Machine-learned interatomic potentials for the syngas conversion on Rhodium

Masterarbeit
aus dem Fachgebiet Theoretische Chemie

von B.Sc. Lena Sauerland
geboren am 31.07.1997 in Erkelenz

für die Masterprüfung in Chemie an der
Ludwig–Maximilians–Universität München

Beginn der Masterarbeit: 15.03.2021
Beim Prüfungsausschuss eingereicht am: 29.10.2021

Erklärung

Ich, Lena Sauerland, versichere, dass ich die vorgelegte Masterarbeit am Fritz-Haber-Institut der Max-Planck-Gesellschaft im Department der Theoretischen Chemie, geleitet von Prof. Dr. Karsten Reuter, in der Gruppe von Dr. Johannes Margraf unter der Anregung und Anleitung von M.Sc. Sina Stocker selbständig durchgeführt und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe.

Berlin, den 29.10.2021

L. Sauerland

Lena Sauerland

Erstprüfer:	Prof. Dr. Hubert Ebert
Zweitprüfer:	Prof. Dr. Karsten Reuter

Abstract

The kinetics and thermodynamics of chemical processes such as heterogeneous catalytic reactions often depend on tremendously complex reaction networks, whose exploration quickly exceeds computational possibilities. The usage of first principle methods to identify and calculate the relevant reaction steps therefore becomes unfeasible, requiring new methods to overcome these challenges. Over the last decade, different machine-learning methods have been developed and applied to chemical problems. These methods range from neural networks to kernel-based methods such as kernel ridge regression or the training of Gaussian approximation potentials (GAPs), which is the machine-learning method used in this work.

Besides the usage of machine-learning to overcome computational barriers, another aspect in the handling of complex reaction networks is their reduction to the most important reaction steps and intermediates. Prerequisite for the reduction of network complexity is the knowledge of the appropriate energy landscape. Finding the global minimum of a chemical system can give deep insights into the relevant conformations for each involved structure bridging the gap to build up the energetic environment of a catalytic reaction.

Therefore, in this work a method is developed to pool forces of both machine-learning and a distinct approach to find the global minima of the involved adsorbates in the syngas conversion on catalytic rhodium surfaces. As part of the work, an iterative training workflow for the training of a GAP is developed. Using this workflow, a system-specific potential is trained for the syngas conversion on Rhodium surfaces. The developed potential is then applied to the global optimization of the involved educts, intermediates and products emerging in this specific system - the syngas conversion on Rhodium.

List of abbreviations

AE atomisation energy

BFGS Broyden–Fletcher–Goldfarb–Shanno

BO Born-Oppenheimer

DFT density functional theory

FPS farthest point sampling

fcc face centered cubic

GAP Gaussian approximation potential

GGA generalized gradient approximation

GPR Gaussian process regression

HF Hartee-Fock

hcp hexagonal closed packed

LDA local density approximation

LSDA local spin density approximation

MAE mean absolute error

ML machine learning

MHM minima hopping method

MD molecular dynamics

PCA principal component analysis

kPCA kernel principal component analysis

PBE Perdew–Burke–Ernzerhof

revPBE revised Perdew–Burke–Ernzerhof

PES potential energy surface

RMSD root mean square deviation

SOAP smooth overlap of atomic positions

vdW van der Waals

Contents

Abstract III

List of abbreviations IV

1 Introduction 1

2 Theoretical background 5

2.1 Description of potential energy surfaces 5
2.1.1 Density functional theory 6
2.1.2 Gaussian approximation potentials 9
2.2 Exploration of potential energy surfaces 15
2.2.1 Local optimization 16
2.2.2 Global optimization by minima hopping 17
2.3 Representation of potential energy surfaces 20
2.3.1 Farthest point sampling 20
2.3.2 Kernel principal component analysis 21

3 Development of a machine-learned interatomic potential 22

3.1 Methods and computational details 22
3.1.1 Workflow development 23
3.1.2 Computational details 25
3.2 Results and discussion 31
3.2.1 Initial training approaches 31
3.2.2 Hyperparameter finetuning 36
3.2.3 Final training approach 39

4 Application of the potential to global optimization 41

4.1 Methods and computational details 41
4.2 Results and discussion 46
4.2.1 Minima hopping from new adsorption sites 46
4.2.2 Kernel principal component analysis 52
4.2.3 Reduced reaction network development 54
4.2.4 Minima Hopping of unknown structures 57
4.2.5 Applicability assessment and outlook 63

5 Summary **65**

Bibliography **67**

Appendices **74**

A GAP training details **74**

B GAP application details **83**

Chapter 1

Introduction

Heterogeneous catalysis is one of the main drivers in industrial chemical processes and also one of the central aspects towards more sustainable chemicals. As most of the chemicals produced and used nowadays are in contact with at least one catalyst during their lifetime, catalytic efficiency plays an important role in the reduction of energy consumption. Therefore, exploring the most active and selective catalysts remains one of the major tasks in designing a sustainable future. [1, 2]

One key reagent in industrial chemistry is synthesis gas (also named syngas) which is a mixture of carbon monoxide and hydrogen that is used to synthesize many basic chemicals and synthetic fuels with the help of catalysts [3]. Whereas several catalysts are able to convert syngas to different products with high activity, many of them lack in catalytic selectivity, characterising the catalysts affinity towards a specific product. Rhodium has been identified as one of the most promising catalysts with selectivity towards ethanol, which is one of the favored products, although the rationale of this selectivity is still deficient. [4] Therefore, computational methods such as density functional theory (DFT) are broadly used to gain theoretical insights about the energetic properties of these systems, represented by their potential energy surface (PES) [5].

However, the accurate determination of the mechanisms leading towards the desired products is a major barrier. The reason is that even simple heterogeneous catalytic reactions depend on tremendously complex reaction networks, whose exploration quickly exceeds computational possibilities when approached with DFT. In particular, not all of the individual steps in the network contribute to the overall mechanism, leading to the need of systematic network reduction approaches differing from traditional chemical intuition. Thermochemistry can facilitate the reduction by calculating reaction energies between educts and products showing up the most likely reaction path and identifying thermodynamically inaccessible network regions with only little impact on the actual mechanism. [6–8]

Nonetheless - even after focusing on the main parts of the network - the simulation of rare events and reaction dynamics on a first principle level by ab-initio molecular dynamics (MD) is impossible. Therefore, empirical potential based MD simulations as well as statistical, coarse-grained microkinetic models like kinetic Monte Carlo simulations have been widely employed in this field. These methods are able to represent more complex, dynamical systems on large timescales, whereby a loss in accuracy has to be accepted. [9]

In order to overcome this loss in accuracy, different machine learning (ML) approaches have been developed and successfully applied to a broad field of chemical problems. These approaches range from neural networks to regression models, whereby particularly the coupling of quantum chemistry and ML is a promising technique, striving for ML-accelerated simulations with chemical accuracy. [10–13] Within the field of regression models, ML interatomic potentials such as Gaussian approximation potentials (GAPs) have been developed and applied to many different chemical systems in literature [14, 15]. In this way, for instance the geometry optimization and dynamics simulation of catalytic systems is enabled at a fraction of cost of a full DFT calculation [16].

To further reduce the computational complexity within a catalytic network, the knowledge of the global minima of the different adsorbates on the catalytic surface is important in order to reveal the most important adsorption sites and conformations as well as their appropriate energies. Therefore, several algorithms have been developed aiming to find the global minimum, including the minima hopping method (MHM). [17]

This work aims to develop a method to find the global minimum of the involved educts, intermediates and products of the syngas conversion on Rhodium in order to develop a reduced reaction network for this catalytic system. For this purpose, the method of GAP is used to train a system-adapted interatomic potential, which is then applied to the minima hopping of single adsorbates on plain Rhodium(111) and stepped Rhodium(211) surfaces.

First, the scientific context and significance of the syngas conversion on Rhodium is outlined. Second, the relevant theoretical background is introduced. Afterwards, the development of the GAP training workflow is depicted. This is followed by the application of the trained potential towards the exploration of the global minimum landscape of the syngas conversion system. At the end, the results are summarized and remaining questions are pointed out.

Syngas conversion on Rhodium

The conversion of syngas into liquid fuels such as ethanol has a considerable potential when striving to create sustainable chemical and energy supply [18]. Syngas is a mixture of carbon monoxide and hydrogen, which is, at present, primarily produced from fossil resources as coal or natural gas and converted into different products required in industrial chemistry. Some of the industrially most relevant processes utilize syngas for example for the synthesis of liquid hydrocarbons in the Fischer-Tropsch process, as a hydrogen supply in the production of ammonia via the Haber-Bosch process or the synthesis of methanol [19].

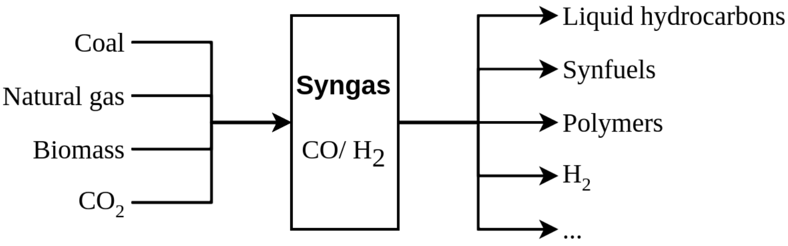


Figure 1.1: Simplified visualization of the industrial production and utilization of synthesis gas.

Besides syngas production from natural, non-renewable feedstock, different production routes arose using renewable feedstock such as biomass [20] or CO_2 [21]. This makes syngas remain an important building block for a sustainable chemical future, further amplified by its ability to convert into various products.

One of the most favored products is ethanol, especially because of its possible utilization as a fuel [22]. As bioethanol synthesis from sugar fermentation suffers from scalability due to its competition with the usage as nourishment, alternative synthesis strategies are demanded [23]. Through coupling of syngas production from renewable feedstock on the one hand and conversion into ethanol and other higher alcohols on the other hand, the production of a biofuel - also called synfuel due to its origin from syngas - becomes possible. This builds up a broad opportunity for partial or even overall substitution of petroleum-derived fuels [24].

However, in order to specifically produce a certain product from syngas conversion, knowledge on the reaction mechanism is required. This knowledge is gained by coupling of experiments and theoretical calculations. It has been early found, that the three most relevant steps in the syngas conversion towards C_{2+} oxygenates are first the CO activation, second the C-C coupling and third the hydrogenation reaction [25]. Thereby, different catalysts dissimilarly promote one or more steps resulting in different favored products.

Rhodium has shown significant selectivity towards C_{2+} oxygenates compared to other catalysts such as copper [26]. Selectivity, in general, is a catalysts affinity towards the production of a specific product. Most of the chemical reactions possibly yield various products, whereas the thermodynamically stable ones prevail. By the usage of catalysts, reaction barriers are lowered and intermediates are stabilized, which reasons the selectivity of different catalysts towards different prevailing products. [27]

In case of the syngas conversion on Rhodium, Yang et al. [4] revealed an intrinsic structure sensitive selectivity of the catalyst. Thus, different Rhodium surfaces show up different selective products. The researchers found a high selectivity of Rh(111) towards C_{2+} products with acetaldehyd instead of ethanol as the prevailing one. However, by addition of co-atoms such as Iron, the selectivity could possibly be shifted towards ethanol [28]. The more active Rh(211) surface did in contrast show up a selectivity towards methane.

For the above named study, Yang et al. refined reduced reaction mechanisms for syngas conversion on Rh(111) and Rh(211) surface facets, which build up the basis of this work's reaction network development. The mechanisms are depicted in appendix A.1.

Chapter 2

Theoretical background

In this chapter, the theoretical background of this work is introduced. In line with the overall goal of this work, the fundamentals required to fit an interatomic potential as a surrogate model to perform global optimizations of the potential energy surface (PES) are focused. Therefore, this introduction is divided into the description, exploration and representation of the PES. For the description of the PES, the methods density functional theory and machine learned interatomic potentials, especially the applied method of Gaussian approximation potentials, are emphasized. The exploration of the PES is divided into local and global optimization techniques including the minima hopping method. In the PES representation, training set generation by farthest point sampling and system visualization and simplification by kernel principal component analysis are focused.

2.1 Description of potential energy surfaces

The PES represents the energetic landscape of a system as a function of parameters, most commonly Cartesian coordinates. The energy can be expressed in terms of the following time-independent, non-relativistic Schrödinger equation, whereby $\hat{H}$ defines the Hamiltonian, $\Psi(\mathbf{r}_i\sigma_i, \mathbf{R}_v)$ the many-body wavefunction, E the energy , $\mathbf{r}_i$ and σ_i the spatial and spin coordinates of the electrons and $\mathbf{R}_v$ the spatial coordinates of the nuclei

$$\hat{H}\Psi(\mathbf{r}_i\sigma_i, \mathbf{R}_v) = E\Psi(\mathbf{r}_i\sigma_i, \mathbf{R}_v). \tag{2.1}$$

Solving this many-body Schrödinger equation yields the respective energy, what quickly becomes a complex, analytically unsolvable task even for small systems. This is reasoned in multiple correlation effects such as electron-electron and electron-nucleus correlation, requiring approximations, for example the Born-Oppenheimer (BO) approximation. As the nuclei's mass remarkably exceeds the electrons' mass, the electronic velocity surpasses the velocity of the nuclei. Thus, the latter are considered stationary, which enables the neglection of the electron-nuclei coupling and the reduction of the problem to its electronic part only. In order to solve the many-body Schrödinger equation within the BO approximation, the Hartee-Fock (HF) method was established. [29]

However, since electron-electron correlations are still neglected in HF theory, a lack in accuracy is the consequence. Therefore, HF nowadays serves as the starting point for other developments approaching a more accurate solution of the many-body problem. In general, two approaches can be distinguished: wavefunction based post-HF methods such as Møller-Plesset perturbation theory or Coupled Cluster on the one hand and density based methods such as density functional theory (DFT) on the other hand. DFT has been established as the workhorse for a broad field of applications including catalysis and surface science due to its lower computational costs compared to wavefunction based methods.

Although DFT facilitates calculations for periodic systems in contrast to wavefunction methods, ab initio molecular dynamics simulations are still very limited in system size and simulation time, when approached with DFT. For this reason, empirical interatomic potentials are still the mainly used physical basis for the dynamics of multi-element systems with heavy and non-systematic losses in accuracy. ML methods such as GAP therefore emerge for the description of the PES, allowing larger systems and longer simulation times than ab initio methods on the one hand, coupled with higher accuracy than empirical potentials on the other hand.

In the following, the methods of DFT and GAP are further elucidated, as these are the methods applied in this work. The introduction to DFT is based on [30–32] and to GAP on [14, 33, 34], unless otherwise specified.

2.1.1 Density functional theory

Density functional theory (DFT) is a method to calculate the electronic structure of a system, first primarily used in solid state physics with broad applications in the chemical context nowadays. In contrast to wavefunction based solutions to the many-body problem, where the wavefunction serves as the central quantity, DFT uses the electron density for this purpose.

According to the first Hohenberg-Kohn theorem [35], the electron density ρ and the external potential v_{ext} of a system are mapping to within a constant:

$$\rho \rightarrow v_{ext} \rightarrow \Psi_0. \tag{2.2}$$

As a consequence, the ground state energy can be expressed as a functional of the electron density

$$E[\rho] = \int \rho(\mathbf{r}) v_{ext} d\mathbf{r} + F_{\mathrm{HK}}[\rho], \tag{2.3}$$

where the Hohenberg-Kohn functional $F_{\mathrm{HK}}[\rho]$ is defined as the sum of the kinetic energy functional $T[\rho]$ and the electron-electron energy $E_{\mathrm{ee}}[\rho]$

$$F_{\mathrm{HK}}[\rho] = T[\rho] + E_{\mathrm{ee}}[\rho]. \tag{2.4}$$

In addition to that, Hohenberg and Kohn [35] were able to prove the validity of the variation principle, which is expressed in their second theorem:

$$E[\rho] \geq E_0^{\text{exact}}. \tag{2.5}$$

Solving this minimization problem by finding the density, that minimizes the energy, finally approaches the ground state energy. However, in the expression of the ground state energy in equation 2.3, only the first part is exactly expressible. In contrast, there is no simple way to determine a suitable expression for the Hohenberg-Kohn functional F_{HK} consisting of the kinetic energy functional T and the electron-electron energy E_{ee}. Though, the knowledge of the Hohenberg-Kohn functional F_{HK} is crucial.

Kohn and Sham [36] approached this problem by specifying and dividing the kinetic energy part of the Hohenberg-Kohn functional T in an exactly expressible part T_{s} and a remaining part $T - T_{\text{s}}$. Also the electron-electron interaction can be parted in two terms: a classical Coulomb term J and unknown non-classical electrostatic contributions E^{ncl}

$$E_{\text{ee}} = J[\rho] + E^{ncl}[\rho]. \tag{2.6}$$

Thus, the exchange-correlation functional E_{xc} can be introduced, which is defined as

$$E_{\text{xc}}[\rho] = (T[\rho] - T_{\text{s}}[\rho]) + (E_{\text{ee}}[\rho] - J[\rho]). \tag{2.7}$$

This finally leads to the following reformulation of the energy as a function of electron density,

$$E[\rho] = T_s[\rho] + \int \rho(\mathbf{r}) v_{ext} d\mathbf{r} + J[\rho] + E_{\text{xc}}[\rho], \tag{2.8}$$

which is the total energy expression in the so-called Kohn-Sham DFT.

As all of the other parts of the total energy are explicitly defined, finding an accurate expression for the exchange-correlation functional E_{xc} is the main task in DFT nowadays. Therefore numerous functionals have been developed to approach this task and an overview is given in the following.

Approach to chemical accuracy

The main developments in DFT nowadays strive to find a more accurate expression to the exchange-correlation functional. Over the years, chemical accuracy was more and more approached, which is visualized by John Perdew's Jacob's ladder [37] in figure 2.1. In contrast to wavefunction based approaches, in DFT improvement is not achieved systematically, but the Jacob's ladder enables sequencing of the different, solitary approximations. Starting from the HF level, numerous non- and semi-empirical functionals have been developed striving for chemical accuracy.

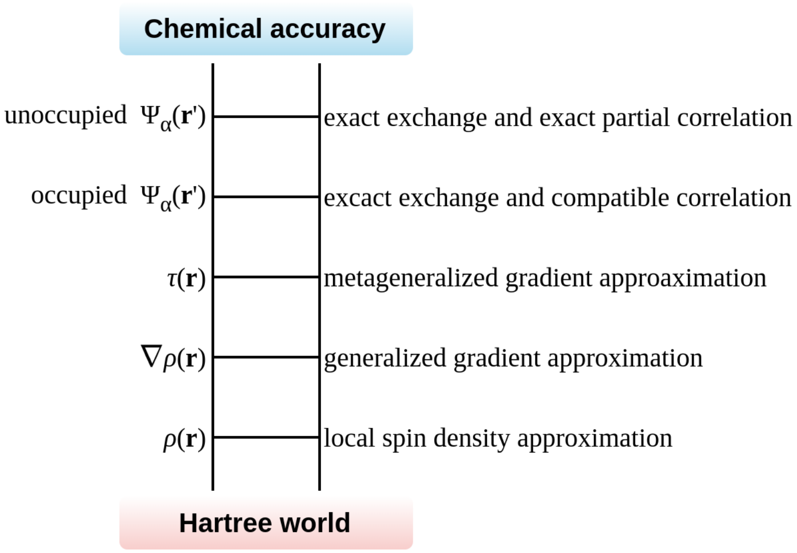

Chemical accuracy
unoccupied $\Psi_\alpha(\mathbf{r}')$
exact exchange and exact partial correlation
occupied $\Psi_\alpha(\mathbf{r}')$
excact exchange and compatible correlation
$\tau(\mathbf{r})$
metageneralized gradient approaximation
$\nabla\rho(\mathbf{r})$
generalized gradient approximation
$\rho(\mathbf{r})$
local spin density approximation
Hartree world

Figure 2.1: Jacob's ladder of exchange-correlation functionals for DFT [37]. Each rung represents a class of exchange-correlation functionals. Starting from the Hartee-Fock level, each rung reaches more and more towards chemical accuracy, which is the overall goal.

The simplest description of the exchange correlation potential only uses the energy density itself and is classified as the local density approximation (LDA), including its extension to the local spin density approximation (LSDA). The generalized gradient approximation (GGA) additionally includes the gradient of the density, which leads to an improvement over LDA. The third rung is build by the metageneralized gradient approximation (meta GGA). In meta GGA, either the second derivative of the density or the kinetic energy density is used further approaching chemical accuracy. The upper two rungs pursue another idea and treat the exchange term explicitly, coming from Hartee-Fock theory. A specific subclass of these are hybrid functionals, which are widely employed. The next step towards chemical accuracy is the introduction of full non-locality.

For all of the above mentioned approximations, various specific functionals have been designed. The revised Perdew–Burke–Ernzerhof (revPBE) functional [38], an extension of the Perdew–Burke–Ernzerhof (PBE) functional, is a GGA functional, which is commonly applied in catalysis and also in this work.

Despite its broad utilization, (semi-)local functionals as PBE lack in predicting non-covalent dispersion interactions. Therefore, several dispersion correction methods have been developed, for example pairwise van der Waals (vdW) corrections [39] or the Lifshitz-Zaremba-Kohn theory for the interactions of atoms and solid surfaces. Ruiz et al. [40] combined these two approaches in the DFT+vdW$^{\text{surf}}$ method. Due to inclusion of the corresponding DFT+vdW$^{\text{surf}}$ parameters, the description of vdW interactions of adsorbates on surfaces can be enhanced. [41] Thus, DFT+vdW$^{\text{surf}}$ parameters are also considered in this work.

Although DFT facilitates the calculation of periodic surface systems, it is still limited in system size and simulation timescale. Therefore, interatomic potentials are widely employed and a new field of potentials is rising, the ML interatomic potentials. An introduction to a special type of ML potentials, GAP, is given in the following.

2.1.2 Gaussian approximation potentials

Gaussian approximation potentials (GAPs) are a special form of ML interatomic potentials, which demonstrate good applicability in the chemical context [42–45]. Due to their potential in the prediction of complex chemical networks [46], they are also applied in this work. In contrast to many empirical force fields, GAP does not require a previous definition of a parametric function, a particular physical model or specific interactions [47]. Instead, it is solely based on the given input data, which is used to approximate the PES by non-linear regression.

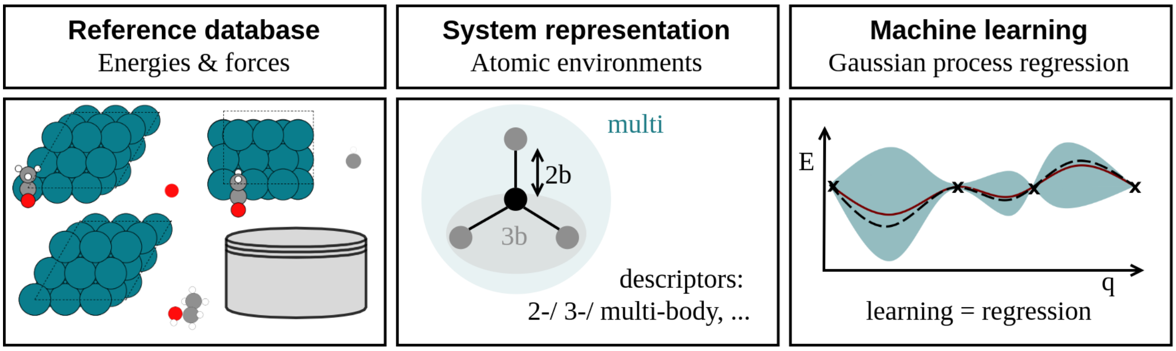


Figure 2.2: Visualization of the GAP approach based on [48, 49]. In order to fit a GAP model, three essential parts are required: 1. the reference database consisting of geometric data with corresponding DFT energies, 2. the system representation by different descriptors (e.g. by 2-, 3- or multi-body descriptors), 3. machine learning part in form of Gaussian process regression.

In more detail, the GAP approach can be divided into three essential components, as depicted in figure 2.2. These are the reference database consisting of chemical geometries with their associated energies and forces, the system representation by their atomic environments and third, the ML part in form of Gaussian process regression (GPR). The three components are subsequently elaborated.

Construction of a training database

The basis of training a GAP is a database consisting of a representative set of configurations with its corresponding quantum mechanical energies and forces obtained via a high level reference method, e.g. DFT calculations. These geometries build up the input data for the subsequent regression, whereby the energies and forces serve as fit properties.

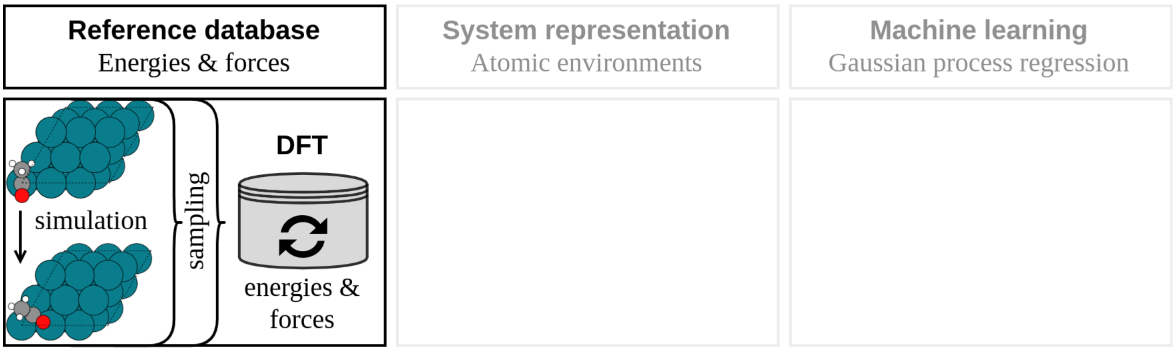


Figure 2.3: The first component of the Gaussian approximation potential (GAP) approach [48,49]: the reference database consisting of geometries with its corresponding DFT energies and forces iteratively updated by generation of new data via a certain simulation method and sampling of the generated data.

The GAP training represents an iterative process as implied by figure 2.3. During the process, the aim is to improve the prediction by an iterative update of the training data and relearning with this data ahead. This requires the iterative generation of new training data unknown to the potential. The data eneration is done by using a certain simulation method, for example MD or optimization, in each iteration and sampling by a chosen sampling technique, for example by farthest point sampling (FPS) later introduced in section 2.3.1. Due to the sampling and choice of new data, the training set is updated and a new training iteration can be started.

Representation of atomic environments

The system representation portrays the second component required to train a GAP, as depicted in figures 2.2 and 2.4. The central measure to represent a system is its total energy in dependency of parameters, for which predicting the total energy is also the target in the GAP approach. The total energy can be divided in short-range and long-range interactions,

$$E = \underbrace{\sum_{d} (\delta^{(d)})^2 \sum_{i \in d} \varepsilon^{(d)}(\mathbf{q}_i^{(d)})}_{\text{short-range}} + \text{long-range contributions}, \tag{2.9}$$

whereby the long-range contributions stem from electrostatics interactions. The short-range contributions consist of local, scaled (δ) energies ε, which are summed over the different descriptor types d. The single local energy contribution $\varepsilon^{(d)}$ is a function of the input configuration represented by its descriptor vector $\mathbf{q}^{(d)}$. It is defined as a linear combination of weighted (α) Kernel functions $K^{(d)}$, measuring the similarity of a given local environment $\mathbf{q}_i^{(d)}$ to the local environment of N_u training configurations $\mathbf{q}_u^{(d)}$:

$$\varepsilon^{(d)}(\mathbf{q}_i^{(d)}) = \sum_{u=1}^{N_u^{(d)}} \alpha_u^{(d)} K^{(d)}(\mathbf{q}_i^{(d)}, \mathbf{q}_u^{(d)}). \tag{2.10}$$

As mentioned above, atomic structures serve as inputs for the ML method. Therefore, one central need in the GAP approach is the representation of the atomic environment, as the energy of an atom is influenced by interactions with its neighborhood. This description should be designed such that it remains invariant to translations and rotations, which excludes the usage of Cartesian coordinates.

Therefore, descriptors, which fulfill these requirements, have been developed. Exemplary descriptors are the 2-body and the smooth overlap of atomic positions (SOAP) descriptor, pictured in figure 2.4. The 2-body descriptor regards all the 2-body contributions of a central atom a to its surrounding atoms within a cutoff radius r_{cut}. In contrast to this, the SOAP descriptor is a multibody descriptor, which does not regard just a single scalar like the 2-body descriptor vector in the squared exponential kernel. It regards all the multibody contributions due to the overlap of atomic positions within a cutoff radius.

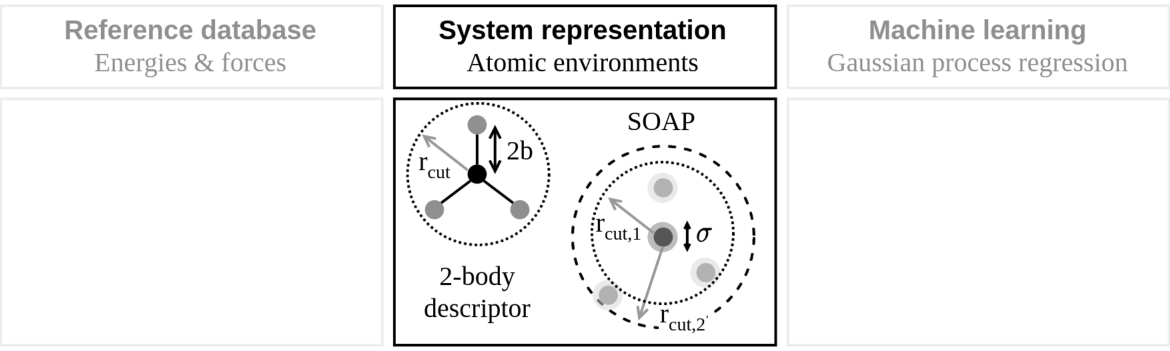


Figure 2.4: The second component of the Gaussian approximation potential (GAP) approach [48, 49]: the system representation by different descriptors, e.g. the 2-body or smooth overlap of atomic positions (SOAP) descriptor.

For these descriptors, kernels or covariance matrices are constructed, which serve as a similarity measure between structures and are capable of representing multivariant atomic environments. Dependent on the descriptor, different types of kernels are typically utilized. In case of the 2-body descriptor, a squared exponential kernel is capable of describing the similarity of two local environments i and u:

$$K^{(2b)}(\mathbf{q}_i^{(2b)}, \mathbf{q}_u^{(2b)}) = \exp\left[-\frac{1}{2}\sum_{\nu} \frac{(q_{\nu,i}^{(2b)} - q_{\nu,u}^{(2b)})^2}{\theta_\nu^2}\right], \tag{2.11}$$

where θ denotes the length scale, which is a smoothness parameter, and ν indexes all the different descriptor vector components. In case of describing 2-body contributions between atoms a and b, the 2-body descriptor vector is then given by the natural distance

r_{ab} between these atoms

$$q^{(2\mathrm{b})} = |\mathbf{r}_b - \mathbf{r}_a| \equiv r_{ab}. \tag{2.12}$$

For SOAP, another approach is followed: SOAP is based on the environment density of an atom a, which is expressed in a Gaussian, to a sum over atoms b with a Gaussian width of σ_{ab} within fixed boundaries of a seceding cutoff function f_{cut}

$$\rho_a(\mathbf{r}) = \sum_b \exp\left[-\frac{(\mathbf{r}-\mathbf{r}_{ab})^2}{2\sigma_{\mathrm{ab}}^2}\right] \times f_{\mathrm{cut}}(r_{ab}). \tag{2.13}$$

For the sake of simplicity, the explanation here is restricted to one atomic species only. The extension due to the expansion to different atomic species, can be found elsewhere [50]. For numerical reasons, the density is expanded in terms of spherical harmonic functions Y_{lm},

$$\rho_a(\mathbf{r}) = \sum_{nlm} c_{nlm}^{(a)} g_n(r) Y_{lm}(\hat{\mathbf{r}}), \tag{2.14}$$

within a local basis of orthogonal basis functions g_n. The expansion coefficients c_{nlm} form the following power spectrum

$$p_{nn'l}^{(a)} = \sqrt{\frac{9\pi^2}{2l+1}} \sum_m (c_{nlm}^{(a)})^* c_{n'lm}^{(a)} \tag{2.15}$$

with limited elements $l \leq$ `lmax` and $n \leq$ `nmax`, whereas `lmax` and `nmax` refer to hyperparameters, which are defined in table 3.2. Building the dot product of the power spectrum and exponentiating with `zeta` ζ in order to sensitise to changes in atomic positions finally yields the definition of the SOAP kernel:

$$\begin{aligned} K^{(\mathrm{SOAP})}(\mathbf{q}_a^{(\mathrm{SOAP})}, \mathbf{q}_u^{(\mathrm{SOAP})}) &= |\sum_{nn'l} p_{nn'l}^{(a)} p_{nn'l}^{(t)}|^{\zeta} \\ &= |\mathbf{q}_a^{(\mathrm{SOAP})} \cdot \mathbf{q}_u^{(\mathrm{SOAP})}|^{\zeta}. \end{aligned} \tag{2.16}$$

Coming back to the expression of the total energy, as defined in equation 2.9, the previously defined kernels can be used to summarize the total energy. In this work, a 2-body descriptor was combined with two SOAP descriptors with different cutoff radii, which by summation of the different kernels for each descriptor [51] yields the final total energy term consolidating this work's approach:

$$\begin{aligned} E = {} & (\delta^{(2\mathrm{b})})^2 \sum_i \sum_u \alpha_u^{(2\mathrm{b})} K^{(2\mathrm{b})}(\mathbf{q}_i^{(2\mathrm{b})}, \mathbf{q}_u^{(2\mathrm{b})}) \\ & + (\delta^{(\mathrm{SOAP}_1)})^2 \sum_j \sum_u \alpha_u^{(\mathrm{SOAP}_1)} K^{(\mathrm{SOAP}_1)}(\mathbf{q}_j^{(\mathrm{SOAP}_1)}, \mathbf{q}_u^{(\mathrm{SOAP}_1)}) \\ & + (\delta^{(\mathrm{SOAP}_2)})^2 \sum_k \sum_u \alpha_u^{(\mathrm{SOAP}_2)} K^{(\mathrm{SOAP}_2)}(\mathbf{q}_k^{(\mathrm{SOAP}_2)}, \mathbf{q}_u^{(\mathrm{SOAP}_2)}). \end{aligned} \tag{2.17}$$

From this functional form of the total energy, predictions of the PES are impossible due to the lack of knowledge on the regression coefficients. The above definitions are rather prerequisites for the ML, which is briefly explained in the following.

Gaussian process regression

At the heart of machine-learning interatomic potentials naturally is the learning. In the GAP method, the supervised learning is done by Gaussian process regression (GPR), which is a non-parametric regression method.

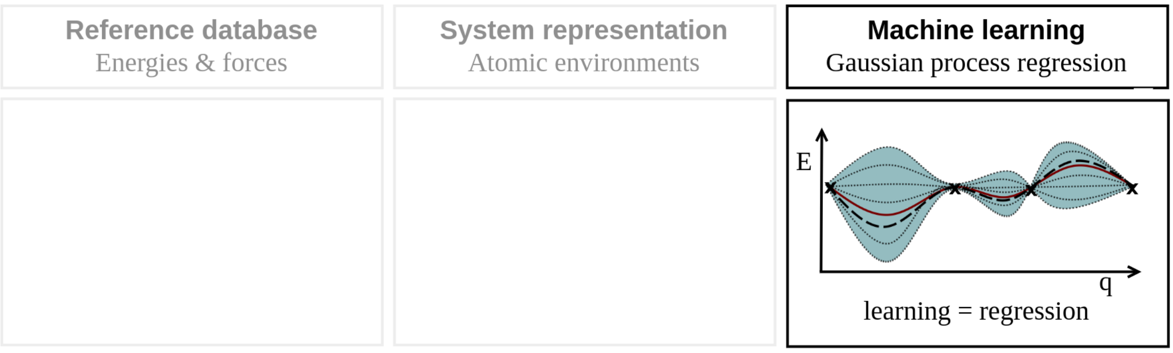


Figure 2.5: The third component of the Gaussian approximation potential (GAP) approach [48, 49]: machine learning by Gaussian process regression (GPR). The solid, red line represents the true potential energy curve, the dotted lines possible fit functions and the dashed line the mean fit function. x are the datapoints and the bar reflects the uncertainty of the prediction.

Splitting the name of the method Gaussian process regression to Gaussian process and regression, gives an explanation to the method itself. In regression, an unknown function y is fitted to a given data set. In case of the well-known linear regression, the functional form is known, resulting in an easy calculation of the linear regression coefficients with the lowest mean-square error. For multivariant data, such as molecular systems and energies, the describing parameters should not be limited to a specific number, and more than one function might be able to represent the data set. Therefore non-parametric regression is required.

A Gaussian process represents the probability distribution of different functions, which are able to fit to the given data set. As indicated in figure 2.5, the uncertainty of the prediction of a GPR is high in regions with no or limited data points and becomes low for regions with many data points. [52]

Coming back to the prediction of the total energy or any other observation $\mathbf{t}$ of an atomic configuration, the covariance is given by the scalar product

$$\mathbf{C} \equiv \mathbf{t} \cdot \mathbf{t}^{\mathrm{T}}, \tag{2.18}$$

whereby the values of the covariance matrix correspond to the kernel or covariance functions $K^{(d)}$ previously defined in equation 2.11 and 2.16. The probability P of the observations $\mathbf{t}$ of the known data point is assumed to be normal distributed ($\mathcal{N}$) with zero mean and a variance given by the covariance $\mathbf{C}$, such that

$$P(\mathbf{t}) = \mathcal{N}(\mathbf{t}; \mathbf{0}, \mathbf{C}) \propto \exp\left(-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{t}\right). \tag{2.19}$$

The probability of an unknown data point y is likewise Gaussian distributed and given by

$$P(y|\mathbf{t}) = \frac{P(\mathbf{t}, y)}{P(\mathbf{t})}. \tag{2.20}$$

Taking the mean of the latter distribution with $\mathbf{k}$ denoting the covariance vector of the predicted value y,

$$\mathbf{k} = y \cdot \mathbf{t}, \tag{2.21}$$

then defines the regression model

$$\overline{y} = \mathbf{k}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{t}, \tag{2.22}$$

whereby $\mathbf{C}^{-1}$ are the coefficients, that have been evaluated by the training. Therefore, it is possible to predict the total energy of unknown configurations from the kernel of given observations $\mathbf{t}$. This makes GPR a powerful and widely applied method in ML. For highly correlated data sets, for example structures with only slight perturbation with similar atomic neighborhoods, 'sparse' approximations become reasonable in order to reduce computational costs. In the GAP approach, sparsification is introduced by setting a number of sparse configurations $\mathtt{n}_{\mathtt{sparse}}$, which are randomly chosen from the overall training set and build the representative atomic neighborhood. [13]

With this enhanced and accelerated description of the PES by the usage of GAP, the consideration of larger systems and the simulation of longer dynamics is facilitated. This enables for example the systematic exploration of the PES, which is focused in the following.

2.2 Exploration of potential energy surfaces

The PES of a system constitutes a 3N-dimensional surface, whose exploration quickly exceeds computational possibilities with increasing N. The three dimensions refer to the spatial coordinates of the N atoms. Thus, the complexity rises with system size and exploring the PES of bigger systems requires the usage of computationally efficient optimization methods. While analysing the PES of a chemical system, illustrated in figure 2.6, many midpoints and stationary points such as saddlepoints and extrema are detected. Thereby, saddle points and minima receive special interest, as the saddle points represent transition states and the minima are (meta-)stable states, whose locations have the greatest impact on a chemical reaction's path. [53]

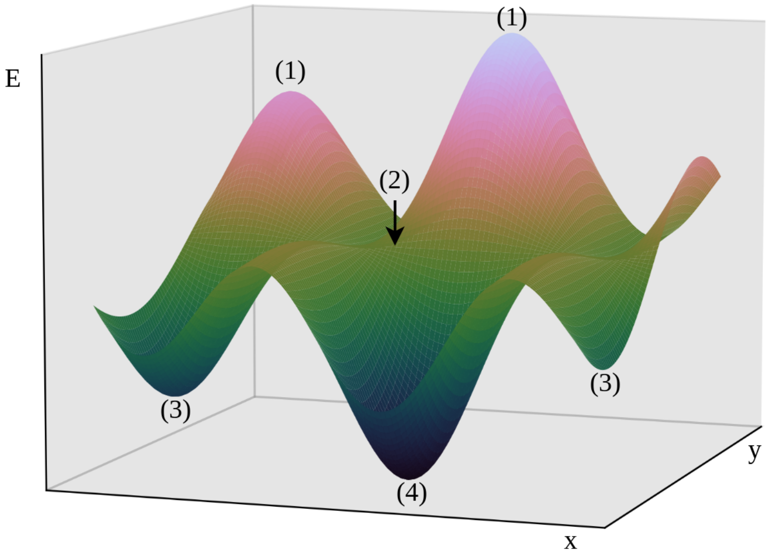


Figure 2.6: Illustration of a potential energy surface (PES) as energy E versus Cartesian coordinates x and y, consisting of (1) maxima, (2) saddle points, (3) local minima and (4) one global minimum.

However, many local but only one global minimum can be found, which is the most stable configuration of the system [54], also referred to as geometric ground state. As the number of local minima rises exponentially with system size [55], finding a local minimum of a system is a rather straightforward task, whereas the demanding one is to find the global minimum [17]. Several methods have been developed coming up with local minima, saddle points, transition states as well as the global minimum. As this work applies local optimization using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm as well as global optimization by the minima hopping method (MHM), these methods are elaborated in the following.

2.2.1 Local optimization

The Broyden–Fletcher–Goldfarb–Shanno (BFGS) method is a quasi-Newton optimization method, which is commonly applied to local optimization problems. Local optimizations in the chemical context mostly indicate geometry optimizations. Starting from an initial structure, the surrounded, most stable configuration is approached, which is the one with the lowest energy. In contrast to global optimization, local optimization strives to quickly find the next optimum, without evaluating the presence of other optima further afield. This distinction is illustrated in the following figure 2.7.

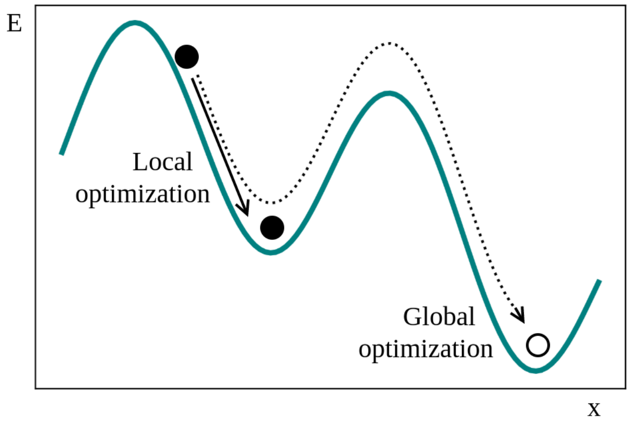


Figure 2.7: Comparison of local (solid arrow) and global optimization (dotted arrow) in a 1D potential energy plot (green curve) as energy E vs. reaction coordinate x.

Local optimization methods can be differentiated in first and second order optimization methods, whereby first means the first derivative of the objective function, for example the potential energy, and second the second derivative. More precisely, second order optimization methods, also classified as Newton optimization methods, make use of the Hessian matrix, which is a square matrix of second-order partial derivatives of a function. By usage of the second order, not only the slope but also the curvature of the objective function is accounted, which provides the direction for optimization and additionally the possible step size. [56]

Quasi-Newton methods circumvent the computationally expensive calculation of the inverse of the Hessian matrix by calculating the first order derivative explicitly and approximating the Hessian. This approximation significantly reduces the computational cost compared to the regular Newton optimization [56, 57]. In case of the BFGS method, the approximation to the Hessian is iteratively updated instead of recalculated, which additionally reduces the computational cost of the BFGS steps. A detailed description of the BFGS approach can be found in the literature [58–61].

The BFGS method serves as the local optimization method in geometry optimizations as part of the MHM method, which is introduced below.

2.2.2 Global optimization by minima hopping

The minima hopping method (MHM) is a non-thermodynamics-based method to explore the PES of a system in order to find its global minimum [62]. As finding the global minimum of a system is a challenging task due to the complexity of the PES, a number of methods has been developed striving to find it. For example, these methods include genetic algorithms [63], basin hopping [64] and simulated annealing [65], whereas all of the methods have its applicability and limitations, especially in system size or revisiting of local minima [66]. A detailed description of the various global minimization techniques exceeds the scope of this work, wherefore the reader is referred to [67–71].

Many global optimization methods, for example stochastic global optimization, are based on thermodynamics, which is achieved by introduction of a Boltzmann distribution. The Boltzmann distribution is designed such that lower energies of a structure result in higher weights in the distribution. Therefore, the global minimum has the highest weight. This can be amplified by lowering the temperature leading to an even higher weight of the global minimum, which is favorable. However, low temperatures increase the probability of basin trapping, because the crossing of high-energy regions is hampered [62]. As the global minimum is surrounded by those high-energy regions, the introduction of a Boltzmann distribution does not guarantee the quick detection of the global minimum. Consequently, the usage of a thermodynamic distribution might lead to failure in finding the global minimum at all [17].

In contrast, the MHM, which is the global optimizer used in this work, makes limited use of thermodynamics. Instead, it aims to quickly explore the low energy regions of the PES through coupling of MD and local geometry relaxations with a distinct feedback mechanism and the dynamical adjustment of parameters. [66]

Figure 2.8 illustrates the MHM algorithm, which is dividable into an inner and an outer part. The method starts with an initial local geometry optimization step with the BFGS method, yielding the initial local minimum M_{cur}. From this local minimum, the inner loop starts with a first escape trial. This escape trial consists of a short MD simulation followed by a local relaxation to the minimum M. If the initial local minimum M_{cur} is the same as the local minimum M found in the escape trial, the kinetic energy is increased by a factor of $\beta_{\mathrm{s}} > 1$ and the inner loop is restarted with random velocities initialed from a Boltzmann distribution.

Otherwise, the newly found local minimum M is proposed to the outer part, reviewing, whether it differs from the current minimum M_{cur}. If the difference in energy of the new minimum $E(M)$ to the energy of the current local minimum $E(M_{\mathrm{cur}})$ is higher than a threshold E_{diff}, the minimum gets rejected and the threshold gets increased by a factor α_{r}. If the energy difference is lower than the threshold E_{diff}, the current minimum gets updated to the new one $M_{\mathrm{cur}} = M$ and the threshold is lowered by $\alpha_{\mathrm{a}} < 1$. This feedback ensures the algorithm's preference towards lower energies.

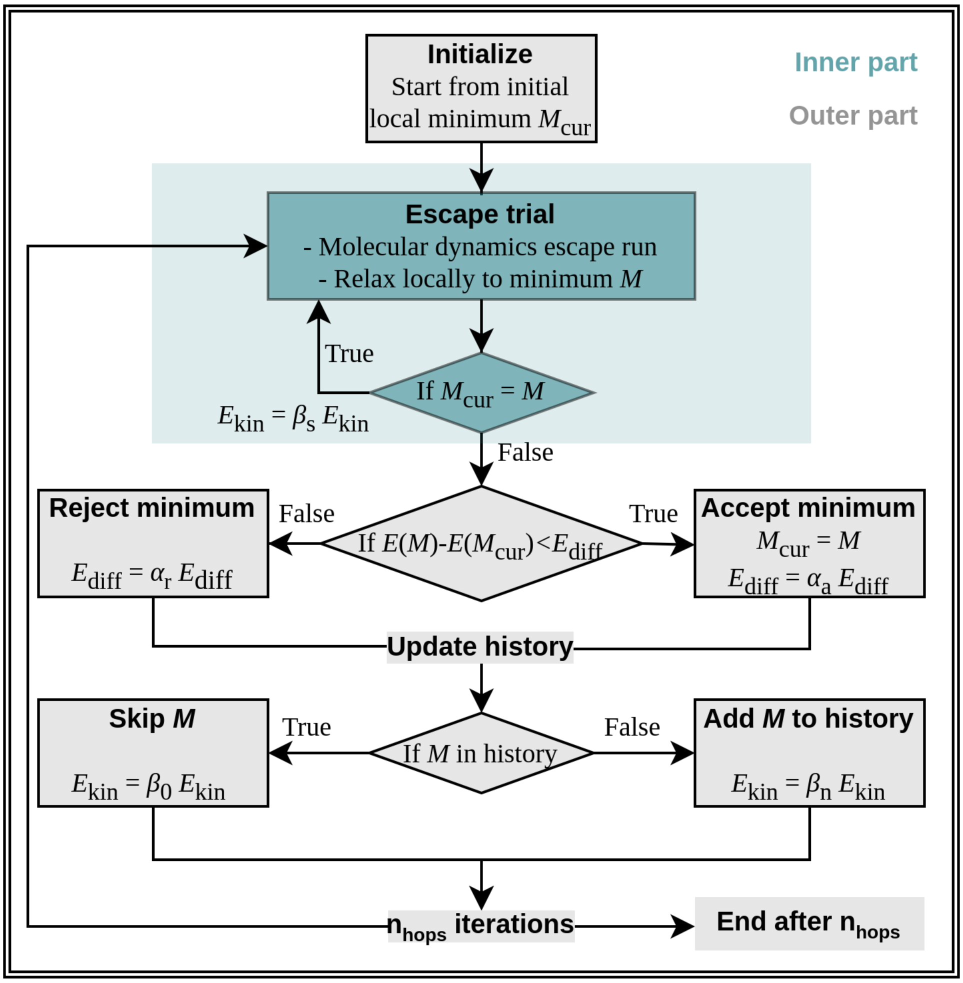

Initialize
Start from initial local minimum M_{cur}
Inner part
Outer part
Escape trial
- Molecular dynamics escape run
- Relax locally to minimum M
True
If $M_{\mathrm{cur}} = M$
$E_{\mathrm{kin}} = \beta_{\mathrm{s}} E_{\mathrm{kin}}$
False
Reject minimum
$E_{\mathrm{diff}} = \alpha_{\mathrm{r}} E_{\mathrm{diff}}$
False
If $E(M)\text{-}E(M_{\mathrm{cur}})<E_{\mathrm{diff}}$
True
Accept minimum
$M_{\mathrm{cur}} = M$
$E_{\mathrm{diff}} = \alpha_{\mathrm{a}} E_{\mathrm{diff}}$
Update history
Skip M
$E_{\mathrm{kin}} = \beta_0 E_{\mathrm{kin}}$
True
If M in history
False
Add M to history
$E_{\mathrm{kin}} = \beta_{\mathrm{n}} E_{\mathrm{kin}}$
n_{hops} iterations
End after n_{hops}

Figure 2.8: Flowchart of the minima hopping method (MHM) algorithm [54]. The MHM consists of an inner part performing molecular dynamics based escape trials and an outer part introducing a feedback mechanism to the visited minima. Explanations of all given variables are given in the text.

Notwithstanding of the outcome, the history of visited minima is updated. If the minimum M is already known in the history, the kinetic energy is increased by β_0 to explore other parts of the PES. If not, the minimum is added to the history and the kinetic energy is lowered by a factor β_{n}. [54] This procedure is repeated for a distinct number of loops n_{hops} and outputs several MD and local relaxation trajectories, the conformations of the found minima as well as their history.

Especially due to the adjustment of the kinetic energy, jumps into different minima basins are enabled and therefore new low energy regions of the PES are explored. During the MD simulation, the kinetic energy is fixed and due to energy conservation only barriers lower than E_{kin} can be overcome [72]. This leads to low energy transitions for adequate E_{kin}. In case the system is not able to get out of the current local minimum, the kinetic energy is increased, allowing also to overcome high barriers and prevent the system from getting trapped [73]. In the outer part of the algorithm, the adaption of E_{kin} allows on the one hand the quick exploration of new regions, in case the minimum already exists in the history. On the other hand, if a new minimum is found, the decrease of the kinetic energy allows to transition over low barriers into even lower minima. This is in accordance to the Bell-Evans-Polanyi principle, which says that exothermic reactions have a low activation energy [17], and therefore transitioning from one minimum over a lower barrier into a new minimum might revel an even lower one.

Hence, the MHM is a powerful method to find the global minimum of a system.

2.3 Representation of potential energy surfaces

In this work, there are two main aspects of representing the PES. The first deals with choosing relevant and representative configurations to build up a database for the training of the GAP, which predicts the PES. Therefore, the statistical method of farthest point sampling (FPS) is applied to generate an appropriate training set. Second, the visualization and analysis of the PES is not straightforward due to its complexity. Therefore, principal component analysis (PCA) can be applied to perform dimensionality reduction, where visualization and analysis of the specific data is re-enabled. An introduction to the theoretical background of these two methods is given in the following.

2.3.1 Farthest point sampling

Farthest point sampling (FPS) is a sampling method widely applied in various parts of data science, for exampe image processing, and likewise in ML. In the latter, it is for example utilized to effectively sample a representative training set from given data. Thereby, every new data point is compared to the already existing data set with distance as a measure.

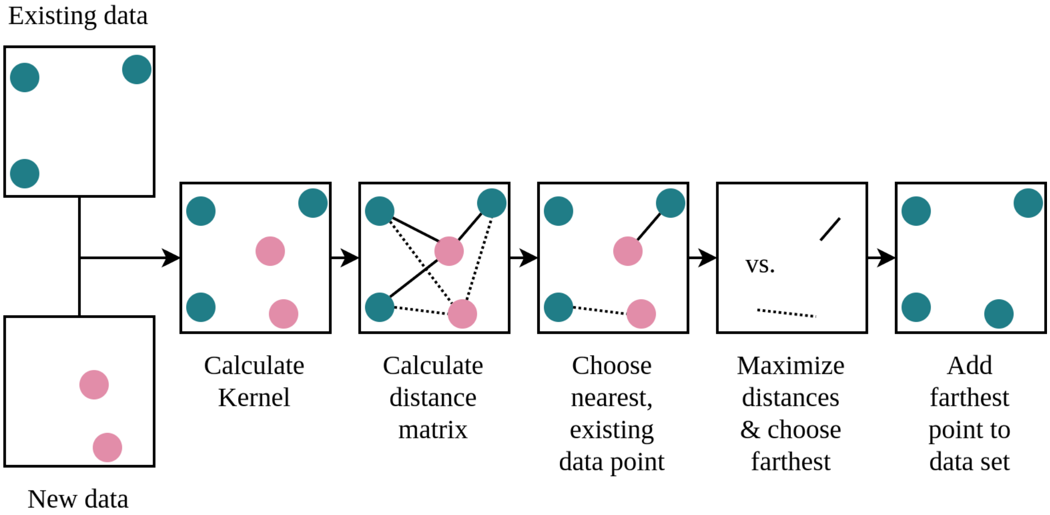

Existing data
New data
Calculate Kernel
Calculate distance matrix
Choose nearest, existing data point
vs.
Maximize distances & choose farthest
Add farthest point to data set

Figure 2.9: Flowchart of the FPS approach adapted to this work [49]. The green/ rose points refer to the existing/ new data and the lines to the distances between new and existing data points with different line types (solid/ dashed) for each new data point.

The approach, adapted to this work, is visualized in figure 2.9. Starting from an existing data set A and a new data set B, a normalized average kernel [7] is calculated, which serves as a similarity measure of the structures within and between the data sets. The average kernel differs from the previously defined kernels (compare section 2.1.2) such that it measures the similarity of structures and not atomic environments. Using this kernel K, the distances between all points from the existing A and new B data sets are computed and summarized in a distance matrix D, defined as

$$D(A,B) = \sqrt{K(A,A) + K(B,B) - 2K(A,B)}. \tag{2.23}$$

For each of the new data points, the nearest point of the existing data is chosen. These nearest distances are maximized and the farthest point, which is the point with the maximized distance towards the existing data, is chosen and added to the training set. Using this approach, the most significant new data points, which differ most from each other, are chosen and an appropriate, diverse training set is constructed. [74]

2.3.2 Kernel principal component analysis

In order to interpret large, multidimensional data sets, principal component analysis (PCA) can be used. In PCA, the data set is transformed into another space in order to reduce dimensionality. By linear projection of the data onto principal components instead of remaining with the initial dimensions, visualization becomes possible and thus patterns and trends might be revealed. [75] Thereby, the principal components are chosen to be perpendicular to each other and target towards the directions of the largest variance.

Kernel principal component analysis (kPCA) is the non-linear subform of PCA, which makes use of the 'kernel-trick'. By usage of kernel functions, linear methods such as PCA or regression, as introduced before, are attuned towards non-linear data sets. [76] As a kernel for example an average SOAP kernel can be used, which is centralized to the number of points N such that

$$\hat{\mathbf{K}} = \mathbf{K} - \mathbf{1}_\mathrm{N}\mathbf{K}\mathbf{1}_\mathrm{N} + \mathbf{1}_\mathrm{N}\mathbf{K}\mathbf{1}_\mathrm{N}, \tag{2.24}$$

whereby a matrix $\mathbf{1}_\mathrm{N}$ is introduced with the same dimensions as the kernel matrix and the value 1/N for every element. By solving the eigenvalue problem with $\mathbf{v}$ denoting the eigenvectors and λ the eigenvalues,

$$\hat{\mathbf{K}}\mathbf{v}_i = \lambda_i \mathbf{v}_i, \tag{2.25}$$

the i^{th} principal component $\mathbf{PC}_i$ can be constructed via

$$\mathbf{PC}_i = \mathbf{K}\mathbf{v}_i. \tag{2.26}$$

Due to this, the data are projected into a lower dimensional space, allowing further analysis and visualization. [7]

All of the previously introduced methods are applied in this work, which is elaborated in the subsequent main parts of this thesis, chapters 3 and 4.

Chapter 3

Development of a machine-learned interatomic potential

As specified in chapter 1, the goal of this work is to first train an interatomic potential using ML to secondly apply this potential to the syngas conversion on Rhodium surfaces. In this chapter, the first part - the development of the system-specific GAP - is enlightened. The chapter starts with an introduction to the methods and its computational details. This is followed by a description and discussion of the results, including both the tested variations as well as the final training.

3.1 Methods and computational details

The development of interatomic potentials is a key approach to enable dynamic simulations of atomistic systems on relevant timescales. Because of the well-investigated inaccuracies of empirical force-fields, other methods emerged, for example ML interatomic potentials. The GAP method, as introduced in 2.1.2, is the class of potentials machine-learned in this work.

Thereby, the research is focused on the development of a GAP tailored to adsorbate-surface systems occurring in heterogeneous catalysis. The specific system of interest is the syngas conversion on Rhodium. Special emphasize is on the alignment of adsorbates on the catalytic surfaces. The goal of this work is to use the GAP to explore the minimum space of the different involved adsorbates coming up with the global minimum of each. Thus, only low-coverage systems - single molecules adsorbed to the surface - are considered. An iterative training approach is applied, coupling the learning of a potential with the generation of new training data by global optimization.

The first step towards the goal of this work is the development of a training workflow for the GAP later applied to the syngas conversion on Rhodium. Therefore, the development of the training workflow is constituted in the following.

3.1.1 Workflow development

The centerpiece of training a GAP is the development of an appropriate training workflow. This work uses an iterative training process, coupling the ML to an iterative training data generation. This enables an iterative update of the training set and due to relearning with this updated set a further improved prediction. This work's iterative GAP training workflow is illustrated in figure 3.1.

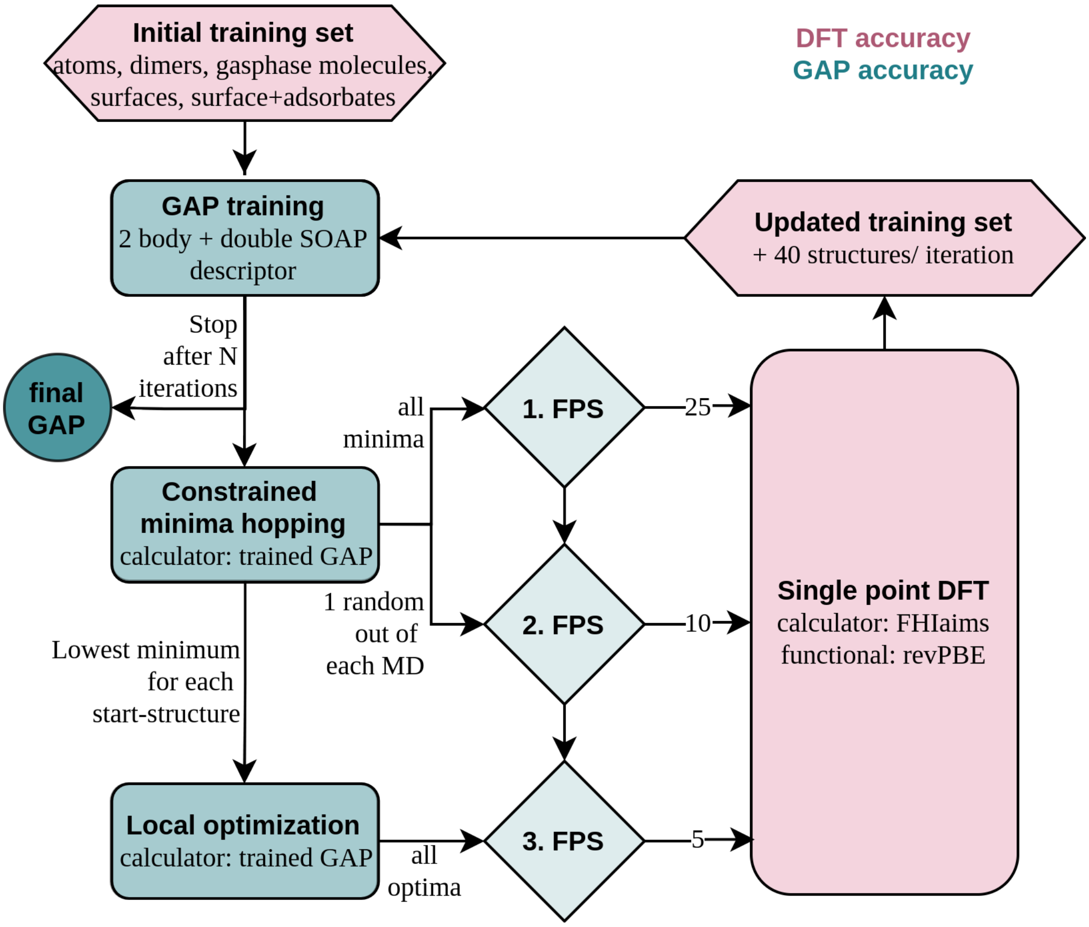


Figure 3.1: Illustration of the GAP training workflow developed and applied as part of this work. The iterative training is split in the training of a potential based on DFT level data on the one hand and the generation of new training data by constrained minima hopping as well as unconstrained local optimization on the other hand. The choice of appropriate structures for updating the training set is based on three consecutive farthest point sampling (FPS) steps.

The first step of training a GAP is the assembly of the initial training data, in accordance to the requirements previously depicted in figure 2.2. The reference database is specific for the system of interest. For our system, it includes the relevant educts, intermediates and products appearing in the syngas conversion on Rhodium. This choice is based on the reaction mechanisms deduced by Yang et al. [4], given in appendix A.1. The molecules are included as optimized gasphase molecules and adsorbates on surfaces. As the special interest of this work is in finding the global minimum of the adsorbates on surfaces, a low-coverage approach is pursued, meaning that one periodic cell only contains a single adsorbate on top of the surface. Single atoms are used as a baseline correction in the GAP code. Therefore, they are included in the training set. To properly consider the different interactions, also dimers are added to the training set. Moreover, the set comprises relaxed, empty Rhodium surfaces, whereas Rh(111) was chosen to represent the plane catalyst surface and Rh(211) to also regard surface steps. All the stated structures are included as geometric data together with their energies and forces on DFT level.

With this data set ahead, the zeroth training iteration is started. The zeroth training iteration is split in two stages, in order to calculate the required fit parameters. In stage one, a 2-body potential with a descriptor cutoff radius of 5Å is fitted and complemented via a baseline potential consisting of the relevant dimer interactions. The baseline potential is depicted in figure A.4. During the second training stage, two SOAP kernels with different cut-off radii, 3 and 6Å, are calculated. The double SOAP approach is pursued in order to consider a broader length scale of interactions. Based on the 2-body potential, the 2-body plus double SOAP (2b+dSOAP) GAP is then fitted.

With this previously trained potential, a constrained minima hopping simulation is started for all the different surface-adsorbate conformations included in the initial training set. Thereby, the Rhodium atoms are fixed and a Hookean spring force is applied to all the bonds in the adsorbate molecule to prohibit bond breaking. This minima hopping step serves to generate new training data in the iterative process. In a second step, the lowest minimum for each structure is chosen and the Hookean constraints are rescinded, allowing the adsorbates to relax and possibly dissociate during a local optimization.

In order to choose the most relevant structures from the data generated by the constrained minima hopping and the unconstrained local optimization, the data are sampled via three consecutive FPS steps. First, all the minima found in the constrained minima hopping are subjoined to the current training set and the 25 farthest conformations are chosen. Second, out of all the MD simulations generated in the MHM, one structure for each adsorbate-surface system is randomly chosen and subjoined to the current training set plus the previously chosen 25 farthest minima. Out of this group, the ten farthest points are chosen. Last, all minima resulting from the unconstrained local optimization are sampled with the current data plus the previously chosen structures and a third FPS is performed outputting the five farthest conformations.

For all of the 40 chosen geometries, the related energies and forces are calculated at the DFT level and these data are added to the training set. With this updated training set, the next training iteration is started, beginning with the GAP fitting.

The process is repeated until sufficient accuracy is obtained. This accuracy is on the one hand assessed by qualitative analysis of the obtained structures. On the other hand, the accuracy is quantified by validation, whereby a recursive validation mechanism is applied. The structures chosen via the three consecutive FPS steps in iteration n also serve as the validation set of iteration n. This validation set is later on added to the current training set, which then builds up the new training set for iteration $n+1$. By doing so, it is ensured, that the validation set is not involved in the data set, which the current GAP in iteration n is trained on.

Using this iterative process, a system-specific GAP is trained. The results of the training are depicted in section 3.2. In the following, the computational realization of the workflow is detailed.

3.1.2 Computational details

This section gives the computational details for the main building blocks of the previously explained iterative GAP training. The different blocks are implemented via the `python` programming language. Overall, three different GAPs have been trained accompanied by a workflow refinement, which are referred to as GAP1, GAP2 and GAP3 in the following.

Settings for DFT reference calculations

DFT calculations are performed for the initial training set as well as the new structures produced via the iterative training approach. The `FHIaims` software [77] is used in the `Atomic Simulation Environment (ASE)` [78]. The revPBE exchange-correlation functional is used with the 'light' default settings as defined by `FHIaims` and a Gaussian smearing of 0.1 eV. In order to model atoms and molecules on a surface, Tkatchenko-Scheffler dispersion corrections with screened vdW interactions are applied. Previous to application, the parameters were approved. The testing can be found in figure A.2.

The surface and surface+adsorbate systems are build with the `CatKit` tool [79]. For the surface and surface+adsorbate systems, periodic boundary conditions are applied and the periodic cell size is set to 3x3x4, consistent to literature [8], with an additional 10Å vacuum layer. A lattice constant of 3.85Å is chosen. This choice is based on a preceding relaxation of bulk Rhodium atoms in a 3x3x4 periodic cell with fixed angles and a force threshold of 10^{-2} eV/Å. The Brillouin zone is sampled with a k-grid of (4x4x1). For these settings, the energy converges, which is visualized in figure A.3.

The DFT singlepoint calculations are performed with a charge density based convergence criterion automatically set depending on the number of atoms in the system. Up to 6 atoms, a value of $10^{-6}\mathrm{e}/\mathrm{a}_0^3$ and in systems with 6 to 60 atoms in one periodic cell, a value of $10^{-6} \cdot n_{\mathrm{atoms}}\ \mathrm{e}/\mathrm{a}_0^3$ is set. The gasphase molecules are optimized with a force threshold of 10^{-2} eV/Å and collinear spin setting set according to table A.4. The upper two layers of the clean Rhodium surfaces are optimized with a force threshold of 10^{-2} eV/Å, whereas the lower two Rhodium layers are fixed as those represent the bulk Rhodium within the previously optimized lattice. The adsorbate+surface systems are optimized with a force criterion of 10^{-1} eV/Å and constraints, such that the Rhodium atoms are fixed and the adsorbate bonds are limited to a length of $1.05 \cdot l_{\text{gasphase}}$ with l_{gasphase} relating the optimized gasphase bondlength for each molecule.

The above defined DFT settings are applied to the initial training set as well as to the DFT singlepoint calculations of the 40 conformations chosen via the FPS steps in each iteration.

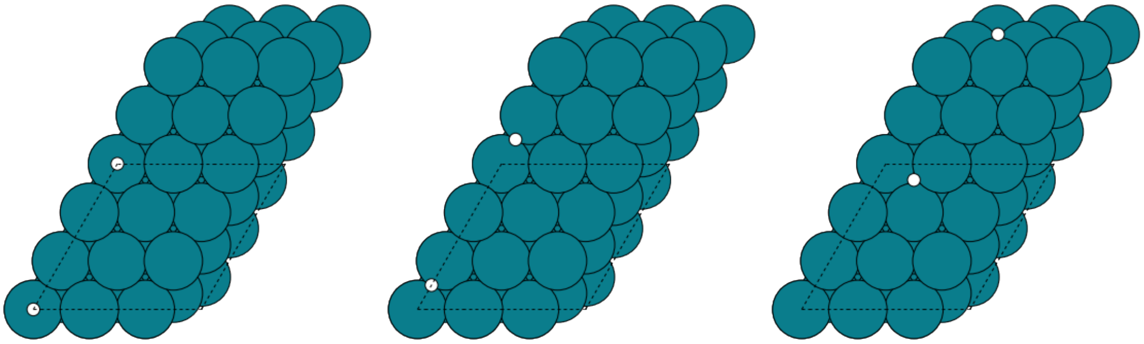
Settings for GAP training

The iterative GAP training is performed using the `QUIP` program package [80] and the `quippy` interface. The inital training set for GAP2 and GAP3 encompasses 175 geometries and specifically involves the components summarized in table 3.1. In the initial training trial GAP1, the training set involves 68 geometries as listed in table A.1.

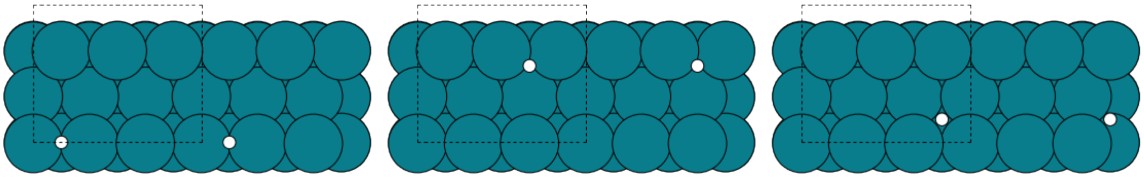
Table 3.1: List and specification of the different components included in the initial training set of GAP2 and GAP3, classified into five groups.

Class	**Components**	**Specification**
atoms	C, O, H, Rh	/
dimers	CC, CO, CH, HH, OH, OO	dimers with varied distances (in Å) $d = (r_{\text{covalent, 1}} + r_{\text{covalent, 2}} + n \cdot 0.1)$ with n ranging from 0 to 4, taken from [81]
gasphase molecules	CO, H_2, H_2O, OH, CH, CH_2, CH_3, CH_4, COH, CHO, CHOH, CHCO, CH_2CO, CH_3CO, CH_3CHO, CH_3CHOH, CH_3CH_2OH	selected optimized gasphase molecules relevant for the syngas conversion on Rhodium
surfaces	Rh(111), Rh(211)	periodic cell consisting of 36 Rh atoms in 4 layers and a 10 Å vacuum layer
surface + adsorbate	single atoms or molecules adsorbed to Rh(111) or Rh(211) surfaces	periodic cell consisting of 1 adsorbate attached to a Rhodium surface on different adsorption sites, as defined in table A.3 and visualized in figure 3.2

For GAP2 and GAP3, three adsorption sites for each Rh(111) and Rh(211) are considered, which are visualized in figure 3.2 and defined in table A.3. The initial training trial of GAP1 just considered one adsorption site per adsorbate as defined in table A.2.



(a) Rh(111), top site **(b)** Rh(111), bridge site **(c)** Rh(111), hollow face centered cubic (fcc) site



(d) Rh(211), step bridge site **(e)** Rh(211), hexagonal closed packed (hcp) terrace site **(f)** Rh(211), 4-fold site

Figure 3.2: Illustration of adsorption sites for Rh(111) (a) to (c) and Rh(211) (d) to (f) surfaces considered for the surface + adsorbate systems in the initial training set.

For the training, multiple hyperparameters need to be preset. Those hyperparameters requiring presetting are listed and described in table 3.2. By the `default energy sigma`, the accuracy of the fit is determined, which is claimed via the `energy percentage` such that

$$\texttt{default energy sigma} = \mathrm{SD}(\texttt{FP}) * \texttt{energy percentage}, \tag{3.1}$$

whereby SD refers to the standard deviation and `FP` is the fit property. For the calculation of the `default energy sigma`, internal fit properties need to be defined. For the 2-body fit, the fit property is the energy per bond and for the SOAP it is the energy per atom. The `energy percentage` then gives the fraction of the standard deviation of this certain fit property taken into account to calculate the `default energy sigma`. As the `default energy sigma` represents the fit accuracy, it is sought that the training and validation errors during the iterative training level with the `default energy sigma`.

The second regularisation parameter, the `force atom sigma`, describes the accuracy of the fitted atomic forces. The `force atom sigma` is scaled, such that higher forces have a higher force atom sigma, according to the following equation

$$\texttt{force atom sigma} = \texttt{sigma}_{\texttt{min}} + \frac{\texttt{C}}{\texttt{A}} \log(1 + \texttt{A} * F + \texttt{A} * f). \tag{3.2}$$

Thereby, $\texttt{sigma}_{\texttt{min}}$ is the lower threshold for the `force atom sigma`, `C` and `A` are additional hyperparameters, F is the normed force and f is the normed force per atom. In order to lower the weighting of the predominating Rhodium atoms, the parameter values for the Rhodium versus the non-Rhodium atoms are differently set. For the Rhodium atoms, $\mathtt{sigma}_{\mathtt{min}}$ is set to 0.1eV/Å, `C` to 0.1 and `A` to 0.01. For the non-Rhodium atoms, the $\mathtt{sigma}_{\mathtt{min}}$ is set to 0.05eV/Å, `C` to 0.05 and `A` to 0.01.

Table 3.2: List and short description of the preset hyperparameters for the GAP training.

Hyperparameter	**Short description**
`default energy sigma`	regularisation parameter of the fit property energy representing the accuracy of the energy data & relative weight of them in fit
`force atom sigma`	regularisation parameter of the fit property force representing the accuracy of the force data & relative weight of them in fit
`atom sigma`	Gaussian smearing width σ_{ab} of atom density for SOAP as given in the density expression in equation 2.13
$\mathtt{n}_{\mathtt{sparse}}$	number of representative points
`theta`	width of gaussians in the squared exponential kernel expression for the 2-body descriptor, as defined in equation 2.11
`zeta`	power, which the SOAP kernel is raised to in order to sensitise to changes in atomic positions (equation 2.16)
`delta`	scaling/ partitioning of kernel per descriptor
`cutoff`	radial cutoff, which represents the highest distance each descriptor takes into account
`cutoff transition width`	distance across which the SOAP kernel is smoothly taken to zero
$\mathtt{l}_{\mathtt{max}}$, $\mathtt{n}_{\mathtt{max}}$	maximum number of radial ($\mathtt{n}_{\mathtt{max}}$) and angular ($\mathtt{l}_{\mathtt{max}}$) indices summed over in the spherical harmonic expansion of the neighbour density in equation 2.14

The initial guess for the hyperparameter values is based on previous work [81]. The zeroth training iteration is splitted in two stages: first, a 2-body potential is fitted and second, a double SOAP potential. This is done, as during the zeroth iteration hyperparameter such as the `delta` values are calculated and taken into account during the following iterations. The `delta`, which is previously described as the scaling of the kernel per descriptor, describes the standard deviation of the Gaussian process. It is calculated in the zeroth training iteration via the following equations:

$$\mathtt{delta}_{\mathrm{2b}} = \frac{1}{3}\,\mathrm{SD}\left(\mathbf{ae} - \mathbf{ae}_{\mathrm{baseline}}\right) \tag{3.3}$$

$$\mathtt{delta}_{\mathtt{SOAP}} = \mathrm{SD}\left(\mathbf{ae} - \mathbf{ae}_{\mathrm{2b}}\right). \tag{3.4}$$

Thereby, SD refers to the standard deviation and **ae** to the vector of the atomisation energies (AEs) per atom for the surface+adsorbate structures involved in the training set, $\mathbf{ae}_{\text{baseline}}$ are the energies calculated by the baseline potential and $\mathbf{ae}_{\text{2b}}$ the energies calculated by the 2-body potential. Note that the heuristics of the `delta` calculation are adjusted during a hyperparameter finetuning, which is detailed in section 3.2.2.

The regularisation parameters `default energy sigma` and `force atom sigma` are based on the given input data, which are iteratively updated. Therefore, these parameters are calculated at the beginning of each training iteration and vary during the iterative process. The other hyperparameter values are given in table 3.3 and are kept the same over all iterations in the three training trials GAP1 to GAP3, if not stated otherwise in the result section 3.2.

Table 3.3: List of initial hyperparameters kept constant during the iterative GAP training.

Hyperparameter	**Unit**	**2-body**	**Double SOAP**	
			1. SOAP	2. SOAP
`atom sigma`	Å	-	0.3	0.6
$\mathtt{n}_{\mathtt{sparse}}$	1	15	2000	2000
`theta`	Å	1.0	-	-
`zeta`	1	-	4	4
`cutoff`	Å	5.0	3.0	6.0
`cutoff transition width`	Å	-	0.5	1.0
$\mathtt{l}_{\mathtt{max}}$	1	-	3	3
$\mathtt{n}_{\mathtt{max}}$	1	-	9	9
`energy percentage`	%	15	1	1
$\mathtt{delta}_{\mathtt{GAP1}}$	eV	0.55	0.22	0.22
$\mathtt{delta}_{\mathtt{GAP2}}$	eV	0.16	0.032	0.032
$\mathtt{delta}_{\mathtt{GAP3}}$	eV	0.16	0.032	0.032

For the third training round, GAP3, a hyperparameter test is performed during the last training iteration. The results build up the final hyperparameter set for the final potential and are given in section 3.2.2.

Settings for constrained minima hopping and unconstrained local optimization

For the generation of new training data, constrained minima hopping and unconstrained local optimizations are performed in the `Atomic Simulation Environment (ASE)` using the previously trained GAP as a calculator.

In every iteration, the constrained minima hopping is simulated for the initial surface+adsorbate geometries, in the following referred to as start geometry and as listed in table 3.1. Similar to the initial geometry optimizations using DFT, as described before, the structures are constrained so that the Rhodium atoms are fixed and the bonds of the adsorbates are restricted to break. This is ensured by applying a Hookean constraint on every occuring adsorbate bond, which responds with a Hookean force with an spring constant of $10\,\mathrm{eV}/\text{Å}^2$ once the bondlength exceeds $1.05 \cdot l_{\text{start geometry}}$. As the bondlength of the start geometry is at the maximum elongated to $1.05 \cdot l_{\text{gasphase}}$, the maximum possible bondlength for the resulting conformations from the constrained minima hopping is restricted to $1.1025 \cdot l_{\text{gasphase}}$.

The minima hopping is repeated for $\mathrm{n_{hops}}$ loops and different start conditions are applied for the three different GAP trainings, which are summarized in table 3.4. For the other parameters, the default values are kept.

Table 3.4: List of MHM conditions for the three different training approaches GAP1, GAP2 and GAP3 with N denoting the iteration number.

Name	**Short description**	$\mathbf{n_{hops}}$	$\mathbf{E_{diff0}}$ **in eV**	$\mathbf{T_0}$ **in K**
GAP1	soft	$5 \cdot 20$	0.5	1000
GAP2	hard	40	5	2000
GAP3	increasing	40	$0.5 + N \cdot 1$ $(E_{\text{diff0,max}} = 5\text{eV})$	$1000 + N \cdot 200$ $(T_{0,\,\text{max}} = 2000\text{K})$

The local optimization is done using the BFGS method as implemented in `ASE` for the global minima of each adsorbate-surface pair found in the prospective minima hopping simulation, evaluated by the GAP potential energy. The Hookean constraints on the adsorbate molecules are repealed. The optimization is conducted with a force criterion set to $5 \times 10^{-2}\mathrm{eV}/\text{Å}$.

3.2 Results and discussion

In this section the results of this work's GAP training are depicted and discussed. First, the initial training approaches with its required adjustments and refinements are enlightened. Afterwards, the final training approach is discussed and related.

3.2.1 Initial training approaches

In this work, three detached training rounds are performed, named GAP1, GAP2 and GAP3. This section summarizes the results of the initial approaches with the hyperparameters given in section 3.1.2. The training progressions of the three rounds are illustrated in figure 3.4. The mean absolute error (MAE) of the AE per atom serves as the pictured error measure.

The first training trial, GAP1, is started with an initial training set size of 68 structures and stopped after eight iterations. At the beginning, the validation and training errors quickly decrease and are already below the regularisation parameter `default energy sigma` in iteration two. Because of this exceed of the `default energy sigma` compared to the training and validation MAE, it is adjusted after four iterations (zero to three). By lowering the `default energy sigma`, it is tested, whether even lower training and validation errors can be achieved. Until this point, this regularisation parameter is calculated by the energy percentage times the standard deviation of all geometries in the current training set. However, due to the inclusion of the AEs of atoms, dimers and gasphase molecules, the distribution of the AEs is broad and unsymmetrical. Therefore, the calculation of the standard deviation is adjusted to that of the distribution of the surface+adsorbate systems only. This adjustment was implemented for iteration four and onwards and also for the entire training rounds GAP2 and GAP3. For GAP1, it results in a sudden decrease of the `default energy sigma` from iteration four to five as illustrated in figure 3.4 (a).

By qualitative analysis, low movement of the adsorbates on the Rhodium surfaces is observed. As a consequence, a low variety of minima and adsorption sites is found during the minima hopping in each iteration. The assumed reason for this is the soft minima hopping conditions and the low variation in the geometries, the minima hopping is started with. Note, that due to an error in the FPS implementation, several same structures are chosen in the subsequent FPS steps and added to the training step in case of GAP1. The error is corrected for the training of GAP2 and GAP3 .

Despite these observations, the first training trial reveals the following: in general, the coupling of GAP training and training data generation by MHM is possible. The structures generated by global optimization receive more chemical quality from iteration to iteration. This improvement in chemical quality is exemplified in figure 3.3. Whereas at the beginning, some non-chemical adsorbate assemblies appeared, this is not the case for the later iterations.

Therefore, the applicability and continuous improvement of the workflow is approved.

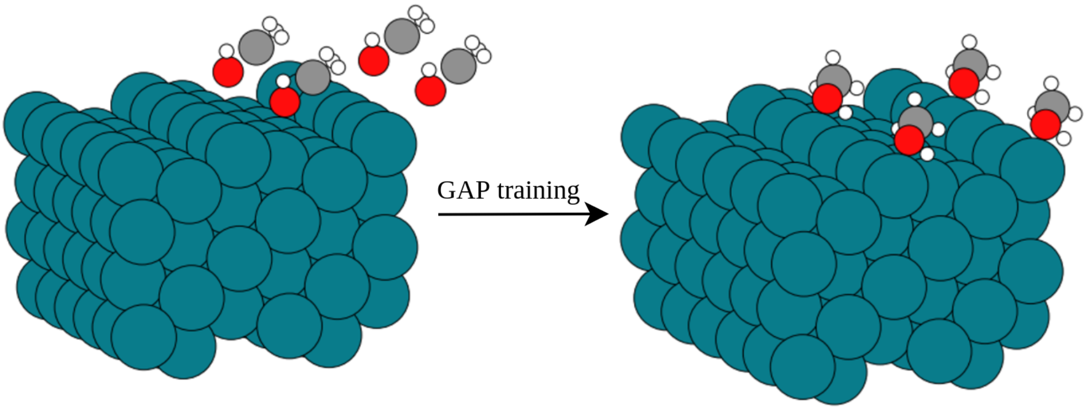


Figure 3.3: Illustration of the quality improvement of the produced minima in the proceeding GAP training iterations. The left structure appears during the first training iteration, whereas the right structure is produced in an advanced iteration.

With these findings, another training round, GAP2, is started. For the training of GAP2, more adsorption sites are considered for both Rh(111) and Rh(211) surface facets. Instead of starting five separate, shorter minima hoppings for each adsorbate-surface pair, different sites for each adsorbate on each surface are taken into account. The initial training set expands to 175 geometries. Moreover, the initial minima hopping conditions are adjusted according to table 3.4, ensuring higher mobility and increasing the possibility to overcome barriers.

The training progress of GAP2 is illustrated in figure 3.4 (b). At the beginning of the iterative training, the training error (MAE of the AE per atom) as well as the `default energy sigma` rise in value. This happens due to the inclusion of structures from the iterative training, produced by the minima hopping and local optimization of structures with the potential training in the zeroth iteration. The quality of the produced structures chosen via the FPS can be underlined by the validation error. Up to the sixth iteration (n_{train}= 415), the validation error lowers and the minima produced via the MHM are qualitatively getting better. Thereafter, the validation error surges. This increase in error is mainly caused by the sampling of exceptional structures via the FPS, which do not comply with the overall training set.

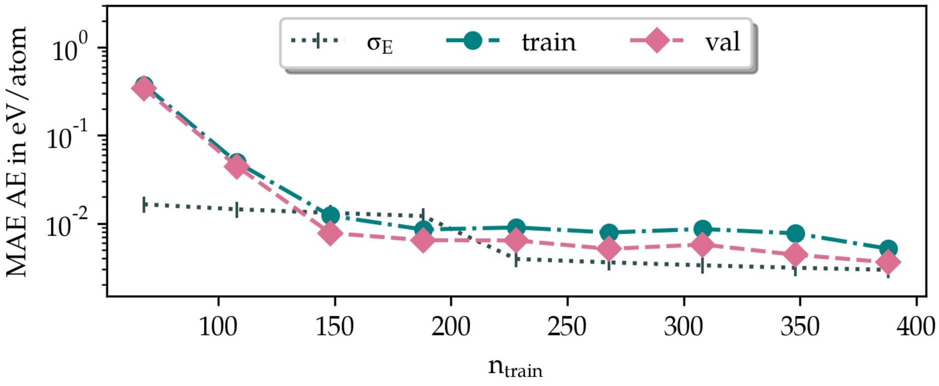
From a chemists point of view, the structures and energies qualitatively improved until iteration six. The training set - to that point - includes the most important structures. Afterwards, 'unchemical' structures are sampled via the FPS, which chooses those structures most different from the existing data set (the farthest points). The reason for this is less attributable to the FPS algorithm, but rather to the hard MHM settings, as those structures mainly appear due to high temperatures and the applied constraints. When the validation error rises, the FPS mainly chooses exceptions and extraordinary transitions occuring in the MDs. Those exceptions for example include structures, for which non-Hydrogen atoms diffuse into the Rhodium layers. Alternatively, for some structures, the adsorbates desorb into the vacuum, diffuse through the vacuum layer and re-adsorb on the ground layer of the above periodic cell, which is built by bulk, non-surface-optimized Rhodium atoms. This especially happens to systems with molecular hydrogen as an adsorbate, which is in accordance to the fact that molecular hydrogen does not adsorb to Rhodium surfaces, but causes errors in the prediction.

As a consequence, different possibilities to restrict those occurrences are considered. Especially, the diffusion into the gasphase is an event, which should not be learned or enhanced by the potential. Most notably, the transition of the entire vacuum should be circumvented. In a first attempt, an additional plane constraint is added during the minima hopping, pushing the adsorbate back towards the surface and therefore restricting the molecule from desorption. This is found to be excessively time consuming in relation to the achievement. Moreover, finding new minima by first slight diffusion into the vacuum, second rearrangement of the adsorbate atoms and third alignment on a new adsorption site should not be restricted.

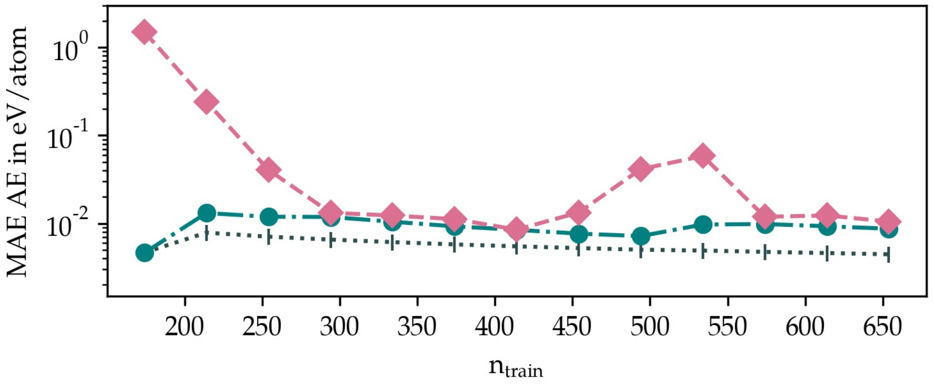
Therefore, the input data to the FPS are filtered. Only those structures, for which all atoms of the adsorbate molecule are within a height range of 14.5 and 24.5 Å (compare table A.3 for the height of the Rhodium surfaces), are considered. This ensures the exclusion of desorbed molecules on the one hand and of atoms moved too far into the bulk Rhodium layers on the other hand. However, as this problem appeared with a rise in minima hopping conditions, a compromise between training round 1 and 2 is developed. Comparing GAP2 to GAP1, a more diverse training set is iteratively built including higher variability in the found minima as well as a higher mobility during the dynamics. The problems include the appearance of extraordinary structures, which challenge the stability of the trained potential.

A new training round, GAP3, is started, which turns out to be the last attempt. The initial training set developed for GAP2 is also considered as the basis of the GAP3 training. In order to ensure the development of a stable potential including higher dynamics and variability, the minima hopping starting conditions for GAP3 were step wise increased from iteration to iteration. According to table 3.4, the conditions start from the soft conditions used in GAP1 and rise up to the hard conditions used in GAP2 with progressing iterations, slowly accustoming the potential to the harder settings. Using this approach, the minima hopping shows a good compromise between movement and a proper variability in chemical adsorption sites without stability problems.

(a) GAP1



(b) GAP2



(c) GAP3

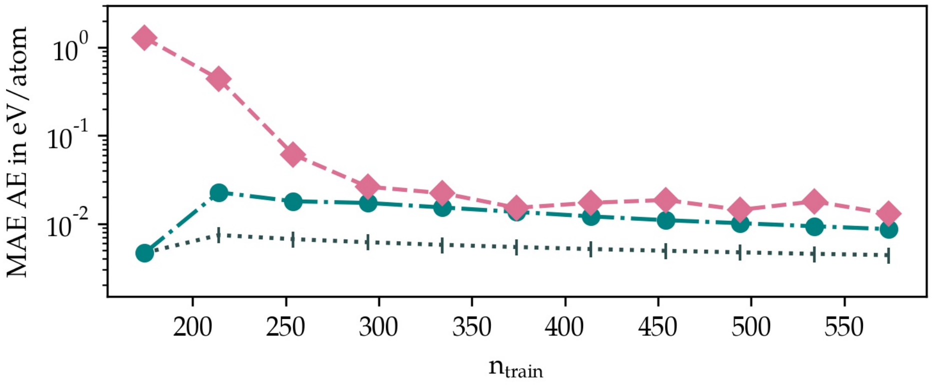


Figure 3.4: GAP training progress for (a) GAP1, (b) GAP2 and (c) GAP3 as mean absolute error (MAE) of the atomisation energy (AE) per atom versus number of training data. The lines interpolate the `default energy sigma` σ_E (black dotted line), the training MAE (teal dashed line) and the validation (val) MAE (rosa dashed line) between the marked iterations. (a) GAP training round 1 (GAP1) from iteration 0 (n_{train}= 68) to 8 (n_{train}= 388). After iteration 3 (n_{train}= 188), the calculation of the `default energy sigma` is adjusted. (b) GAP training round 2 (GAP2) from iteration 0 (n_{train}= 175) to 12 (n_{train}= 655). (c) GAP training round 3 (GAP3) from iteration 0 (n_{train}= 175) to 10 (n_{train}= 575).

The training progress of GAP3 is pictured in figure 3.4 (c). Compared to GAP2, the validation error reaches slower the level of the training error. A reason for this might be the adjustment of the initial minima hopping conditions, whereby from iteration five onwards (n_{train}= 375) the initial conditions were kept constant. Afterwards, the training error and `default energy sigma` further slightly decrease and the validation error fluctuates with decreasing tendency.

While comparing the absolute level of the `default energy sigma` and the training and validation error, a gap can be observed. Until now, the MAE of the AE per atom is calculated taking into account every surface+adsorbate structure in the respective training or validation set. These structures include the initial DFT optimized input structures, as well as the constrained minima, random MD structures and unconstrained local minima. The highest errors are observed for those structures produced via the MD simulations or the unconstrained local optimizations.

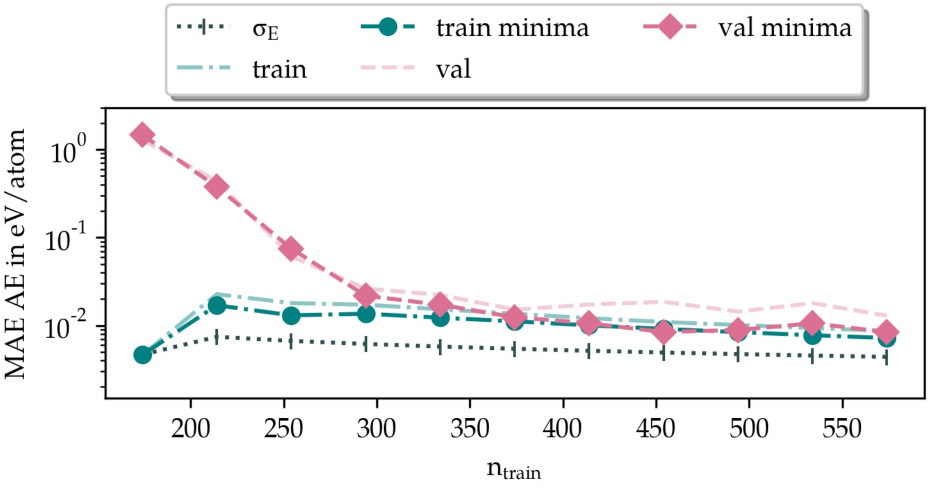


Figure 3.5: GAP training progress for GAP3 as mean absolute error (MAE) of the atomisation energy (AE) per atom versus number of training data. The marker and the corresponding interpolation refer to the `default energy sigma` σ_E as well as the training MAE and the validation (val) MAE of the minima only. The lighter lines indicate the errors calculated for all surface+adsorbate training structures, as pictured in figure 3.4(c).

In contrast, this work mainly focuses on the minima. Therefore, the error measure is adjusted such that just the errors of the minima included in the respective training or validation set are calculated. Regarding these changes, the training and validation errors approach more and more the `default energy sigma`, as illustrated in figure 3.5. In this figure, the training and validation errors calculated for all surface-adsorbate structures are directly compared to the errors for the minima only. Overall, the errors respecting the minima only undercut the errors for all the surface-adsorbate structures. Therefore the prediction accuracy is even higher than expected from the error analysis of all surface-adsorbate structures.

Moreover, the minima produced via the iterative approach qualitatively correspond to the expectations. The variety and mobility is high and only at the beginning of the iterative process, some 'unchemical' structures appeared. Additionally, the GAP3 seems to be stable in the high temperature range.

To conclude, three different potentials, namely GAP1 to GAP3, are trained. Thereby, the process is continuously improved until sufficient performance, which is monitored by quantitative and qualitative criteria. GAP3 shows the best compromise between stability of the potential and variety of the produced minima, as well as low training and validation errors near to the `default energy sigma`, which represents the accuracy of the fit. Therefore, GAP3 and the corresponding generated training set is taken as the input for a hyperparameter refinement and the final training approach, for which the results are elucidated in the following.

3.2.2 Hyperparameter finetuning

This work's initial hyperparameter choice is based on prior research findings [81], as detailed in section 3.1.2. With this initial parameter set, the GAP training workflow and refinement is developed. For the final potential, accuracy can be increased by fine-tuning the hyperparameters. This is done by repeating iteration number ten of GAP3 with varied hyperparameters.

As each tested variation requires a new fit to monitor the effect, not a holistic hyperparameter grid search is performed, but a strategic variation. Overall five parameters are varied: the `default energy sigma`, $\mathtt{l}_{\mathtt{max}}$, $\mathtt{n}_{\mathtt{max}}$, the `delta` of the 2-body and the `delta` for the two SOAPs. The tested hyperparameter sets are summarized in table 3.5.

Starting from the initial hyperparameter set (0), which corresponds to the hyperparameters applied in iteration number ten of the GAP3 training, first the `default energy sigma` is step-wise lowered and the other hyperparameters are kept constant. This results in an immediate decrease of the training error, but just a slight decrease of the validation error, which can additonally be seen in figure A.5 (1) and (2).

As a next step, the test is repeated with increased hyperparameters $\mathtt{l}_{\mathtt{max}}$ and $\mathtt{n}_{\mathtt{max}}$, which are defined in table 3.2. This corresponds to the sets (3) to (5). The adjustment additionally slightly decreases the training error with only low or in case of set (5) negative effect on the validation error. Thus, the training and validation errors diverge, which is unfavorable and leads to overfitting. As none of the tested hyperparameter sets improve both the training and validation errors, the further hyperparameter finetuning is proceeded with the initial values for the `default energy sigma`, $\mathtt{l}_{\mathtt{max}}$ and $\mathtt{n}_{\mathtt{max}}$.

In order to increase the effect on the validation error, the `delta` values, which correspond to the scaling of the kernel, are varied. As the `delta` values are calculated in the zeroth iteration for each training round (GAP1 to GAP3), new `delta` values are also calculated for the zeroth iteration of GAP3, but with new heuristics to calculate it. First, the heuristic for the calculation of the 2-body `delta` is adjusted. Hitherto, the 2-body `delta` is calculated as the standard deviation of the energy of the surface-adsorbate systems divided by the number of bonds in the system, as defined in equation 3.3.

Table 3.5: Overview of the hyperparameter variation. Eleven different hyperparameter sets ((0) to (10)) are tested, whereas (0) refers to the initial hyperparameter set of GAP3. The five hyperparameters `default energy sigma` σ_E, $\mathtt{l_{max}}$, $\mathtt{n_{max}}$ as well as the `delta` of the 2-body descriptor and the SOAP are stepwise varied. Note that the depicted values refer to one SOAP only. For the double SOAP approach, 2$*$ the listed `delta` values are considered.

Set	**Default energy sigma in eV**	$\mathbf{l_{max}}$	$\mathbf{n_{max}}$	**Delta in eV**	
				2-body	**SOAP**
(0)	0.004	3	9	0.16	0.032
(1)	0.002	3	9	0.16	0.032
(2)	0.001	3	9	0.16	0.032
(3)	0.004	4	12	0.16	0.032
(4)	0.002	4	12	0.16	0.032
(5)	0.001	4	12	0.16	0.032
(6)	0.004	3	9	0.51	0.030
(7)	0.004	3	9	0.51	0.060
(8)	0.004	3	9	0.51	0.12
(9)	0.004	3	9	0.51	0.24
(10)	0.004	3	9	0.45	0.25

The `delta` value of the hyperparameter set number (6) is calculated as the average (instead of standard deviation) of the surface-adsorbate systems' energy divided by the number of atoms (instead of the number of bonds) in the system. This results in an increase of the 2-body `delta` value with low effect on the training and validation error compared to the initial set (0). Therefore, the `delta` for the two SOAPs, which is calculated by the standard deviation of the energy of all surface-adsorbate systems divided by the number of atoms, is step-wise doubled from set (6) to (9). This results in a step-wise decrease in both training and validation error. For set (7), the training and validation error as well as the `default energy sigma` level to the same value. As the `default energy sigma` represents the fit accuracy, a leveling of the errors with the `default energy sigma` is targeted during the training. In this step-wise value lowering, the set (9) achieves the lowest training and validation errors.

Similar values are achieved by calculating the 2-body and SOAP `deltas` taking into account the whole initial training set including the dimers and gasphase molecules and just excluding the single atoms. The effect of this adjustment of the `deltas` can be observed in figure A.5 (10). For hyperparameter set (10), the overall lowest training and validation errors are achieved.

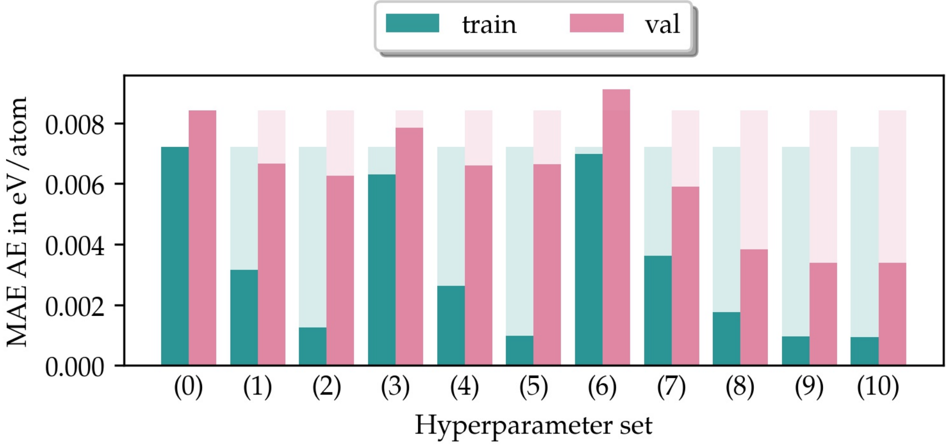


Figure 3.6: Illustration of the improvement achieved by varying the hyperparameter set. The initial hyperparameter set (0) is compared to the other sets (1) to (10) with the training and validation mean absolute error (MAE) of the atomisation energy (AE) per atom as a measure. The dark bars represent the errors of the related set, the underlying light bars repeat the errors of the initial hyperparameter set (0).

The figure 3.6 summarizes the improvement of the errors achieved by the hyperparameter finetuning. In the figure, the different hyperparameter sets are compared to the initial set (0) with the MAE of the AE per atom for the minima included in the training or validation sets as a measure.

To conclude, the variation of the hyperparameters `default energy sigma`, $\mathtt{l}_{\mathtt{max}}$ and $\mathtt{n}_{\mathtt{max}}$ results in a decrease of the training error but only low effect on the validation error. Just the variation of the `delta` values, especially an increase of the SOAP `delta` clearly lowers the validation error. By increasing the `delta` of the SOAP, the weighting of the SOAP is increased in comparison to the 2-body descriptor. As the SOAP is a multibody descriptor, the multibody contributions are now increasingly considered, which might reason the decrease of the errors. Both the lowest training and validation errors are yield by the hyperparameter set (10).

The further studies and final training iteration are conducted with the hyperparameter sets (0) and (10). This is reasoned by comparability in case of set (0) and the lowest overall errors in set (10).

3.2.3 Final training approach

The final training approach assembles the previously explained results. The GAP3 from iteration ten as well as the training and validation set from iteration 10 are used as a basis for the final GAP training. In a first step, the training set for the final GAP is cleaned up. All the structures with adsorbate atoms outside the height range of 16.0 and 23.6 Å are removed, ensuring that the set only includes adsorbate atoms, which are neither in the below Rhodium layers nor too far in the vacuum layer. Consequently, 44 surface-adsorbate structures are removed from the final training set, which now includes 571 geometries.

The final training approach is performed with the initial hyperparameter settings as summarized in table 3.3 with adjustments in the `delta` values as resulted from the hyperparameter testing described in the previous section 3.2.2. For comparison, two of the eleven tested hyperparameter (HP) sets are used in the final training iteration. We define one final potential, which is named finalGAP and trained using the hyperparameter set (10) as detailed in table 3.5. Moreover, the other variation is named initialGAP, whereas initial denotes the usage of the initial hyperparameter set (0).

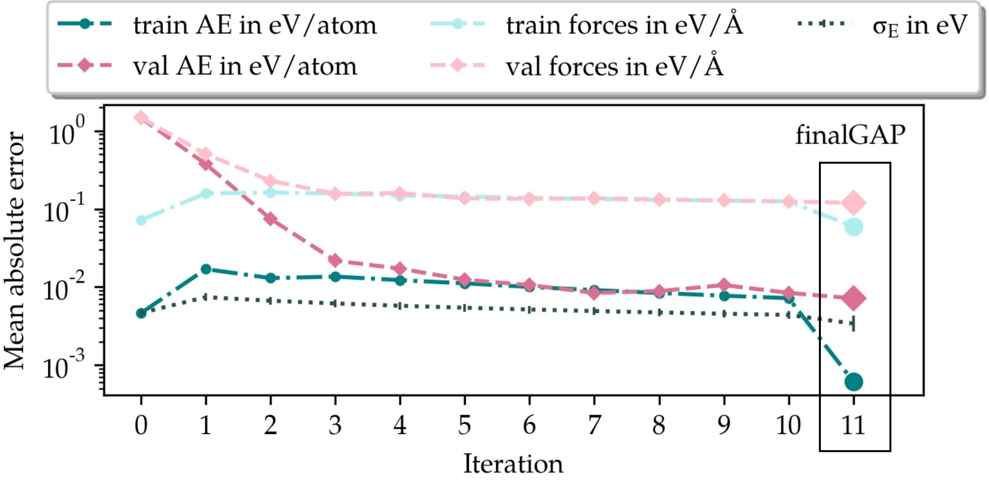


Figure 3.7: GAP learning curve of the final training depicted the mean absolute error (MAE) of either forces or atomisation energy (AE) per atom versus iterations. The final GAP (iteration 11) is based on the plotted iteration 0 to 10 of GAP3 with adjusted hyperparameters. The marker and the corresponding interpolation refer to the `default energy sigma` σ_E as well as the training MAE and the validation (val) MAE of the minima only.

Figure 3.7 overviews the overall training leading to the final potential. In addition to the previously defined error measure - the MAE of the AE per atom-, a second error measure is depicted: the MAE of the forces, as the forces are the second fit parameter in the GAP training, which is monitored throughout the training. The final training iteration, iteration eleven, aligns with the previous iterations. Also the new introduced force errors quickly balance with proceeding iterations. Due to the hyperparameter finetuning, the validation and especially the training error of the final potential is significantly lowered compared to the preceding iterations.

The final GAP leads to a high correlation between the trained GAP energy and the DFT energy, which is emphasized by figure 3.8.

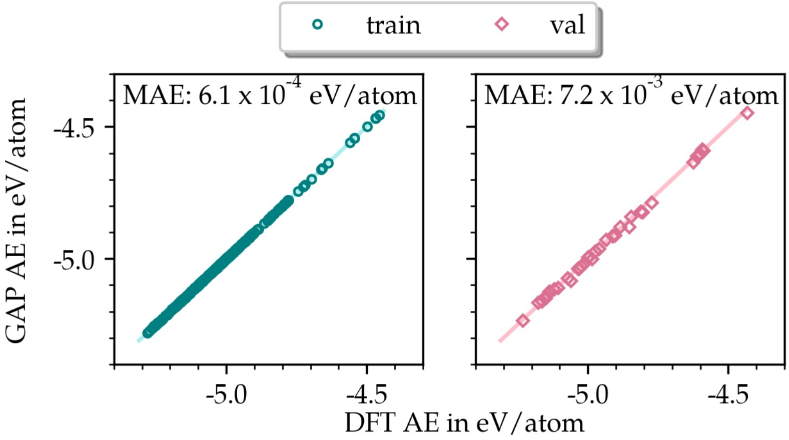


Figure 3.8: Correlation of the GAP versus the DFT atomisation energy (AE) per atom for the final GAP training. The left plot visualizes the correlation of the training data set with a mean absolute error (MAE) of 6.1 × 10^{-4} eV and the right plot this of the validation with a MAE 7.2 × 10^{-3} eV. The solid lines depict the intended full correlation of the energies.

To summarize, three different training rounds have been overall performed in this work: GAP1, GAP2 and GAP3. Throughout these training rounds, the iterative training workflow has been designed and continuously refined. Especially the variation of training data has been improved by the adjustment of the initial training set as well as the minima hopping conditions. The initial hyperparameters for the ML serve as an appropriate set throughout the development. In the hyperparameter finetuning, the accuracy of the predictions could be further improved. With these finetuned hyperparameters, the final GAP is trained, which application represents this work's next step.

Thus, the subsequent chapter 4 details the application of the developed finalGAP to the syngas conversion on Rhodium surfaces.

Chapter 4

Application of the potential to global optimization

In accordance to the overall goal of this work, this chapter portrays the application of the previously trained interatomic potential (see chapter 3) to the global optimization of different adsorbates on Rhodium surfaces. It is divided into two parts: first, the strategy and methodological details are stated and second, the results are presented and discussed.

4.1 Methods and computational details

The goal of this work is the development of a fast and accurate interatomic potential for the prediction of minimum structures of different adsorbates on catalytic surfaces. As an examplary system, the syngas conversion on Rhodium is chosen.

As introduced in section 2.1, dynamical simulations of large, periodic systems over longer timescales are not realizable when approached with ab-initio quantum chemical methods. Accordingly, global optimizations are most commonly approached with empirical force fields, which possess massive losses in computational accuracy. With the rising field of ML interatomic potentials an alternative to the usage of classical force fields appeared. As those potentials, for example GAP, typically provide more accurate predictions than empirical force fields, the GAP approach is also pursued in this work.

In the previous chapter, the development of the training workflow for the ML interatomic potential as well as the final GAP are detailed. The section thereby focuses on the computational improvement of the potential. Besides that, the applicability of the developed potential needs to be tested and verified. This is the focus of this chapter.

Therefore in a first instance, production runs are performed. The trained GAP is used as the underling potential to conduct minima hopping runs for different, new systems. In a first production run, the potential is applied to new surface-adsorbate systems with surfaces and adsorbate molecules already known to the potential, but varied adsorption sites. During the iterative training, the minima hopping is conducted with the same start geometries and sites in every iteration, as detailed in section 3.1. Thus, the first production run tests the applicability of the potential towards new adsorption sites. In a second production run, the potential is applied to the minima hopping of new, unknown start geometries taken from literature [4] in order to test its applicability limits.

By performing production runs, only the general, qualitative applicability of the potential towards different systems can be approved. However, a quantitative evaluation of the found minima is not possible. In order to quantitatively evaluate the applicability of the potential towards the minima hopping of unknown start geometries, the testing is followed by an analysis of the found minima. This is done by comparison to and additional optimizations with DFT.

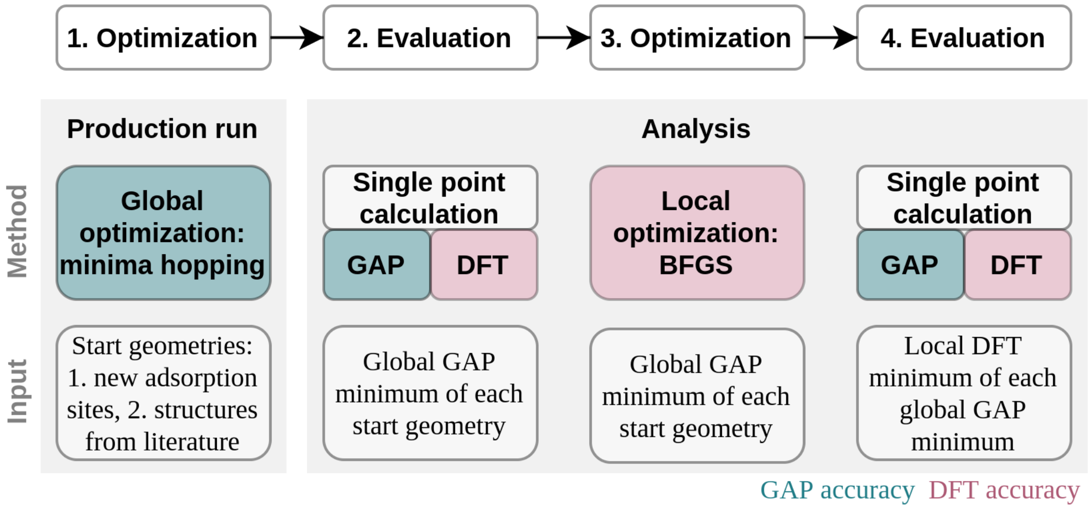

1. Optimization
2. Evaluation
3. Optimization
4. Evaluation
Production run
Analysis
Method
Input
Global optimization: minima hopping
Single point calculation
GAP
DFT
Local optimization: BFGS
Single point calculation
GAP
DFT
Start geometries: 1. new adsorption sites, 2. structures from literature
Global GAP minimum of each start geometry
Global GAP minimum of each start geometry
Local DFT minimum of each global GAP minimum
GAP accuracy
DFT accuracy

Figure 4.1: Illustration of the GAP application procedure. The trained potential is used to perform global optimization production runs (step 1). The output of the production runs is then analysed by evaluation of the energies of the produced global minima (step 2). Afterwards, the global minima are re-optimized using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method (step 3) and again evaluated (step 4). The evaluations and additional local optimization are performed using DFT.

The applicability and evaluation analysis of this work's trained potential towards the global minimum search for the syngas conversion on Rhodium can be summarized in four steps as visualized in figure 4.1. In the process, optimizations and evaluations alternate. First, the start geometries are globally optimized by the MHM using the ML potential. Afterwards, the global minima are evaluated by singlepoint calculations using both GAP and DFT. This ist followed by step 3, a local DFT BFGS optimization of the global minima and again, an evaluation of the discovered local minima of the global GAP minima in step 4. More details on the single steps can be found below.

In addition to the described procedure, further analyses of the found global minima are performed. On basis of the reaction mechanisms (compare section A.1) developed by Yang et al [4], a reduced reaction network for the syngas conversion on either Rh(111) or

Rh(211) surfaces is developed. Moreover, kPCA is conducted in order to reveal patterns and trends in-between the revealed global minima.

In the following, the computational settings for the above described methods are detailed.

Computational details

The application part of this work is implemented via the `python` programming language in the `Atomic Simulation Environment (ASE)` [78]. For the production runs (step 1 in figure 4.1), global optimizations are performed with the minima hopping algorithm as part of the `ASE` package. The minima hopping is performed for n_{hops} = 40 steps with an initial temperature T_0 of 2000 K and an initial difference energy E_{diff0} of 5 eV (compare section 2.2.2). These values correspond to the values applied in the final training. The trained GAP is used as the underling potential for the minima hopping and the systems are constrained in the same way as during the iterative training (compare section 3.1.2).

Two distinct production runs are performed, hereafter referred to as first and second production run. The production runs differ in the start systems for the minima hopping as well as the applied interatomic potential. The systems considered in the first production run sum up to 342 structures. As adsorbates, the optimized gasphase molecules as well as single atoms also included in the initial training set are used. Those adsorbates are attached to four different Rh(111) and 14 different Rh(211) adsorption sites, which is detailed in table B.1. In contrast to the surface-adsorbate systems included in the initial GAP training set, more adsorption sites are considered and the systems of the production run are not locally optimized with DFT prior to the minima hopping.

For the second production run, the low coverage systems of the study of Yang et. al are taken into account [4]. In summary, 33 systems are selected. The considered adsorbates equal those in the first production run, whereas differences are in the Rhodium surfaces. In the work of Yang et al., for the low coverage system one type of Rh(111) surface and two types of Rh(211) with different periodic cell sizes and lattice constants (3.86 Å and 3.866 Å compared to 3.85 Å in the first run) are used. Moreover, the adsorption sites and conformation of the adsorbates on the surfaces differ from those regarded in the first production run. Not every adsorbate is added to every surface. Thus, the surface-adsorbate pairs of the second production run are specified in table B.3.

The previously detailed systems for both production runs can be summarized as follows:

- Adsorbates: C, H, CO, H_2, H_2O, OH, CH, CH_2, CH_3, CH_4, COH, CHO, CHOH, CHCO, CH_2CO, CH_3CO, CH_3CHO, CH_3CHOH, CH_3CH_2OH
- Surfaces for the first run: Rh(111) (four considered adsorption sites) or Rh(211) (14 considered adsorption sites) in a 3x3x4 fcc periodic cell with a lattice constant of 3.85 Å

- Surfaces for the second run: Rh(111) or Rh(211) in a 3x3x4 fcc periodic cell with a lattice constant of 3.86 Å, as well as Rh(211) in a 3x2x3 fcc periodic cell with a lattice constant of 3.866 Å

As underling potentials for the minima hopping, the initialGAP is used in the first production run and the finalGAP, as specified in 3.2.3, in the second run.

The analysis of the explored minima (steps 2 to 4 in figure 4.1) are parted in evaluation, optimization and again evaluation. In both production runs, the analysis is performed for a certain selection of minima. In case of the first run, the global minimum of each minima hopping start geometry as well as the found minima with a maximum difference of 0.5 eV to the global minimum are selected. The choice is based on the GAP potential energies. In case of the second production run, the global minima are chosen on basis of both GAP and DFT energy. Therefore, for each start geometry, two structures for further analysis are chosen.

Table 4.1: Summary of the potentials used for the minima hopping, selection and evaluation analysis of the first (1^{st}) and second (2^{nd}) production run.

	Minima hopping	**Selection**	**Evaluation**
1^{st}	initialGAP	global GAP minima + 0.5 eV minima range evaluated by GAP potential energy	DFT, initialGAP, finalGAP
2^{nd}	finalGAP	global minima both evaluated by DFT and GAP potential energy	DFT, finalGAP

For the selected structures, singlepoint calculations (step 2: evaluation) are performed using the finalGAP or initialGAP and DFT. The DFT calculations are set up similar to those in the previous chapter (compare section 3.1.2). Afterwards, the structures are locally optimized (step 3: optimization) using the BFGS method with DFT revPBE using the `FHIaims` package within `ASE` and a force convergence criterion set to f_{max}=0.05 eV/Å. Those local minima of the previously globally optimized GAP minima are then reevaluated (step 4: evaluation) with energy singlepoint calculations, using the finalGAP or initialGAP and DFT. Additionally, the root mean square deviation (RMSD) of atomic positions is calculated for the structures before and after the local DFT optimizations, such that

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||a_i - b_i||^2}. \tag{4.1}$$

Thereby a and b denote the structures before and after the optimizations and the sum is only taken over the n adsorbate atoms in the systems without the Rhodium atoms, as their positions are fixed. The previous description of the used potentials for the first and second production run and analysis are summarized in table 4.1.

The output minima of the first production round are additionally further analysed by a network analysis on the one hand and kPCA on the other hand. The reduced catalytic network is designed using the `igraph` network analysis software [82]. The depicted energies correspond to the formation energies E_f. Those energies are calculated in accordance to literature [8], such that:

$$E_\mathrm{f} = E_\mathrm{surf+ads} - E_\mathrm{surf} - \sum_{i \in \{C,H,O\}} n_i \mu_i. \tag{4.2}$$

Thereby, $E_\mathrm{surf+ads}$ is the electronic free energy of the surface-adsorbate system on DFT revPBE level, E_surf the energy of the empty surface, n_i the absolute number of elements n of each atomic species i and μ_i are the electronic potential energies, which are defined as

$$\begin{aligned} \mu_\mathrm{O} &= E_\mathrm{H_2O} - E_\mathrm{H_2} \\ \mu_\mathrm{C} &= E_\mathrm{CO} - E_\mathrm{O} \\ \mu_\mathrm{H} &= E_\mathrm{H_2}/2. \end{aligned} \tag{4.3}$$

The above energies E in formula 4.3 correspond to the electronic free gasphase energies of the molecules. For the formation energy of the adsorbate molecules in the gasphase, the following equation 4.2 is adjusted, such that:

$$E_\mathrm{f} = E_\mathrm{molecule} - \sum_{i \in \{C,H,O\}} n_i \mu_i. \tag{4.4}$$

In addition to the network analysis, kPCA of the discovered global minima is performed using the kPCA functionality from the `MLinCRS` code collection [83]. For the kPCA, an average kernel is calculated, which equals the kernel used in the FPS as part of the iterative GAP training. The underlying descriptor is a SOAP with a cutoff radius of 3 Å, an atom sigma of 0.3 Å and $\mathrm{l_{max}}$ and $\mathrm{n_{max}}$ set to 3 and 9 similar to the GAP training. For the plotting, only the first principal components are used. For interpretation, the dimensionality reduction is additionally performed with `Automatic Selection And Prediction tools for materials and molecules (ASAP)` [84] and visualized via the `projection viewer` [85,86].

With the methods and computational details described in this chapter, the application of the ML interatomic potential is investigated. The results are presented and discussed in the following.

4.2 Results and discussion

In this section, the results of the application of the previously trained interatomic potential to the syngas conversion on Rhodium are stated and discussed. First, the results for the application of the potential towards the minima hopping starting from new adsorption sites of the already known systems are given. This is followed by in-depth analysis of the discovered global minima: reduced reaction networks are developed and a kPCA is performed. The subsequent section 4.2.4 provides the results of the application of the trained potential towards the minima hopping of unknown structures. At the end, the overall applicability is assessed, limitations are pointed out and an outlook is given.

4.2.1 Minima hopping from new adsorption sites

In this section, the results of the first production run, as introduced in section 4.1, are stated. This results section is structured along the production run and analysis procedure, as summarized in figure 4.1.

Step 1: Optimization

The first optimization step - the global optimization by minima hopping for the overall 342 surface-adsorbate start systems - produces more than 6000 minima. These minima show a high variability, which can be illustrated by the following figure 4.2. The figure shows the minima ranges for all the 342 start geometries, defined as the difference of the minima to the global minimum of each structure. Thereby, the teal graph illustrates the difference energy of the minima obtained by the global optimization using the initialGAP.

One minima hopping with $n_{\text{hops}} = 40$ steps, outputs approximately 20 minima per start geometry. Thereby, an energy difference of up to 4.4 eV towards the global MHM minimum is obtained. This energy difference accords to the preset initial energy difference $E_{\text{diff, 0}}$ of 5 eV for the minima hopping feedback mechanism (compare section 2.2.2 for details on the parameters).

As the aim of this work is to develop an interatomic potential to reduce the computational complexity, not all of the discovered minima are further evaluated and re-optimized with DFT, but only those minima within a 0.5 eV energy range from the global minimum. This 0.5 eV range is also displayed in figure 4.2. All the minima above this boundary are considered for further analysis and sum up to approximately 1600 surface-adsorbate minima. In the proceeding analysis, this minima selection is first evaluated and then locally optimized using DFT.

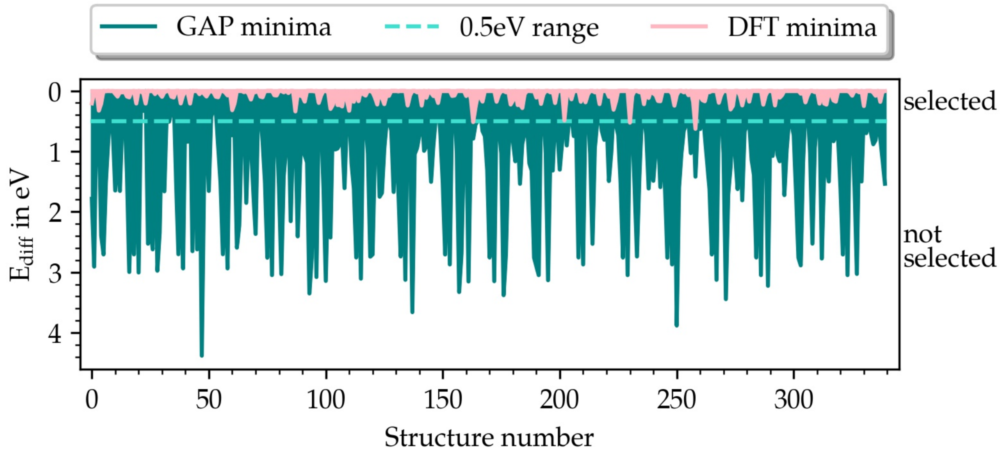


Figure 4.2: Illustration of the minima ranges resulted from global optimization using GAP and the subsequent local DFT optimization. The ranges are depicted as the absolute potential energy difference (calculated by the initialGAP) of the lowest minimum to the other minima for each structure. The GAP minima within a energy difference range of 0.5eV are selected for further analysis.

Step 2: Evaluation

In figure 4.3, the first evaluation is depicted. In the evaluation, the energies calculated by GAP and DFT are compared. The left figure shows the relative frequency of structures as a distribution of the absolute difference between the GAP and the DFT AE. Although these structures are solely optimized with the ML initialGAP, the deviations of GAP and DFT AE are in a range of -0.7 to 0.4 eV.

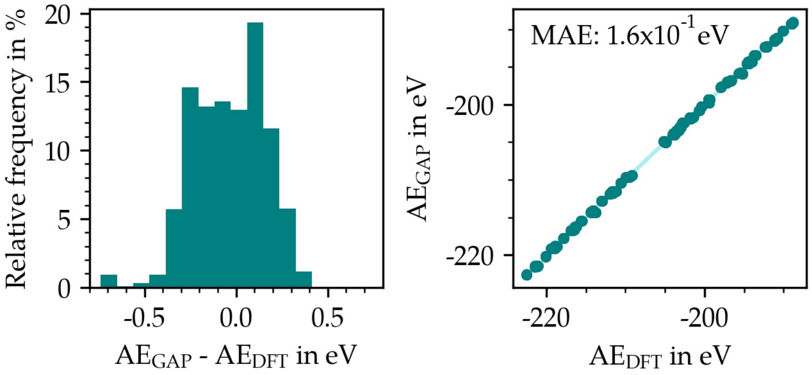


Figure 4.3: Evaluation of the minima obtained via minima hopping using the initialGAP as a calculator. Left, the distribution of the deviation between GAP and DFT atomisation energy (AE) is depicted. Right, the correlation between the two energies is shown.

This good agreement between GAP and DFT AE can be further emphasized by the right plot of figure 4.3. In this plot, the correlation of the GAP and the DFT energy is shown. Overall, a high correlation between GAP and DFT AE is achieved before the DFT optimization. The MAE of the AE accounts 1.6×10^{-1} eV (or 4.1×10^{-3} eV/atom) and is therefore in the same order of magnitude as the validation error during the iterative GAP training.

Step 3: Optimization

In order to further validate the quality of the minima discovered by global optimization using the ML initialGAP, the selected minima are re-optimized. This optimization is conducted as local BFGS optimization using DFT, as the the ML potential is also trained on DFT energies and forces.

By DFT re-optimization, insufficiencies in the global optimization using the initialGAP can be revealed. In order to quantify the optimization, two measures are considered. On the one hand, the difference in the AE before and after the local DFT optimization is investigated in order to reveal the energetic changes due to the local optimization. On the other hand, the RMSD, as defined in equation 4.1, quantifies the geometric changes of the adsorbates.

The energetic changes of the selected minima before and after the local BFGS optimization using DFT are investigated from singlepoint calculations using the initialGAP, the finalGAP and DFT. This investigation is depicted in figure 4.4.

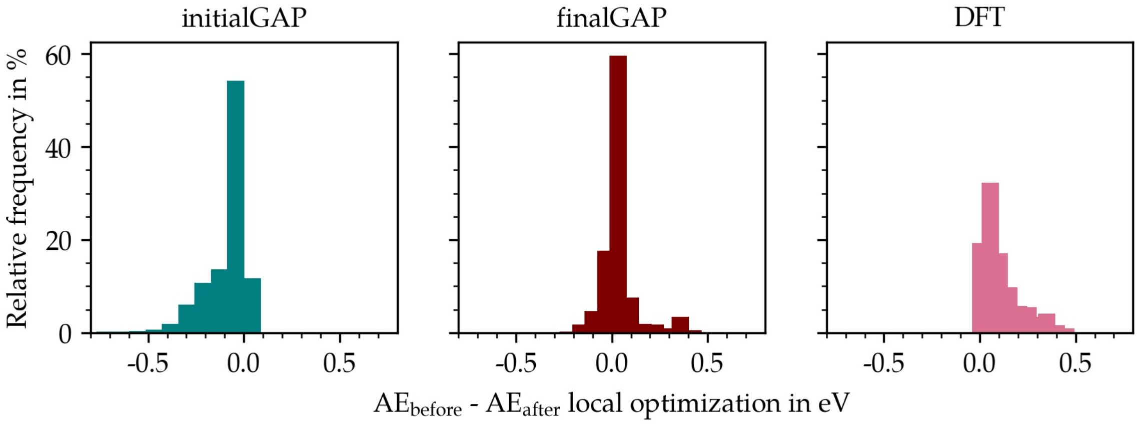


Figure 4.4: Frequency distribution of the selected minima as a function of the difference between the atomisation energy (AE) before and after the local DFT optimization. The difference is evaluated by singlepoint calculations using the initialGAP (left), the finalGAP (middle) and DFT revPBE (right).

When evaluated with the initialGAP, as pictured in the left subfigure, the energy difference shows an unilateral distribution and only declines due to the DFT optimization. For more than 50 percent of the structures, the DFT optimization has no big influence on the

AE. For the rest of the cases, the predicted AE diminishes. This is on the one hand favorable, as the initialGAP is the underling potential for the previous minima hopping and therefore, the minima hopping algorithm works properly. On the other hand, this reveals flaws of the initialGAP, as the global GAP minima can be further optimized by local DFT optimizations.

This can be further emphasized by the DFT evaluation of the minima before and after the local optimization in the right plot of figure 4.4. The optimization potential by DFT BFGS optimization reaches up to approximately 0.5 eV and again, a unilateral distribution is obtained. This comes as no revelation, as the BFGS algrithm is designed such that the energy is minimized and therefore the difference of the energy before and after the optimization is positive.

In the middle plot of figure 4.4, the frequency distribution of the AE difference before and after local optimization evaluated by the finalGAP is mapped. The distribution is more symmetrical than the distribution of the initialGAP. The depicted energy difference is the difference between the global minima obtained with the initialGAP and the locally reoptimized minima with DFT. Therefore, none of the structures are optimized with the depicted finalGAP, it is solely used as an additional calculator for the evaluation of the global GAP and the local DFT minima. The distribution reveals, that the finalGAP approaches more to the DFT energies than the initialGAP.

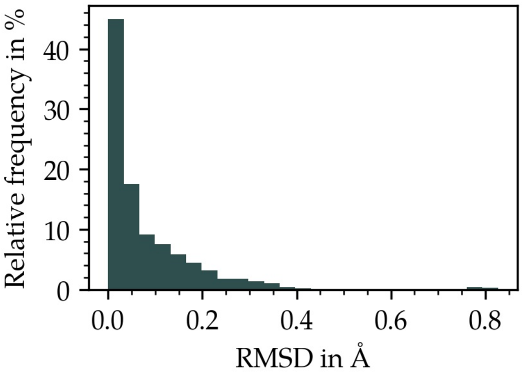


Figure 4.5: Frequency distribution of the selected minima as a function of the RMSD of the adsorbates in the adsorbate-surface systems before and after the local DFT optimization.

The second measure, the geometric changes expressed in form of the RMSD, is displayed in the previous figure 4.5. Nearly 45 percent of the minima re-optimized with DFT show very low displacement in a range of 0 to 0.03 Å due to the optimization. The distribution spreads to a maximum RMSD of 0.4 Å with one outlier at 0.8 Å. As the RMSD is a measure difficult to conceive, figure 4.6 exemplarily shows a global GAP minimum structure before and after the local DFT optimization with an RMSD of 0.2 Å. During the local re-optimization, the hydrocarbon backbone of the adsorbate slightly moves and rotates along the carbon-carbon bond. The RMSD of 0.2 Å is in the higher displacement range yielded due to the local optimization.

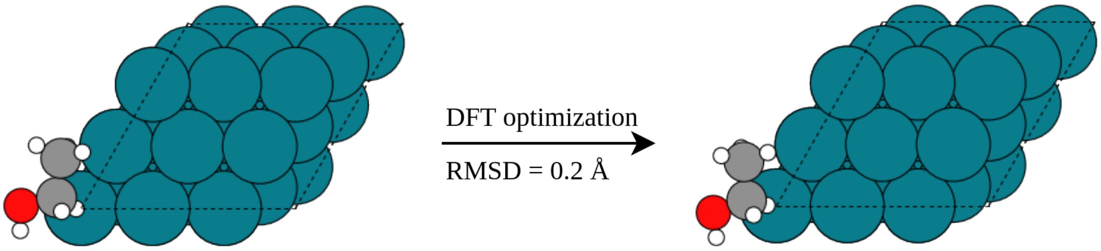


Figure 4.6: Example of the displacement of a global GAP minimum before (left) and after the local DFT re-optimization. The RMSD of the depicted structures accounts 0.2 Å.

Additionally, when redescribing the minima ranges depicted in figure 4.2, the selected 0.5 eV GAP minima range mostly does not diverge through the local re-optimization. The depicted DFT minima ranges are in most cases below the 0.5 eV boundary, with just three exceptions out of 342 systems. This additionally reinforces the observation, that energetically alike minima are obtained by the local re-optimization.

With the energetic and geometric reference ahead, the overall quality of the the global minima discovered by the preceding minima hopping using the initialGAP as a potential can be assessed as adequate.

Step 4: Evaluation

The last analysis step for the first production run is again an evaluation. The energies of the global GAP minima further optimized in a local DFT optimization, are evaluated. Just like in the first evaluation before the local re-optimization, the GAP and DFT energies are compared in figure 4.7. In the upper part, the comparison for the initialGAP, which is also applied in the minima hopping, is shown. The lower part additionally pictures the comparison with the finalGAP, which is solely used for additional evaluation.

In the left half of the figure, the frequency distribution of the difference in the AE of the GAP and DFT is depicted. In case of the initialGAP, which is shown in the upper left subfigure, the distribution is slightly shifted towards right. A reason for this shift might be the underrating of the DFT minima by the initialGAP, which is also previously revealed in figure 4.4. As the initialGAP is trained towards structures obtained by minima hopping, the comparison reveals the differences in predictive power for it's own global GAP minima (step 2) and the other structures, for example the DFT minima. Overall, the deviation between the GAP and DFT energies are in the range of -0.7 to 0.7 eV.

In case of the finalGAP, which is shown in the lower left subfigure, the difference is distributed in a range of approximately -0.5 to 0.3 eV and is not shifted to a certain side. This indicates an even higher accordance of the finalGAP to DFT.

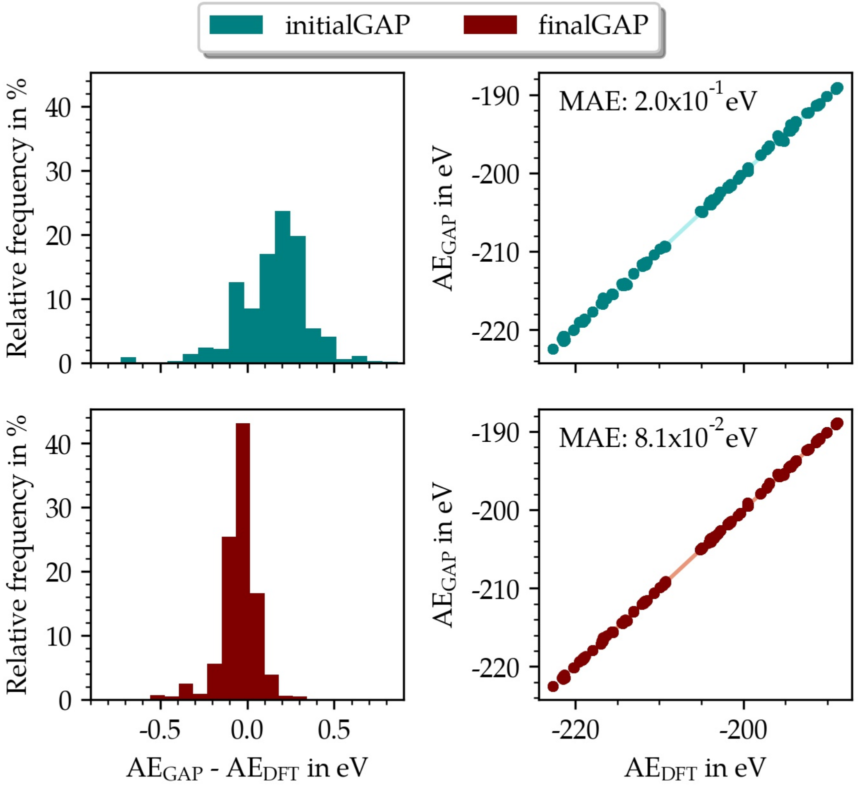


Figure 4.7: Evaluation of the minima obtained via local re-optimization using DFT revPBE as a potential. Left, the distribution of the deviation between GAP and DFT atomisation energy (AE) is depicted. Right, the correlation between the two energies is shown. The upper teal part of the figure refers to the evaluation by the initialGAP, the lower dark red part to the finalGAP.

The correlation between the GAP and DFT AE is likewise high, which is depicted in the right part of figure 4.7. The MAEs level with the MAE of the energies before the local optimization (step 2) and also the validation errors of the GAP training in section 3.2. The MAE of the finalGAP even undercuts the MAE of the initialGAP with a value of 2.0×10^{-1} versus 8.1×10^{-2} eV. Therefore, the second stage of the application test - the second production run of unknown systems - is performed using the finalGAP.

Beforehand, the results of further analysis for the first production run are given, beginning with the kPCA outcomes.

4.2.2 Kernel principal component analysis

In order to further analyse the minima obtained by the minima hopping, kPCA is performed. As introduced in section 2.3.2, kPCA allows to reveal patterns in multidimensional data sets. Here, more than 1500 minima, which are discovered and selected for further optimization in the first production run, are analysed. The following figure 4.8 depicts the results of the kPCA in coordinates of the first two principal components PC1 and PC2. The data points are colored based on the DFT AE of the appurtenant minima.

In the upper part of the figure, the kPCA for the surface-adsorbate systems is pictured including both the Rh(111) and Rh(211) surfaces. In the lower part of the figure, the analysis is divided to the different surface types.

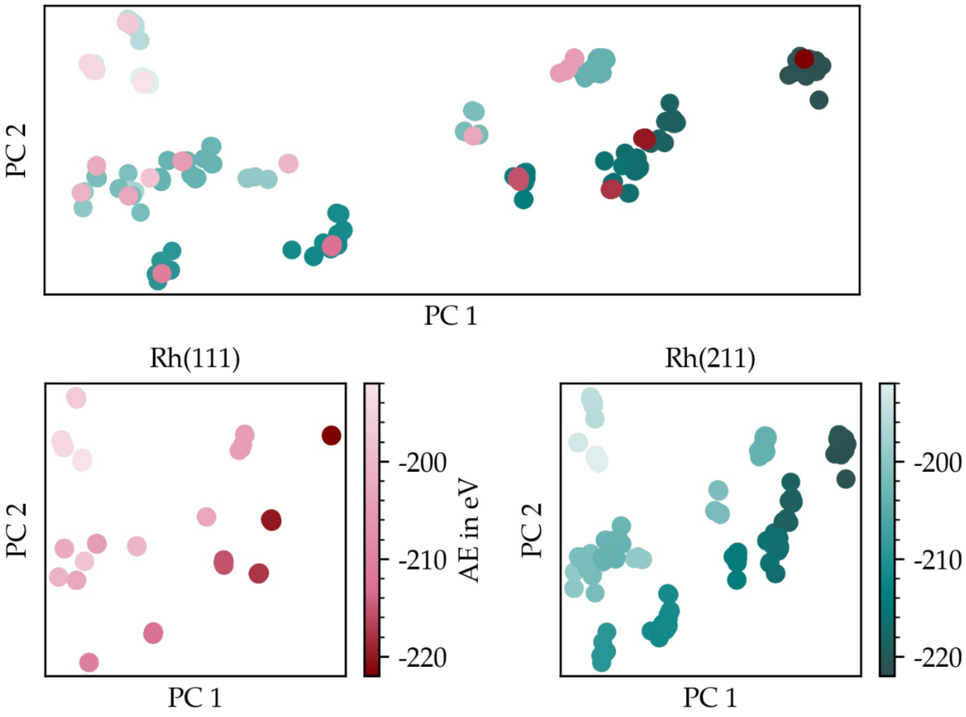


Figure 4.8: Illustration of the kPCA of the MHM GAP minima re-optimized with DFT. The first two principal components PC1 and PC2 are the coordinates, the coloring corresponds to the atomisation energy (AE) calculated by DFT. In the upper figure, the kPCA is summarized for both the Rh(111) and Rh(211) systems. The lower figures differentiate the two surface types.

The systems arrange in distinguishable clusters with similar energies. In the upper left edge of the plots, the systems with the highest AE are located and in the upper right edge the systems with the lowest AE. By comparison of the kPCA for the Rh(111) and Rh(211) systems, it can be observed, that the systems are more centered in case of the Rh(111) systems. The Rh(211) systems are more distributed and some clusters overlap.

Further analysis of the kPCA reveals that each cluster can be allocated to a certain adsorbate. This allocation is indicated in figure 4.9. Two trends are observed: first, the clusters organize in three distinguishable domains along the second principal component (PC2) with either no or one or two carbon atoms. Second, the number of hydrogen atoms increases from the lower left to the upper right part of the illustration (along PC1).

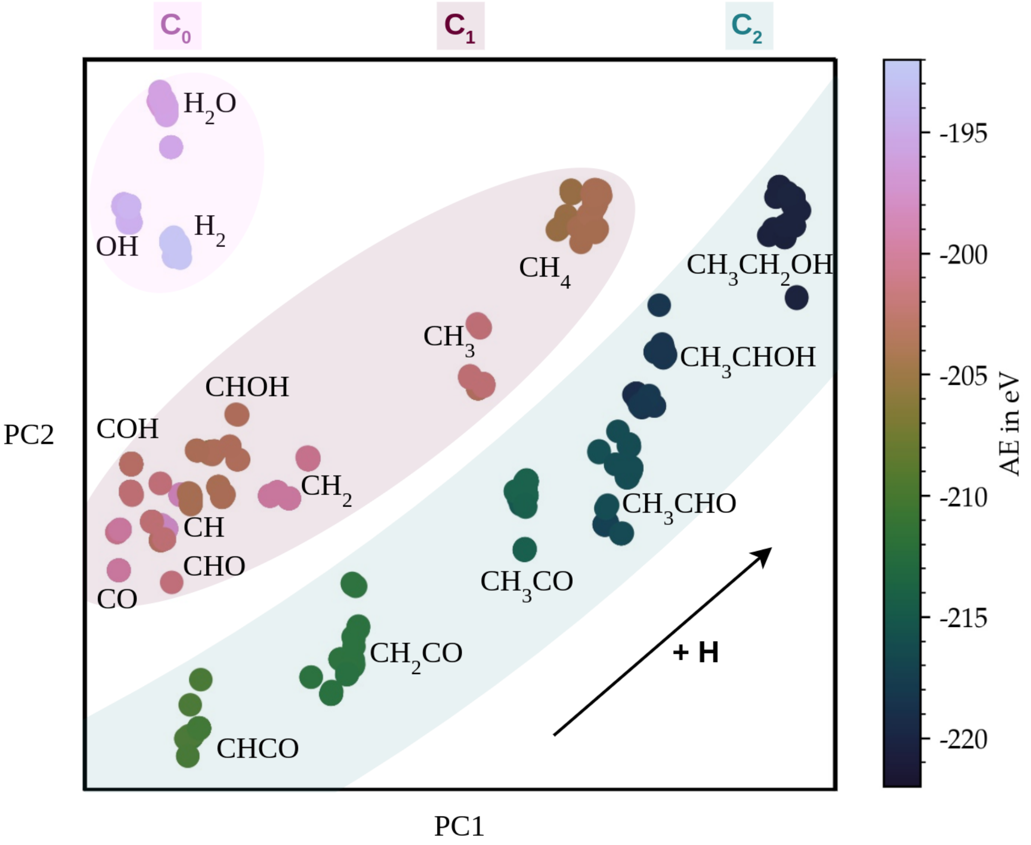


Figure 4.9: Interpretation of the kPCA with adsorbates allocated to the different clusters. PC1 implies the number of H atoms and PC2 the number of C atoms. The coloring corresponds to the atomisation energy (AE) calculated by DFT.

A reason for the overlapping and broaded clusters especially in case of the Rh(211) is the availability of multiple adsorption sites. In contrast, on the plane Rh(111) surface a lower variety in conformations is observed and therefore also a higher distinction of the adsorbates is enabled. However, the energies on the Rh(111) and Rh(211) surfaces are in a similar range for each cluster. In order to further unveil the differences for the Rh(111) and Rh(211) surface systems, reduced reaction networks are developed. The results are given in the following.

4.2.3 Reduced reaction network development

As this work's approach and choice of systems is based on the reduced reaction mechanisms for the syngas conversion on Rhodium developed by Yang et al., the mechanisms are also the starting point of this work's reaction network development. In figure 4.10, the developed reduced networks are illustrated.

In the networks, the selected adsorbates emerging during the syngas conversion on either Rh(111) (upper network) or Rh(211) (lower network) surfaces are depicted. The networks are colored in terms of the DFT formation energies for the adsorbates on the surfaces, calculated according to equation 4.2. Note that a low coverage approach is followed in this work. The energies correspond to the found global minima of the single adsorbates on the Rhodium surfaces. Therefore no reactions are simulated and the evaluation is solely based on thermochemistry without the consideration of reaction barriers.

Section 4.2.1 reveals, that the difference between the initial GAP minima and the GAP minima further optimized with DFT is small. As even more accurate predictions and lower energies are achieved by the GAP minima further optimized with DFT, the formation energies of those structures are calculated and pictured in the networks. The energies are additionally summarized in table B.2.

The networks start from the educts hydrogen and carbon monoxide. By first analysis of the networks, it can be observed, that adsorbed hydrogen is a central component in the conversion, required in most of the ensuing reaction steps. Overall, three production routes are depicted: the production of water, methane and ethanol. Especially, the side production of water protrudes because of it's comparatively low formation energy and the energy difference to the underlying educts.

The main difference in the networks is the CO-activation step and it's influence on the subsequent reactions. Whereas in case of Rh(111), the CO activation proceeds via hydrogenation of the carbon atom towards CHO, for Rh(211) the oxygen is hydrogenated resulting in COH. From these intermediates, different routes towards CH are followed. In both cases, especially the side production of water protrudes because of it's comparatively low formation energy and the energy difference to the underlying educts. The CH intermediate is either further hydrogenated up to methane or reacts towards CHCO via a C-C coupling step. This C-C coupling step is one of the most relevant steps in the syngas conversion towards C_{2+} oxygenates and is accompanied by a step in the formation energy.

To further analyse the change in formation energy, a stability diagram for the reaction routes towards methane and ethanol is developed. Thus, the progression of the formation energy along the reaction coordinate is given in the figure 4.11. In the upper part of the figure, the production routes on Rh(111) and Rh(211) are comparatively plotted. As the hydrogen adsorption follows the mechanism of the dissociative adsorption, the gasphase H_2 formation energy is plotted. In the lower two subfigures, the gasphase formation energies are compared to the adsorbate formation energies on either Rh(111) or Rh(211).

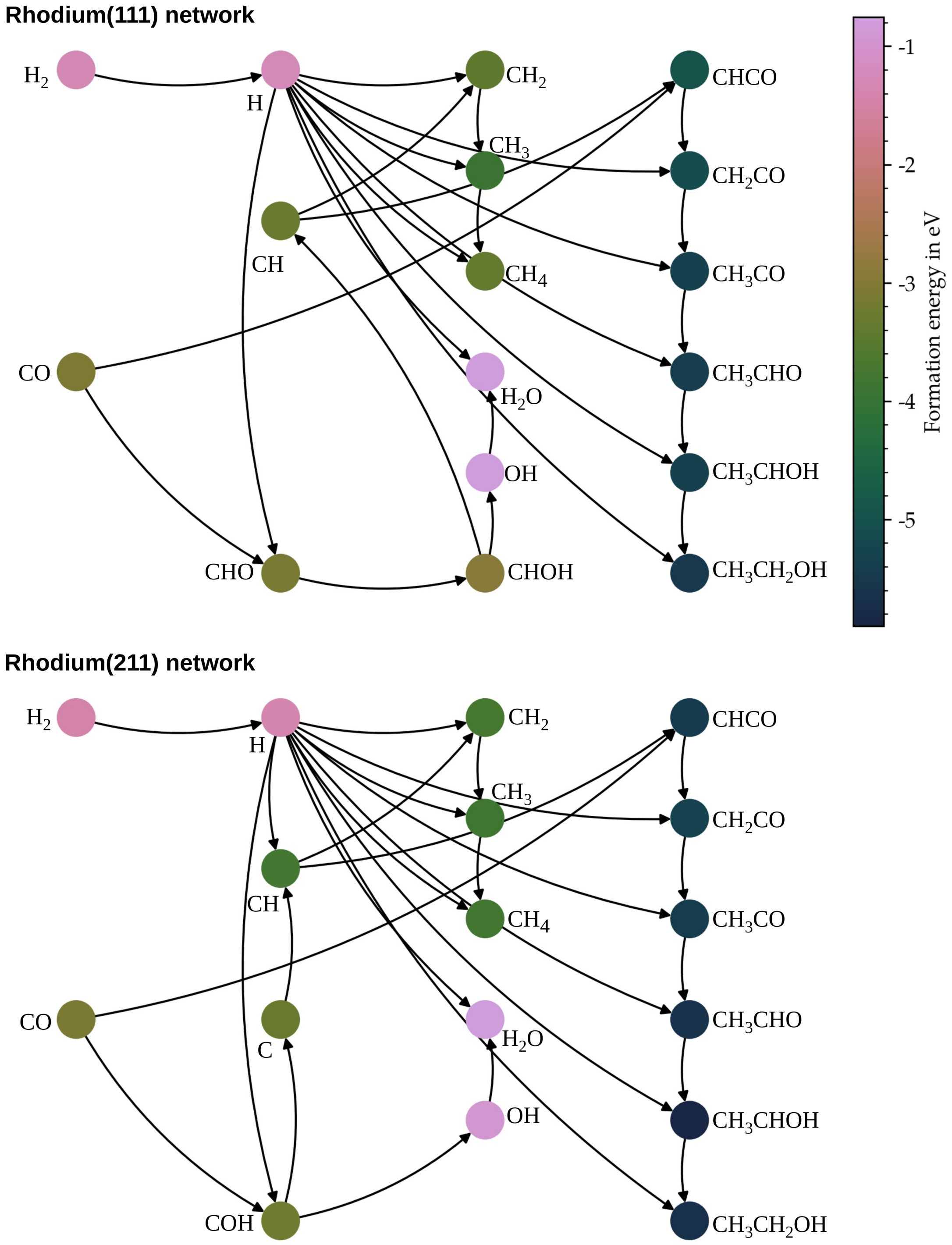


Figure 4.10: Reduced reaction networks for the syngas conversion on Rhodium surfaces based on [4]. On the top, the Rh(111) network is depicted, on the bottom the Rh(211) network. The nodes are colored in terms of the formation energies on DFT level of the re-optimized global GAP minima for each considered adsorbate.

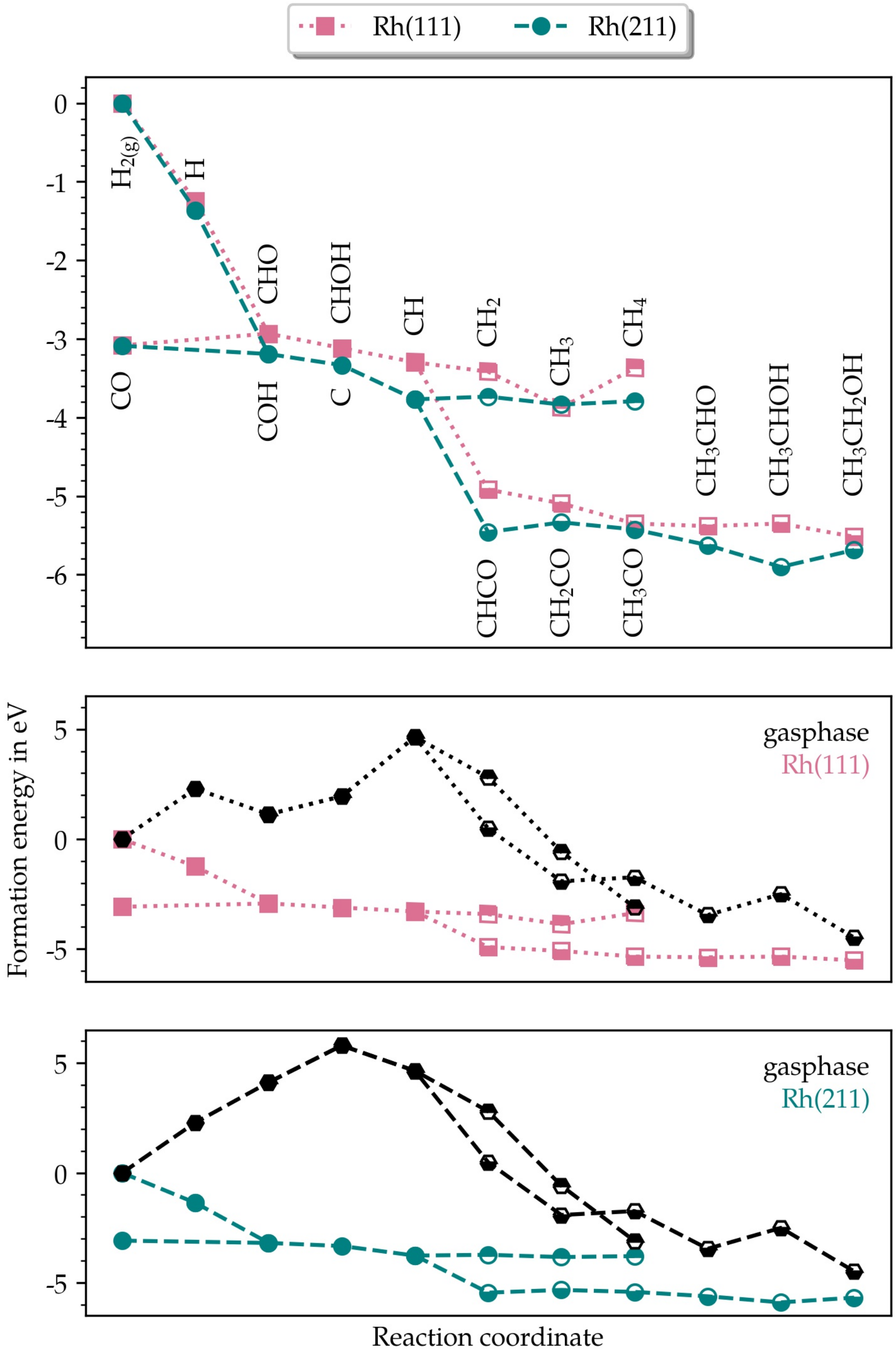


Figure 4.11: Stability diagrams of two possible production routes for the syngas conversion on either Rh(111) or Rh(211) surfaces: methane and ethanol formation. The stability diagrams are depicted as the formation energy versus the reaction coordinate. In the upper plot, the stability of the adsorbates on the Rh(111) and Rh211 is compared. The lower two plots compare the gasphase formation energy to the adsorbate formation energies on either Rh(111) or Rh(211).

Overall, the adsorbate formation energies on the different surfaces are in a similar range all along the production routes. Compared to the gasphase formation energies, all the adsorbates are tremendously stabilized due to the adsorption. The step in formation energy for the C-C coupling revealed in the network, is also conspicuous in the stability diagram. Another salient state is the CH_3 intermediate formation energy on the Rh(111) surface, which is lower than the previous CH_2 intermediate and the following CH_4 product. Such fluctuating steps also occur in Rh(211) reaction route towards CH_3CH_2OH. In the literature, a selectivity of Rh(111) surface facets towards CH_3CHO and of Rh(211) towards CH_4 is reported [4]. This seems to dissent to the pathway depicted in this work's stability diagram, as for example the formation energy of CHCO is lower for the Rh(211) surface, whereas Rh(111) has the higher selectivity towards this C_{2+} production route. A possible reason for this might be in the reaction barriers, which are not considered in this work's evaluation.

However, further analysis of the reaction mechanisms and the detailed adsorbate formation energies is beyond the focus of this work as this production run is initially performed to examine the applicability of the developed ML interatomic potential. Therefore, as a last step towards this examination, the developed GAP is applied to new, unknown structures. The results are given in the following.

4.2.4 Minima Hopping of unknown structures

In order to further validate the applicability of the trained potential, the finalGAP is applied to the minima hopping of unknown structures, taken from the studies of Yang et al. [4]. The term unknown structures refers to out of sample structures, whereby the main difference to the structures considered in the first production run is in the lattice constant and the consideration of systems with a smaller periodic cell, as detailed in section 4.1. This second production run as well as its analysis are resulted in this section.

The production run varies from the first one, detailed in section 4.2.1, in terms of considered systems and general extent. The first production run delivers the result, that the trained GAP can be used to quickly sample low energy minimum structures for adsorbates and surfaces already known to the potential with high accuracy compared to DFT. The focus of this second production run is therefore not on the generation of even more minima, but on testing limitations and boundaries of the potential.

The global optimization of the second production run is performed for 33 selected start geometries and produces 664 minima during the minima hopping. From the produced minima, a certain selection is further analysed, whereby the selection heuristics differ from the first production run: in this second production run, all the minima produced by the minima hopping are evaluated by the finalGAP as well as by DFT and each the global GAP-evaluated and global DFT-evaluated minima are chosen for further analysis and local re-optimization.

The following figure 4.12 shows the analysis of the DFT-evaluated global minima before local re–optimization. The upper subfigure shows all global minima for each of the 33 start geometries and the corresponding absolute energy difference between GAP and DFT. Here, the global minima are selected according to their lowest DFT energy after the minima hopping. The lower left plot illustrates the distribution of the energy difference. The lower right plot directly compares the absolute GAP and DFT AE.

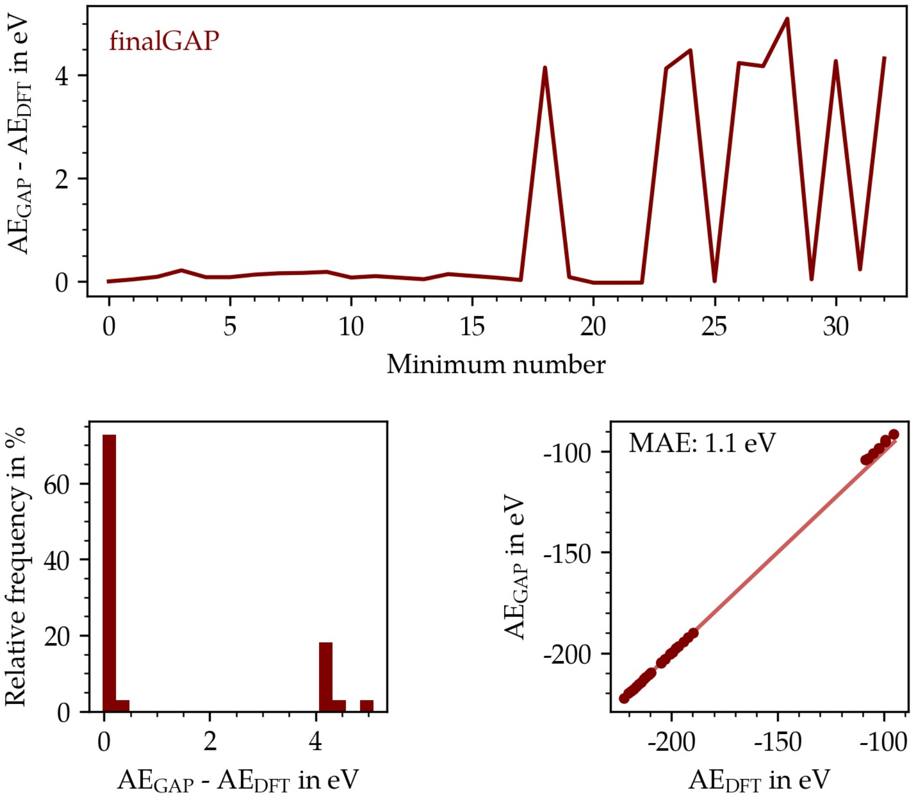


Figure 4.12: Analysis of the DFT-evaluated minima obtained via minima hopping using the finalGAP. In the top, the absolute difference between GAP and DFT AE is depicted for each minimum. Bottom left, the distribution of the deviation between GAP and DFT AE is given. Bottom right, the correlation between the two energies is shown.

Especially in the upper plot, the high deviation between GAP and DFT energy of some minima is apparent. The highest deviation is approximately 5 eV. In addition, the high difference in the deviations is notable and comes out even more in the lower left distribution. Approximately 75 percent of the start geometries show a low energy difference. However, there are several outliers. This also influences the correlation of the energies resulting in a high MAE of 1.1 eV. Especially the structures with an AE around -100 eV deviate.

The precise analysis of the high deviations reveals, that just those structures show up a high deviation, which are build with a 3x2x3 periodic cell for the Rh(211) surface facets.

Leaving out these structures build by a 3x2x3 periodic cell, which all show an absolute deviation bigger than 4 eV, entails an immediate improvement of the prediction. This is illustrated in the following figure 4.13.

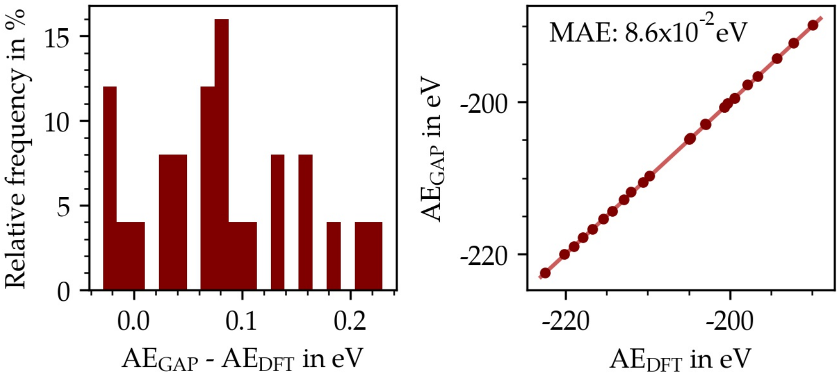


Figure 4.13: Difference between GAP and DFT AE as relative frequency distribution on the left and correlation the two energies on the right. The analysis is limited to the minima with an absolute energy deviation below 4 eV.

By excluding the outliers, the MAE is improved to 8.6×10^{-2} eV. Thus, it is assumed, that the varied cell size is the reason for the high errors. In the next subparagraph, the reasons for the high deviations due to the change of the cell size are further discussed and a possible approach to higher accuracy is suggested.

Similar to the first production run, the MHM minima are further optimized using local BFGS optimization on DFT level. By doing so, the quality of the minima produced using the trained potential is approved. In figure 4.14, the energetic as well as the geometric change resulting from the local optimization is illustrated. The energetic change is depicted as the frequency distribution of the absolute AE difference before and after the local DFT optimization. In the left figure, the energy shift due to the optimization is evaluated by singlepoint calculations using the finalGAP. In the middle, the same evaluation is performed by DFT singlepoint calculations. The right figure shows the RMSD of the structures before and after the local optimization.

In accordance to the first production run starting from known systems (compare section 4.2.1, step 3), the local DFT re-optimization for unknown structures only slightly changes the minima obtained by the MHM using the GAP. Energetically, the highest change is approximately 0.2 eV, geometrically about 0.22 Å. As in the analysis of the first production run, the local DFT optimization energetically improves the minima when evaluated by DFT and impairs when evaluated by the finalGAP. Overall the energy differences as well as the RMSDs are commensurable to the first production run it can be resulted that the optimization steps as part of the MHM already provide a good energy convergence. As also the previously detected outliers are part of this evaluation, the quality of the minima obtained by the MHM using the finalGAP is precluded as a reason for the insufficient energy prediction for the outliers.

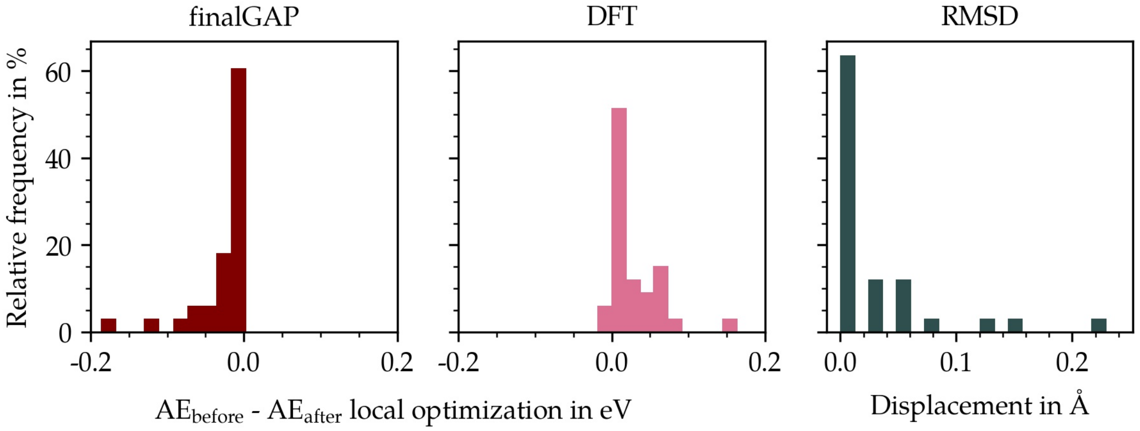


Figure 4.14: Frequency distribution of the DFT-evaluated MHM minima. The left and middle plot show the distribution as a function of the difference between the AE before and after the local optimization, evaluated by singlepoint calculations using the finalGAP (left) and DFT (middle). The right figure shows the distribution as a function of the RMSD of the adsorbates in the adsorbate-surface systems before and after the local optimization.

Other possible reasons for the high errors for the outliers as well as an approach to higher prediction accuracy is given in the following.

Approach to higher accuracy: additional training iteration

The previous analysis of the minima obtained by MHM of unknown structures taken from literature using the finalGAP brings up limitations in the applicability of the finalGAP for certain structures. It is identified, that the outliers are those structures constructed with a 3x2x3 periodic cell. The reasons for the high deviations and an approach to higher accuracy are discussed here. As the outliers are identified by the comparison of the DFT and GAP energies, both DFT as well as GAP are considered as error sources.

This work's GAP training approach is based on DFT reference calculations. All the input structures to the training are constructed with a 3x3x4 periodic cell and the DFT calculations are performed with a k-grid of (4x4x1), which resulted from a convergence test for the 3x3x4 periodic surfaces (compare figure A.3). Therefore, the potential is not trained towards other periodic cell sizes and the DFT convergence is not approved to cell size changes. Since in general, the required k-grid is dependent on the periodic cell size, the influence of a k-grid variation is tested for an example system. However, no significant improvement of the energy due to k-point grid adaption is observed and the largest energy difference to the current k-point grid was in the order of 10^{-2} eV. Thus, the k-point grid can be eliminated from the sources of error.

The other, more probable source of error is in the prediction by the GAP. The main distinctions are the different lattice constants and the cell size. For the unknown structures, a lattice constant of 3.86 Å for the 3x3x4 and 3.866 Å for the 3x2x3 cell is used, which slightly differs from this work's lattice constant of 3.85 Å. Moreovere, the 3x2x3 cells are smaller in two directions: the y-direction as well as the z-direction, which both have influence on the prediction. In the y-direction, the adsorbate molecules come closer to each other. In the z-direction, one bulk Rhodium layer is completely removed. These both lead to changes in the atomic environments of the atoms included in the system and therefore might lead to predictability limits.

In order to approach and test this hypothesis, clean Rhodium surfaces are constructed similar to those of the unknown structures and their energy is calculated by the finalGAP and DFT in comparison. For the clean 3x3x4 surfaces with a lattice constant of 3.86 Å, an absolute deviation of 3.0×10^{-2} eV between the DFT AE and the finalGAP AE is found. Thus, the slight difference in the lattice constant does not spawn a predictability limit. Though, the deviation quickly enlarges for the 3x2x3 surface with a lattice constant of 3.866 Å. Here, the absolute difference between GAP and DFT AE accounts 4.1 eV. Therefore, the main reason for the high deviations for some structures in the second production run is ascribed to the influence of the cell size on the energies of Rhodium atoms.

This influence is simply not considered by the finalGAP, as the training is conducted with one type of periodic cells only. However, to approach a higher accuracy, the 3x2x3 structures from the unknown data set are added to the training set of the finalGAP and the potential is retrained with otherwise similar settings. More precisely, the 11th iteration of the third training round (GAP3, compare section 3.2.3) with the hyperparameter set (10) is repeated with a training set expanded by eight 3x2x3 structures. The potential resulting from the 11th iteration of GAP3 with finetuned hyperparameters is previously named finalGAP, therefore the retrained potential is now defined as 'retrained finalGAP'.

To evaluate the retraining, the energies of the start geometries, GAP- and DFT-evaluated MHM minima and locally re-optimized minima are recalculated with the retrained finalGAP. As the start geometries are directly added to the training set, an improvement for those is evident. Therefore, especially the influence of the retraining on the further optimized structures is analysed in the following. Figure 4.15 shows the effect of the retraining on the energy deviation between GAP and DFT. Thereby, the minima, which were first obtained by minima hopping using the finalGAP, secondly selected as global minima by DFT evaluation and thirdly locally reoptimized with DFT, are analysed.

In direct comparison, as depicted in the upper plot of figure 4.15, the retraining significantly improves the deviation of the GAP to the DFT AE. Analysing the predictions by the retrained finalGAP even further, the improvement can be confirmed. As can be seen in the lower left subfigure, the energy difference distribution considerably narrower, with most of the structures having deviation less than 0.5 eV. This is accompanied by an improvement of the MAE due to the retraining, lowered by one order of magnitude in contrast to the finalGAP.

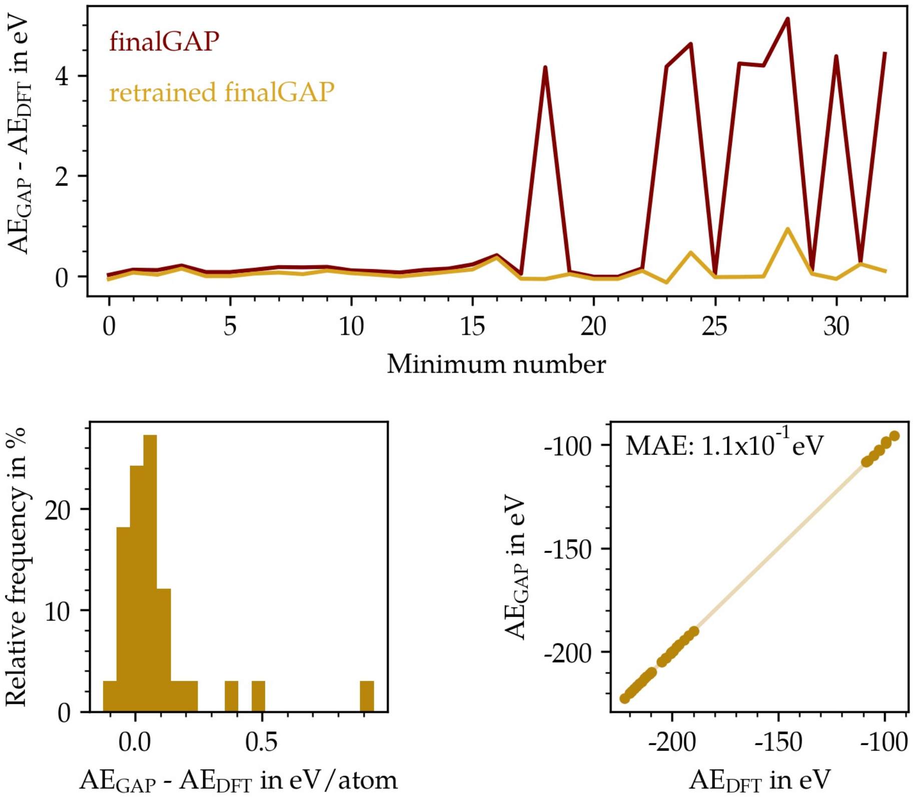


Figure 4.15: Analysis of the locally DFT optimized global minima evaluated by the retrained finalGAP. In the top, the absolute difference between GAP and DFT AE is depicted for each minimum, evaluated by both the finalGAP and the retrained finalGAP. Bottom left, the distribution of the deviation between the retrained finalGAP and DFT AE is given, bottom right the correlation between the two energies.

However, this error is still above the error of the first production run. One possible reason for this might the interplay of SOAP cutoff versus periodic cell size. The higher errors for the retrained finalGAP are ascribable to systems, where a big part of the adsorbate is within the SOAP cutoff radius of the adsorbate in the adjacent cell. Although a certain overlap also occurs for bigger cell sizes, this is confined to the overlap of single atoms only. Therefore, long-ranged adsorbate-adsorbate contributions appear not to be learned during the GAP training. Nevertheless, this is in conformity with this work's low coverage approach.

In the following, the applicability of the developed GAP is conclusively assessed and an outlook is given.

4.2.5 Applicability assessment and outlook

This work's ML interatomic potential is applied to the minima hopping of the involved structures in the syngas conversion on Rhodium surfaces. In two production runs, the potentials applicability is tested. In a first, broad production run, the application towards new adsorption sites is attempted. Secondly, the potentials limitations are assayed by the application towards new, unknown structures. The results of this production runs are previously stated and assessed here.

Overall, the generation of minima applying the trained GAP to the minima hopping of adsorbates on Rhodium surfaces can be appraised as adequate. The potential quickly allows to sample minimum structures of various adsorbates on the different Rhodium surfaces, which only need slight local re-optimization if at all. Limitations especially occur with surface-adsorbate structures not involved in the training set. The second application test reveals, that changing the periodic cell deteriorates the predicted energy accuracy. Including the structures with new characteristics such as a varied periodic cell in the training set and retraining the potential however leads to an immediate improvement of up to one order of magnitude.

Beyond this work, the developed GAP is already applied to sample minimum adsorption geometries for various adsorbates part of an extended reaction network for the syngas conversion on Rhodium [87]. During this expanded sampling, it appears that for those adsorbates included in the training set, the transferability is high. However, new adsorbates lead to high errors. Those observations are in alignment with this work's results and the errors are in a similar range as those observed for the 3x2x3 surfaces. However, the sampled adsorption structures serve as a good starting point for further analyses, for example more accurate DFT calculations or microkinetic simulations.

Therefore, it can be generally stated, that for known adsorbates and surfaces the energy prediction by the trained GAP has a high accuracy. This accuracy is not sustained when applying the potential towards new systems. In such cases, an additional training iteration is suggested to increase the accuracy. With the current settings and training set size, one additional training iteration takes approximately two hours and improved the predictions in this work's applicability test by one order of magnitude, which represents a valuable cost-trade ratio. Nonetheless, this work's GAP is able to sample minima for unknown start geometries and the proposed structures can be qualitatively evaluated as chemical.

The training workflow yields a system-specific GAP with good stability towards high temperatures and temperature changes, due to the iterative training with rising minima hopping conditions. The applicability of the workflow is estimated to not being limited to the syngas conversion on Rhodium only, but also to be adaptable to other heterogeneous catalytic reaction networks. An obvious, prospective expansion of the workflow could be the inclusion of a higher variety of surface layers in the early training iterations. Additionally, the periodic cell size could be increased, which is valuable for larger adsorbate molecules. Due to size extensitivity of the GAP, the increase in cell size is expected to be viable. In case of the syngas conversion on Rhodium, a higher variability of adsorbates could additionally be considered, which is already ongoing [87]. Moreover, adsorbate-adsorbate interactions could be supplementally included, which is already done in other work [81].

It should be noted, that the potential is trained and tested towards the applied simulation methods only. These methods include the constrained global optimization using the MHM, which performs MD simulations as well as local optimizations, as well as unconstrained BFGS optimizations. Therefore, no statement can be made to the applicability towards other simulation methods. In future work, a worthwhile expansion could be build about the exploration of transition states also. This would enable a more detailed insight into the relevant adsorbate states and reaction barriers.

Retrospectively, less overall training iterations and an earlier adaption of the training towards the specific scientific problem of interest could have even increased the training efficiency. Therefore, this is suggested as a general take for further developments.

Chapter 5

Summary

The goal of this work was to develop a method to discover the global minimum space of the involved educts, intermediates and products of a heterogeneous catalytic reaction network. As a model system, the syngas conversion on Rhodium has been chosen. Although syngas is one of the key reagents in industrial chemistry, the mechanisms leading towards specific products are not fully disclosed and especially the rational of the selectivity of different catalysts towards specific products is still deficient.

Yang et al. [4] were able to formulate reduced reaction mechanisms for the syngas conversion on Rhodium(111) and Rhodium(211) surface facets, which served as a basis of this work's approach. In order to discover the reaction network of the syngas conversion on Rhodium, this work aimed to develop a fast and accurate interatomic potential for the prediction of minimum structures of the different appearing adsorbates on the catalytic Rhodium surfaces. Therefore, the overall goal of this work was divided into two subgoals: the development of a machine learning (ML) interatomic potential on the one hand and the application of the developed potential on the other hand.

In the first part - the development of a ML interatomic potential - a Gaussian approximation potential (GAP) has been trained. As part of this work, an iterative GAP training workflow has been developed, which is tailored to the global optimization of the adsorbates involved in the syngas conversion on Rhodium using the minima hopping method (MHM). Overall three different training rounds have been performed, whereby the training workflow was refined until a good compromise between stability of the potential, computational accuracy compared to DFT and chemical quality of the produced minima has been achieved. For the final potential, a hyperparameter finetuning allowed to even increase the accuracy of the potential.

In the second part of this work - the application of the potential to global optimization - the previously trained GAP was applied to the minima hopping of new start geometries. In a first attempt, the potential has been applied to the global optimization of the same surface and adsorbate conformations as considered in the GAP training but with additional adsorption sites. This first application served as a general production run testing the overall applicability of the potential and sampling various minima. The overall quality of the produced minima compared to DFT was adjudged as adequate. With the produced global minima and its related energies, the development of reduced reaction networks for the syngas conversion on both Rhodium(111) and Rhodium(211) has been enabled.

In a second attempt, the potential has been applied to unknown start geometries taken from literature, whereby the main distinction to the previously tested geometries was in the periodic cell size and the lattice constant. This second production run was performed in order to test the potential towards out of sample structures and to discover limitations. It turned out, that the potential's transferability is high for those structures already included in the training set, but high errors appeared for structures with new, unknown characteristics as for example a varied periodic cell size. A retraining of the potential with an expanded training set by structures with the new characteristics however lead to an immediate decrease of the error. Due to this quick and high adaptability, the workflow as well as the potential serves as a good starting point for further customization towards other specific scientific problems.

To conclude, this work brought up a method to train a system-tailored, ML interatomic potential for the accelerated minima hopping of adsorbates in heterogeneous catalytic systems, examined on the syngas conversion on Rhodium.

Bibliography

[1] C. M. Friend and B. Xu, "Heterogeneous catalysis: A central science for a sustainable future," *Acc. Chem. Res.*, vol. 50, pp. 517–521, 2017.

[2] T. Bligaard and J. K. Nørskov, "Heterogeneous catalysis," in *Chemical Bonding at Surfaces and Interfaces* (A. Nilsson, L. G. Pettersson, and J. K. Nørskov, eds.), ch. 4, pp. 255–321, Elsevier B.V., 2008.

[3] C. Du, P. Lu, and N. Tsubaki, "Efficient and new production methods of chemicals and liquid fuels by carbon monoxide hydrogenation," *ACS Omega*, vol. 5, pp. 49–56, 2020.

[4] N. Yang, A. J. Medford, X. Liu, F. Studt, T. Bligaard, S. F. Bent, and J. K. Nørskov, "Intrinsic selectivity and structure sensitivity of Rhodium catalysts for C_{2+} oxygenate production," *Journal of the American Chemical Society*, vol. 138, pp. 3705–3714, 2016.

[5] F. Studt, "Grand challenges in computational catalysis," *Frontiers in Catalysis*, vol. 1, no. 658965, pp. 1–4, 2021.

[6] J. T. Margraf and K. Reuter, "Systematic enumeration of elementary reaction steps in surface catalysis," *American Chemical Society Omega*, vol. 4, pp. 3370–3379, 2019.

[7] S. Stocker, G. Csányi, K. Reuter, and J. T. Margraf, "Machine learning in chemical reaction space," *Nature Communications*, vol. 11, no. 5505, pp. 1–11, 2020.

[8] Z. W. Ulissi, A. J. Medford, T. Bligaard, and J. K. Nørskov, "To address surface reaction network complexity using scaling relations machine learning and DFT calculations," *Nature Communications*, vol. 8, no. 14621, pp. 1–7, 2017.

[9] M. Andersen, C. Panosetti, and K. Reuter, "A practical guide to surface kinetic Monte Carlo simulations," *frontiers in Chemistry*, vol. 7, no. 202, pp. 1–24, 2019.

[10] M. Rupp, "Machine learning for quantum mechanics in a nutshell," *International Journal of Quantum Chemistry*, vol. 115, pp. 1058–1073, 2015.

[11] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Physical Review Letters*, vol. 98, no. 146401, pp. 1–4, 2007.

[12] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost," *Chemical Science*, vol. 8, pp. 3192–3203, 2017.

[13] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," *Physical Review Letters*, vol. 104, no. 136403, pp. 1–4, 2010.

[14] V. L. Deringer and G. Csányi, "Machine learning based interatomic potential for amorphous carbon," *Physical Review B*, vol. 95, no. 094203, pp. 1–15, 2017.

[15] S. Wengert, G. Csányi, K. Reuter, and J. T. Margraf, "Data-efficient machine learning for molecular crystal structure prediction," *Chemical Science*, vol. 12, pp. 4536–4546, 2021.

[16] J. Timmermann, F. Kraushofer, N. Resch, P. Li, Y. Wang, Z. Mao, M. Riva, Y. Lee, C. Staacke, M. Schmid, C. Scheurer, G. S. Parkinson, U. Diebold, and K. Reuter, "IrO_2 surface complexions identified through machine learning and surface investigations," *Physical Review Letters*, vol. 125, no. 206101, pp. 1–6, 2020.

[17] S. Goedecker, "Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems," *The Journal of Chemical Physics*, vol. 120, pp. 9911–9917, 2004.

[18] J. Kang, S. He, W. Zhou, Z. Shen, Y. Li, M. Chen, Q. Zhang, and Y. Wang, "Single-pass transformation of syngas into ethanol with high selectivity by triple tandem catalysis," *Nature Communications*, vol. 11, no. 827, pp. 1–11, 2020.

[19] C. Du, P. Lu, and N. Tsubaki, "Efficient and new production methods of chemicals and liquid fuels by carbon monoxide hydrogenation," *ACS Omega*, vol. 5, no. 1, pp. 49–56, 2020.

[20] R. Rauch, J. Hrbek, and H. Hofbauer, "Biomass gasification for synthesis gas production and applications of the syngas," *WIREs Energy and Environment*, vol. 3, no. 4, pp. 343–362, 2013.

[21] S. Hernández, M. A. Farkhondehfal, F. Sastre, M. Makkee, G. Saraccob, and N. Russoa, "Syngas production from electrochemical reduction of co2: current status and prospective implementation," *Green Chemistry*, vol. 19, pp. 2326–2346, 2017.

[22] V. Subramani and S. K. Gangwal, "A review of recent literature to search for an efficient catalytic process for the conversion of syngas to ethanol," *Energy Fuels*, vol. 22, pp. 814–839, 2008.

[23] M. Vohra, J. Manwar, R. Manmode, S. Padgilwar, and S. Patil, "Bioethanol production: Feedstock and current technologies," *Journal of Environmental Chemical Engineering*, vol. 2, no. 1, pp. 573–584, 2014.

[24] V. Subramani and S. K. Gangwal, "A review of recent literature to search for an efficient catalytic process for the conversion of syngas to ethanol," *Energy & Fuels*, vol. 22, pp. 814–839, 2008.

[25] M. lchikawat and T. Fukushima, "Mechanism of syngas conversion into c2-oxygenates such as ethanol catalysed on a si02-supported rh-ti catalyst," *Journal of The Chemical Society, Chemical Communications*, pp. 321–323, 1985.

[26] M. Gupta, M. L. Smith, and J. J. Spivey, "Heterogeneous catalytic conversion of dry syngas to ethanol and higher alcohols on Cu-based catalysts," *ACS Catalysis*, vol. 1, no. 6, pp. 641–656, 2011.

[27] R. Schlögl, "Heterogeneous catalysis," *Angewandte Chemie International Edition*, vol. 54, no. 11, pp. 3465–3520, 2015.

[28] Y. Liu, F. Göeltl, I. Ro, M. R. Ball, C. Sener, I. B. Aragão, D. Zanchet, G. W. Huber, M. Mavrikakis, and J. A. Dumesic, "Synthesis gas conversion over Rh-based catalysts promoted by Fe and Mn," *ACS Catalysis*, vol. 7, no. 7, pp. 4550–4563, 2017.

[29] A. Szabo and N. S. Ostlund, *Modern Quantum Quemistry - Introduction to Advanced Electronic Structure Theory*. Mineola, New York: Dover Publications, Inc., 1996.

[30] W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory*. Wiley-VCH Verlag GmbH, 2 ed., 2001.

[31] J. A. Keith, J. Anton, P. Kaghazchi, and T. Jacob, "Modeling catalytic reactions on surfaces with Density Functional Theory," in *Modeling and Simulation of Heterogeneous Catalytic Reactions: From the Molecular Process to the Technical System* (O. Deutschmann, ed.), ch. 1, Wiley-VCH Verlag GmbH, 1 ed., 2012.

[32] N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals," *Molecular Physics*, vol. 115, no. 19, pp. 2315–2372, 2017.

[33] A. Bartók-Partáy, *The Gaussian Approximation Potential*. Berlin Heidelberg, Germany: Springer Theses, 2010.

[34] A. Bartók and G. Csányi, "Gaussian approximation potentials: A brief tutorial introduction," *International Journal of Quantum Chemistry*, vol. 115, pp. 1051–1057, 2015.

[35] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Physical Review*, vol. 136, no. 3B, pp. 864–871, 1964.

[36] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Physical Review*, vol. 140, no. 4A, pp. 1133–1138, 1965.

[37] J. P. Perdew and K. Schmidt, "Jacob's ladder of density functional approximations for the exchange-correlation energy," *AIP Conference Proceedings*, vol. 577, pp. 1–20, 2001.

[38] Y. Zhang and W. Yang, "Comment on "generalized gradient approximation made simple"," *Physical Review Letters*, vol. 80, no. 4, p. 890, 1998.

[39] A. Tkatchenko and M. Scheffler, "Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data," *Phys. Rev. Lett.*, vol. 102, p. 073005, Feb 2009.

[40] V. G. Ruiz, W. Liu, E. Zojer, M. Scheffler, and A. Tkatchenko, "Density-functional theory with screened van der Waals interactions for the modeling of hybrid inorganic-organic systems," *Physical Review Letters*, vol. 108, no. 146103, pp. 1–5, 2012.

[41] V. G. Ruiz, W. Lui, and A. Tkatchenko, "Density-functional theory with screened van der Waals interactions applied to atomic and molecular adsorbates on close-packed and non-close-packed surfaces," *Physical Review B*, vol. 93, no. 035118, pp. 1–17, 2016.

[42] S. Fujikake, V. L. Deringer, T. H. Lee, M. Krynski, S. R. Elliott, and G. Csányi, "Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures," *The Journal of Chemical Physics*, vol. 148, no. 241714, pp. 1–10, 2018.

[43] F. Maresca, D. Dragoni, G. Csányi, N. Marzari, and W. A. Curtin, "Screw dislocation structure and mobility in body centered cubic fe predicted by a gaussian approximation potential," *npj Computational Materials*, vol. 4, no. 69, pp. 1–7, 2018.

[44] C. W. Rosenbrock, K. Gubaev, A. V. Shapeev, L. B. Pártay, N. Bernstein, G. Csányi, and G. L. W. Hart, "Machine-learned interatomic potentials for alloys and alloy phase diagrams," *npj Computational Materials*, vol. 7, no. 24, pp. 1–9, 2021.

[45] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules," *Chemical Reviews*, vol. 121, no. 16, pp. 10073–10141, 2021.

[46] J. Xu, X.-M. Cao, and P. Hu, "Perspective on computational reaction prediction using machine learning methods in heterogeneous catalysis," *Physical Chemistry Chemical Physics*, vol. 23, pp. 11155–11179, 2021.

[47] Y. Mishin, "Machine-learning interatomic potentials for materials science," *Acta Materialia*, vol. 214, no. 116980, pp. 1–16, 2021.

[48] V. L. Deringer, M. A. Caro, and G. Csányi, "Machine learning interatomic potentials as emerging tools for materials science," *Advanced Materials*, vol. 31, no. 1902765, pp. 1–16, 2019.

[49] S. Stocker. personal communication, 2021.

[50] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," *Physical Chemistry Chemical Physics*, vol. 18, no. 20, p. 13754–13769, 2016.

[51] G. Csányi, M. J. Willatt, and M. Ceriotti, "Machine-learning of atomic-scale properties based on physical principles," in *Machine Learning Meets Quantum Physics* (K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller, eds.), vol. 968 of *The Lecture Notes in Physics*, Springer, Cham, 2020.

[52] J. Wang, "An intuitive tutorial to gaussian processes regression," 2021.

[53] E. G. Lewars, "The concept of the potential energy surface," in *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*, ch. 2, pp. 9–49, Springer, 3 ed., 2016.

[54] M. Amsler, "Minima hopping method for predicting complex structures and chemical reaction pathways," in *Handbook of Materials Modeling* (W. Andreoni and S. Yip, eds.), ch. 117, pp. 2791–2810, Springer Nature Switzerland AG, 2020.

[55] F. H. Stillinger, "Exponential multiplicity of inherent structures," *Physical Review E*, vol. 59, pp. 48–51, 1999.

[56] R. Battiti, "First- and second-order methods for learning: Between steepest descent and newton's method," *Advances in Difference Equations*, vol. 638, pp. 1–24, 2020.

[57] S. K. Mishra, G. Panda, S. K. Chakraborty, M. E. Samei, B. Ram, and M. E. Samei, "On q-bfgs algorithm for unconstrained optimization problems," *Advances in Difference Equations*, vol. 638, pp. 1–24, 2020.

[58] C. G. Broyden, "The convergence of a class of double-rank minimization algorithms 1. general considerations," *IMA Journal of Applied Mathematics*, vol. 6, pp. 76–90, 1970.

[59] R. Fletcher, "A new approach to variable metric algorithms," *The Computer Journal*, vol. 13, pp. 317–322, 1970.

[60] D. Goldfarb, "A family of variable-metric methods derived by variational means," *Mathematics of Computation*, vol. 24, pp. 23–26, 1970.

[61] D. F. Shanno, "Conditioning of quasi-newton methods for function minimization," *Mathematics of Computation*, vol. 24, pp. 647–656, 1970.

[62] S. Goedecker, "Global optimization with the minima hopping method," in *Modern Methods of Crystal Structure Prediction* (A. R. Organov, ed.), ch. 6, pp. 131–146, Weinheim, Germany: WILEY-VCG Verlag GmbH & Co.KGaA, 2011.

[63] D. E. Goldberg, *Genetic Algorithms in Search, Optimization & Machine Leaning*. Boston, MA: Addison-Wesley Publishing Company, inc., 1989.

[64] D. J. Wales and J. P. K. Doye, "Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms," *The Journal of Physical Chemistry A*, vol. 101, no. 28, pp. 5111–5116, 1997.

[65] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[66] S. Goedecker, "Minima hopping: Searching for the global minimum of the potential energy surface of complex molecular systems without invoking thermodynamics," *arXiv: Materials Science*, 2004.

[67] P. M. Paraldos, D. Shalloway, and G. Xue, "Optimization methods for computing global minima of nonconvex potential energy functions," *Journal of Global Optimization*, vol. 4, pp. 117–133, 1994.

[68] R. V. Pappu, R. K. Hart, and J. W. Ponder, "Analysis and application of potential energy smoothing and search methods for global optimization," *Journal of Physical Chemistry B*, vol. 102, pp. 9725–9742, 1998.

[69] P. M.Pardalos, H. Romeijn, and H. Tuy, "Recent developments and trends in global optimization," *Journal of Computational and Applied Mathematics*, vol. 124, pp. 209–228, 2000.

[70] C. A. Floudas and C. E. Gounaris, "A review of recent advances in global optimization," *Journal of Global Optimization*, vol. 45, no. 3, 2009.

[71] A. Žilinskas and A. Zhigljavsky, "Stochastic global optimization: A review on the occasion of 25 years of Informatica," *Informatica*, vol. 27, no. 2, pp. 229–256, 2016.

[72] M. Sicher, S. Mohr, and S. Goedecker, "Efficient moves for global geometry optimization methods and their application to binary systems," *The Journal of Chemical Physics*, vol. 134, no. 044106, pp. 1–7, 2011.

[73] B. Schaefer, S. Mohr, M. Amsler, and S. Goedecker, "Minima hopping guided path search: An efficient method for finding complex chemical reaction pathways," *The Journal of Chemical Physics*, vol. 140, no. 214102, pp. 1–13, 2014.

[74] R. K. Cersonsky, B. A. Helfrecht, E. A. Engel, S. Kliavinek, and M. Ceriotti, "Improving sample and feature selection with principal covariates regression," *Machine Learning: Science and Technology*, vol. 2, no. 035038, pp. 1–16, 2021.

[75] J. Lever, M. Krzywinski, and N. Altman, "Principle component analysis," *Nature Methods*, vol. 14, no. 7, pp. 641–642, 2017.

[76] M. Tipping, "Sparse kernel principal component analysis," in *Advances in Neural Information Processing Systems* (T. Leen, T. Dietterich, and V. Tresp, eds.), vol. 13, MIT Press, 2001.

[77] "FHI-aims The ab initio materials simulation package." Website: `https://fhi-aims.org/`, 2021. latest accessed on August 25,2021.

[78] "Atomic simulation environment." Website: `https://libatoms.github.io/GAP`, 2021. latest accessed on October 5,2021.

[79] "CatKit: Catalysis Kit." Website: `https://github.com/SUNCAT-Center/CatKit`, 2018. latest accessed on August 25,2021.

[80] A. Bartok-Partay, N. Bernstein, G. Csanyi, and J. Kermode, "GAP and SOAP documentation." Website: `https://libatoms.github.io/GAP`, 2019. latest accessed on September 22,2021.

[81] S. Stocker. unpublished research, 2021.

[82] "igraph - The network analysis package." Website: `https://igraph.org/`, 2003 – 2020. latest accessed on October 11,2021.

[83] S. Stocker, "Code collection for KRR ML models." Website: `https://zenodo.org/record/4025972#.YWk9rOpBzXQ`, 2020. latest accessed on October 15,2021.

[84] B. Cheng, "Automatic Selection And Prediction tools for materials and molecules." Website: `https://bingqingcheng.github.io/index.html#`, 2020. latest accessed on October 14,2021.

[85] C. Kunkel, S. Wengert, and T. K. Stenczel, "Projection viewer." Website: `https://github.com/chkunkel/projection_viewer`, 2020. latest accessed on October 14,2021.

[86] B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, and G. Csanyi, "Mapping materials and molecules," *Accounts of Chemical Research*, vol. 53, no. 9, pp. 1981–1991, 2020.

[87] H. Jung. unpublished research, 2021.

Appendix A

GAP training details

A.1 Reduced reaction mechanisms for the syngas conversion on Rhodium

The following reduced reaction mechanisms are taken from [4] and used as a basis for this work. Based on the appearing educts, intermediates and products, the training set of this work's GAP training is constructed. In the mechanisms, $\star_\mathrm{f}$, $\star_\mathrm{s}$ and $\star_\mathrm{h}$ refer to the fourfold, step and hydrogen reservoir sites, which are the three sites considered by Yang et al. [4]. In this work, more sites are considered as detailed in sections 3.1.2 and 4.2.1.

Mechanism on Rh(111) surface

$$\mathrm{H_{2(g)}} + 2\,\star_\mathrm{h} \longleftrightarrow 2\,\mathrm{H}^{\star h} \tag*{\{1\}}$$

$$\star_\mathrm{s} + \mathrm{CO_{(g)}} \longleftrightarrow \mathrm{CO}^{\star s} \tag*{\{2\}}$$

$$\mathrm{CO}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{H{-}CO}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CHO}^{\star s} + \star_\mathrm{h} \tag*{\{3\}}$$

$$\mathrm{CHO}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{HCO{-}H}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CHOH}^{\star s} + \star_\mathrm{h} \tag*{\{4\}}$$

$$\mathrm{CHOH}^{\star s} + \star_\mathrm{s} \longleftrightarrow \mathrm{CH{-}OH}^{\star s} + \star_\mathrm{s} \longrightarrow \mathrm{CH}^{\star s} + \mathrm{OH}^{\star s} \tag*{\{5\}}$$

$$\mathrm{CH}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{CH{-}H}^{\star s} + \star_\mathrm{h} \tag*{\{6\}}$$

$$\mathrm{CH}_2^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{H{-}CH}_2^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CH_3}\star_\mathrm{s} + \star_\mathrm{h} \tag*{\{7\}}$$

$$\mathrm{OH}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{H{-}OH}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{H_2O_{(g)}} + \star_\mathrm{s} + \star_\mathrm{h} \tag*{\{8\}}$$

$$\mathrm{CH}^{\star s} + \mathrm{CO}^{\star s} \longleftrightarrow \mathrm{CH{-}CO}^{\star s} + \star_\mathrm{s} \longrightarrow \mathrm{CHCO}^{\star s} + \star_\mathrm{s} \tag*{\{9\}}$$

$$\mathrm{CHCO}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{H{-}CHCO}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CH_2CO}^{\star s} + \star_\mathrm{h} \tag*{\{10\}}$$

$$\mathrm{CH_2CO}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{H{-}CH_2CO}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CH_3CO}^{\star s} + \star_\mathrm{h} \tag*{\{11\}}$$

$$\mathrm{CH_3CO}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{H{-}CH_3CO}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CH_3CHO}^{\star s} + \star_\mathrm{h} \tag*{\{12\}}$$

$$\mathrm{CH_3CHO}^{\star s} \longleftrightarrow \mathrm{CH_3CHO_{(g)}} + \star_\mathrm{s} \tag*{\{13\}}$$

$$\mathrm{CH_3CHO}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{CH_3CHOH{-}H}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CH_3CH_2OH}^{\star s} + \star_\mathrm{h} \tag*{\{14\}}$$

$$\mathrm{CH_3CH_2OH}^{\star s} \longleftrightarrow \mathrm{CH_3CH_2OH_{(g)}} + \star_\mathrm{s} \tag*{\{15\}}$$

Mechanism on Rh(211) surface

$$\mathrm{H_2} + 2\star_\mathrm{h} \longleftrightarrow 2\,\mathrm{H}^{\star h} \tag*{\{16\}}$$

$$\star_\mathrm{s} + \mathrm{CO_g} \longleftrightarrow \mathrm{CO}^{\star s} \tag*{\{17\}}$$

$$\mathrm{CO}^{\star s} + \star_\mathrm{f} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{CO{-}H}^{\star f} + \star_\mathrm{h} + \star_\mathrm{s} \longrightarrow \mathrm{COH}^{\star f} + \star_\mathrm{h} + \star_\mathrm{s} \tag*{\{18\}}$$

$$\mathrm{COH}^{\star f} + \star_\mathrm{s} \longleftrightarrow \mathrm{C{-}OH}^{\star f} + \star_\mathrm{s} \longleftrightarrow \mathrm{C}^{\star f} + \mathrm{OH}^{\star s} \tag*{\{19\}}$$

$$\mathrm{C}^{\star f} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{C{-}H}^{\star f} + \star_\mathrm{h} \longrightarrow \mathrm{CH}^{\star f} + \star_\mathrm{h} \tag*{\{20\}}$$

$$\mathrm{CH}^{\star f} + \mathrm{H}^{\star h} + \star_\mathrm{s} \longleftrightarrow \mathrm{CH{-}H}^{\star s} + \star_\mathrm{h} + \star_\mathrm{f} \longrightarrow \mathrm{CH_2}^{\star s} + \star_\mathrm{h} + \star_\mathrm{f} \tag*{\{21\}}$$

$$\mathrm{CH_2}^{\star s} + \mathrm{H}^{\star h} \longrightarrow \mathrm{CH_2{-}H}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CH_3}^{\star s} + \star_\mathrm{h} \tag*{\{22\}}$$

$$\mathrm{CH_3}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{CH_3{-}H}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CH_{4(g)}} + \star_\mathrm{s} + \star_\mathrm{h} \tag*{\{23\}}$$

$$\mathrm{O}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{O{-}H}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{OH}^{\star s} + \star_\mathrm{h} \tag*{\{24\}}$$

$$\mathrm{OH}^{\star s} + 2\,\mathrm{H}^{\star h} \longleftrightarrow \mathrm{H{-}OH}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{H_2O_{(g)}} + \star_\mathrm{s} + \star_\mathrm{h} \tag*{\{25\}}$$

$$\mathrm{CO}^{\star s} + \mathrm{CH}^{\star f} \longleftrightarrow \mathrm{CH{-}CO}^{\star s} + \star_\mathrm{f} \longrightarrow \mathrm{CHCO}^{\star s} + \star_\mathrm{f} \tag*{\{26\}}$$

$$\mathrm{CHCO}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{H{-}CHCO}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CH_2CO}^{\star s} + \star_\mathrm{h} \tag*{\{27\}}$$

$$\mathrm{CH_3CO}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{H{-}CH_3CO}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CH_3CHO}^{\star s} + \star_\mathrm{h} \tag*{\{28\}}$$

$$\mathrm{CH_3CHO}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{CH_3CHO{-}H}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CH_3CHOH}^{\star s} + \star_\mathrm{h} \tag*{\{29\}}$$

$$\mathrm{CH_3CHO}^{\star s} \longleftrightarrow \mathrm{CH_3CHO_{(g)}} + \star_\mathrm{s} \tag*{\{30\}}$$

$$\mathrm{CH_3CHOH}^{\star s} + \mathrm{H}^{\star h} \longleftrightarrow \mathrm{CH_3CHOH{-}H}^{\star s} + \star_\mathrm{h} \longrightarrow \mathrm{CH_3CH_2OH_{(g)}} + \star_\mathrm{h} + \star_\mathrm{s} \tag*{\{31\}}$$

A.2 Additonal information on the initial training set

For the training of GAP1 - the first training trial - another initial training set is used than for the subsequent rounds GAP2 and GAP3 (compare table 3.1). Therefore the following table A.1 summarizes the components involved in the initial training set of GAP1.

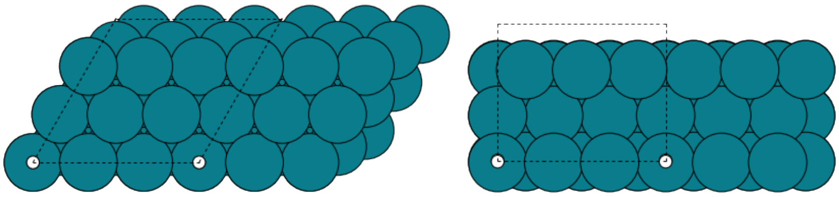
Table A.1: List and specification of the different components included in the initial training set of GAP1, classified into five groups.

Class	Components	Specification
atoms	C, O, H, Rh	/
dimers	CC, CO, CH, HH, OH, OO	dimers with varied distances (in Å) $d = (r_{\text{covalent, 1}} + r_{\text{covalent, 2}} + n \cdot 0.1)$ with n ranging from 0 to 4, taken from [81]
gasphase molecules	CO, H_2, CH_4, COH, CHO, CHOH, CH_2O, CH_2OH, CH_3O, CH_3OH	selected optimized gasphase molecules relevant for the syngas conversion on Rhodium
surfaces	Rh(111), Rh(211)	periodic cell consisting of 36 Rh atoms in 4 layers and a 10 Å vacuum layer
surface + adsorbate	single atoms or molecules adsorbed to Rh(111) or Rh(211) surfaces	periodic cell consisting of 1 adsorbate attached to a Rhodium surface on a specific adsorption site, as defined in table A.2 and visualized in figure A.1

The training of GAP1 just considers one adsorption site (the top site) as a starting point for the iterative training data generation via minima hopping applying the potential of the respective iteration. This adsorption site is pictured in figure A.1 and the site coordinates are given in table A.2.

Table A.2: Site coordinates of the considered surface adsorption sites of Rh(111) and Rh(211) for GAP1. The sites are considered for the construction of surface-adsorbate structures for the initial training set of GAP1.

Surface	Adsorption site	Site coordinates in Å		
		x	y	z
Rh(111)	top (a)	0.0000000	0.0000000	16.6683956
Rh(211)	step top (b)	0.0000000	0.0000000	18.6446576



(a) Rh(111), top site **(b)** Rh(211), step top site

Figure A.1: Illustration of the adsorption sites for (a) Rh(111) and (b) Rh(211) surfaces considered for the surface + adsorbate systems in the initial training set of GAP1.

During the training of GAP2 and GAP3, three adsorption sites are considered for each Rh(111) and Rh(211) with the site coordinates as given in the following table A.3. These sites are additionally pictured in figure 3.2.

Table A.3: Considered adsorption sites for the initial training set of GAP2 and GAP3.

Surface	**Adsorption site**	**Site coordinates in Å**		
		x	y	z
Rh(111)	top	0.0000000	0.0000000	16.6683956
	bridge	0.6805903	1.1788169	16.6683956
	hollow fcc	0.6805903	1.1788169	16.6683956
Rh(211)	step bridge	1.3611806	0.0000000	18.6446576
	terrace hcp	5.4447222	3.7046642	17.5968203
	4-fold	6.8059028	1.1113993	17.8587796

A.3 DFT convergence tests and additional settings

Spin settings

As part of the GAP training, DFT reference calculations are performed. For the initial training set, gasphase reference energies are calculated with DFT, set up with collinear spin settings. The spin settings for these gasphase reference calculations are summarized in table A.4.

Table A.4: Spin settings for DFT calculations of the listed molecules in gasphase

Molecule	**Initial magnetic moment**
CO, H_2, H_2O, CH_2, CH_4, CH_2CO, CH_3CHO, CH_3CH_2OH, $CHOH$	0
OH	[1, 0]
CH	[1, 0]
CH_3	[1, 0, 0, 0]
$CHCO$	[0, 1, 0, 0]
CH_3CO	[1, 0, 0, 0, 0, 0]
CH_3CHOH	[1, 0, 0, 0, 0, 0, 0, 0]
COH	[1, 0, 0]
CHO	[1, 0, 0]

The DFT reference calculations for the surface+adsorbate structures are performed without collinear spin settings.

Hirshfeld parameter test

In order to model atoms and molecules on surfaces, the DFT calculations are set up with additional Tkatchenko-Scheffler dispersion corrections with screened vdW interactions to model atoms and molecules on surfaces. For the non-surface atoms, Tkatchenko-Scheffler DFT+vdW corrections are applied and for interaction of the surface Rhodium atoms with non-surface atoms DFT+vdW$^{\text{surf}}$ corrections, whereby the Rh-Rh interactions are ignored. Previous to application, the parameters have been validated, which is called as hirshfeld test in the following.

During the hirshfeld test, the interaction of selected molecules, CO and CH_4, with both Rh(111) and Rh(211) surface facets are tested. The distance of the adsorbate molecules to the surfaces is increased and the adsorption energy is monitored. The adsorption energy E_{ad} is calculated with the following equation:

$$E_{\text{ad}} = E_{\text{surf+ads}} - E_{\text{surf}} - E_{\text{gas}}, \tag{A.1}$$

whereby $E_{\text{surf+ads}}$, E_{surf} and E_{gas} refer to the electronic free energies of surface+adsorbate, clean surface and gasphase molecule respectively.

Three different parameter settings are tested: 'none' referring to the calculation of energies without any additional vdW correction, 'default' referring to the default vdW parameter values as part of the `FHIaims` package and 'set' referring to screened vdW interactions with parameter values taken from literature [41]. The test is performed in three steps:

1. Optimize gasphase molecules for the three conditions (none, default, set)
2. Optimize adsorbate (optimized gasphase molecule) on Rh surface for all conditions
3. From optimized adsorbate on surface: increase the distance of the adsorbate to the Rh surface (0 to 5Å)

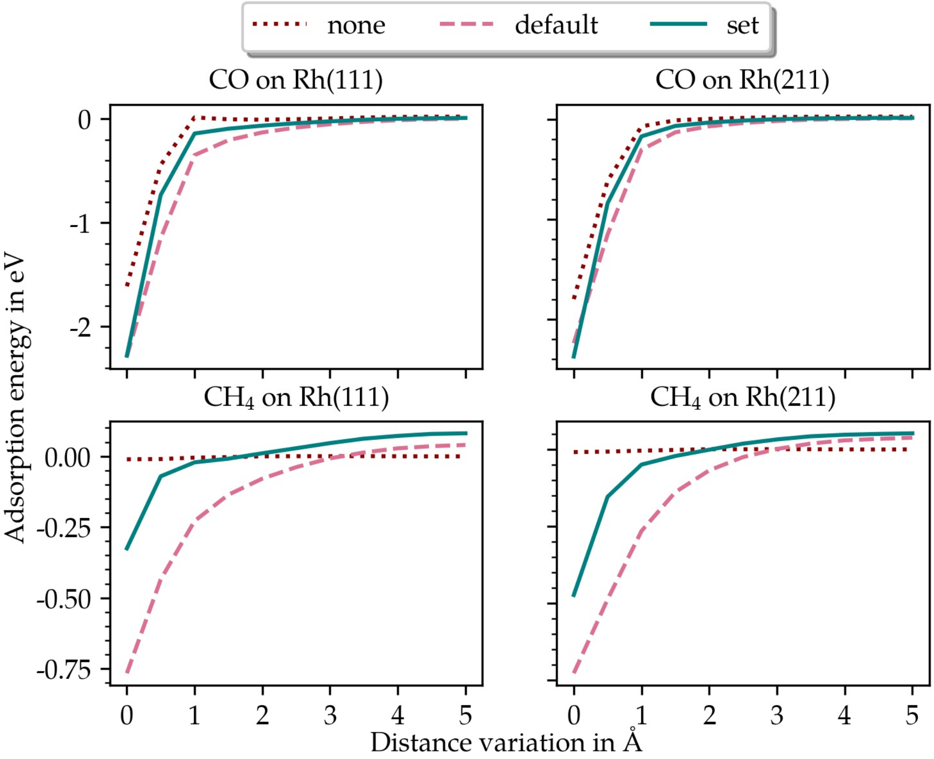


Figure A.2: Effect of the application of DFT+vdW$^{\text{surf}}$ parameters on the DFT adsorption energy of CO or CH_4 on Rh(111) and Rh(211) surfaces. The figure shows the adsorption energy as a function of the distance variation starting from an optimized adsorption geometry without ('none'), with default vdW parameters ('default') and with parameter values [41] ('set').

Figure A.2 shows the effect of the DFT+vdW$^{\text{surf}}$ parameters on the adsorption energies. For CO, the adsorption energies do not differ widely for the different parameter settings. For CH_4, however the need of the additional setting of the parameters is revealed. Without setting additional parameters ('none'), the adsorbate desorbs from the surface and diffuses into the gasphase. That is why the adsorption energy for CH_4 is 0 eV for the 'none' settings, since the molecule does not adsorb to the surface at all without setting additional parameters. For the 'default' settings, the adsorption energies are overestimated. Therefore, the DFT calculations as part of this work are conducted with 'set' parameters, meaning additional Tkatchenko-Scheffler dispersion corrections with screened vdW interactions.

Convergence test for k-point grid

This work's DFT calculations are performed with periodic boundary conditions. The surfaces are built with a periodic cell size of 3x3x4. Thus, DFT reference calculations are calculated with a distinct number of k-points. In order to set the appropriate number of k-points, a k-point grid convergence test has been performed.

Figure A.3 shows the AE, since this is also the energy considered in the GAP training, for 11 different k-point grids. The AE converges for a k-grid of (4x4x1) with a convergence criterion is set to 0.05 eV. Thus, the DFT reference calculations as part of the GAP training are calculated with this converged k-point grid.

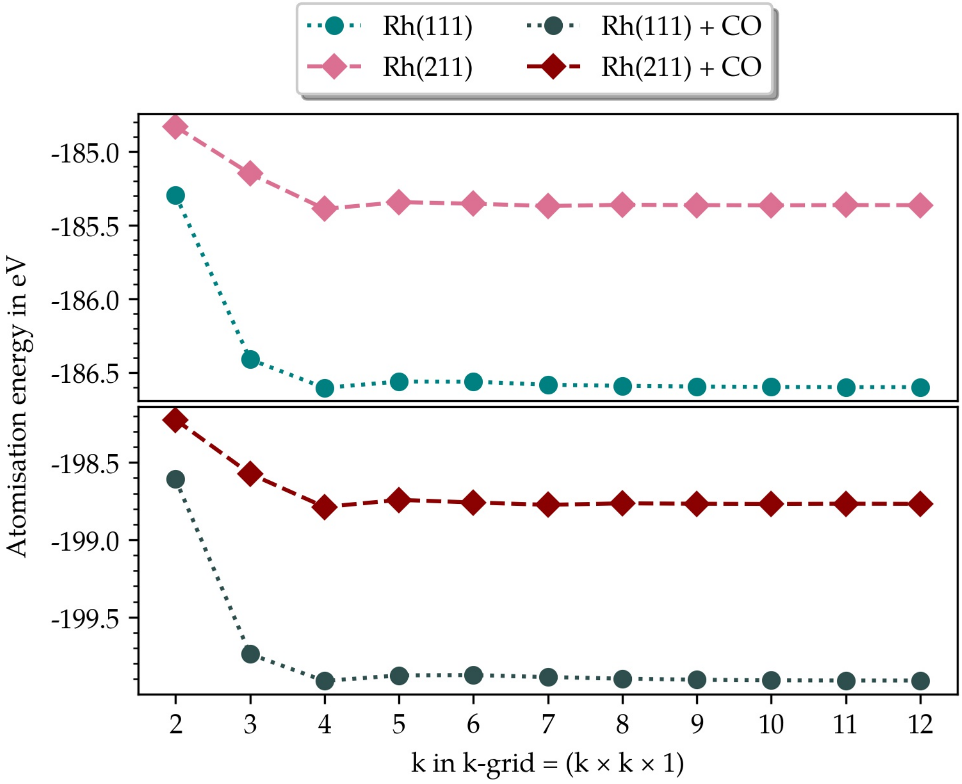


Figure A.3: k-grid convergence test for Rh(111) and Rh(211) surfaces without (upper plot) and with adsorbed CO (lower plot). The convergence test is based on DFT calculations and pictured as atomisation energy (AE) versus k-grids.

A.4 GAP training baseline potential

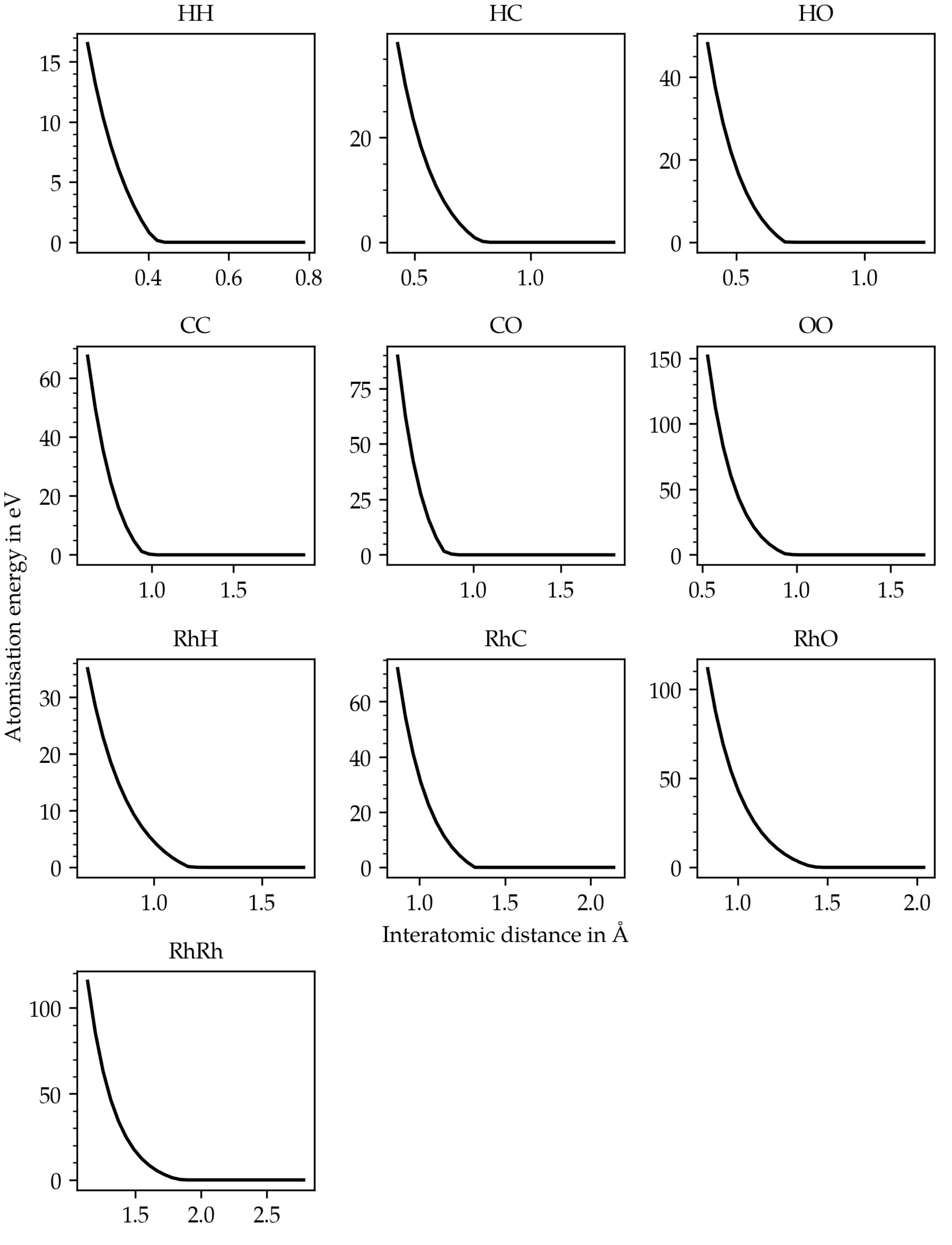


Figure A.4: Illustration of the comprehensive dimer interactions considered by the baseline potential. The baseline is used as a basis underling potential for the training of this work's GAP and taken from [81].

A.5 Hyperparameter finetuning

For the final GAP, a hyperparameter finetuning is performed, as resulted in section 3.2.2. The following figure additionally emphasizes the effect of the different hyperparameter sets on the training and validation errors.

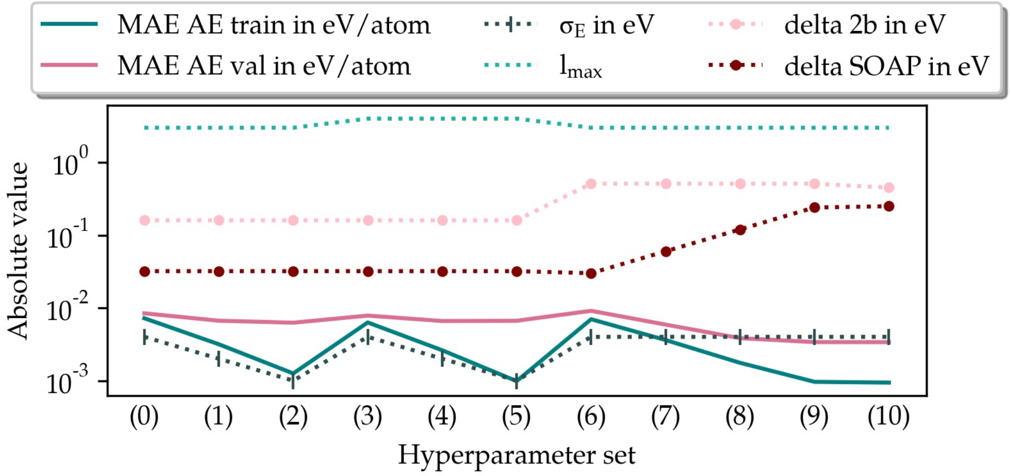


Figure A.5: Effect of different hyperparameter sets (0) to (10) on the training and validation mean absolute error (MAE) of the atomisation energy (AE) per atom (solid lines). The depicted hyperparameters (dashed lines) include the `default energy sigma` σ_E, `l`$_{\texttt{max}}$ (`n`$_{\texttt{max}}$ excluded for simplification) as well as the `deltas` of the 2-body descriptor and the first SOAP only (for both SOAPs, the `depicted` delta value is multiplied by 2).

Appendix B

GAP application details

B.1 Additional details on the first production run

In the first production run as part of the application testing of this work's GAP, the potential is applied to the minima hopping of known adsorbates on known surfaces, but started from new adsorption sites. The considered adsorption sites for both the Rh(111) and Rh(211) surface facets are summarize in table B.1.

Table B.1: Considered adsorption sites for the first production run.

Surface	**Site coordinates in Å**		
	x	y	z
Rh(111)	0.0000000	0.0000000	16.6683956
	0.6805903	1.1788169	16.6683956
	0.6805903	1.1788169	16.6683956
	5.4447222	6.2870237	16.6683956
Rh(211)	0.0000000	2.2227985	17.0729016
	1.3611805	4.4455971	17.8587796
	0.0000000	0.0000000	18.6446576
	6.1253125	3.3341978	17.4658406
	0.6805903	5.5569963	18.2517186
	0.0000000	4.4455971	17.8587796
	4.0835417	2.2227985	17.0729016
	1.3611806	0.0000000	18.6446576
	5.4447222	1.1113993	17.8587796
	1.3611806	5.9274628	18.3826982
	4.0835417	2.9637314	17.3348610
	5.4447222	3.7046642	17.5968203
	5.4447222	5.1865299	18.1207389
	6.8059028	1.1113993	17.8587796

Based on the produced minima via the first production run, reduced reaction networks are developed. The following energies, $E_{f,min}$, are used for the coloring of the networks and are calculated via equation 4.2.

Table B.2: Formation energies in eV of DFT reoptimized GAP minima used in the reaction networks.

Adsorbate	**Rh(111)**		**Rh(211)**	
	$E_{f,\text{ min}}$	$E_{f,\text{ mean}}$	$E_{f,\text{ min}}$	$E_{f,\text{ mean}}$
CO	-3.0832	-3.0832	-3.0884	-3.0584
H2	-1.3118	-1.1959	-1.4440	-1.2989
H	-1.2475	-1.2216	-1.3667	-1.3004
H_2O	-0.5736	-0.5663	-0.7534	-0.5192
OH	-0.2907	-0.2748	-0.9013	-0.8553
CH	-3.2984	-3.2621	-3.7697	-3.7540
CH_2	-3.4179	-3.4152	-3.7352	-3.7235
CH_3	-3.8767	-3.8763	-3.8359	-3.7867
CH_4	-3.3723	-3.3618	-3.7937	-3.5889
CHCO	-4.9182	-4.9166	-5.4589	-5.3064
CH_2CO	-5.0942	-5.0693	-5.3353	-5.0624
CH_3CO	-5.3506	-5.2937	-5.4256	-5.1597
CH_3CHO	-5.3832	-5.3730	-5.6251	-5.3975
CH_3CHOH	-5.3502	-5.3336	-5.9021	-5.4808
CH_3CH_2OH	-5.5193	-5.4981	-5.6865	-5.4434
CHOH	-2.9333	-2.8725	-3.0213	-2.8389
COH	-2.9003	-2.8220	-3.1929	-3.1788
CHO	-3.1181	-3.1142	-3.1064	-3.0721
C	-2.2433	-2.1472	-3.3344	-3.3341

B.2 Additional details on the second production run

In the second production run - the out of sample testing -, unknown structures from the literature [4] are optimized via minima hopping with this work's trained potential. Thereby, the following low-coverage surface-adsorbate systems are considered:

Table B.3: Considered structures for the second production run.

Surface	Periodic cell	Lattice constant	Adsorbates
Rh(111)	3x3x4	3.860Å	H, H_2, CO, OH, H_2O, CH, CH_2, CH_3, CH_4, CHO, CHOH, CHCO, CH_2CO, CH_3CO,CH_3CHO, CH_3CHOH, CH_3CH_2OH
Rh(211)	3x3x4	3.860Å	CO, CHCO, CH_3CO, CH_3CHO, CH_3CHOH
Rh(211)	3x3x4	3.866Å	O, C, CH, CH_2, CH_3, OH, COH