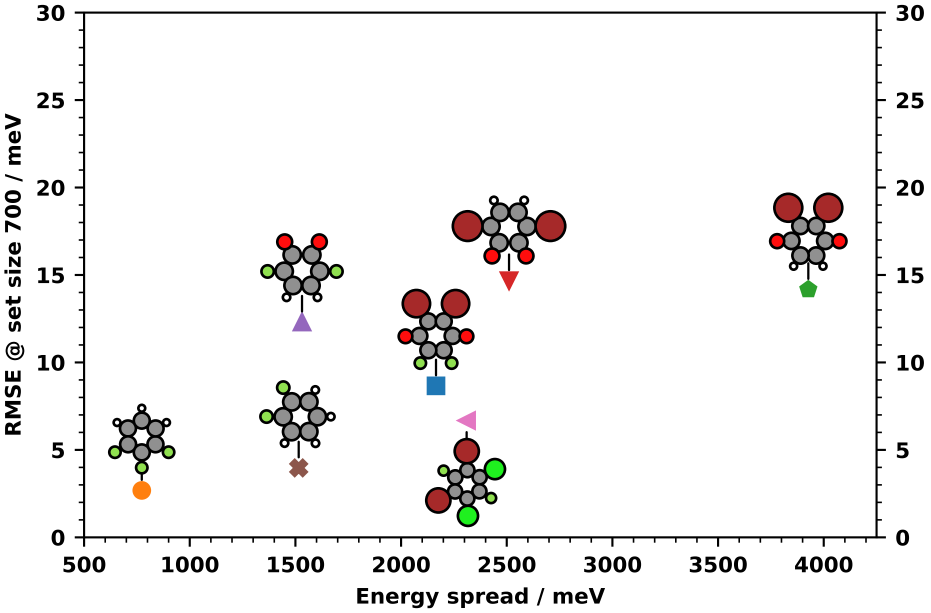


Figure 3.12.: RMSE versus energy spread of test set for all level 2 systems. All displayed prediction uncertainties are based on training the model with 700 configuration.

3.5. Level 3: Influence of substituent patterns

3.5.1. Objective and test systems

The third set of test systems shall answer the question, what, if any, influence the type of substituent pattern has on the learning capabilities. The thought process behind this level is as follows: As is known from organic chemistry, the relative arrangement of two (or more) substituents to a benzene ring influences the electrostatic potential of the resulting molecule. This, in turn, impacts macroscopic properties of the corresponding substance, such as its melting point, and has an influence on the bulk/surface structures that the material will form.

In theory, there a two aspects of substitution patterns that are of interest in the context of SAMPLE: 1) the type of molecular symmetry that they induce and 2) their effect on the PES of surface polymorphs. The first aspect is important due to the findings of level 1 (see section 3.3). The second aspect is based on the assumption that some substitution patterns will cause a more heterogeneous, or rougher, PES than others, which, at least in theory, increases the difficulty for the SAMPLE algorithm.

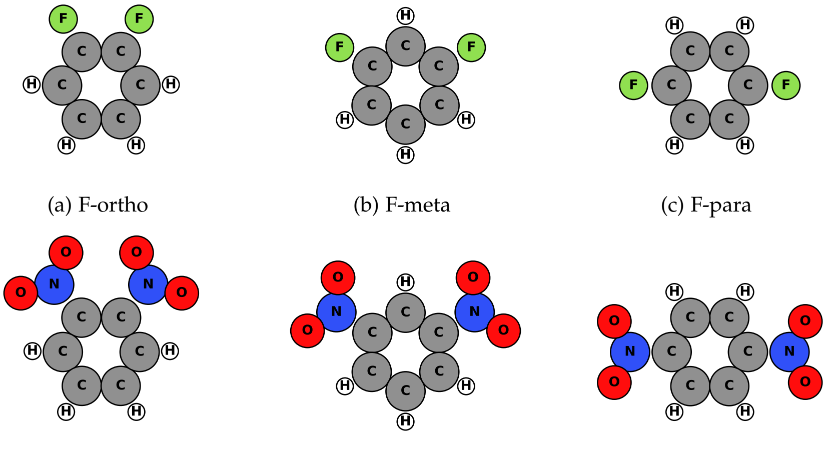
Now that the intent behind this level is clear, we need to design suitable test systems. The choice falls on two types of benzene derivatives, one with two fluorine atoms and the other with two amine groups (-NO_2). For both types, variants with an *ortho*, *meta* and *para* pattern are constructed, giving a total of 6 test systems. The geometries of the corresponding test molecules are shown in figure 3.13.

3.5.2. Results and findings

For this level we will skip showing all learning curves in favor of comparing the prediction performances at a fixed training set size, again 700 configurations. The double exponential fits based on equation 2.12 were carried out nonetheless and the estimated model errors plus the rest of the optimized fit parameters can be found in table 3.4.

Figure 3.14 shows the aforementioned results, separated into two subsets based on the used substituent group, F and NO_2, respectively.

Looking at the bottom half of figure 3.14, we see that the type of substituent pattern has no influence on the prediction accuracy for the F-substituted molecules. The RMSEs lie around 4 meV in all cases, with variations of less than 0.3 meV between the different patterns. Since we assume the uncertainty of the underlying DFT calculations to be on the order of $\approx$5 meV, these results show that in all three cases,



(a) F-ortho (b) F-meta (c) F-para

(d) NO_2-ortho (e) NO_2-meta (f) NO_2-para

Figure 3.13.: Overview of all level 3 test molecules, labeled by their identifier tag. Remark: test molecule *F-ortho* is also used for level 2 (see 3.4), where it is referred to as *tm7*.

we reached the lower limit in terms of prediction errors that SAMPLE is reasonably able to achieve.

Going on to the top half of figure 3.14, which features the results for the NO_2-substituted systems, we see a rather different picture. Here, it is noticeable that with an RMSE of 10.4 meV the molecule with the *ortho* pattern can be predicted with measurably lower accuracy than the *para* and *meta* systems with prediction uncertainties of 7.9 meV (*para*) and 8.2 meV (*meta*), respectively.

In a separate test, it is investigated if the number of feature vector entries per atom species pair, also known as the pairwise feature dimension (see 1.2.4), might have an influence on the results of the NO_2 subset. For this, the corresponding learning process is done three more times, once with all feature dimensions N_{AB} equal to 1, once with all N equal to two and once with half the original (maximal) feature dimensions. The resulting learning curves can be seen in figure 3.16 and show that as long as one uses more than one feature entry per pair, the learning performance is not influenced by the feature dimensions N_{AB}.

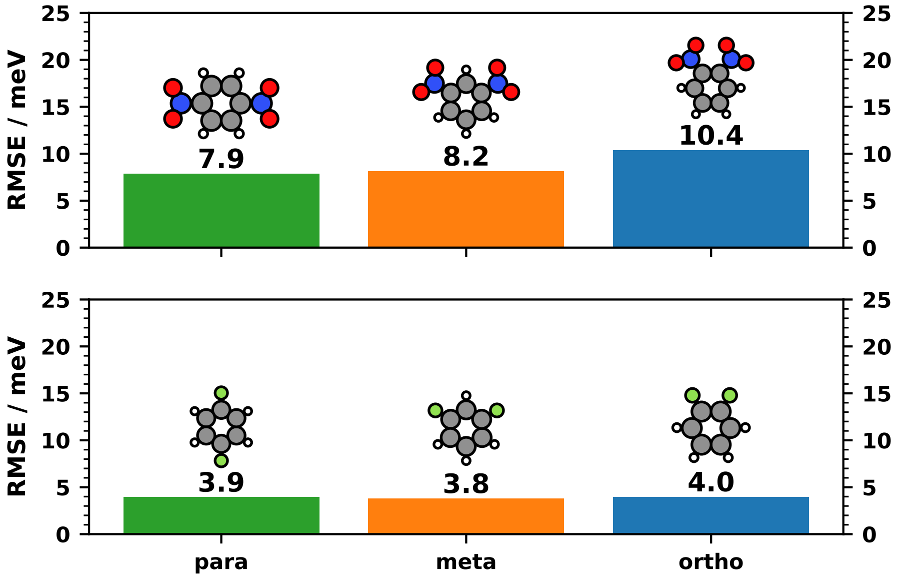


Figure 3.14.: Comparison of RMSE @ set size 700 for all level 3 systems

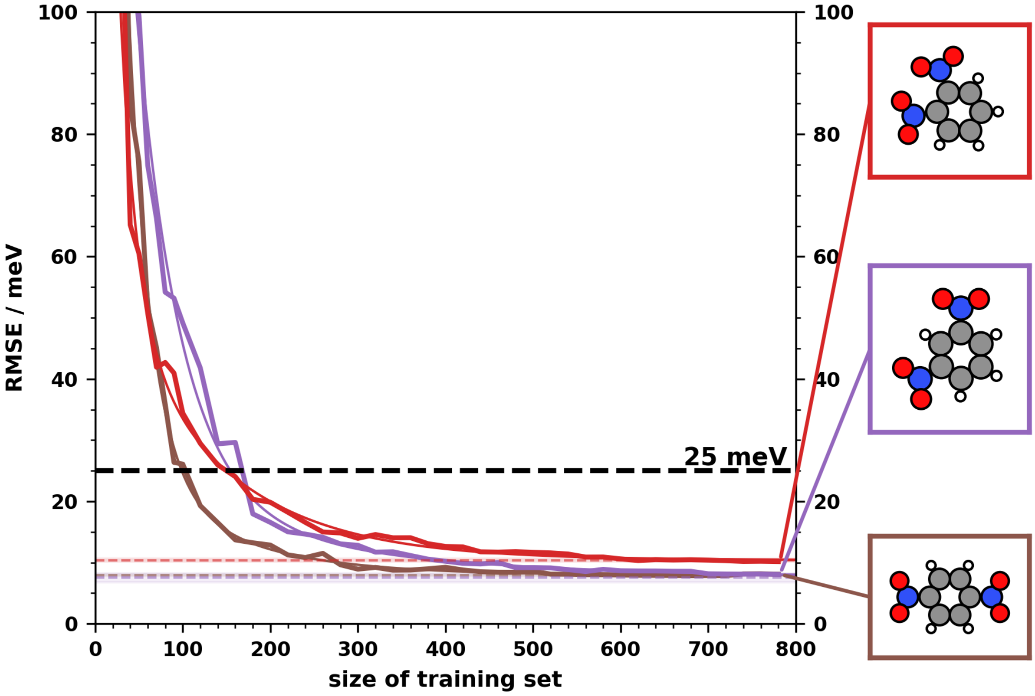


Figure 3.15.: Learning curves for NO_2 subset. Measured data is plotted as thick lines, with corresponding double exponential fits superimposed as thin lines. The heights of the estimated δE_{sys} are signified by dashed horizontal lines. It appears that an ortho pattern (red, top geometry) is slightly harder to learn than the other two (meta, para).

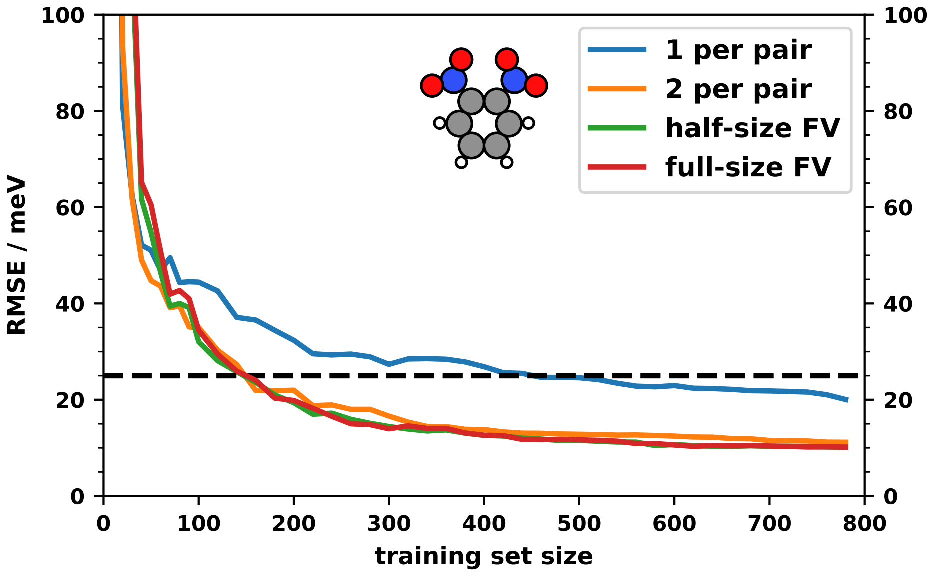


Figure 3.16.: Learning curves for test molecule *NO_2-ortho* as a result of using feature vectors with varying feature dimension. The blue line corresponds to the least detailed type of feature vector $\mathbf{f}$ with only one entry per species pair (A, B) $(N_{AB} = 1)$, while the yellow curve results from using $N_{AB} = 2$. The red line represents the most detailed $\mathbf{f}$, where each pair is assigned the maximum number of entries N_{AB}^{max}. For the yellow curve, we have $N_{AB} = N_{AB}^{max}/2$ for all pairs.

Table 3.4.: Optimized parameters for double-exponential fit model 2.12, based on level 3 learning curves. Listed are the average value for each parameter (mean) and the corresponding standard deviation (std).

		F-para	F-meta	F-ortho	NO_2-para	NO_2-meta	NO_2-ortho
δE_{sys} / meV	mean	3.76	3.7	3.91	7.93	7.51	10.4
	std	0.13	0.1	0.12	0.12	0.70	0.2
K_1 / meV	mean	258	290	240	335	203	230
	std	12	30	15	15	13	40
λ_1	mean	26.0	22	26.5	25.7	51	19
	std	0.8	2	1.5	1.2	5	3
K_2 / meV	mean	6.6	27	14	24	15	50
	std	1.2	5	3	5	7	5
λ_2	mean	190	112	135	114	230	121
	std	30	11	20	14	100	9

3.6. Level 4: Functional groups

3.6.1. Objectives

The goal for this array of test systems is to determine whether functional groups affect SAMPLE's prediction accuracy, and if so, why. In general, we want to find answers to the following questions:

- Which type of functional group can be learned well, which ones badly?
- Does the number of different functional groups in a molecule have an effect?
- Are there general trends?

In theory, we would expect that the more (different) functional groups we put on the molecule, the harder it should be to successfully describe all interactions between the molecules since the electrostatics are assumed to get more complicated.

In order to investigate the objectives above in more detail, a total of 16 test systems is designed for this level. The general approach to choosing the test systems is as follows:

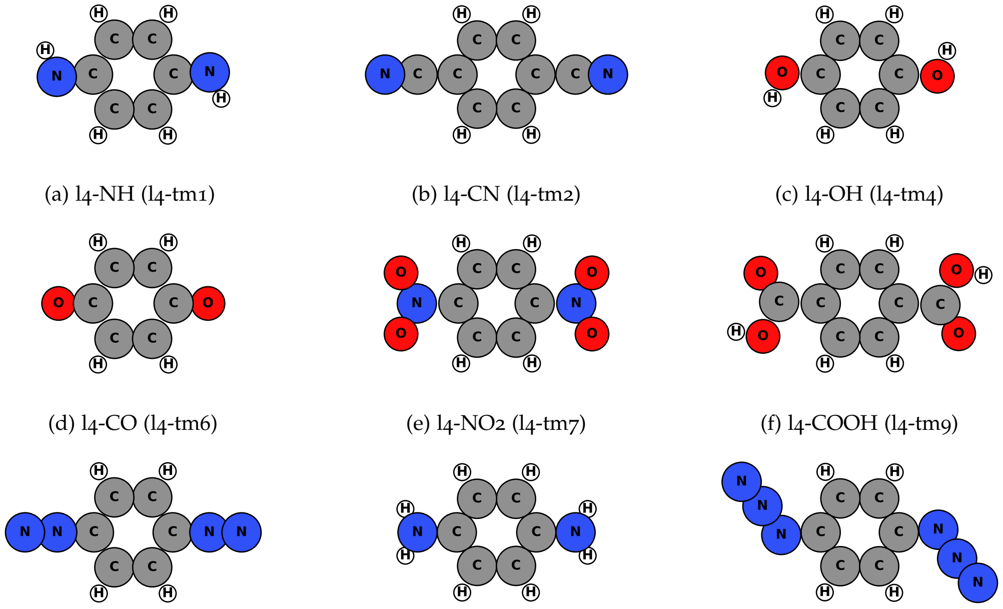
1. Choose a set of different *functional groups*, i.e. oxygen and/or nitrogen centered substituent groups
2. For each functional group, construct a test molecule where that functional group is the only type of substituent (besides hydrogen atoms). Maximize the molecular symmetry of the test molecule.
3. Construct test molecules with one or more different functional groups

Due to the large number of test systems, the discussion of the results is split into two parts which form the content of sections 3.6.2 and 3.6.3.

3.6.2. Type of functional group

First we look at systems where the test molecules only contain one type of functional group[5]. In total, this restriction yields a subset which consists of 9 systems, meaning we can compare 9 different types of functional groups. Table 3.6 presents a more detailed account of these *single-group systems* including, for instance, the names of the

[5] Hydrogens which are bonded to one of the central C-atoms are not identified as functional groups



(a) l4-NH (l4-tm1)

(b) l4-CN (l4-tm2)

(c) l4-OH (l4-tm4)

(d) l4-CO (l4-tm6)

(e) l4-NO2 (l4-tm7)

(f) l4-COOH (l4-tm9)

(g) l4-N2 (l4-tm11)

(h) l4-NH2 (l4-tm12)

(i) l4-N3 (l4-tm13)

Figure 3.17.: Compilation of all level 4 test molecules with only one type of functional group. Names inside brackets stem from the original enumeration of all level 4 systems, which also includes the multi-group systems (shown in figure 3.20).

involved functional groups and their structure. In addition to that, the test molecule geometries for these 9 systems are depicted in figure 3.17.

Now on to the results. Figure 3.18 features the learning curves for all single-group systems and the corresponding double exponential fit functions. The optimized fit parameters and their standard deviations are listed in table 3.5.

It is found that while the acceptance threshold of 25 meV can be reached for each type of functional group, there are significant differences in terms of learning speed and highest achievable prediction accuracy. For system *l4-NH2* (gray), SAMPLE only needs data from just 20 training configurations to reach γ_{accept} and only 80 to decrease its prediction accuracy down to 5 meV. A calculated model error of $\delta E_{sys}(NH2) = 0.74(1)$ meV is by far the lowest model error of all tested systems. It is approximately 8 meV lower than the one of the next best systems, whose final RMSEs all hover slightly under 10 meV.

A skeptic might now say that the extraordinary learning performance for *l4-NH2* must be too good to be true. Unfortunately, it seems as if in this particular case, the skeptic is at least partly correct, because when we calculate energy spreads for each system, that is we determine the range of monolayer formation energies for DFT dataset and compare them, as is shown in figure 3.19, we find a possible explanation

Table 3.5.: Optimized parameters of double-exponential fit functions 2.12, based on single group learning curves. Listed are the average value for each parameter (mean) and its standard deviation (std).

	δE_{sys} / meV		K_1 / meV		λ_1		K_2 / meV		λ_2	
System	mean	std	mean	std	mean	std	mean	std	mean	std
l4-NH_2	0.74	0.01	36	1	32.4	1.2	2.0	0.5	135	20
l4-N_3	7.7	0.2	224	20	33	3	88	7	142	6
l4-N_2	8.89	0.11	284	20	22.2	1.2	13	3	116	20
l4-NO_2	9.0	0.2	875	50	23	1	29	6	130	20
l4-CO	9.5	0.2	475	30	19	1	16	3	130	20
l4-COOH	10.0	1.4	230	30	70	12	40	20	250	100
l4-OH	11.4	0.3	240	70	9	2	76	3	180	7
l4-NH	12.4	0.2	330	40	19	2	68	3	152	6
l4-CN	16.0	0.2	293	13	27.3	1.4	28	5	119	14

for this behavior. Namely, that the energy spread for the *l4-NH2* is just around 250 meV, which is significantly smaller than for all the other single group systems, which typically have energy spreads of 1000 meV. Based on this, *l4-NH2* appears to be the least interacting system, with neither highly repulsive nor strongly attractive interactions. This suggests a shallow, smooth PES, which would facilitate learning. However, seeing that the system with the second smallest energy spread, *l4-OH*, ranks third from last in terms of model error, raises questions whether a different effect might contribute to the singular nature of the *l4-NH2* curve.

Focusing on the other test systems, we notice that the majority of learning curves eventually meet up into two clusters, one of which ends at around 10 meV, consisting of systems *l4-N3*, *l4-N2*, *l4-NO2* and *l4-CO*. The learning curves of the other cluster start out way higher, but the separation shrinks from around 15 meV at 300 configurations down to less than 5 meV.

While the molecules belonging to cluster 1 do not seem to be related in any obvious way, the case of cluster 2 does raise interest, because all corresponding test molecules include either an OH- or and NH-group. Both functional groups feature an unbound electron pair. This could mean that, at least near the mentioned groups, the electrostatics should be similar.

The test system *l4-CN* appears to be an outlier in the opposite direction to *l4-NH2*, as its prediction uncertainty remains highest, with an estimated δE_{sys} of 16.0(2) meV. Remarkably, the acceptance threshold is actually reached quite early for this system, at around 160 training configurations, much earlier than for the systems in cluster 2.

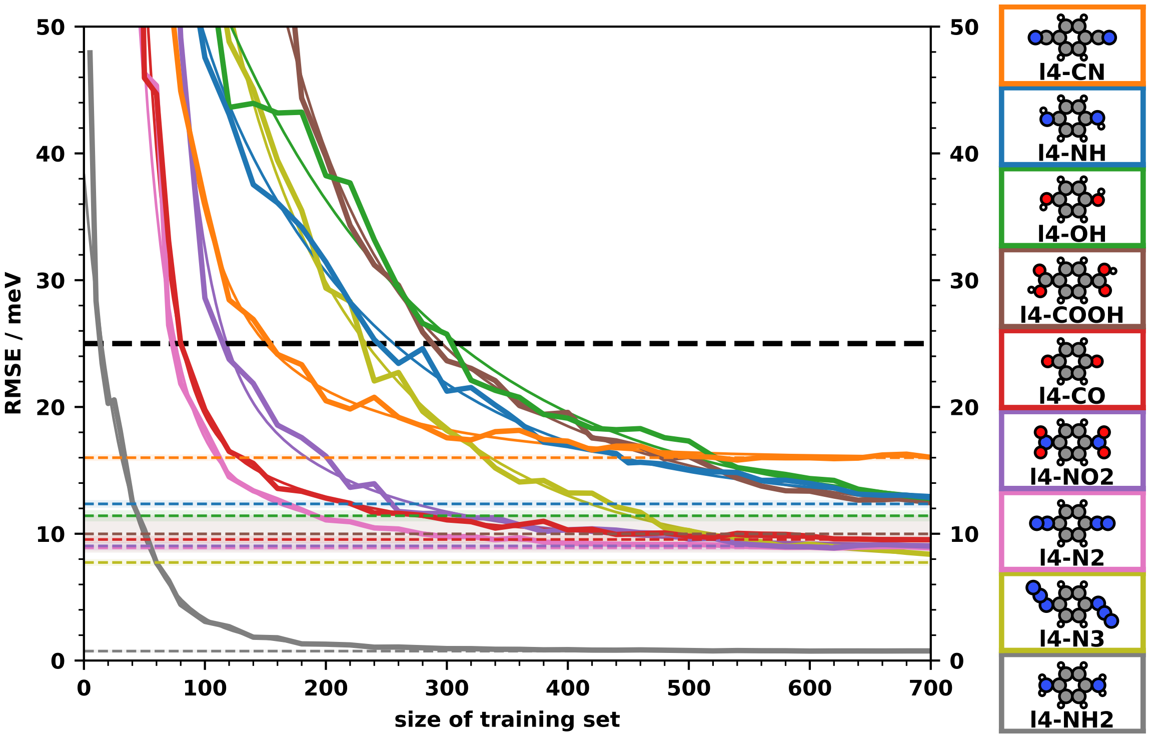


Figure 3.18.: Collection of learning curves for all test molecules with just one type of functional group. Double exponential fits are shown as thin curves. Estimated model errors δE_{sys} and their 1-σ confidence intervals are marked by colored dashed horizontal lines and semitransparent vertical spans. The thick black dashed line marks the acceptance threshold of 25 meV.

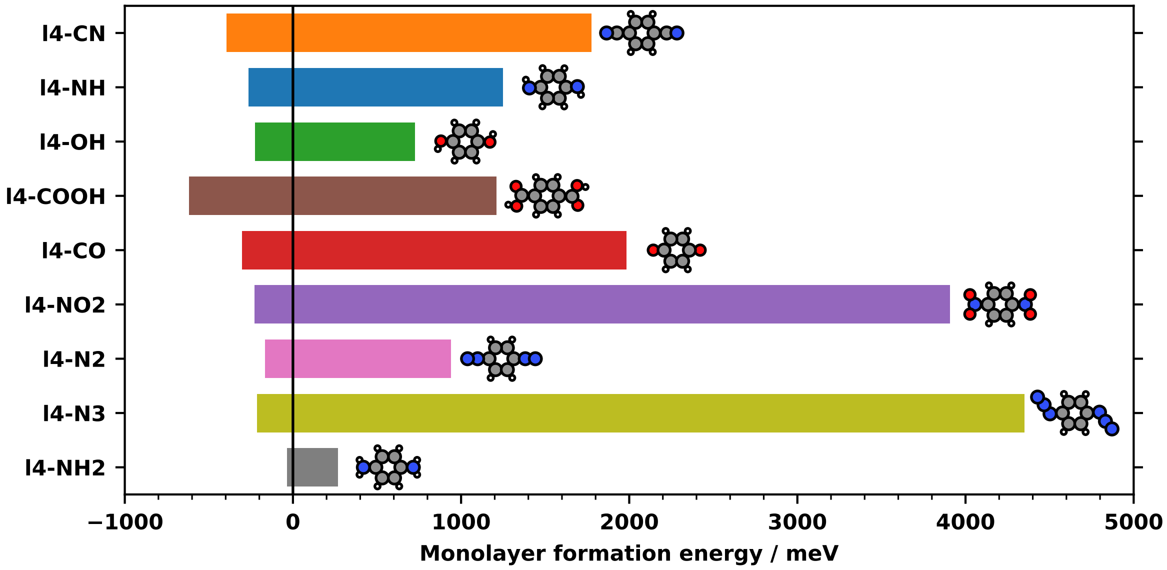
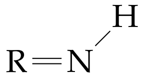
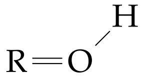
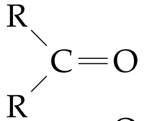
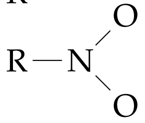
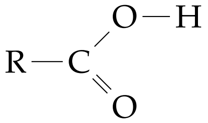
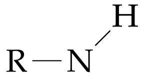
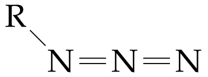


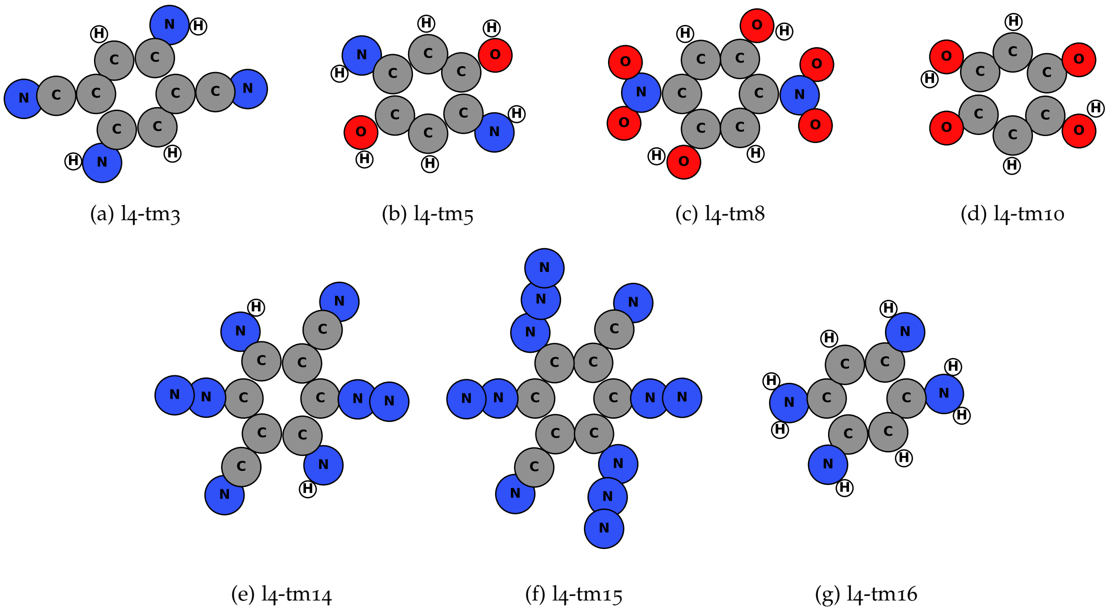
Figure 3.19.: Range of E_{MLF} of test set configurations for all single-group systems

Table 3.6.: Chart of single-group system properties. Includes information on the corresponding functional group. In the depicted chemical structure formulas, R stands for the connection to the rest of the molecule.

System identifier	Functional group identifier	Structural formula	Name of functional group
l4-NH	NH		imine-group
l4-CN	CN	$R-C\equiv N$	cyano-/cyanide-group
l4-OH	OH		hydroxy-group
l4-CO	CO		carbonyl-/ketone-group
l4-NO2	NO_2		nitro-group
l4-COOH	COOH		carboxyl-/carbonic acid group
l4-N2	N_2	$R=N=N$	diazo-/diazonium group
l4-NH2	NH_2		amino-group
l4-N3	N_3		azide-group

3.6.3. Combining multiple functional groups

This section builds upon the results for the single-group systems (see 3.6.2), adding test systems with more than one functional group. The geometries of these additional test molecules are depicted in figure 3.20.



(a) l4-tm3 (b) l4-tm5 (c) l4-tm8 (d) l4-tm10

(e) l4-tm14 (f) l4-tm15 (g) l4-tm16

Figure 3.20.: Overview of all level 4 test molecules with multiple types of functional group.

We begin by examining the learning curves of just these additional systems. They are plotted in figure 3.21. Although not shown in figure 3.21, double exponential fits were also carried out for these multi-group systems. The corresponding optimized fit parameters are listed in table 3.7.

Looking at figure 3.21, we see that the learning curves can be separated broadly into two groups based on their appearance. The first group includes systems *l4-tm10* and *l4-tm8*. System *l4-tm10* shows a lower required test set size ($x^{l4-tm10}_{required} \approx 140$) than *l4-tm8* ($x^{l4-tm8}_{required} \approx 200$), but in both cases, the prediction uncertainty ends up at around 5 meV.

The learning curves for the remaining multi-group systems all cluster together and lie 10 meV apart at most. Their estimated systematic errors cover a range of 10 meV to 15 meV and a training set size of between 250 and 350 configurations was needed for γ to reach the acceptance threshold. SAMPLE performed worst for system l4-tm5 (OH and NH substituents) and was only marginally better for the systems with nitrogen functional groups. Ranked by increasing learning performance, these were the systems *l4-tm15*, *l4-tm14*, *l4-tm3* and *l4-tm16*.

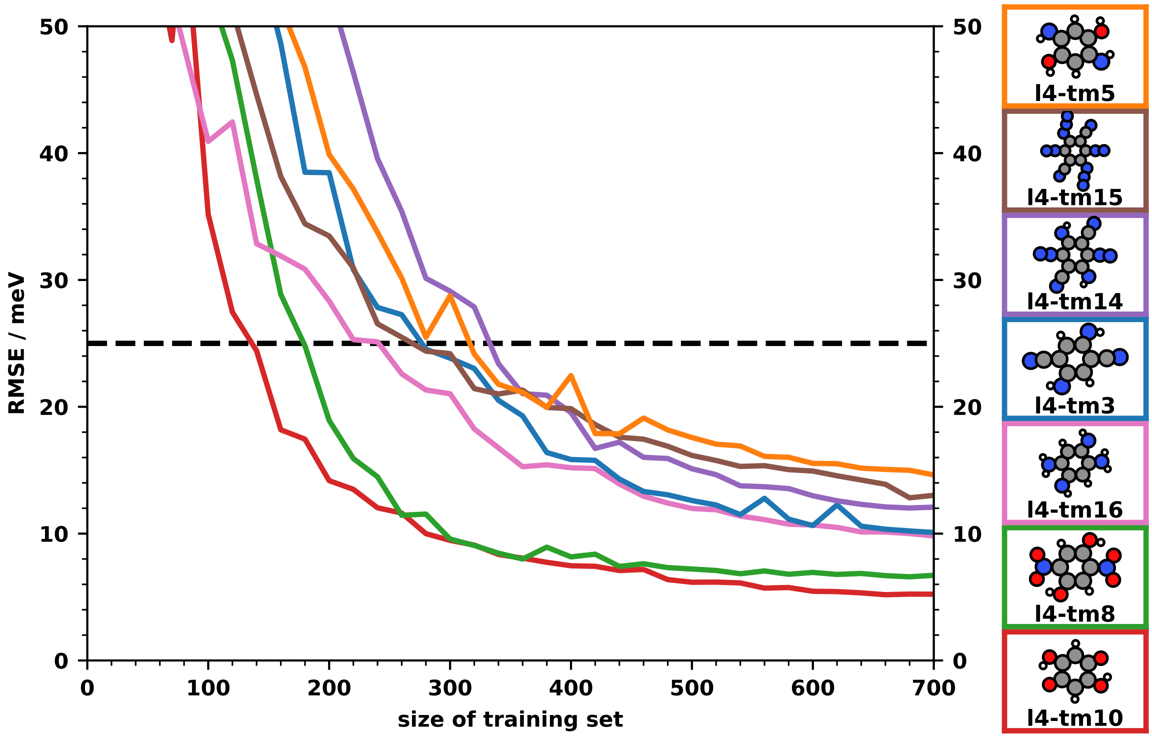


Figure 3.21.: Learning curves for all test systems with more than one type of functional group. Plots of the double exponential fit functions were omitted for sake of clarity. The corresponding optimized fit parameters are tabulated below in table 3.7. Geometries of the test molecules are depicted in the column on the right.

Table 3.7.: Fit parameters for multi-group systems. As all attempts to estimate the learning curves for the *l4-tm8* system with a double exponential fit function failed, it was decided to fall back to fitting the tail of the learning curve (set size ≥ 300) with $\gamma(x) = \delta E_{sys} + K_2 \exp\{-x/\lambda_2\}$ instead.

		l4-tm3	l4-tm5	l4-tm8	l4-tm10	l4-tm14	l4-tm15	l4-tm16
δE_{sys} / meV	mean	9.8	14.4	5.4	4.6	10.3	9.2	8.5
	std	0.4	0.3	0.4	0.4	1.2	0.9	0.8
K_1 / meV	mean	190	390	-	158	420	183	103
	std	25	60	-	14	20	12	15
λ_1	mean	37	13	-	45	38	46	18
	std	7	2	-	5	2	4	3
K_2 / meV	mean	115	132	11	23	150	40	57
	std	20	6	2	8	20	4	3
λ_2	mean	137	125	290	190	143	290	185
	std	12	4	70	40	14	40	15

Nitrogen systems

To finish the discussion of multi-group systems, we will look at one case were we take one test molecule with three different functional groups, *l4-tm15*, and compare its learning curve to those of the three test molecules that each carry one of *l4-tm15*'s functional groups. These three molecules are *l4-CN*, *l4-N2* and *l4-N3*. This combination of systems was chosen because it excludes functional groups with hydrogens or oxygens, such as NH_2, OH or NH and focuses nitrogen-based groups instead. The resulting comparison can be seen in figure 3.22.

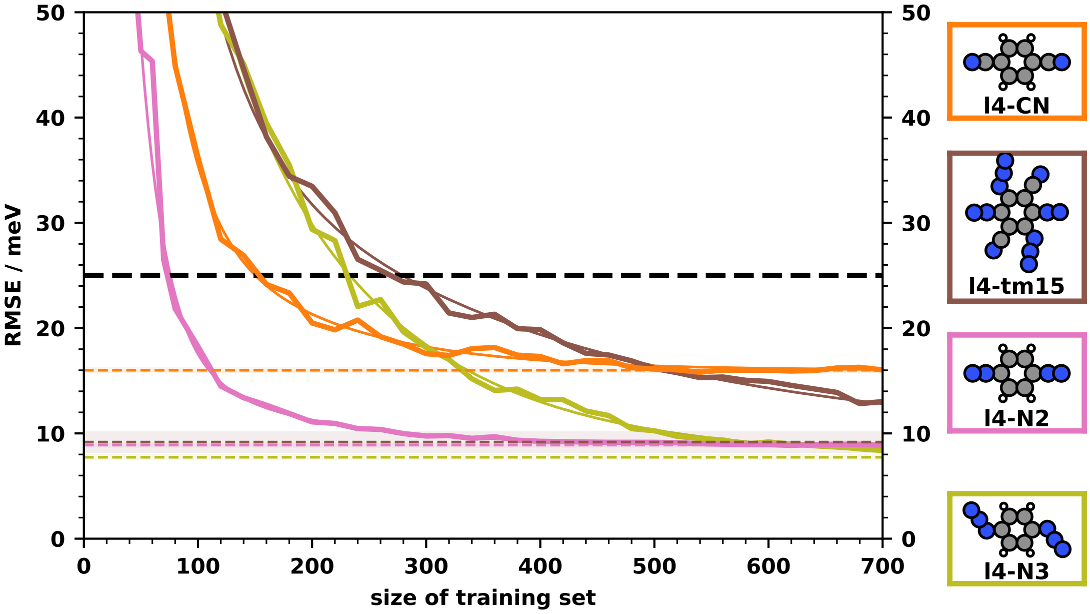


Figure 3.22.: Learning curves for single and multigroup nitrogen systems. Double exponential fit functions are drawn as thin lines. Model errors are denoted by dashed horizontal lines.

We observe that for the majority of the investigated set size range, i.e. from the beginning up to approximately 500 configurations, the multi-group system *l4-tm15* is the hardest one to learn with SAMPLE, as it shows a relatively slow decline in RMSE and subsequently requires the highest number of configurations to reach γ_{accept}. However, above a training set size of about 500 configurations, it is surpassed, in the negative way, by the learning curve of *l4-CN*, which essentially runs flat in this interval. With regard to the reason behind this behavior a clear prediction cannot be made based on this data, but it can be hypothesized that the learning performance of SAMPLE for the system *l4-tm15* is somewhat of a mix/superposition of the learning curves of the involved single-group systems. Whatever causes the learning curve of *l4-CN* to remain high, it seems reasonable to assume that the same factor also hinders training for multi-group system *l4-tm15*, since it basically contains *l4-CN*.

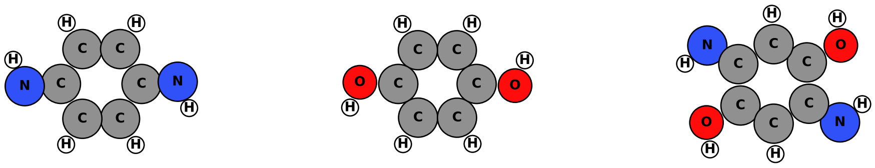
3.7. A look at transfer learning with SAMPLE

Based on the findings from levels 1 to 4, which are discussed at length in sections 3.3 to 3.6, we know that the number of training points that are needed to converge the energy model and achieve good prediction accuracies can be substantial. This might not be a huge problem when the test systems are small and the substrate is omitted, as is the case here, but when working with actual production systems, generating training data can become much more costly. The amount of available compute resources then puts a practical limit on the number of different systems that can be treated.

In the light of these facts, it would be very advantageous if training data and predictions for one system could be (re)used to predict the properties of another system. This transfer of information/knowledge from one system to the other is at the core of what is known as *transfer learning*.

So far, it has not been tested, whether SAMPLE can be used to facilitate transfer learning. The purpose of this section is to remedy this by investigating whether it is possible to predict properties of a test molecule based on prediction data from two other, albeit closely related molecules within the current framework of SAMPLE.

The systems that are chosen for this test are shown in figure 3.23 and fall into two categories: *source systems* and the *target system*. The former set is formed by test molecules *tm-NH* and *tm-OH*, which each feature functional groups in a para configuration. While *tm-NH* carries two imine groups (-NH), molecule *tm-OH* features two hydroxy groups (-OH).



(a) NH (b) OH (c) NH+OH

Figure 3.23.: Test systems used for a demonstration of transfer learning with SAMPLE. Data from the *source systems*, *tm-NH* (a) and *tm-OH* (b), is used to predict properties of the *target tm-OH+NH* (c).

The *target system*, molecule *tm-NH+OH*, combines both types of functional groups that are found in the source systems. The relative orientation of imine and hydroxy groups in the target system represents the energetically most favorable arrangement.

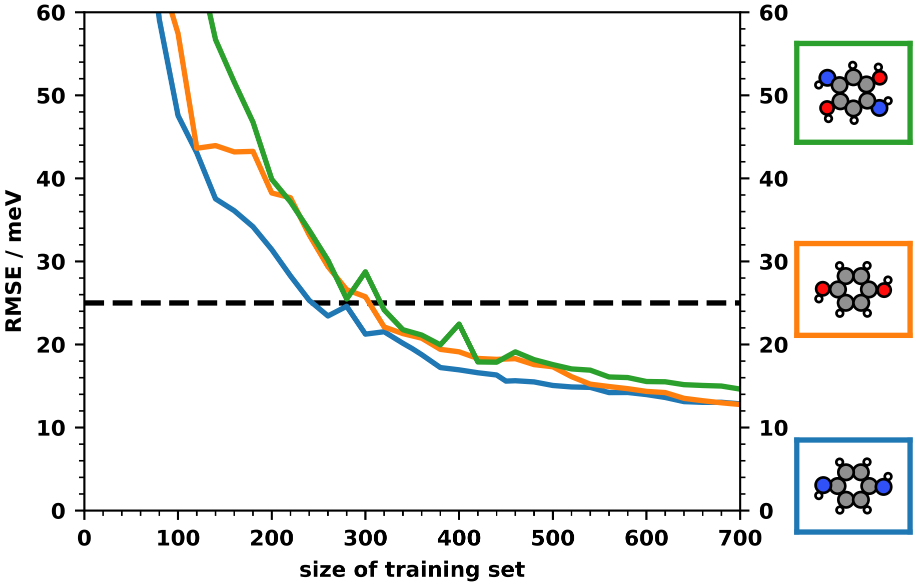


Figure 3.24.: Learning curves for source (*tm-NH*, *tm-OH*) and target (*tm-NH+OH*) systems used in the transfer learning test. The dashed, black horizontal line marks the acceptance threshold (25 meV).

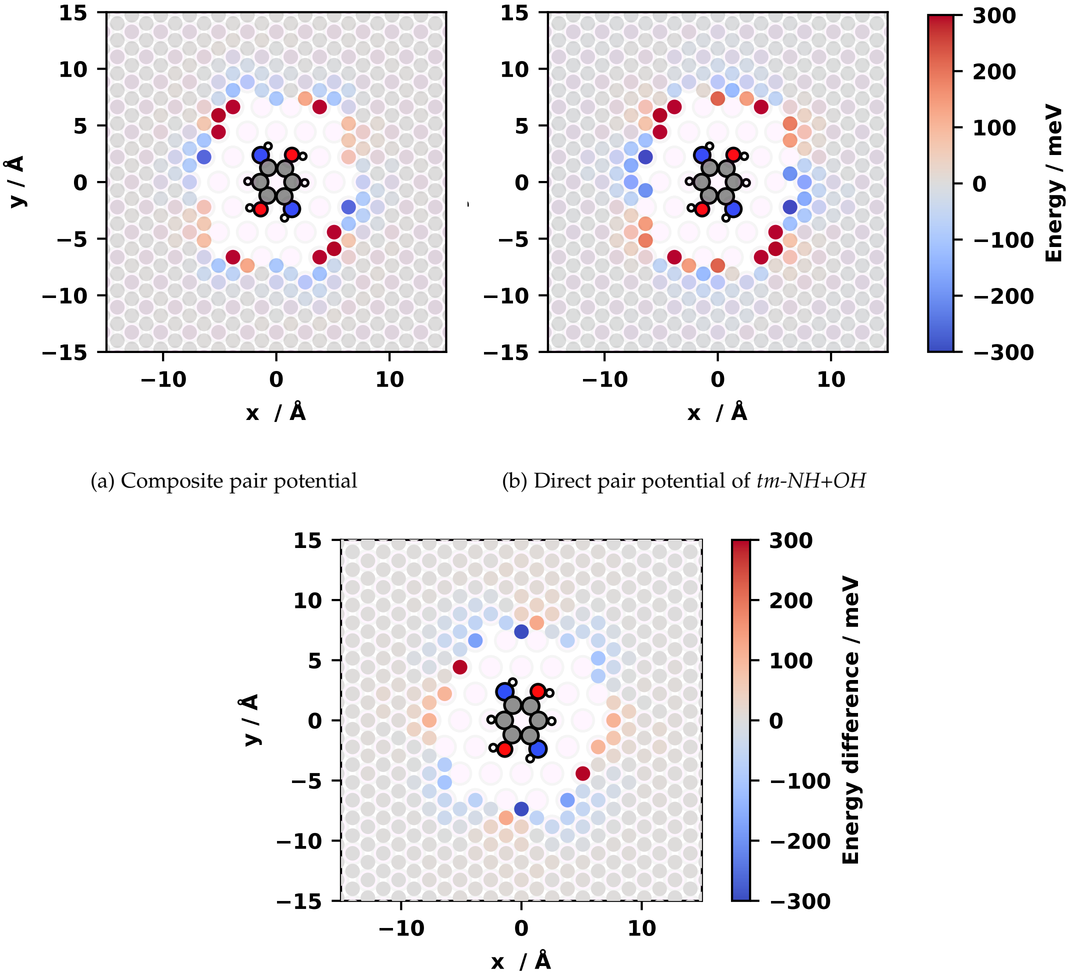
As preparation for the actual test, SAMPLE was applied to all three systems, using the hyperparameters and lattice settings as described in sections 2.3 and 2.4, respectively. The models for each system were then trained using the approach described in section 2.2.2. From the resulting learning curves, shown in figure 3.24, we see that learning performance is similar for all three systems, with RMSEs hitting the acceptance threshold of 25 meV above a training set size of 340 and reaching values between 14 meV and 16 meV for the biggest training sets (800 configurations).

Based on the above, this setup seems to be a good candidate to try transfer learning. The property which we are going to focus on are pair potentials, i.e. a mapping of the strength of the interaction between two molecules as a function of their relative position. Information on how these pair potentials are calculated is given in section 2.2.3.

Out of all local geometries of both *tm-NH* and *tm-OH*, we choose those that best match the target system in terms of orientation. Then we predict pair potential maps, separately for *tm-NH* and *tm-OH*, based on the chosen local geometries. These are then summed up point-by-point, with a small restriction: for consistency, we only include those data points that appear in both pair potentials. The reason, why not all positions are featured in both maps is grounded in two factors: (A) The minimal distance threshold for (N, H) is larger than for (O, H) and (B) the selected *tm-NH* geometries are rotated by 60° with respect to the *tm-OH* geometries. In combination, these factors cause some positions of the second molecule to be dismissed as being "too close" to the center molecule for *tm-OH*, but not for *tm-NH* and vice versa.

This carefully assembled pair potential is then compared against the pair potential of the target system. This comparison is shown graphically in the upper row of figure 3.25, with the composite pair potential on the left and the target pair potential on the right. Looking at the results, we see that, at least qualitatively, the composite potential can reproduce the target potential reasonably well.

When comparing the two pair potentials quantitatively, however, it is noticeable that point-wise differences can locally be rather large, i.e. up to 300 meV, as illustrated in figure 3.25d.



(a) Composite pair potential

(b) Direct pair potential of *tm-NH+OH*

(d) Difference between composite potential and direct pair potential of the target system

Figure 3.25.: Comparisons between composite and direct NH+OH pair potential

4. Summary and Conclusion

The primary focus of this thesis was to analyze the learning performance of the SAMPLE method with regard to complex molecular interactions. In oder to achieve this, the method was applied to four different classes of test systems, each designed to cover specific aspects of molecular interactions. The specific test systems consisted of free-standing monolayers of small organic test molecules, each derived from benzene by adding one or more substituents.

The performance of the algorithm was analyzed in terms of the accuracy with which it could predict system properties such as the formation energies of the tested monolayers. It was tracked how much training data was needed to push the root mean square error of said monolayer formation energy down to an acceptable level (γ_{accept}), which was set to 25 meV per molecule. In addition to that, it was attempted to estimate the systematic model error δE_{sys} of SAMPLE by fitting the measured RMSE vs. training set curves with a custom fit function and extrapolating them to infinite amounts of training data.

Carrying out this testing methodology for all systems, it was found that SAMPLE was always able to achieve the acceptance threshold of 25 meV per molecule. The amount of training data needed to reach γ_{accept} also stayed below 500 training configurations, which was deemed acceptable. Thus, the SAMPLE method can be seen as a robust structure search algorithm under the criteria for robustness defined in this work.

The first set of test systems featured molecules with different symmetries and only halogen atoms as substituents. The results showed that molecules with lower molecular symmetry were generally harder to learn, due to the higher number of fit parameters that they contain.

In the second tranche of test systems, also hydrogen atoms were allowed as substituents to look at the influence of hydrogen bonds. Comparing the observed learning performance with that of similar systems without hydrogens led to the conclusion that the existence of hydrogen bonds increases the learning difficulty. While the cause for this behavior is not fully understood, it is thought to be connected to the roughness of the potential energy surface of the individual test molecules that form the monolayer, with hydrogen-carrying molecules sporting a comparably rough potential energy surface due to the inclusion of substituent atoms with high as well as with low electronegativity.

With the third class of test systems it was investigated whether substitution patterns had an effect on the learning performance. It featured systems for two types of substituents, F and NO_2, which were arranged in ortho-, meta- and para-patterns. It was found that, when correcting for effects attributable to symmetry differences, the type of pattern made no significant difference.

The last batch of test systems focused on functional groups with the aim to determine which type of functional group performs best and whether or not trends can be deduced. The results for this class of test systems unfortunately proved much harder to interpret and classify and no clear tendencies or trends could be established.

Additionally, the large number and variety of systems also made it possible to try out transfer learning in a proof-of-concept style experiment, by trying to recreate the pair potential of a target system from the potentials of two source systems. This was successfully demonstrated, highlighting a potential route to make SAMPLE much more efficient, especially once it gains wide-spread use and shared databases of training data become available.

On a more general level, the large number of test sets needed to explore all 4 areas of interest underscored the importance of a systematic and structured workflow. This encompassed both the design of the test sets and the choice of the model hyperparameters. Designing the testing methodology also required a significant amount of consideration towards the question of how to best measure learning performance. This concerned both the test set selection process and the different ways to interpret learning curves.

Coming to a conclusion and summing up all of the above, this thesis finds SAMPLE to be a robust and reliable structure prediction algorithm, which has a lot of potential for use and further development.

Appendix

Appendix A.

Determination of hyperparameters

A.1. Calculated minimal distance thresholds

Table A.1.: Minimal distance thresholds. Calculated via equation 2.15 with parameters $\beta_{attr} = 0.55$ and $\beta_{rep} = 0.65$, with the exception of d_{min}^{OH}, which was chosen manually in accordance with the corresponding minimal distance sweep, which is depicted in figure A.2.

Atom A	Atom B	d_{min}^{AB} / Å	Atom A	Atom B	d_{min}^{AB} / Å
Br	Br	2.41	Cl	F	2.09
Br	C	2.31	Cl	H	1.62
Br	Cl	2.34	Cl	N	2.15
Br	F	2.16	Cl	O	2.13
Br	H	1.68	F	F	1.91
Br	N	2.21	F	H	1.47
Br	O	2.19	F	N	1.96
C	C	2.21	F	O	1.94
C	Cl	2.24	H	H	1.56
C	F	2.06	H	N	1.51
C	H	1.89	H	O	1.35
C	N	2.11	N	N	2.02
C	O	2.09	N	O	2.00
Cl	Cl	2.27	O	O	1.98

A.2. Minimal distance threshold sweeps

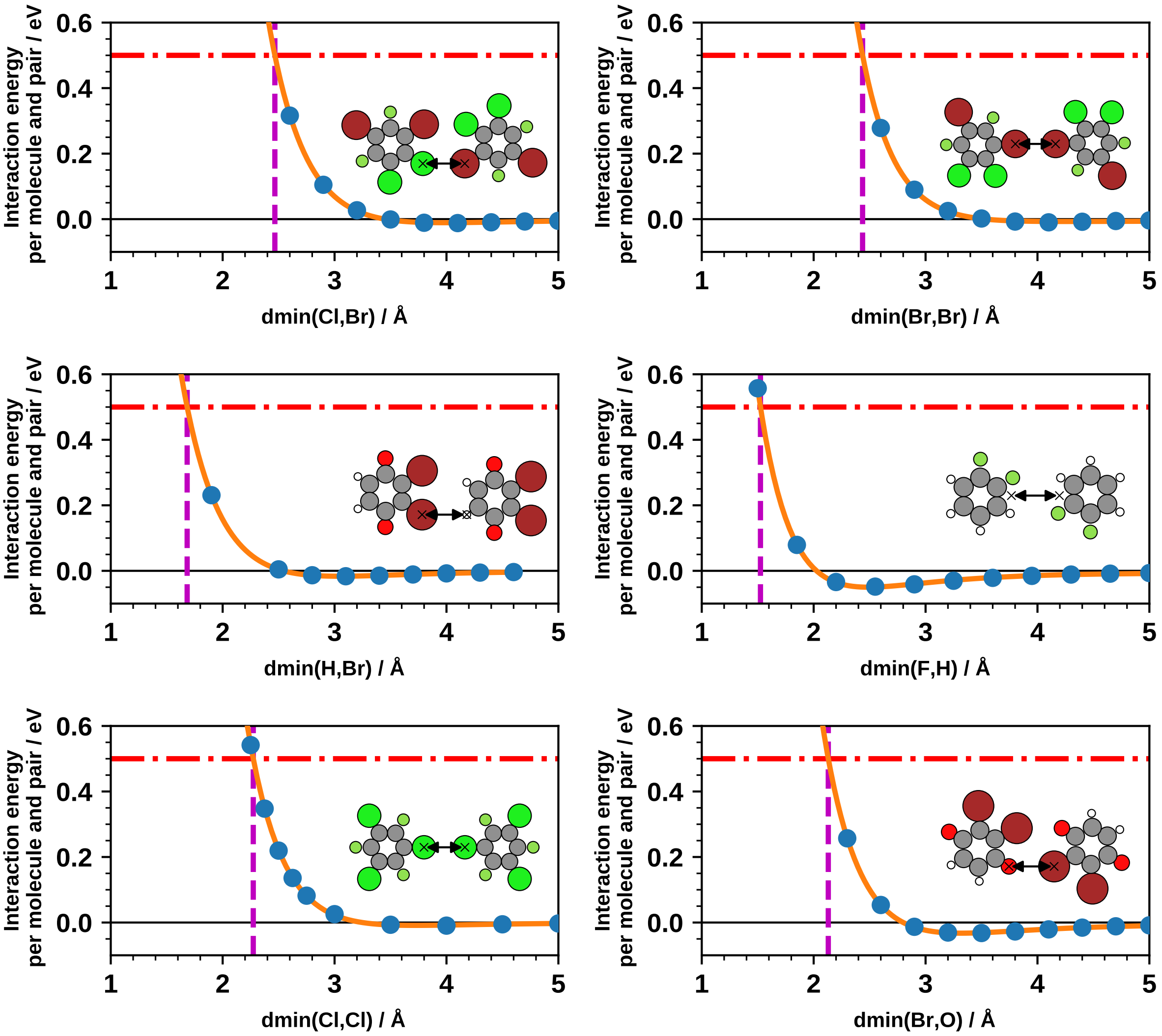


Figure A.1.: Distance threshold sweep curves, page 1

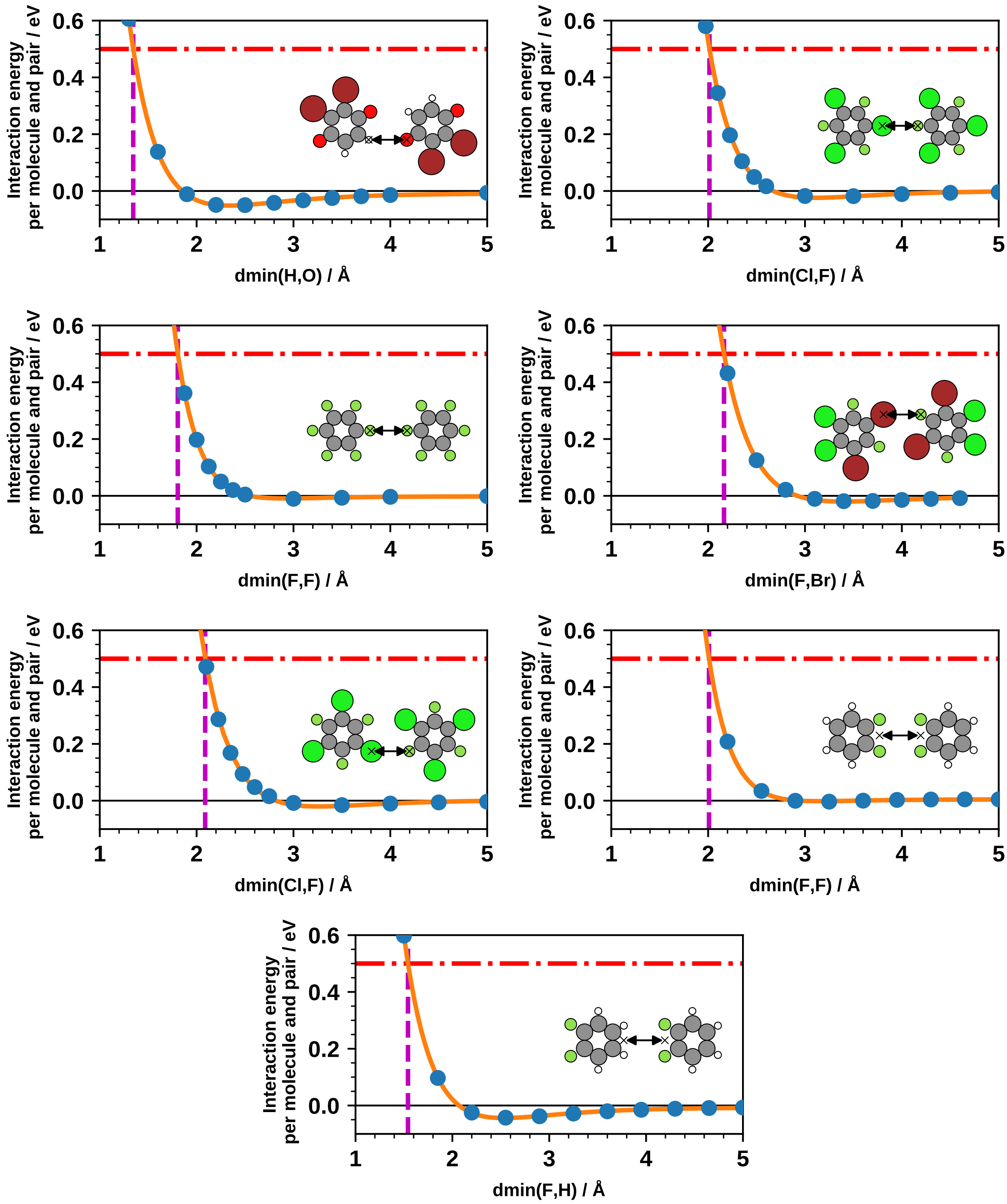


Figure A.2.: Distance threshold sweep curves, page 2

Appendix B.

Listings

B.1. FHIaims control file for periodic calculations

```
# General Settings:
xc     pbe
spin   none
charge 0.0
relativistic  atomic_zora scalar
occupation_type gaussian 0.01
k_grid <Nx> <Ny> 1

# Convergence criteria:
sc_accuracy_forces    0.001
sc_accuracy_etot      1e-06
sc_iter_limit     100

# Mixer:
mixer pulay
charge_mix_param 0.3
preconditioner none

# Corrections:
compensate_multipole_errors  .true.
vdw_correction_hirshfeld     .true.
use_dipole_correction    .true.
```

For each calculation, the placeholders `<Nx>` and `<Ny>` are substituted for the actual number of k-points in the x- and y-direction, which depend on the size of the corresponding unit cell.

B.2. FHIaims control file for non-periodic calculations

```
# General Settings:
xc     pbe
spin   none
charge 0.0
relativistic  atomic_zora scalar
occupation_type gaussian 0.01

# Convergence criteria:
sc_accuracy_forces   0.001
sc_accuracy_etot     1e-06
sc_iter_limit     100

# Mixer:
mixer  pulay
charge_mix_param 0.3
preconditioner none

# Corrections:
compensate_multipole_errors  .true.
vdw_correction_hirshfeld     .true.
```

Acronyms and Symbols

Acronyms

RMSE root mean square error

DFT density functional theory

SAMPLE **S**urface **A**dsorbate Poly**m**orph **P**rediction with **L**ittle **E**ffort

PES potential energy surface

GPR Gaussian process regression

FHIaims Fritz Haber Institute *ab initio* molecular simulations

PBE Perdew-Burke-Ernzerhof

fcc face-centered cubic

LG local geometry

SILG symmetry-inequivalent local geometry

KS Kohn-Sham

SC supercell

SCF self-consistent field

HP hyperparameter

FV feature vector

Symbols

E_{MLF} monolayer formation energy

γ prediction uncertainty