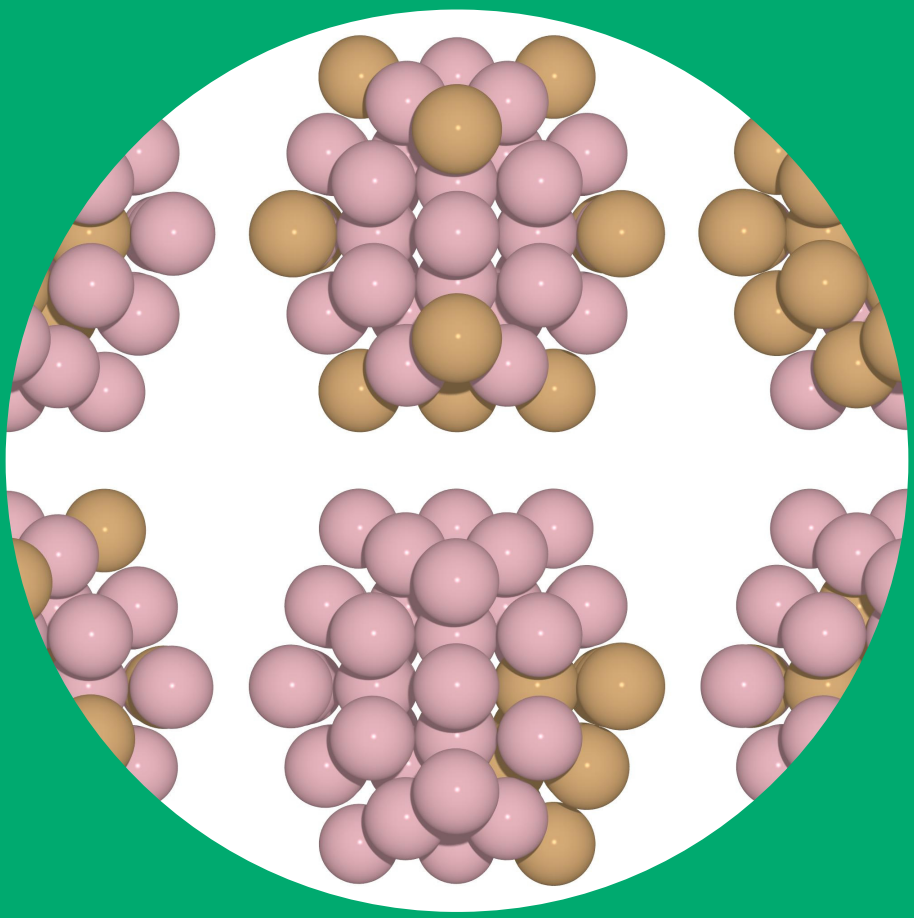
Department of Applied Physics

Efficient screening of nanoclusters as catalysts for the hydrogen evolution reaction

Marc Jäger



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 130/2020

Efficient screening of nanoclusters as catalysts for the hydrogen evolution reaction

Marc Jäger

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, remote connection Zoom link https://aalto.zoom.us/j/61477725193, on 30th September 2020 at 11:00.

Aalto University
School of Science
Department of Applied Physics
Surfaces and Interfaces at the Nanoscale (SIN)

Supervising professor
Professor Adam Stuart Foster, Aalto University, Finland

Preliminary examiners
Dr. Thomas Hammerschmidt, Ruhr University Bochum, Germany
Professor Thomas Bligaard, Technical University of Denmark, Denmark

Opponent
Professor Thomas Bligaard, Technical University of Denmark, Denmark

Aalto University publication series
DOCTORAL DISSERTATIONS 130/2020

ISBN 978-952-64-0016-7 (printed)
ISBN 978-952-64-0017-4 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0017-4

Unigrafia Oy
Helsinki 2020

Finland

Printed matter
4041-0619

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Marc Jäger

Name of the doctoral dissertation
Efficient screening of nanoclusters as catalysts for the hydrogen evolution reaction

Publisher School of Science

Unit Department of Applied Physics

Series Aalto University publication series DOCTORAL DISSERTATIONS 130/2020

Field of research Engineering Physics

Manuscript submitted 10 June 2020 **Date of the defence** 30 September 2020

Permission for public defence granted (date) 7 September 2020 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Heterogeneous catalysis is a key component in modern industry as catalyst breakthroughs improve existing or accommodate the emergence of new technologies. For instance, catalyzing the splitting of water for energy storage purposes efficiently and cheaply is a potentially disrupting innovation. Nanoclusters have the potential to replace existing catalysts due to their catalytic behaviour at the nanoscale. However, experimental testing is often slow and expensive, and focuses on gradual improvements of known catalysts, prohibiting the discovery of novel materials. Computer simulations offer a method to design a new catalyst from scratch, allowing nanoclusters to be screened efficiently for their catalytic activity.

Existing screening methods are often designed for simple infinite surfaces, neglecting the shape and size effects of nanoclusters. The large search space of catalyst screening at the nanoscale also poses a challenge to computational screening methods. This dissertation deals with the development of new methods for efficient screening of nanoclusters, explicitly capturing size and shape effects. In particular, machine learning (ML) approaches were used to reduce computational cost and a large part of the work was devoted to the benchmarking of descriptors as a key step in ML. New nanocluster-adsorbate tools were developed, these are the efficient exploration of nanocluster configurations, the exclusion of redundant adsorption sites and a DFT-ML loop. The screening workflow was automated and connected to a database allowing for screening and management of large sets of data. The workflow was verified on the simple hydrogen evolution reaction, a key reaction to electrolytic water splitting, and a bimetallic dataset containing several compositions of Ti, Co, Fe, Ni, Cu and Pt was screened.

The implementation of new tools was kept modular and the programming aspect of the work is captured in three packages which are all publicly available and benefit other computational materials science researchers. The developed tools are not restricted to the model reaction; they are kept general so that they can be applied to other catalytic reactions on nanoclusters of arbitrary shapes and sizes.

Keywords nanoclusters, catalysis, hydrogen evolution reaction, Rational Catalyst Design, computational materials science, automation, machine learning

ISBN (printed) 978-952-64-0016-7 **ISBN (pdf)** 978-952-64-0017-4

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2020

Pages 129 **urn** http://urn.fi/URN:ISBN:978-952-64-0017-4

Preface

Catalysis research is a vast and wonderful field of which I have merely scratched the surface during my doctoral studies. I would first and foremost like to thank Prof. Adam Foster for welcoming me into his Surfaces and Interfaces at the Nanoscale (SIN) group with open arms and mind, despite only being a physicist of the electron shell. But the work would not have been so successful without Prof. Ingo Krossing's invaluable lessons in quantum chemistry during my master's, teachings of a solid foundation in catalysis and life lesson of chemical intuition enabling me to additionally handle liquid matters.

I am thankful to Dr. Thomas Hammerschmidt and Prof. Thomas Bligaard, for acting as pre-examiners and for their insightful observations regarding this dissertation. I feel honored that the latter agreed to be the opponent at my public defense.

The opportunity to work in a big European project ("have a break, have a CritCat") also posed the problem of interdisciplinary and combinatorial information sharing. Prof. Jaakko Akola, David Gao and Olli Ahlstedt made the numerous project meetings and travels enjoyable and scientifically significant. It was a pleasure to sometimes share my hopes and dreams with them. Furthermore, Marc Heggen's and Yuri Kolen'ko's hospitality made my research visits productive and memorable.

I would like to thank all former (Ville Haapasilta, Eero Holmstrøm, Martha Zaidan, Tiziana Musso, Lidija Zivanovic, John Tracey, Prokop Hapala, Stavrina Dimosthenous, Simiam Ghan, Lei Yang) and current members (Ygor Morais Jaques, Orlando Silveira Júnior, Ondrej Krejci, Fabio Priante, Niko Oinonen, Fedor Urtev, Lauri Kurki) of the SIN group by name, because I am grateful for getting the chance to socialize and work with each of them, and explicitly not to abuse this sentence as a mnemonic device. My work was particularly catalyzed by the following diligent SINners, co-authors and friends. To Eiaki Morooka who started as a project partner but became a great friend. To Filippo Federici Canova who shattered my unrealistic expectations with his brutal honesty. To Yashasvi Ranawat whom I witnessed growing in the past years from having no ID

to being indeed a great researcher. To Lauri Himanen who turned me into a better coder (from zero to mediocre). To you goes life-long gratitude.

The Saints of the Department of Applied Physics provided the much-needed counterbalance to an otherwise workaholic environment. Dorothea Golze, Kunal Gosh, Marc Dvorak, Jari Järvi, Miguel Caro and Azeema Saeed enriched my work life in the corridor with semi-scientific chats, extended coffee breaks and epidemic laughter. I greatly appreciated that Prof. Patrick Rinke helped me improve my scientific writing. Milica Todorovic always had helpful advice and funny anecdotes up her sleeve. Rina Ibragimova quickly rose to become one of the most esteemed scientists and presidents the corridor had ever seen. Petri Hirvonen patiently tested my Finnish between doors, after work at the solid-liquid interface and at low-friction powdered surfaces. Last but not least, *Prof.* Annika Stuke greatly increased the prestige of the department by founding the LIT group which became well-known even beyond computational materials science. I stand in her debt and I personally thank her mother for raising such a brave and visionary woman.

Many new friends in Finland have turned this chapter in my life into a win-Finn situation. With their help, I could find the right work-life balance to ensure that I would not graduate too early. I am grateful that I found friends one could rely on in times of need. Especially, I am very happy to have met Ahmed and Lucas whom I shared many memorable moments with. I was sad to see Ahmed, the best flatmate, leave to the North. I am much obliged to my remote friends Hanife, Jan Philip, Marina and Martin who bared with me throughout all these years.

To my parents (aka *Hallo Papa, Hallo Mama*), who have made my education possible and encouraged me to pursue a PhD. They were always there in difficult times. Their support is appreciated beyond the nanoscale. Vera, the best calendar ever exist(ed) and whom Outlook can only envy, I profoundly thank for introducing me to the world of off-the-shelf molecules.

Espoo, August 31, 2020,

Marc Jäger

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Lauri Himanen, Marc Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z.Gao, Patrick Rinke, Adam S.Foster. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, Volume 247, Article number 106949, 12 pages, February 2020.

II Marc Jäger, Eiaki V. Morooka, Filippo Federici Canova, Lauri Himanen, Adam S. Foster. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials*, Volume 4, Number 37, 8 pages, July 2018.

III Marc Jäger, Filippo Federici Canova, Eiaki V. Morooka, Adam S.Foster. Efficient machine-learning-aided screening of hydrogen adsorption on bimetallic nanoclusters. Submitted to *ACS Combinatorial Science*, May 2020.

Author's Contribution

Publication I: "DScribe: Library of descriptors for machine learning in materials science"

The author co-implemented and tested descriptors, contributed to the continuous integration of the package, provided parts of the documentation and commented on the manuscript.

Publication II: "Machine learning hydrogen adsorption on nanoclusters through structural descriptors"

The author created the data, performed machine learning and wrote the manuscript.

Publication III: "Efficient machine-learning-aided screening of hydrogen adsorption on bimetallic nanoclusters"

The author created and analyzed the data, contributed large parts of the nanocluster-adsorbate tools, programmed and ran the workflow, and wrote the manuscript.

Abbreviations

AiiDa	Automated Interactive Infrastructure and Database for Computational Science
ACSF	Atom-centered symmetry functions
BEP	Bell-Evans-Polanyi principle
BOSS	Bayesian optimization structure search
CP2K	Car-Parrinello 2000
DFT	Density functional theory
FPS	Farthest-point sampling
GGA	General-gradient approximation
HER	Hydrogen evolution reaction
KRR	Kernel ridge regression
KS	Kohn-Sham
LDA	Local-density approximation
MAE	Mean absolute error
MBTR	Many-body tensor representation
MC	Monte-Carlo
ML	Machine learning
PBE	Perdew–Burke-Ernzerhof functional
PES	Potential energy surface
PGM	Platinum group metals
SOAP	Smooth overlap of atomic positions

Symbols

$\ddagger$	Transition state
$*$	Active site of a catalyst
E	Total energy
E_0	BEP constant energy shift
E_a	Activation energy
E_H	Hartree energy
E_{xc}	Exchange correlation functional
$E_\Theta L$	Expected loss given parameters Θ
f	d-band filling
g_{nl}	Radial basis function
$\hat{\mathbf{H}}$	Hamiltonian operator
h_{KS}	Kohn-Sham Hamiltonian operator
$\mathbf{K}$	Kernel matrix
k	Reaction rate
k_B	Boltzmann constant
$\mathbf{I}$	Identity matrix
n	Electron density
n_0	Ground state electron density
$\mathbf{P}^\xi(\mathbf{x})$	SOAP power spectrum at position x with configuration ξ
R_c	Cutoff radius
S	Entropy
T	Temperature or kinetic energy
V	Coupling matrix element
V_{ext}	External potential
Y_{lm}	Spherical harmonic function

α	BEP pre-factor
β	Coupling parameter
γ	Kernel length-scale
ΔE_r	Reaction energy
ΔE_{ads}	Adsorption energy
ΔG_{ads}	Gibb's free adsorption energy
ϵ_d	d-band center
ϵ_d^w	d-band center plus half the d-band width
ϵ_H	Energy of the free hydrogen 1s-orbital
ϵ_i	Eigenvalue of the ith Kohn-Sham state
ϵ_u	Maximum of the d-band Hilbert transform
λ	Kernel regularization parameter
ν	Arrhenius pre-exponential factor
ν_{ext}	Single-particle external potential
ρ	Atomic density
$\mathbf{\Psi}$	Wave-function
ψ	KS wave-function

1. Introduction

Most physical products, of which rare elements dictate the price of e.g. consumer electronics, are born out of raw materials. Therefore, minimizing the use of raw material while maintaining product quality reduces manufacturing costs. For instance, materials for coatings can be broken down to smaller and smaller grains to enable thin and even layering. Likewise, catalyst materials are made more effective by maximizing the surface area. As catalytic converters in cars use expensive noble metals they are sparsely dispersed on porous materials. The reduction of particle size to a scale of a few nanometers has lead to material improvements in e.g. paints, filters, lubricants and catalysts.

In studying the behaviour of nanoparticles of about 10 to 1000 atoms, we first tend to describe them using familiar structures, either molecules of a few atoms or practically infinite crystals [1]. Yet those structures of 1 to 3 nm, so called nanoclusters, do not submit to the molecule or crystal category but lie in an intermediate position with regard to their properties. Compared to crystals, nanoclusters display remarkably different magnetic, optical, and reactivity properties [2, 3, 4, 5, 6].

Properties change because of three main reasons: (i) atoms on the surface dominate (large surface-to-volume ratio), (ii) atoms on surface edges have low coordination, (iii) and the core measures only few atomic lengths [7]. The core is sometimes under strain and differently packed than in a crystal. Nanoclusters can take on a myriad of sizes and shapes, and can be composed of a multitude of elements with different compositions. In the realm of possible structures, new unique sets of properties wait to be explored but the field is no longer just of academic interest [8, 9]. With the advent of new manufacturing methods to control the size of nanoclusters, hope rose for fine-tuning their properties, and design materials with improved characteristics [9, 10, 11, 12, 13].

Heterogeneous catalysis is one of the fields which can benefit greatly from the exploration of nanoclusters. Catalysts are substances which enable or speed up certain reactions, but do not get consumed. They are called heterogeneous if they are in a different phase (usually solid) than the reac-

tants (usually liquid or gaseous). This dissertation addresses nanoclusters in the field of heterogeneous catalysis. In particular, it focuses on the electrolytic splitting of water. Water splitting can be used to store excess electricity from the electric grid in the energy carrier hydrogen. Electricity prices are expected to get more volatile in the future, hence it is a viable option to produce hydrogen when supply is high and reverse the reaction at high demand, stabilizing the electric grid [14]. Next to electricity prices, the catalyst material, platinum, contributes substantially to the technology cost [15]. Platinum is a rare and expensive noble metal which has been deemed a critical element due to its industrial importance and high supply risk [16, 17]. Therefore, reduction in its use, or replacement, is of high economic and political interest.

This dissertation approaches the search for new catalysts from the theoretical side. In computational materials science we rely on simulations instead of experiments. Simulations have many advantages in vast and relatively unexplored fields. In fact, computational screening is much cheaper than experimental screening, especially in catalysis where experiments tend to take long and require elaborate setups. On top of that, it is easier to uncover underlying trends with a series of simulations in a zero-noise environment. Computational screening is indeed a good way to sieve out promising catalyst candidates to eventually select those to study more in-depth experimentally. However, the search space of nanoclusters is so large that an efficient screening approach is paramount.

Machine learning has become a popular tool in data-driven materials science. If many nanoclusters of interest have a similar structure, machine learning can help interpolate between them, as a result predicting their properties in batches, by only explicitly simulating a few. This saves computational resources and enables screening of larger datasets.

Before the actual application of machine learning, the preparation of the data is a necessary step and can cause the prediction process to succeed or fail. Machine learning takes the structure, or other known information about the nanocluster in order to predict properties of interest. The success of machine learning depends on how the structure is represented and provided as input. A large part of the dissertation is concerned with how to transform relevant structural information into machine-readable input, also known as structural descriptors.

Finally, in order to obtain and manage large databases of nanoclusters, automation plays an important role in making the results consistent and reducing chances for human error. Human interaction can be greatly reduced by automating the whole workflow from the generation and exploration of nanoclusters, via the submission of simulation jobs to the machine learning prediction.

1.1 Research Objectives

This dissertation addresses the problem of a large nanocluster search space and demonstrates methods to efficiently screen it. A large part of my work pertains to the development of methods, centered around machine learning, to enable the simulation of large databases. The rest is dedicated to the application of those methods on catalysis topics of high current interest. The target nanoclusters and the test catalytic reaction were chosen on that regard; the hydrogen evolution reaction is simple yet key to many energy storage applications [18, 19, 20].

The first study is a library of descriptors which I contributed to. By packaging several state-of-the-art descriptors, it facilitates applying machine learning to materials science problems with ease. As such, there is a lower threshold for scientists who are new to machine learning to start applying these descriptors in their research. The second study leverages the implementations of descriptors for the library by applying them to a benchmark on nanocluster adsorption. It was the first extensive machine learning work on nanoclusters, proving that adsorption energies can be efficiently learned. The last study builds on the insights acquired from the previous studies. It expands the toolbox culminating in a library devised to build cluster-adsorbate structures. The large configuration space is addressed and machine learning is applied in new ways on nanoclusters as part of an automated workflow. It enables the creation of large databases of nanoclusters of arbitrary shapes, sizes and compositions.

The following chapter presents concepts of catalysis in more detail, providing the theoretical background of the principles and theories that guide surface-molecule interaction. The third chapter introduces the concept of descriptors and touches on its link to density functional theory. The fourth chapter provides a nanocluster application of machine learning, that is benchmarking of descriptors and hydrogen adsorption energy prediction on the systems MoS_2 and AuCu. The fifth chapter features the tools from generating nanoclusters to efficient screening of adsorption energy distributions which were ultimately combined to an automated workflow. It was demonstrated on a dataset of bimetallic nanoclusters for the hydrogen evolution reaction. In the last chapter, the work is summarized and an outlook to future challenges and opportunities in theoretical nanocatalysis is given.

2. Rational Catalyst Design

In this chapter I explain the foundation of theoretical catalysis excluding kinetic theories. They stretch from the early concepts of catalysis to d-band theory. Different levels of approximation of how active a catalyst is are introduced. The chapter closes with the introduction of the electronic properties that guide the search for promising catalysts.

2.1 Concepts of catalysis

A catalyst is a substance which takes part in a chemical reaction but eventually returns to its initial state where it can react again, hence it is not consumed by the reaction. By definition a catalyst also changes the path of a reaction favourably so that the thermodynamic equilibrium is reached faster (usually many orders of magnitude faster) than without its participation. Additionally, catalysts can enable reactions that would not otherwise take place under normal conditions. For instance, the splitting of the stable nitrogen molecule would happen in air only through lightning strikes but Haber and Bosch developed a process that fixates atmospheric nitrogen (N_2) to produce ammonia (NH_3) on an iron catalyst surface [21]. How does a catalyst enable the otherwise impossible? In order for N_2 to take part in any reaction, its strong triple bond will have to break. It either breaks completely into two N-radicals by, e.g. a high-energy lightning source, or it breaks step-wise as it happens on the surface of iron. The latter requires less activation energy, E_a, the energy necessary to transition N_2 to NH_3.

The reaction pathway of an educt A to a product B is best illustrated through the analogy of a journey in a mountainous area. The height of the mountain is analogous to the energy of the system. The valleys of low potential energy represent the (meta-)stable states such as A and B. On a journey from valley A to its neighbouring valley B, mountain peaks of high energy are states which will not be traversed since the alternative leading through a mountain pass (saddle point) requires the least energy.

This is the highest point on the path representing the transition state $A^{\ddagger}$ determining the activation energy of the reaction. It is called activation energy because the educt requires at least that amount of additional energy for the reaction to take place. In the example of splitting the nitrogen molecule, the transition state lies where the bond between the nitrogen atoms is elongated and almost broken. If the transition state lies too high making the reaction via the direct path unfeasible, an indirect path could lead through lower-lying mountain passes (transition states) eventually reducing the activation energy. A catalyst enables such a detour illustrated in figure 2.1. The reaction path with the participation of a catalyst passes

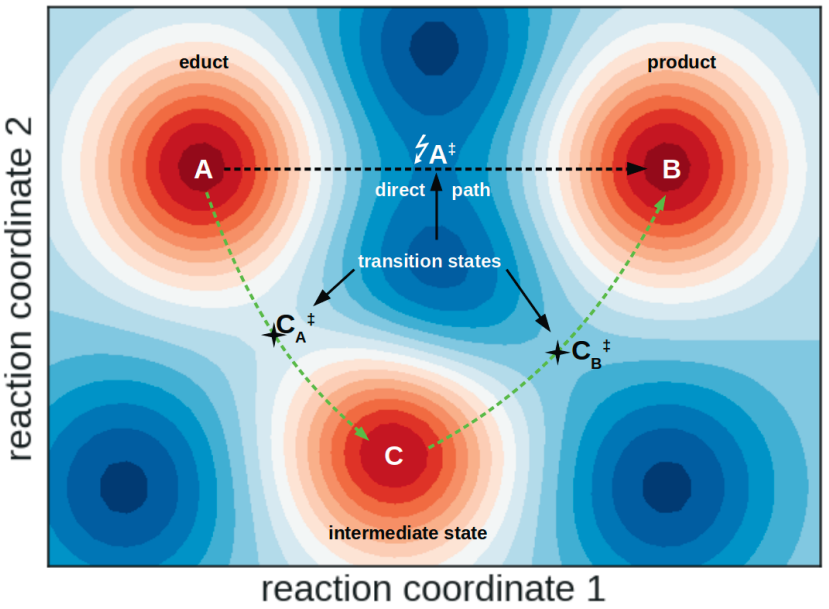


Figure 2.1. Illustration of reaction pathways without (black dashed line) and with (green dashed line) a catalyst. The axes represent reaction coordinates, the meaning of which depends on the reaction. The x-axis for instance could represent the elongation and breaking of a bond and the y-axis could represent the proximity to the catalyst. The educt A, product B and intermediate state C involving the catalyst are all local minima (valleys). The transition states (mountain passes) have different energy-levels: $A^{\ddagger}$ lies the highest so that the direct path is energetically unfavourable. The reaction would likely go through $C_A^{\ddagger}$ and $C_B^{\ddagger}$. The rate-determining step is on the product-side because $C_B^{\ddagger}$ lies higher than $C_A^{\ddagger}$.

via, at least, another valley called the intermediate state. In certain cases, the intermediate state is stable and can be isolated, while in others the reaction continues so quickly that it is difficult to observe. In the simplest catalytic reaction, A + * $\longrightarrow$ A* $\longrightarrow$ B + *, the asterisk denoting the catalyst binding site, there are two transition states. The highest-lying transition state, $C_B^{\ddagger}$, in this example, requires the most activation energy, thus the second step determines the rate of the reaction.

A heterogeneous catalyst is a catalyst in a different phase from its reactants. The fixation of nitrogen is a prime example of heterogeneous catalysis with high industrial importance. The formation of NH_3 from H_2 and N_2 in the Haber process can be broken down into the following steps:

$$\begin{aligned}
\mathrm{N_2} + 2\,* &\longrightarrow \mathrm{N_2}* \\
\mathrm{N_2}* + * &\longrightarrow 2\,\mathrm{N}* \\
\mathrm{H_2} + 2\,* &\longrightarrow \mathrm{H_2}* \\
\mathrm{H_2}* + * &\longrightarrow 2\,\mathrm{H}* \\
\mathrm{N}* + 3\,\mathrm{H}* &\longrightarrow \mathrm{NH_3}* + 3\,* \\
\mathrm{NH_3}* &\longrightarrow \mathrm{NH_3} + * \\
\hline
3\,\mathrm{H_2} + \mathrm{N_2} &\longrightarrow \mathrm{NH_3}
\end{aligned} \tag{2.1}$$

The second step determines the rate of the reaction, since the breaking of the triple bond N≡N, although aided by the catalyst, requires the most activation. Similarly, in the fourth step, the hydrogen molecule H_2 is split on the catalyst surface, however, the single bond breaks at a much lower activation energy.

The reaction rate k, determined by the energy level of the highest transition state, is temperature dependent and is given by the Arrhenius law [22]:

$$k(T) = \nu e^{-\frac{E_a}{k_B T}} \tag{2.2}$$

where ν is the pre-exponential factor and k_B is the Boltzmann constant. Changes in activation energy or temperature T lead to exponential changes in reaction rates. A series of experiments is required to determine the activation energy. It can be approximated by simulations, yet it becomes a daunting task the more complex the reaction is. The question arises if there is a way to express the reaction rate without having to explicitly calculate E_a. It turns out that, although the absolute reaction rate remains unknown, the relative reaction rate in a family of similar reactions can be derived. Since the energies of the minima (valleys) and the energies of the saddle points (mountain passes) on the energy landscape are not independent of each other, it is sufficient to know the relative energy levels of educts and products.

2.1.1 Bell-Evans-Polanyi principle

The reaction energy, $\Delta E_r = \sum E_p - \sum E_e$, is defined as the difference between the total energies of products and educts. Brønsted, Bell, Evans and Polanyi showed that E_a depends linearly on the reaction energy for a family of reactions [23, 24, 25, 26, 27, 28].

$$E_a = E_0 + \alpha \Delta E_r \tag{2.3}$$

where E_0 is a constant energy shift and α is a constant specific for a family of reactions. The Bell-Evans-Polanyi (or Brønsted-Evans-Polanyi) principle connects a kinetic quantity (E_a, see 2.2) with a thermodynamic one (ΔE_r). It allows one to make relative statements about reaction rates purely from knowing thermodynamic properties. As a result, computational materials science leverages such a principle to reduce simulation time considerably.

2.1.2 Sabatier principle

The Sabatier principle as a qualitative catalytic concept states that a substrate should neither bind too weakly nor too strongly on the catalyst surface [29]. If the binding is too weak, the reaction path via adsorption on the catalyst is slow or unfavourable. If the binding is too strong, the resulting intermediate state can poison the catalyst because the last step, the desorption of the product, slows down the whole reaction. Hence, an efficient catalyst needs to keep the adsorption and desorption reactions in balance. Together with the BEP it is possible to quantitatively predict the optimal energy-level of an intermediate state.

Let A + * $\longrightarrow$ A* $\longrightarrow$ B + * be an energy-neutral reaction. Furthermore, let the adsorption A + * $\longrightarrow$ A* and desorption A* $\longrightarrow$ B + * belong to the same family of reactions. Then the highest activation energy can be minimized as follows.

$$\begin{aligned}
\Delta E_r =& \Delta E_r^{ads} + \Delta E_r^{des} = 0 \\
\Delta E_r^{ads} =& -\Delta E_r^{des} \\
\min_{\Delta E_r^{ads}} \max(E_a^{ads}, E_a^{des}) =& \min \max(E_0 + \alpha \Delta E_r^{ads}, E_0 + (1-\alpha)\Delta E_r^{des}) \\
=& E_0 + \min \max(\alpha \Delta E_r^{ads}, (1-\alpha)\Delta E_r^{ads}) \\
=& E_0 \text{ with } E_r^{ads} = 0
\end{aligned} \tag{2.4}$$

Hence, the optimal energy-level of the intermediate state is aligned with the energy-level on both educt and product sides. It can be generalized to non-energy-neutral reactions, where the intermediate state lies optimally half-way in energy between educts and products.

In conclusion, the adsorption or desorption energy should be around zero for a good catalyst. Unfortunately, this does neglect entropy as a driving force which is not an easily-accessible quantity. The reaction energy ΔE_r should therefore be replaced by the reaction free energy ΔG_r. Accordingly, a good catalyst fulfills [30, 31]:

$$\Delta G_{ads} \approx 0 \tag{2.5}$$

As the adsorption free energy ΔG_{ads} is directly correlated with catalytic activity, it is thus called a descriptor [30]. A detailed description of the

Gibb's free energy $G = E + pV - TS$ is out of the scope of this work. At constant pressure and temperature the chemical potential of a reactant i is the partial molar Gibb's free energy:

$$\mu_i = \left(\frac{\delta G}{\delta N_i} \right)_{T,P} \tag{2.6}$$

The chemical potential has to drop from educts via intermediate states to products in order for a reaction to take place. It should be noted that the change in entropy S has a considerable impact even at room temperature. Since calculating the true ΔG_{ads} is challenging, it is usually approximated [32, 33]. Despite the existence of other, more expensive, simulation methods to determine ΔG_{ads}, even a constant shift in a family of reactions yields good results [32, 34].

2.1.3 Hydrogen evolution reaction

An example of a simple reaction with one intermediate state is the hydrogen evolution reaction (HER). It is the cathodic half reaction of electrolytic water splitting into hydrogen and oxygen [35, 36].

$$\begin{aligned} 4\,\mathrm{e^-} + 4\,\mathrm{H^+} &\longrightarrow 2\,\mathrm{H_2} \\ \underline{2\,\mathrm{H_2O} &\longrightarrow 4\,\mathrm{H^+} + \mathrm{O_2} + 4\,\mathrm{e^-}} \\ 2\,\mathrm{H_2O} &\longrightarrow 2\,\mathrm{H_2} + \mathrm{O_2} \end{aligned} \tag{2.7}$$

As the reaction is reversible, the simple reaction arrows should be replaced by $\rightleftharpoons$ to be precise. However, since the hydrogen oxidation reaction is part of a different application, I noted only the direction of interest for simplicity. HER involves the following steps with the last two competing with each other [37, 38]:

$$\text{Volmer step: } \mathrm{H^+} + \mathrm{e^-} + * \longrightarrow \mathrm{H*} \tag{2.8}$$

$$\text{Heyrovsky step: } \mathrm{H*} + \mathrm{H^+} + \mathrm{e^-} \longrightarrow \mathrm{H_2} + * \tag{2.9}$$

$$\text{Tafel step: } 2\,\mathrm{H*} \longrightarrow \mathrm{H_2} + 2\,* \tag{2.10}$$

The reactions assume an acidic environment, although HER can also be conducted in base with analog reactions. At first, in the Volmer step a proton adsorbs on the catalyst surface which is provided with excess electrons by applied voltage. The second proton either reacts directly from the liquid phase (Heyrovski step) or reacts only after adsorption on the surface to form H_2 (Tafel step). The product molecule desorbs quickly from the surface. Regardless of the reaction pathway, Volmer-Heyrovsky or

Volmer-Tafel, the intermediate state H* is always part of the mechanism. The adsorption energy is defined as:

$$\Delta E_{ads} = E(\mathrm{H}^*) - E(^*) - \frac{1}{2}E(\mathrm{H_2}) \tag{2.11}$$

The reaction is endergonic, meaning that it will require energy in the form of an electrode potential in order to take place. The ideal catalyst would facilitate the reaction after lifting the free energy level of the educt side to match the product side. Evidently, the intermediate state would have to level off, too, as explained by the Sabatier principle ($\Delta G_{ads} = 0$). In practice, the electrode potential even with the best catalyst needs to be higher to set the reaction in motion. This excess potential is called overpotential and is an experimental measure of catalytic activity.

HER is only one side of the water splitting process (see 2.7). Since the oxygen evolution reaction leads through 3 intermediate states, it is a more complex reaction and ultimately more challenging to find a good catalyst via simulations [39]. However, the Sabatier principle can be applied here as well, for each intermediate state independently. The above reactions and their reverse counterparts are key to mobile and stationary energy storage applications.

Hydrogen adsorbed on catalytic surfaces is not only of interest in HER. The Haber process (see 2.1) is only one of many reactions adsorbing hydrogen, other industrial examples are methanol production from syngas and desulfurization. In my work, the hydrogen evolution reaction was the starting point to develop new methods for catalysis on nanoclusters. Due to its simplicity and participation in many industrial applications, it is a prime candidate to showcase new screening methods for nanocatalysts.

2.2 d-band theory

As we exploited the BEP and the Sabatier principle, we could circumvent determining transition states of catalytic reactions and solely rely on thermodynamic properties. Is there a way to reduce the computational cost even further? Is it possible to make predictions on the catalytic activity of a material without simulating any surface-adsorbate interactions? After all, the adsorbate remains the same, as does the termination of the surface. The answer is yes, but at the price of losing accuracy. The d-band theory links the electronic structure of a surface with the adsorption strength of atoms or small molecules [30, 40, 41]. Since it cannot reach the accuracy of adsorption simulations, it is nowadays used in a more qualitative manner to explain trends in reactivity [42].

As transition metals compose the vast majority of heterogeneous catalysts, the electrons highest in energy which dictate the reactivity usually occupy d-orbitals. In the periodic table section of transition metals, going

from scandium to zinc, from yttrium to cadmium and from lanthanum to mercury, 10 electrons subsequently fill 5 d-orbitals. Atomic orbitals are spatial representations of the states an electron or a pair of electrons can exist in around an atom. They can be used to calculate the probability of finding an electron at a certain point in space around an atom. When atoms come in close proximity to each other, their orbitals overlap and are forced to change their shapes, extending the reach of electrons to multiple atoms. If the resulting force is attractive, it creates bonds which can either be localized between two atoms or delocalized. In molecules, from two up to several tens of atoms share electrons with each other. In crystals the delocalization of electrons becomes so large that the mathematical description of orbitals transitions into infinite bands of electronic states, whereas the d-band forms through the overlap of atomic d-orbitals. The metallic bonds are extended throughout the whole crystal as electrons can move practically freely among nuclei. Nanoclusters are a particular case, in which molecular orbitals are still resolved. However, extending over the whole nanocluster, their energy levels are so dense that they constitute a quasi-continuum, hence can still be described as bands. It is important to keep in mind that the finite nature of the system changes the response of the d-band to an external perturbation (e.g. adsorption) than it would otherwise be expected in a crystalline material.

Let us consider the adsorption of hydrogen on a family of transition metal surfaces of the same termination. Hammer and Nørskov derived that the change in adsorption energy δE_d due to the interaction with the metal d-band can be approximated by freezing the electron density of the metal and the adsorbate [43]:

$$\delta E_d \approx -2(1-f)\frac{V^2}{\epsilon_d - \epsilon_H} + \beta V^2 \tag{2.12}$$

where ϵ_H is the energy of the hydrogen 1s-orbital prior to the interaction with the d-band and ϵ_d is the d-band center. The above relation uncovers trends in the periodic table of elements. Going from left to right the filling of the d-band f increases. As the d-band becomes filled by more than half, anti-bonding states get occupied which weaken the adsorption. The coupling matrix element V on the other hand increases from right to left as well as going down in the periodic table. It normally strengthens the adsorption, yet on metals with fully filled d-bands a higher V weakens it [40]. The above mentioned filling of the d-band and the coupling strength directly influence how the d-band energy-levels are distributed.

2.3 Electronic descriptors predict adsorption energy

The energy distribution of the d-band determines the chemical reactivity with substrates such as hydrogen. Although equation 2.12 can make

useful adsorption energy predictions, it depends on f, V and ϵ_d. A single property correlating strongly with the adsorption energy would make the d-band theory more intuitive. Such a property can be termed an electronic descriptor, three of which shall be introduced shortly.

We have already encountered the first descriptor. Small changes in the d-band center ϵ_d correlate linearly with the adsorption energy [42, 44, 45]:

$$\delta E_d \approx \frac{V^2}{|\epsilon_d - \epsilon_H|^2} \delta\epsilon_d \tag{2.13}$$

This law also holds well for early transition metals, because an increased V counters the emptying of bonding metal d-states [30, 40, 41].

The second descriptor was devised since deviations occurred in alloys with almost filled d-orbitals (d^9 or d^{10}) where the peak of the hydrogen-surface anti-bonding state is near the Fermi-level [46, 40]. The deviations are better accounted for by the quantity $\epsilon_d^w = \epsilon_d + \frac{W_d}{2}$ which corrects the d-band center by half of the d-band width W_d [47].

Lastly, the maximum of the d-band Hilbert-transform ϵ_u as a direct measure of the d-band edge correlated best with adsorption energies on late transition metals [48]. Since ϵ_u describes the shape of the d-band at the Fermi-level it predicts the energy-level, and hence the filling of the hydrogen-surface anti-bonding state, well. The Hilbert-transform of the d-band has a theoretical foundation in the strong chemisorption limit of the Newns-Anderson model where the minimum and the maximum determine the positions of adsorbate-metal bonding and anti-bonding states [49, 50].

While the three electronic descriptors are intuitive, they are prone to information loss. The d-band theory in general does not account for geometric effects either. In the fifth chapter, we will revisit ϵ_d, ϵ_d^w and ϵ_u to consider the limitations of the d-band theory.

2.4 Paradigm shift in designing catalysts

As the materials discovery process evolved from empirical science towards theories to explain and predict experimental evidence, the field of heterogeneous catalysis developed its concepts (Sabatier and BEP) and theories (d-band, kinetic theories) alongside [22, 51]. When simulations started in the 1950's in the advent of computational technology, simulations of catalytic systems were challenging due to their complexity. Once numerical solutions of quantum mechanical systems (see section 3.1 Schrödinger equation) were implemented, they facilitated atomically precise computation of structures and their properties. Since quantum mechanical simulations required, in principle, only the atomic structure and no empirical parameters, they were coined *ab initio* simulations. Throughout the 1970's and 80's computational power exploded enabling *ab initio* simulations on flat surfaces which culminated in the 90's in systematic studies

of transition metals and adsorption thereon [41, 52, 53, 54, 55]. To date, surface reactions on complex systems with stepped edges, large supercells or co-catalyst interfaces are daunting but manageable, even molecular dynamics simulations at *ab initio* level are possible [56]. Such detailed studies make sense in particular cases in order to support experimental findings, e.g. uncovering the active site or the rate-limiting step. They take a supportive role guiding experiments to build on and improve known catalysts. Detailed studies however fall short in predicting a new catalyst as it would be like searching for a needle in a haystack. Screening of a large range of potentially new catalysts is at present not possible with a full kinetic and thermodynamic analysis.

As materials science turns toward big data, surface science and in particular heterogeneous catalysis needs effective screening methods to guide the search for potential catalysts [51, 57, 58, 59, 60, 61]. As introduced above, the concepts in catalysis allow us to correlate relatively cheap simulations such as the adsorption energies, with catalytic activity. Thanks to the d-band theory and electronic descriptors, it is even possible to make qualitative statements about the catalytic activity of a material without the simulation of a single surface-adsorbate interaction. Inexpensive properties correlating with catalytic activity are indispensable for screening of large databases. As data driven catalysis is just beginning and theoretical catalysis has not provided any major breakthrough in the last 20 years, empirical catalysis supported by detailed simulations is still the modus operandi [38].

Thanks to advances in analytical techniques from infrared spectroscopy in the 1950's to resolution of surfaces at atomic precision with the scanning tunnel-microscopy, experimental knowledge has always been ahead of theoretical knowledge [22]. This acknowledges that to date rational catalyst design relies first and foremost on experimental insight. Large-scale computational screening efforts may however emancipate the simulation side from a supporting and explaining role to a predicting and guiding role on the road towards new catalysts [35, 59, 62].

3. Descriptors

In the previous chapter, I presented a few properties which are cheaper to compute than and correlate with catalytic activity. We can refer to ϵ_d, ϵ_d^w, ϵ_u and ΔE_{ads} as one-dimensional descriptors which are intuitive and based on catalytic concepts. However, they compress information about the system into one quantity. Since machine learning (ML) generally requires multiple input features for accurate predictions, ML-descriptors should not consist of a single property. In materials science, a few methods were proposed to engineer descriptors with multiple features from information at the atomic level, for example electronegativity, electron affinity, ionization potential and atomic radii [63, 64, 65, 66, 67, 68]. Nevertheless, encoding structural information is key where small changes in structure matter. Although several structural properties such as bond lengths, bond angles and coordination numbers [69, 70, 71], correlate with ΔE_{ads}, they do not predict it well enough on their own. How are structural features transformed optimally for machine learning? This chapter provides an overview of common structural descriptors. In Publication I, I participated in testing several descriptors for machine learning while incorporating them in the python library DScribe. These descriptors only rely on types and positions of atoms in a system and can be fed to any machine learning model. Figure 3.1 illustrates the role of the descriptor in a workflow going from structure to predicting a property such as ΔE_{ads}. The generation of descriptors is an auxiliary step in machine learning, but determines the success or failure of it. The preparation of the input is just as important as the ML model itself. Since structural features are transformed into a different representation, it could also be regarded as an unsupervised learning method (see below section 3.3).

3.1 Bypassing the Schrödinger equation

Machine learning provides a shortcut from atomic structures to their properties, however, it requires reference data for training. The refer-

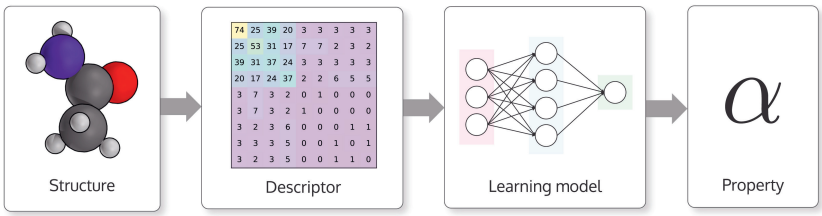


Figure 3.1. In order to predict a property with machine learning, the Cartesian coordinates and atom types of a structure are first transformed into a descriptor. It is a numerical fingerprint encoding relevant structural information. This transformation is computationally cheap and done for both training and test sets. The machine learning model then both trains and predicts with the descriptor features as inputs. Reproduced from Publication I.

ence is usually provided by *ab initio* simulations, which are based on the Schrödinger equation. Given the atom positions $\mathbf{R}$ and nuclear charges $\mathbf{Z}$, the Schrödinger equation yields the total energy E:

$$\hat{\mathbf{H}}(\mathbf{R}, \mathbf{Z})\Psi(\mathbf{R}, \mathbf{Z}) = E\Psi(\mathbf{R}, \mathbf{Z}) \tag{3.1}$$

As the nuclei are fixed in the Born-Oppenheimer approximation, the Hamiltonian $\hat{\mathbf{H}}$ and the wave function Ψ are functions of $\mathbf{R}$ and $\mathbf{Z}$. In practice, the Schrödinger equation is solved numerically with different degrees of approximation. For instance, coupled cluster theory with singlet, doublet and perturbed triplet excitations CCSD(T) often referred to as the *ab initio* gold standard generally reaches chemical acccuracy of about 0.04 eV [72, 73, 74]. Since CCSD(T) is currently unfeasible for large numbers of nanocluster simulations, accuracy needs to be traded off against computation speed. Density Functional Theory (DFT) is a popular tool among computational materials scientists which is fast and its accuracy is referred to as about 0.1 eV [75]. A short introduction is given in the next section.

Assuming the property of interest, in particular ΔE_{ads}, as defined by a set of total energies E (see 2.11), is simulated accurately enough to draw meaningful conclusions, then machine learning can supplement simulation data at most at the accuracy of the reference simulations. Since the computational cost of machine learning is negligible compared to *ab initio*, it is a powerful tool to enhance the data manifold.

In summary, solving the Schrödinger equation and machine learning are not two competing ways to arrive from structure to property, but complementary. The *ab initio* simulations provide the property for training the machine learning model. Machine learning on the other hand is a computationally cheap surrogate model increasing the data set size at an accuracy loss. Hence, the Schrödinger equation can only partly be bypassed.

3.2 Density Functional Theory

Hohenberg and Kohn formulated an alternative path to the Schrödinger equation for *ab initio* simulations. They proved that the ground state electron density n_0 uniquely determines the external potential V_{ext} of the atomic nuclei and that a density functional exists for the energy E which minimizes E at n_0 [76].

The Hohenberg-Kohn theorem states that such a functional exists, yet does not disclose its form. After more than 50 years, the exact functional is still unknown. In need of approximations to the exact solution, Kohn and Sham introduced a fictitious system of non-interacting electrons [77]. The Kohn-Sham approach is a reformulation of the problem, allowing one to make practical approximations to the density functional. At first, equation 3.1 is replaced by the single-electron Kohn-Sham equation:

$$h_{KS}\psi_i = \epsilon_i\psi \tag{3.2}$$

where h_{KS} is the Kohn-Sham Hamiltonian. The electrons in this picture are considered non-interacting. The hard-to-solve problem of electrons interacting with each other, equivalent with not knowing the exact density functional, is forwarded to the exchange-correlation functional E_{xc}.

$$E = T + E_H + \int n(r)\nu_{ext}(r)dr + E_{xc} \tag{3.3}$$

where the kinetic energy T, the Hartree energy E_H and the external potential $\nu_{ext}(r)$ are all non-interacting and reappear as their single-electron analoga (lower-case letters) in the Kohn-Sham Hamiltonian:

$$h_{KS} = t_{KS} + \nu_H + \nu_{ext} + \nu_{xc} \tag{3.4}$$

Since both $\nu_{xc}[n(r)]$ (of unknown form) and $\nu_H = \int \frac{n(r')}{|r-r'|}dr'$ depend on the electron density

$$n(r) = \sum_{i}^{occ} |\psi(r)|^2 \tag{3.5}$$

we are posed with a problem of several unknown variables. Supposed we found the form of the functional $\nu_{xc}[n(r)]$ or a sufficient approximation to it, the set of equations 3.2, 3.4 and 3.5 can be solved iteratively. Given an initial guess of the electron density, the KS-Hamiltonian (3.4) is computed. Next, the KS-equation (3.2) yields the wave function which in turn can be used to calculate another electron density (3.5). With every iteration, the electron density gets closer to its ground state. The loop is stopped once n and h_{KS} are self-consistent, meaning the difference in change between two consecutive steps is negligible. This indicates that the ground state n_0 is reached.

As with the Hohenberg-Kohn functional, the exact exchange-correlation functional is unknown. However, there are approximations to the exchange-correlation functional which work well in practice. For our problems in heterogeneous catalysis, we are concerned that the functional needs to be sufficiently flexible to describe both surface and adsorbate. The local density approximation (LDA) for instance is not good enough at estimating bonding energies and distances [78, 79]. It systematically overestimates the former and underestimates the latter because of the strong assumption of a homogeneous local electron density. By allowing density gradients as in generalized-gradient approximation (GGA), the overbinding in LDA is corrected and agreement with experiment is improved [80]. The investigated nanoclusters mostly have metallic behaviour which needs to be mirrored. This is the case for the functional by Perdew–Burke-Ernzerhof (PBE) as shown in this study on transition metals [79]. The long-range surface-adsorbate interaction can be taken into account by van-der-Waals correction. If the only property of interest is adsorption energies DFT-accuracy is sufficient. However, if one wants to derive electronic descriptors e.g. band structure, one needs to be careful. LDA as well as GGA tend to underestimate the band gap significantly.

Density Functional Theory (DFT) offers a good trade-off between speed and accuracy. As implemented in CP2K, the computation speed scales effectively linearly with respect to system size [81, 82, 83] which makes large-scale calculations of nanoclusters feasible. DFT calculations in Publications II and III were carried out with the CP2K software using the PBE-functional [84]. The double-ζ valence plus polarization MOLOPT-SR-DZVP acted as the basis set [85], including norm-conserving Goedecker-Teter-Hutter (GTH) pseudopotentials [86, 87, 88]. Grimme's DFT-D3 dispersion model with Becke-Johnson damping corrected for Van-der-Waals interactions [89, 90]. The DFT calculations in Publications II and III are part of this dissertation.

3.3 Machine learning and the need for structural descriptors

The field of machine learning can be divided into two subfields, unsupervised and supervised learning [91]. As unsupervised learning is devoid of labelled data, it transforms a feature vector x into a different representation for various purposes. Although, there is no correct representation, it is helpful to detect patterns in the data. For instance, the dimension of x can be reduced with methods such as linear discriminant analysis, principle component analysis or T-distributed stochastic neighbour-embedding. It is helpful for visualizing the data distribution. Otherwise, unsupervised learning can detect clustering of data as well as outliers and it can assert that data are identically distributed.

Supervised learning on the other hand relies on the labelling of data by a property y^{true}. A machine learning model is an arbitrary function $f_\Theta(\mathbf{x}) = y^{pred}$ mapping a feature vector x to a property y. In this dissertation, x is usually a structural descriptor and y is the adsorption energy. Since the choice of parameters Θ determine how well y^{pred} predicts y^{true}, the task in supervised learning revolves around finding the optimal parameters to reduce the expected prediction error. It becomes a minimization task with respect to Θ of the expected loss:

$$E_\Theta L(\mathbf{y^{true}}, \mathbf{y^{pred}}) = \frac{1}{N} \sum_i |y_i^{true} - y_i^{pred}| \tag{3.6}$$

where N are the number of labelled datapoints. Here the loss function is the mean absolute error (MAE), but there are other common loss functions such as the mean squared error or the Kullback-Leibler divergence.

Care should be taken as to what data is used for minimization or fitting. The expected loss states the error on the already seen data, but does not tell how well the machine learning model generalizes on unseen data. For that purpose, all available labelled data is split into a training and a test set. While f_Θ is optimized, it is usually observed that the training error drops, the more complex the model function is. After all, there are no limitations to f and it can have any number of parameters. At the latest once f_Θ perfectly fits all training points, the data is over-fitted and generally will not predict unseen data accurately. The model will also fluctuate strongly with respect to what training data it sees, introducing a high variance. This problem can be countered by reducing the complexity of the model. An excessive reduction in model complexity, however, can lead to under-fitting of the data. As an example, the function $f = 0$ without parameters introduces a high bias. Eventually, the complexity of the model should reflect the complexity of the underlying feature-property relation. How do you find the optimal bias-variance trade-off, a model which is neither under-complex (high bias) nor over-complex (high variance)? To that end, a part of the training set is put aside for validation. The parameters are then optimized on the reduced training set and the model complexity is chosen based on the performance of the validation set. In practice, parameters not optimizable by the loss function can tune model complexity in a family of model functions. They are called hyper-parameters, those can also include parameters for the generation of structural descriptors. A robust validation method is k-fold cross-validation. The training set is first divided into k equal-sized partitions. The model is then validated on every partition separately and the validation errors are averaged.

Once and only when the best model is found, it is finally tested on the test set in order to get an estimate on the generalization error (expected error on unseen data). If the step of validation is omitted and the test set is taken for that purpose or part of the test set was seen at any time during the

process of optimization, the generalization error will be underestimated. This specifically includes the choice of descriptor and machine learning model.

Let us consider the input features for machine learning. One might be tempted to feed raw structure data as input, such as Cartesian coordinates (list for each atom) or Z-matrix (list of distances and angles) form. However, the same structure can be expressed in many ways, for instance the list ordering can be changed at will (permutation) and the structure can be translated and rotated in space. This at first does not seem problematic, but results in similar structures having potentially very different representations. Since machine learning algorithms need to efficiently interpolate between data, they require a compact and continuous feature space [92]. Learning from Cartesian coordinates is inefficient since besides the underlying target correlation, rotational, translational and permutational invariances have to be learned as well. This demonstrates the need for a structural descriptor which should fulfil the following criteria [92, 93]:

- compact - redundant features should be minimised. Especially the descriptor should be invariant with respect to rotation, translation and permutation of atom indexing.
- unique - there should be one unambiguous way to construct a descriptor for any given structure
- non-degenerate - structures with different relevant properties do not have identical descriptor features. If chiral molecules collapse into the same representation it can be neglected for most properties. Otherwise, systems with identical representations, but different properties introduce noise. This might be the case if the descriptor is local, but there are non-negligible long-range effects.
- continuous in the spanned feature space - similar structures (e.g. two adsorption sites on a symmetric surface) yield similar representations. Machine learning models in general assume smoothness in feature space.
- general - any imaginable combination and arrangement of atoms in space should be representable including small and large molecules, clusters and periodic systems
- fast - The overall goal is to save computational time. Hence the time needed in order to compute the descriptor should be much lower than to compute the property of interest with DFT.

The DScribe package contains general unit tests to check for rotational, translational and permutational invariances for all descriptors. The other criteria are harder to test and are currently not covered.

3.4 Descriptors in DScribe

In the past years, many structure representations for machine learning in materials science have emerged [94, 95, 96, 97, 98, 99, 100]. Several common descriptors are currently implemented in the Dscribe package: Coulomb matrix [101], Ewald sum matrix [93], Sine matrix [93], Many-body tensor representation (MBTR) [92], local MBTR, Atom-centered symmetry functions (ACSF) [102] and Smooth overlap of atomic positions (SOAP) [103]. Of those, four descriptors were tested on nanoclusters in Publication II: Coulomb matrix, ACSF, MBTR, and SOAP.

The descriptors are introduced briefly below. A descriptor can be global, encompassing the whole structure, or local, centered around a position in space, e.g. an adsorption site.

The Coulomb matrix is a global descriptor which contains the pairwise coulomb repulsion of the atomic nuclei in a symmetric matrix:

$$C_{ij} = \begin{cases} \frac{1}{2} Z_i^{2.4}, & \text{for } i = j \\ \frac{Z_i Z_j}{|\mathbf{R_i} - \mathbf{R_j}|}, & \text{for } i \neq j \end{cases} \tag{3.7}$$

where R_i is the position and Z_i is the nuclear charge of the atom i. The diagonal elements encode the nuclear charge or atomic type and are a polynomial fit to the potential energy of atoms.

The Ewald sum matrix and Sine matrix are two possible periodic extensions of the Coulomb matrix. In the former, the Ewald technique is used to sum up the infinite electrostatic interactions. This sum converges only with a neutralizing background. In the latter, the interaction terms inspired by the coulomb repulsion are replaced by a sine function which reduces the computational cost of the descriptor.

ACSF as a local descriptor is a set of symmetry functions encoding distances and angles of a center atom i to neighbouring atoms. The local environment within the radius R_c is represented by pairwise and triplet atomic interactions of atoms therein. The cutoff function $f_c(R_{ij})$ attributes less weight to neighbour atoms j further away from the center.

$$f_c(R_{ij}) = \begin{cases} 0.5[\cos(\frac{\pi R_{ij}}{Rc}) + 1] & \text{for } R_{ij} \leq R_c \\ 0 & otherwise \end{cases} \tag{3.8}$$

The pairwise symmetry functions G_i^1, G_i^2 and G_i^3 encode the radial distribution around atom i:

$$G_i^1 = \sum_j f_c(R_{ij}) \tag{3.9}$$

$$G_i^2 = \sum_j e^{-\eta(R_{ij} - R_s)} \cdot f_c(R_{ij}) \tag{3.10}$$

$$G_i^3 = \sum_j \cos(\kappa R_{ij}) \cdot f_c(R_{ij}) \tag{3.11}$$

where η, R_s and κ are hyper-parameters. The triplet symmetry functions G_i^4 and G_i^5 encode angular distributions θ_{ijk} of pairs of neighbour atoms j and k.

$$\begin{aligned} G_i^4 =& 2^{1-\zeta} \sum_{j,k\neq i}^{all} (1 + \lambda \cos\theta_{ijk})^\zeta \cdot e^{-\eta(R_{ij}^2 + R_{ik}^2 + R_{jk}^2)} \\ & \cdot f_c(R_{ij}) \cdot f_c(R_{ik}) \cdot f_c(R_{jk}) \end{aligned} \tag{3.12}$$

$$\begin{aligned} G_i^5 =& 2^{1-\zeta} \sum_{j,k\neq i}^{all} (1 + \lambda \cos\theta_{ijk})^\zeta \cdot e^{-\eta(R_{ij}^2 + R_{ik}^2)} \\ & \cdot f_c(R_{ij}) \cdot f_c(R_{ik}) \end{aligned} \tag{3.13}$$

where ζ and λ are another set of hyper-parameters. In practice, the descriptor is composed of several symmetry functions of each type with different sets of parameters, in order to enhance the representation of the local environment.

MBTR is a global descriptor which divides all combinations of the relative orientations of k atoms to each other by atomic type. The many-body factor k usually goes up to 3, but can go higher. The resulting features are arranged in a tensor. The k-body terms are metrics such as (inverse) distances and angles. The number of occurrences are counted and then distributed on a 1D-grid with a grid point x:

$$f_k(x, z) = \sum_I^{N_a} w_k(I) \mathcal{D}(x, g_k(I)) \tag{3.14}$$

with the index tuple $I = (i_1, \ldots, i_k) \in \{1, \ldots, N_a\}^k$. The summation is done separately for tuples of atoms N_a belonging to a set of atomic types. $\mathcal{D}$ is a normal distribution, w_k is a weighting function attributing less weight to atoms further away from each other, and g_k is a spatial function with a metric unique for every k. The resulting $(k+1)$-dimensional tensors are concatenated in practice, forming a size-consistent descriptor.

Since SOAP was the best-performing descriptor in Publication II and was also utilized in Publication III, I will describe it in more detail here. The local descriptor SOAP represents the density of gaussian-distributed atomic positions in space by coefficients of orthonormal basis functions [103, 104]. It describes the local environment around an adsorption site up to a certain radial cutoff R_c. Not only can adsorption sites among different nanoclusters be compared, but also the similarity of whole structures can be assessed by matching all local-environment pairs across structures [104].

The features at a specific configuration ξ around a chemical environment x are computed by the powerspectrum $\mathbf{P}$:

$$\mathbf{P}^{\xi}(\mathbf{x}) = p^{\alpha\beta\xi}_{nn'l}(\mathbf{x}) = \sum_{m} c^{\alpha\xi}_{nlm}(\mathbf{x}) c^{*\beta\xi}_{n'lm}(\mathbf{x}) \tag{3.15}$$

All unique combinations of atom species α and β are concatenated to give the full descriptor. The coefficients c represent the atomic density decomposed by a set of spherical harmonics Y_{lm} and a set of radial basis functions g_{nl}.

$$\rho^{\alpha\xi}(\mathbf{r}) = \sum_{nlm} c^{\alpha\xi}_{nlm}(\mathbf{x}) g_{nl}(r) Y_{lm}(\Omega) \tag{3.16}$$

The coefficients are solved by:

$$c^{\alpha\xi}_{nlm}(\mathbf{x}) = \int_0^{R_c} dr\, r^2 \int_S d\Omega\, g_{nl}(r) Y_{lm}(\Omega) \rho(r, \Omega) \tag{3.17}$$

The atomic density ρ is given by a gaussian-smoothed distribution:

$$\rho^{\alpha\xi}(\mathbf{r}) = \sum_{i} e^{-(\mathbf{r} - \mathbf{r}_i^{\alpha\xi})^2/2}, \tag{3.18}$$

where $\mathbf{r}_i$ is the position of atom i in the structure. In practice, a maximum radial basis value $n = n'$ and a maximum angular value l are chosen so that the machine learning accuracy is converged.

3.5 Impact of library of descriptors

The DScribe package is available as open-source software online [105] along with documentation [106]. The user only communicates with a python interface, whereas some heavy descriptor evaluations are written in C or C++. Continuous integration ensures that current features do not break while new features are added.

Machine learning in materials science has become popular in the last decade. At first, following the workflow in figure 3.1, a combination of descriptor and machine learning model must be chosen. It is not trivial to know the best descriptor for a machine learning model in advance, which means in practice that several options should be tested [100]. Several machine learning frameworks in materials science have incorporated both steps [107, 108, 109, 110]. In contrast, to facilitate testing of descriptors against any machine learning model we chose to decouple them and focus only on the descriptor module in our package, as it is the key first step to most machine learning tasks. This deliberately excludes methods where the machine learning is entwined with the structure representation step such as graph neural networks (GNN) [100, 111, 112, 113, 114]. A GNN mimics the structure incorporating the descriptor step into the shape of the neural network.

DScribe brings several advantages to computational materials science: it can be used in combination with both unsupervised and supervised learning methods. Descriptors can be easily exchanged for the purpose of feature importance weighting. For instance, in order to explain which part of a structure is responsible for a property, searching for distinctive features can be easier in some descriptors. And lastly, it has an easy-to-use python interface consistent throughout all implemented descriptors, so that benchmarking and switching between descriptors require only few code changes.

As computational materials science becomes data-driven, many research areas might start to use machine learning as a tool. Since easy-to-use machine learning packages such as scikit-learn already exist [115], DScribe lowers the threshold for new researchers by providing off-the-shelf descriptors. Since some physical properties are determined locally such as the adsorption energy and some are defined by the whole structure such as atomization energy, DScribe supports local and global descriptors. As descriptors in materials science are an active research field [92, 93, 100, 101, 102, 104], the library is open to external contributions.

The DScribe package in its initial state, although not fully launched, was tested extensively and used to facilitate machine learning on nanocluster adsorbates in Publication II. Later on, it was a key ingredient to the workflow automation in Publication III. Apart from that, to the date of writing, DScribe has been applied to several materials science works [116, 117, 118, 119, 120].

4. Machine learning screening adsorption on nanoclusters

The elements and structures of adsorption sites determine their adsorption energy ΔE_{ads} (see equation 2.11) which in turn is a descriptor for catalytic activity. The adsorption energy of a single site is relatively cheap to compute by DFT but since several diverse sites populate a nanocluster, screening all of them for many nanoclusters quickly becomes too expensive. Especially on distorted, irregular surfaces, the adsorption sites can be similar but hard to classify with heuristic methods. This chapter explores how machine learning can interpolate ΔE_{ads} both on a single cluster and on multiple clusters simultaneously and which descriptor does the job most efficiently.

Machine learning has recently gotten traction in heterogeneous catalyst research [70, 111, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130], yet studies on nanoparticles and nanoclusters are still scarce [69, 71, 131, 132, 133, 134]. As was mentioned in the last chapter, it is important to find the best combination of descriptor and machine learning model. Faber et al compared the accuracy of several machine learning models [100]: gated graphs, graph convolutions, kernel ridge and random forest regression. As both graph convolutions and gated graphs are types of graph neural networks, they could not be combined with descriptors in chapter 3. In Publication II, I tested both kernel ridge and random forest regression, whereas the latter can be found in the supporting information [135].

I chose kernel ridge regression (KRR) as the most suitable machine learning model for the following reasons. First, KRR outperformed random forest regression by a large margin [135]. Secondly, in the study of Faber et al, it reached good accuracies throughout the tested properties and was sometimes the best model [100]. And lastly, it was the ML model of choice for the inventors of two descriptors: MBTR and SOAP [92, 103, 104].

4.1 Kernel ridge regression

In conventional regression tasks such as neural network [136] or polynomial regression, sets of parameters are trained or fitted on labelled data. The trained machine learning model neither has memory of, nor requires the original data to make predictions on the new data. By contrast, KRR is a memory-based ML method which encodes training data in a symmetric kernel matrix $\mathbf{K}$.

$$\mathbf{K} = \begin{bmatrix} K(\mathbf{x_1}, \mathbf{x_1}) & \dots & K(\mathbf{x_1}, \mathbf{x_N}) \\ \vdots & \ddots & \\ K(\mathbf{x_N}, \mathbf{x_1}) & & K(\mathbf{x_N}, \mathbf{x_N}) \end{bmatrix}, \tag{4.1}$$

The size of $\mathbf{K}$ is equal to the number of training points N. Each entry is a similarity metric $K(\mathbf{x_i}, \mathbf{x_j})$ of two feature vectors $\mathbf{x_i}$ and $\mathbf{x_j}$. The similarity metric has to be a symmetric positive semi-definite kernel function, for example the radial basis function:

$$K(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|_2) \tag{4.2}$$

where γ is a hyper-parameter optimizable by cross-validation. In order to predict a property y of a new data point $\mathbf{x_{pred}}$, the matrix needs to be inverted. Inversion is the speed bottle neck of KRR as the computational cost scales cubically with the amount of training data. Once the matrix is inverted, y is given by

$$y(\mathbf{x_{pred}}) = \mathbf{k_{pred}}^T (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y_{train}}, \tag{4.3}$$

where the regularization parameter λ is another hyper-parameter and $\mathbf{y_{train}}$ is the property vector of the training set. As λ controls the model-complexity it is also optimized during cross-validation. In order to compute the kernel vector $\mathbf{k_{pred}}$, the training data is required once more:

$$\mathbf{k_{pred}} = \begin{bmatrix} K(\mathbf{x_{pred}}, \mathbf{x_1}) \\ \vdots \\ K(\mathbf{x_{pred}}, \mathbf{x_N}) \end{bmatrix} \tag{4.4}$$

KRR is best-suited for medium-sized datasets; other ML methods would be more suitable in terms of computation time for datasets larger than 10000 points. As KRR transforms the feature space into an N-dimensional space of datapoint relations, it is only capable to interpolate and not extrapolate.

4.2 Benchmark

Finding the best ML representation for an adsorbate-surface interaction requires benchmarking several descriptors on diverse test systems. The

adsorption energy might be fully local or there might be global effects, too. The accuracies with respect to ΔE_{ads} of the structural descriptors Smooth Overlap of Atomic Positions, Many-Body Tensor Representation, Atom-Centered Symmetry Functions and Coulomb Matrix were compared against each other. The 2D-material molybdenum disulphide (MoS_2) is a promising catalyst for the HER as it forms nanoclusters of triangular and hexagonal shapes and can be terminated with different amounts of sulphur on its edges [137, 138, 139]. The alloy gold-copper (AuCu) is exemplary for bimetallic alloys where atoms can be arranged in a multitude of configurations. The two test systems MoS_2 and AuCu covered an array of structural patterns representative of other nanoclusters, such as symmetry and near-symmetry of adsorption sites on the one hand, but distorted asymmetric surfaces on the other hand, sharp and smooth edges, as well as several element compositions at the edges. Figure 4.1 shows learning curves (error against training set size) for 4 different benchmark systems, whereas the descriptor and kernel hyper-parameters were optimized via cross-validation. Figures 4.1a and 4.1b show prediction accuracies of the potential energy surfaces (PES) of single MoS_2 and AuCu clusters.

PES in general describe the energy of a system, with respect to atomic positions. The nanocluster was kept rigid, moving only the hydrogen atom, resulting in a 3-dimensional PES. For small molecular adsorbates, the PES has 3 more dimensions because the molecules rotate relative to the surface. For molecules with internal degrees of freedom, the amount of dimensions quickly explodes.

In figures 4.1a and 4.1b, with only 200 to 300 training points, the PES of both single MoS_2 and AuCu clusters could be reconstructed. At that point, an MAE of 0.1 eV (DFT-accuracy) could be reached by all descriptors except for the Coulomb Matrix. The learning rates were similar, however SOAP was most accurate at larger training set sizes. The error could further be reduced by instead of picking training points randomly from the surface, they were ordered by their similarity. The most distinct points were picked first, points identical to any other previous point were ordered last. The algorithm farthest point sampling (FPS) starts with a random first point, then iteratively determines the point that maximizes the minimum distance to all previous points [140]. FPS can speed up ML especially for highly symmetric structures like MoS_2 (D_{3h}) even if the symmetry is not perfect. FPS harmonizes a training set which otherwise would not be identically distributed in feature space, reinstating an important machine learning assumption [91].

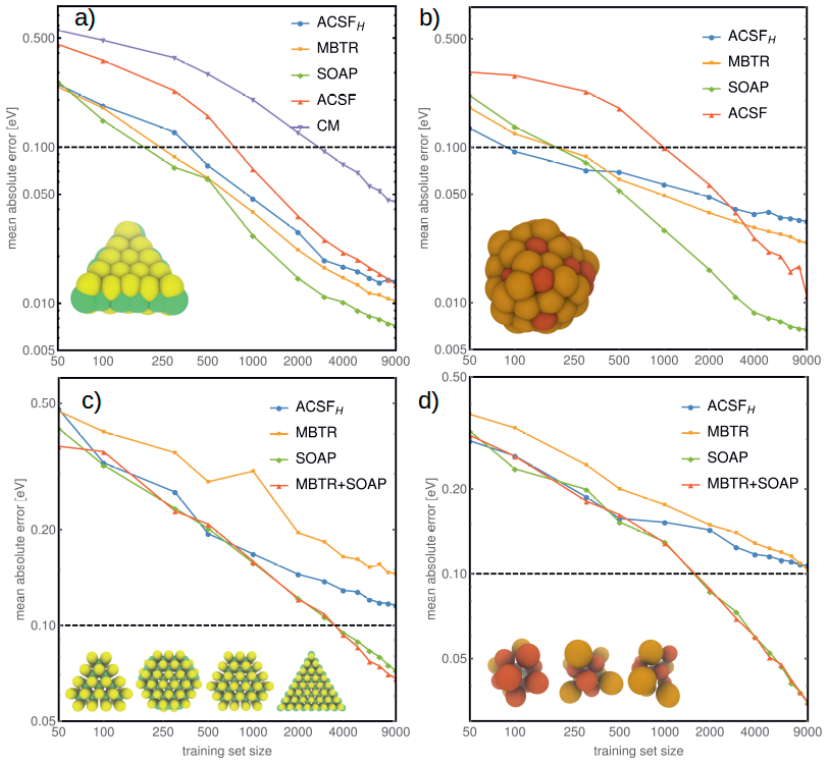


Figure 4.1. Learning curves with mean absolute error (MAE) plotted against training set sizes. The benchmark shows the accuracy of the descriptors CM, SOAP, MBTR, ACSF and ACSF$_H$ in KRR to predict adsorption energies. ACSF$_H$ are symmetry functions centered around hydrogen. Figures a) and b) show results on datasets of single MoS_2 and AuCu clusters. Figures c) and d) show the outcome when the model was trained on several MoS_2 and AuCu cluster structures simultaneously. Reproduced from Publication II.

4.3 Simultaneous screening of nanocluster PES

Machine learning the PES cannot only be done on a single nanocluster at a time, but on multiple clusters simultaneously. Figures 4.1c and 4.1d show results of 91 MoS_2 and 24 AuCu clusters, respectively. The descriptor SOAP reached DFT accuracy whereas MBTR and ACSF did not, even at a training set size of 9000. Predicting the PES in batches has synergistic effects, since the required training points per cluster are reduced significantly.

As MBTR is a global descriptor, it contributes new information to the local descriptor SOAP. Although the nanoclusters of MoS_2 were of different size and shapes, the added information turned out to be irrelevant for ML. Testing the combined descriptor SOAP+MBTR (red curve in figures 4.1c and 4.1d) revealed that it did not surpass the accuracy reached with SOAP. It could be concluded that global information is not required and differences in adsorption energies on different nanoclusters could be explained by the local environment of the adsorbate. This does not mean that there are no long-range effects like the size of the nanocluster, but that they can be learned locally from small changes in bond distances and angles.

4.4 Interpolation and extrapolation

Figure 4.2a shows a parity plot of the predicted PES against the calculated PES of a single MoS_2 cluster. The sampling is not so dense in the high- and low-energy regions. Since the adsorption sites are the positions of the local minima on the PES, only the meta-stable states constitute the adsorption energy distribution. Figure 4.2b illustrates the accuracy of predicted adsorption site energies of multiple MoS_2 clusters.

The MAE was slightly higher at 0.13 eV, possibly due to less sampling in the low-energy region. Yet, the predicted adsorption energy distribution agrees well with the calculated distribution. In summary, the combination of SOAP and KRR allows for an accurate prediction of adsorption energies along with the position of the hydrogen adsorbate. In order to demonstrate the limits of this method, the training data of small AuCu clusters (13 atoms) was used to predict the PES of a larger $Au_{40}Cu_{40}$ cluster. The parity plot is shown in figure 4.3. The agreement was not good, in particular a large part of the high-energy region was underestimated. This indicated that some local environments were not represented in the training set, hence KRR had to extrapolate far in feature space. This highlights the importance of picking training points in view of the prediction targets. For efficient ML, the range of nanoclusters should be determined before gathering any training data. The presented approach proved successful for diverse sets of nanoclusters, yet it revealed two disadvantages. Scanning

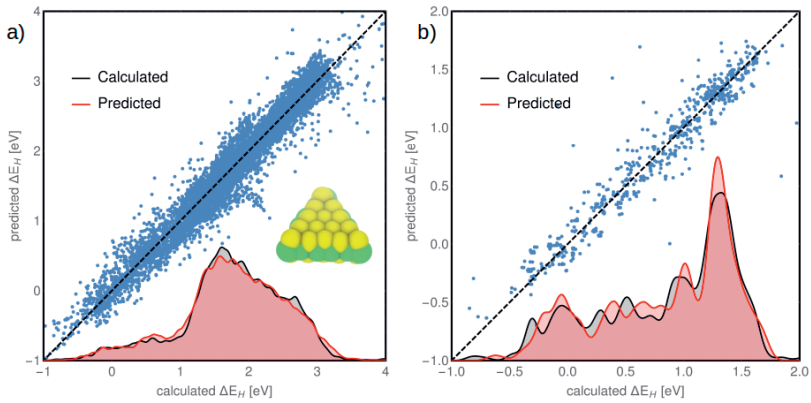


Figure 4.2. Parity plots of predicted against calculated adsorption energies showcasing the predictive power of scarcely sampled datapoints on multiple MoS_2 clusters (training set). a) The test set consists of a dense potential energy scan on a single MoS_2 cluster. b) The test set consists of local minima on several frozen MoS_2 clusters. The histogram of predicted (red) and calculated (black) energy distributions agree very well. The outliers of the scatter plots tended to be stronger in figure b). Reproduced from Publication II.

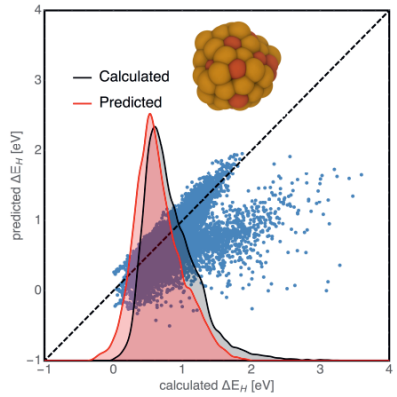


Figure 4.3. Parity plot of predicted against calculated adsorption energies showing the limitations of extrapolation. The training set consisted of small AuCu clusters, but was tested on a several times larger $Au_{40}Cu_{40}$ cluster. The histogram of the calculated (black) energy distribution is shifted slightly towards higher energies compared to the predicted (red) histogram. The scatter plot features 2 groups of datapoints indicating that some sites were not properly represented in the training set. Reproduced from Publication II.

the full PES becomes more expensive with molecular adsorbates and impractical with larger molecules. Furthermore, the only part of the PES, more precisely the low-energy region of the local minima, eventually matters to predict the energy of adsorption sites. The results show room for improvement in the selection of training points.

5. Challenges on automation of screening nanocluster surface interaction

The search space for nanocluster catalysts is daunting due to the fact that the element composition and size are practically free variables. In addition, it is also important to find stable shapes and configurations. In chapter 4 I established that machine learning aids the screening of nanoclusters as it makes the computation of ΔE_{ads} cheaper. The descriptor SOAP and machine learning model KRR proved to work in conjunction with each other. Nevertheless, several challenges remain in efficiently exploring the large search space. Publication III suggests an automated screening method which is capable of handling the four search dimensions: composition, configuration, size, and shape. It was tested on bimetallic combinations of 6 elements at different compositions and configurations, keeping size and shape constant due to computational limitations. This chapter is an overview of the challenges involved in automating a workflow from the generation of nanocluster configurations to the prediction of ΔE_{ads}.

5.1 Nanoclusters

Although there are numerous nanocluster classes such as transition metal dichalcogenides (e.g MoS_2), I will introduce the class of nanoalloys, nanoclusters of more than one (transition) metal. It is hard to find the most stable structures of nanoalloys because of the number of possible isomers. Given a fixed shape and composition, a bimetallic nanocluster $A_{N_A}B_{N-N_A}$ has $\frac{N!}{N_A!(N-N_A)!}$ configurations, although some can be equivalent due to symmetry. The total number of structures of a bimetallic nanocluster of fixed shape but any composition is 2^N. Even for a small nanocluster $A_{13}B_{42}$ of 55 atoms, the number of configurations amount to $\frac{55!}{13!\ 42!} \approx 1.45\ 10^{12}$ and summing up the compositions of A_xB_{55-x} amounts to $2^{55} \approx 3.60\ 10^{16}$. Even with the reduction of configurations by symmetry, exhausting the search space in order to find the global minimum configuration is out of reach. In section 5.3 I introduce a practical screening approach.

Nanoclusters can have crystalline and non-crystalline shapes. While

crystalline clusters have closed-packed cores, the non-crystalline clusters such as icosahedra and decahedra have cores which are under strain. Instead, non-crystalline nanoclusters exhibit facets which minimize the surface energy. As the strain has a destabilizing effect and scales with the volume of the core, icosahedral shapes transition at increasing size into closed-packed shapes such as octahedra [141, 142]. Figure 5.1 illustrates a typical transition between shapes from icosahedra via decahedra and octahedra to Wulff-shapes. Icosahedra, decahedra and octahedra are

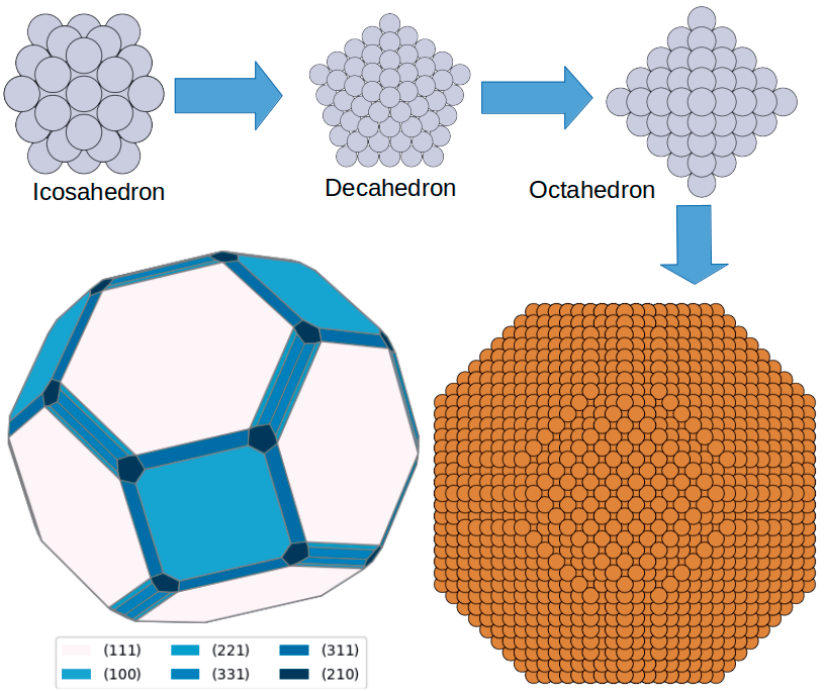


Figure 5.1. Illustration of an icosahedron, decahedron and octahedron, followed by a Wulff-shape construction of a copper cluster. The facets with Miller indices (111) and (221) dominate the surface. Facets of higher surface energy are small, however, they get formed once the nanocluster is big enough and can even dominate the catalytic activity.

terminated by facets with Miller indices (111). However, decahedra and octahedra usually have capped corners to attain a more spherical shape like icosahedra. The core of truncated octahedra is already closed-packed and is congruent with Wulff-clusters at small size. Wulff proposed that the distance h_i from the surface of a specific plane i to the center of the cluster is proportional to the surface energy γ_i [22].

$$h_i \propto \gamma_i \tag{5.1}$$

The Wulff-construction favors low-energy facets while the area of facets with higher surface energy are usually negligible at small nanocluster

sizes. Once the nanocluster grows into a nanoparticle however, they can dominate catalytic activity despite their small contribution to the total surface area [143]. It depends on the composition which cluster shape is most stable but there are methods to estimate cross-over sizes [141, 142].

Two metals in a nanoalloy can either mix or segregate internally [144]. The degree of mixing is determined by the bond-strengths between atoms of the same or different atomic type. An example of a segregated nanoalloy is a core-shell nanocluster where a core consisting of one element is surrounded by a shell of a second element. The driving force behind this phenomenon is the difference in surface energy, but also the difference in atomic radii plays a role [144]. For instance, in nanoclusters with a strained core such as icosahedra, the smaller atoms tend to occupy the core [145, 146]. In mixed nanoalloys, the elements can either be ordered or randomly arranged. In-situ effects often change the observable configurations. If the binding-energy to molecules at the interface is considerably stronger by one element, those atoms tend to segregate to the surface [144, 147].

The finite size of nanoalloys leads to a discretization of the electron bands. The confinement of the s-electrons of gold, for instance, is effective at room temperature up to a nanocluster size of roughly 2 nm [1]. At what size exactly the confinement effects start is hard to measure and most likely depends on the element composition [7]. Even without those effects, nanoparticles are usually more reactive than plain surfaces. This is due to the high surface-to-volume ratio. Since the nanoparticle shapes are determined by the Wulff-construction, simulations are usually carried out in parallel on periodic slabs of several terminations instead of explicitly simulating the whole nanoparticle at once.

5.2 Workflow automation

Although several high-throughput screening for the HER had been reported, they were carried out on periodic slabs, not on nanoclusters [57, 60, 125]. Consequently, it was necessary to devise a specialized workflow. Since there are several classes of promising catalysts such as single-atom alloys, high-entropy alloys and metal carbides, nitrides, phosphides and chalcogenides, a good workflow should be capable of handling various structure types [148, 149, 150, 151]. Though, for a practical screening instance it is important to limit the search to a manageable size.

At first, relevant elements should be selected. As the hydrogen evolution reaction is catalyzed by platinum in industry it provides a helpful reference. Furthermore, the elements iron, cobalt, nickel, copper and titanium form stable icosahedral nanoclusters [152] and have catalytic potential for the HER [39, 153, 154, 155]. Hence, the above 6 elements reduced the search complexity to only icosahedral nanoclusters. The search was

further constrained to a nanocluster size of 55 atoms since they are the smallest icosahedra with a core and do not distort significantly upon hydrogen adsorption. The choice of elements, shape and size can be regarded as input variables for the workflow sketched in figure 5.2.

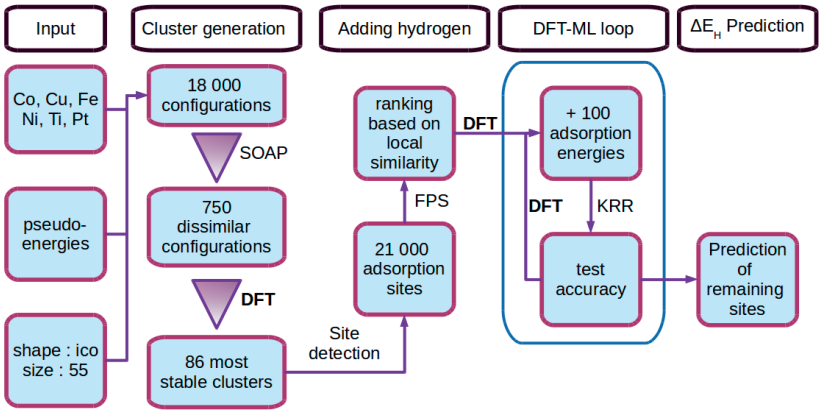


Figure 5.2. The sketch of the automated workflow shows steps from cluster generation and selection, adsorption site detection and ranking to the prediction of adsorption energies in a loop with DFT. Reproduced from Publication III.

The workflow automates the completion of 3 tasks: the nanocluster generation, the adsorption site detection and the subsequent submission of jobs in a DFT-ML loop. DFT jobs are submitted in batches while the machine learning accuracy is tested in each loop until the error is below a certain threshold. The whole workflow, especially the latter task requires an architecture to manage computation and storage of simulations. In materials science the most popular tools are *atomate* [156], *AiiDa* [157, 158] and *Fireworks* [159]. *Fireworks* is an open-source workflow manager with a python interface and a comprehensive documentation. It was developed and tested during the Materials Project [160]. For this project, I programmed the workflow in *Fireworks* and made it publicly available [161].

The first two tasks, the generation of nanoclusters and the detection of adsorption sites are exclusive to nanoclusters. The challenges to automate them are discussed below.

5.3 Efficient exploration of nanocluster configurations

Without prior knowledge of how atoms in a cluster are arranged, the configurations are most efficiently explored if they are picked maximally different from each other. The clusters were selected based on the global SOAP similarity metric [104]. For one exemplary composition out of 75,

figure 5.3 shows the 10 most dissimilar $Cu_{13}Co_{42}$ structures.

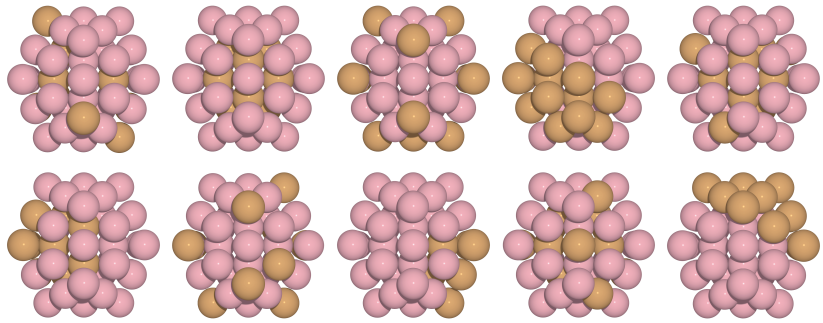


Figure 5.3. 10 example nanoclusters of the composition $Cu_{13}Co_{42}$. They were generated with the target of maximizing their dissimilarity. They feature a core-shell, an ordered and several segregated clusters. Reproduced from Publication III.

The configuration generation method uses Monte-Carlo (MC) with varying pseudo-energies to construct nanoclusters of a given shape.

$$E = \sum_{i}^{n} E_{c,i} * c_i + \sum_{i}^{n} E_{x,i,j} \tag{5.2}$$

where the pseudo-energies E_c and E_x represent core attraction and interaction between neighbouring atoms. Neighbouring atoms are defined by Voronoi-tesselation. The pseudo-energies are depicted in figure 5.4. All experimentally observable nanocluster classes emerge such as core-shell, ordered, segregated and randomly configured clusters [144]. Since finding the global minimum configuration is too computationally demanding, instead the most stable classes of nanoclusters are determined for every composition.

After the 10 most dissimilar structures for every composition had been optimized by DFT, their stabilities were compared to each other. The core-shell trends as well as the miscibility generally agreed well with the literature [145, 162]. The intensity of exploration can be increased at will and prior knowledge of miscibility and segregation can be reflected in the range of pseudo-energies for the MC algorithm. It is even possible to systematically approach the most stable configuration by Bayesian optimization of pseudo-energies [91, 163].

5.4 Nanocluster-adsorbate tools

The most common adsorption sites are top, bridge, 3-fold hollow and 4-fold hollow [22]. On regular periodic slabs they can be determined with heuristic methods. Since nanoclusters have edges and vertices, an objective definition of these sites is important. Due to significant distortions, it is

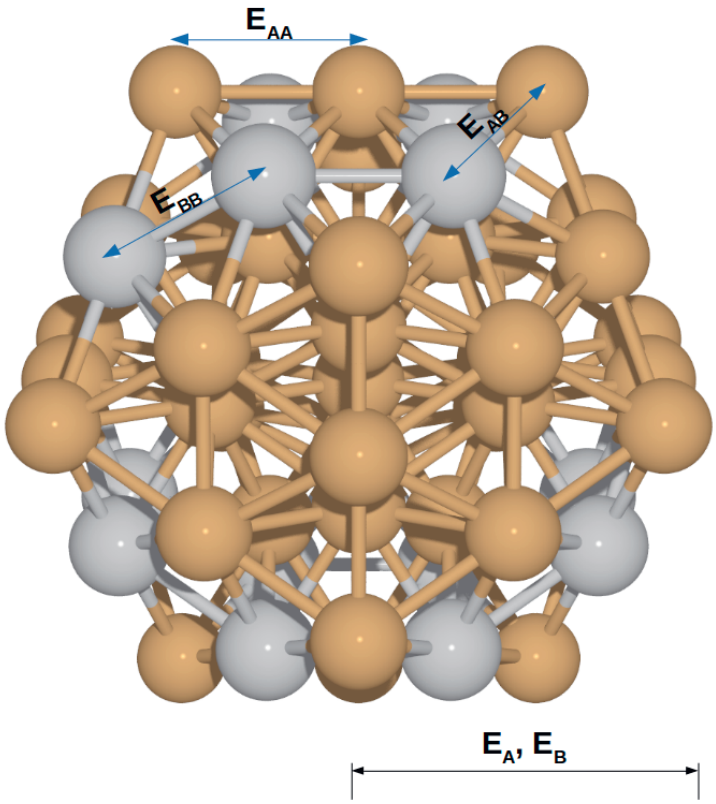


Figure 5.4. Pseudo-energies of an icosahedral bimetallic nanocluster. The core-pseudo energies E_A and E_B mimic surface segregation and the pseudo-energies E_{AA}, E_{AB} and E_{BB} characterize interaction between atoms which leads to phenomena of mixing or element segregation emerging.

sometimes hard to tell which atoms belong to the surface. The solution is an objective method for adsorption site detection and classification depicted in figure 5.5.

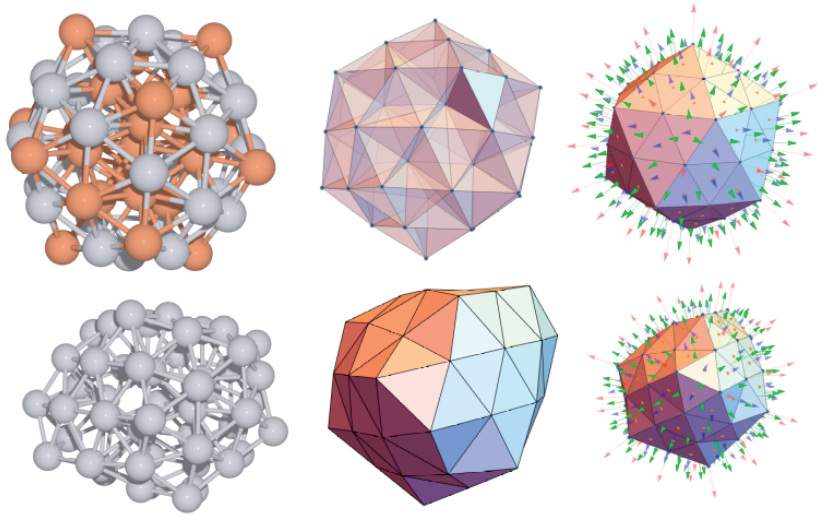


Figure 5.5. Illustration of the surface detection algorithm. The nanocluster is first divided into tetrahedra with the Delaunay algorithm. The outermost triangles then determine the surface atoms. Normal vectors of the triangles define the direction of top (red), bridge (green) and hollow (blue) sites. The algorithm works not only on highly symmetric clusters such as icosahedra (top), but also on distorted clusters (bottom). Reproduced from Publication III.

To demonstrate that the method works on any given shape, the example below shows a platinum nanocluster with a reduced core.

The surface detection algorithm proceeds as follows. The volume of the nanocluster is first tetrahedralized by the Delauny-algorithm. The open triangular faces then not only represent the surface, but also determine 3-fold hollow sites and, given their edges and vertices, subsequently bridge and top sites. The algorithm robustly arrives at uniquely-defined positions for adsorbates to bind on nanoclusters of arbitrary shapes and does not require visual checks which is paramount for large datasets. The tools to generate nanoclusters and detect adsorption sites are gathered in the python package Cluskit, along with other functionality regarding surface-adsorbate structures [164].

5.5 Machine learning accuracy

In chapter 4, the adsorption energies were inferred by predicting the whole PES and hence find the local minima. In Publication III, I chose a different, arguably more efficient route. The structures of training points on top,

bridge and hollow sites were optimized to find ΔE_{ads} corresponding to the closest local minima. The machine learning model was then trained on the local environments of the *initial* positions with labels of the *relaxed positions*. Thus, for new data, the adsorption energies were predicted skipping the process of finding local minima entirely. This alleviates the need to compute and predict high-energy parts of the PES. A similar approach had been reported in the literature before [165].

The adsorption sites on the most stable bimetallic clusters were ranked by farthest point sampling (FPS) as in chapter 4 to avoid computing redundant sites. The DFT-ML loop ran until an accuracy of 0.11 eV MAE was reached at a training set size of less than 10 % of the full dataset. Machine learning was in this case able to speed up the screening process by an order of magnitude. A detailed analysis of the errors revealed that there was still room for improvement in future studies. First, the learning rates differed depending on the elements constituting the adsorption sites. A Bayesian optimization approach which detects the slowest-learning element compositions could potentially increase the learning rate further [166]. Second, hydrogen tended to traverse from the initial guess to neighbouring adsorption sites, and that drift had a small effect (about 0.02 eV) on the predictive power of the model. A model which classifies initial guesses into stable and drifting adsorption sites could reduce the error even further.

Figure 5.6 summarizes adsorption energy distributions per composition.

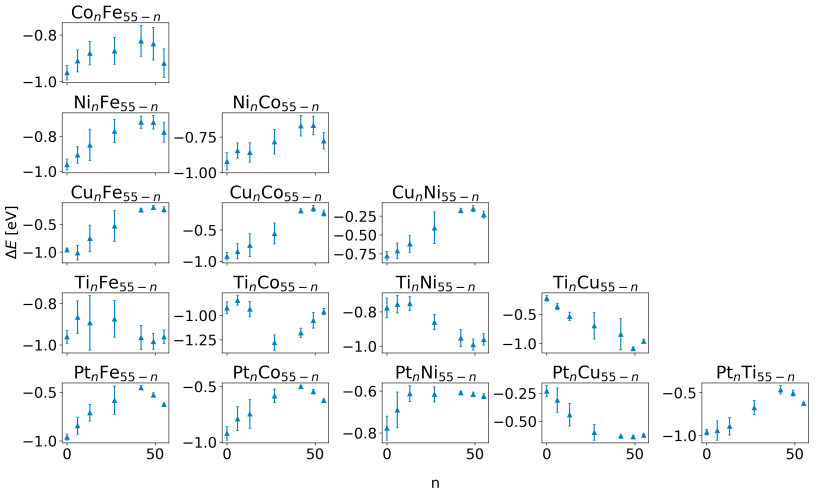


Figure 5.6. Hydrogen adsorption energy predicted per composition of bimetallic combinations of Fe, Co, Ni, Cu, Ti and Pt. The distributions are characterized by their means and standard deviations. Reproduced from Publication III.

Taking platinum as a reference point of a good catalyst, the ordering of adsorption energies of pure elements agreed well with experiments [167]. However, the agreement was poor compared to bimetallic catalytic

activities [168, 169]. The reason(s) for poor agreement can lie on both the computational and experimental side. On the one hand, experiments always introduce noise due to experimental conditions that are hard to control. Additionally, a consistent series of experiments is scarce. Lastly, the subjects of the experiments are usually nanoparticles of different sizes, in stark contrast to simulated nanoclusters of controlled size. On the other hand, simulations try to approximate reality with imperfect models and they also rely on theoretical assumptions to infer catalytic activity from ΔE_{ads}. It is a *descriptor* of catalytic activity and it neglects the equilibrium coverage and entropic contributions.

Comparing the adsorption energy distributions to other computational datasets of periodic slabs, it became apparent that they were shifted to lower energies [57, 170]. However, the shift was not constant. This highlights the importance of the structural model for capturing nanocluster effects.

5.6 Electronic descriptors evaluation on nanoclusters

The d-band center ϵ_d, the d-band center plus half the d-band width ϵ_d^w and the maximum of the d-band Hilbert-transform ϵ_u were introduced in chapter 2. They are descriptors of the local density of states of the d-band. They form a linear relation with ΔE_{ads} of top sites on periodic slabs [41, 47, 48]. Figure 5.7 summarizes the correlation of the electronic descriptors with nanocluster adsorption energies.

The correlation coefficient R indicates how good the linear relationship on nanoclusters is. While the correlation of ϵ_d and ϵ_d^w is weak, the correlation of ϵ_u is moderate. Constraining the adsorption energies to only those sites consisting of one element, the correlation improves slightly and is even stronger when restricted to only pure nanoclusters. As descriptors of a single quantity, ϵ_d, ϵ_u and ϵ_d^w are prone to information loss. Yet, the correlation of ϵ_u with ΔE_{ads} is strong enough that it could be used semi-quantitatively to pre-screen nanoclusters. It could filter out nanoclusters in ranges where the catalytic activity is expected to be low. This process has the potential to speed up the discovery of catalysts for several reactions at the same time, since it circumvents simulations of nanocluster-adsorbate interaction altogether. It is not possible to make predictions of catalytic activity with electronic descriptors but it is helpful in narrowing down the search space. Apart from that electronic descriptors could be used as a property to check the stability of nanocluster configurations.

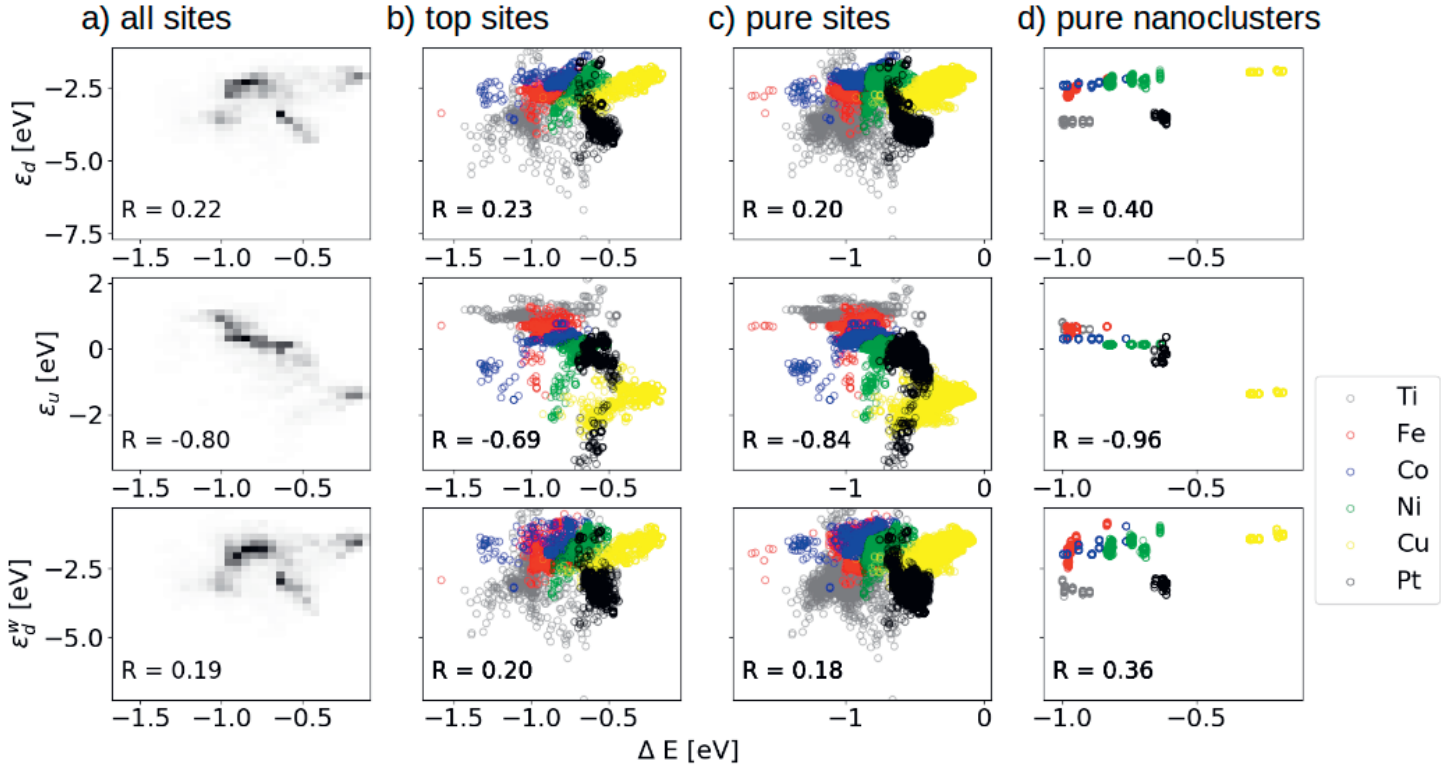


Figure 5.7. The correlation coefficient R measures how well electronic descriptors against adsorption energies fit a linear law. Investigated descriptors are ϵ_d, ϵ_u and ϵ_d^w in the rows from top to bottom. The adsorption sites have been restrained in the columns from left to right: a) all adsorption sites, b) all top sites, c) only pure adsorption sites (top, bridge and hollow sites made up of a single atomic type) and d) only adsorption sites from pure nanoclusters. The datapoints have been assigned to elements by colour where applicable. Reproduced from Publication III.

6. Conclusions and perspective

Industrial heterogeneous catalysts have improved steadily throughout a century, lead by experimental scientists and engineers. Catalytic activity is expensive to determine as it is an ensemble property of surface and adsorbates. That is reflected by the difficulty to simulate catalytic activity accurately. Rational Catalyst Design faces challenges to approximate complex real systems with means simple enough for computational screening methods. In this dissertation I addressed some of these challenges with a focus on the field of nanoclusters.

In a brief introduction into the concepts behind heterogeneous catalysis I explained that kinetic properties can be replaced by purely thermodynamic properties. The d-band theory let us catch a glimpse that the interaction between catalyst surface and adsorbate need not be explicitly simulated. In theory, catalytic activity can be predicted by properties of the catalyst surface only.

In light of the above, the first study focused on structural descriptors that can describe the surface without the adsorbate itself. I participated in the implementation of representations of atomistic systems as input for machine learning in materials science. I presented an overview of descriptors and introduced Smooth Overlap of Atomic Positions (SOAP) in more detail. The work resulted in an open-source python library which has been embraced by the materials science community.

The second study provided the first benchmark of structural descriptors for nanocluster systems. It proved that machine learning adsorption energies from sole knowledge of the local environment is possible. As a useful reference for future work, the descriptor-ML combination SOAP-KRR was deemed most efficient. Methods were established to make screening of adsorption energies on nanoclusters more efficient.

Finally, I developed an automated workflow to screen large datasets of nanoclusters. It is publicly available and supports any catalytic reaction on any nanocluster class. A bimetallic dataset containing several compositions of Ti, Co, Fe, Ni, Cu and Pt was generated. New methods were proposed to address the large search space with respect to composition, configura-

tion, size and shape. The maximum of the d-band Hilbert-transform was identified as a potential pre-screening property on the nanocluster level. The workflow was demonstrated on the hydrogen evolution reaction and the results were compared to experiments. The results were not satisfying and underlined that there was still a large gap between simple screening simulations and experiments.

Rational Catalyst Design is still far from maturity. The goal in data-driven discovery is to circumvent simulating nanocluster-adsorbate and adsorbate-adsorbate interactions explicitly. Thus, it is important to develop and test descriptors to help reach that goal. I presented several descriptors in this dissertation. The electronic descriptors ϵ_d, ϵ_d^w and ϵ_u are intuitive descriptors. They are not well-suited for machine learning but rather to understanding catalytic trends. The presented structural descriptors are less intuitive, however they are efficient fingerprints for machine learning. Both types of descriptors have their place. The former help in improving the theoretical foundation of catalysis, the latter help in making data-driven catalysis predictions as accurate as possible.

Several levels can be tackled to improve computational screening of nanoclusters. In a screening funnel, potential candidates can be sieved out on different levels of computation complexity. The levels lower in the funnel require more computational resources as depicted in figure 6.1.

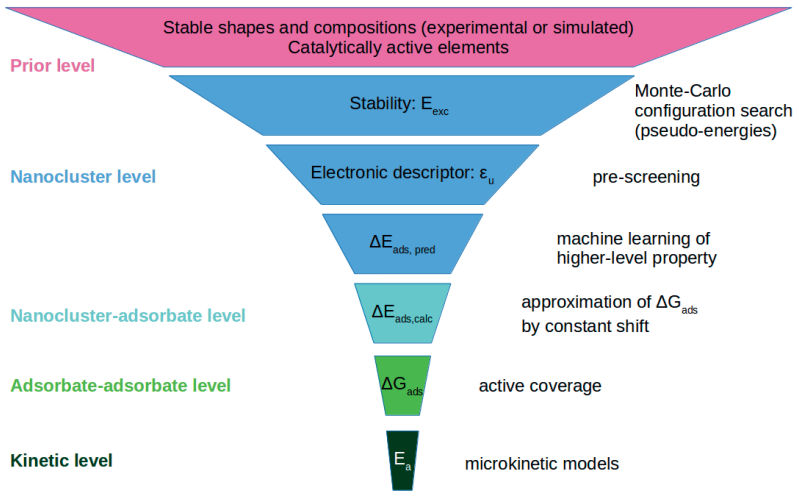


Figure 6.1. Nanocluster screening funnel. The lower the level in the funnel the more computational resources are required. Hence, it is paramount to filter out nanoclusters as early as possible.

I addressed the levels up to the nanocluster-adsorbate level. Pre-screening

nanoclusters with respect to stability and electronic properties is important before simulating nanocluster-adsorbate interactions. The adsorbate-adsorbate level is not subject of this dissertation, but is currently being addressed in collaboration. For instance, I implemented an automated workflow to determine the optimal nanocluster coverage which is currently subject to testing [161]. It is paramount for future studies to make simulations at the adsorbate-adsorbate level more efficient, since it is the only way to estimate catalytic activity well and verify theoretical predictions without relying on experiments.

In the future, it would be highly beneficial to have automatic simulation feedback loops. On the nanocluster-adsorbate level, screening potential energy surfaces could be made more efficient with Bayesian Optimization Structure Search (BOSS) [163] or Gaussian Process Regression [171]. To generate feedback between different levels (e.g ΔE_{ads} and nanocluster selection) an ML optimizer needs to be hooked to the automated workflow. A framework which would allow that is Rocketsled [172]. It is based on Fireworks (as the developed workflow) and chooses instances of workflows (with different inputs) on the fly based on an optimization engine. As summarized in figure 5.2 and explained in detail in chapter 5, these inputs can range from elements, shapes, sizes to pseudo-energies. The target property could be stability and ϵ_u for pre-screening or ΔE_{ads} and ΔG_{ads} at higher computation complexity.

Ultimately, experiments guide the design of new catalysts. A smoothly working feedback loop between simulations and experiments lies far in the future. However, this shall not inhibit collaboration between computational materials scientists and experimentalists. It can be useful to restrain the search space to a starting point of manageable size. For instance, elemental compositions and nanocluster shapes should be chosen based on experimental evidence. Consistent series of experiments on the catalytic activity of nanoclusters are still scarce. There are a few unpublished experiments with controlled size and composition which could be a useful benchmark. However, there are not enough data to machine learn with the target property of experimental catalytic activities.

The goal of efficient exploration aside, what can be done to improve our understanding in nanocatalysis? The proposed workflow can help unveil trends with respect to nanocluster shape and size if the search space is chosen accordingly. A sensible next step would be the compilation of a larger dataset encompassing different common shapes such as icosahedral, decahedral and Wulff as well as sizes up to 2 nm. The trends in electronic descriptors and adsorption energies could shed light on the influence of the local and the global structure. Nanocluster effects of core strain, size, and low coordination of edges could be resolved quantitatively. Such analysis could result in refined structurally resolved electronic descriptors.

References

[1] Jin, R. Quantum sized, thiolate-protected gold nanoclusters. *Nanoscale* **2**, 343–362 (2010).

[2] Zhou, K. & Li, Y. Catalysis based on nanocrystals with well-defined facets. *Angewandte Chemie - International Edition* **51**, 602–613 (2012).

[3] Fan, Z., Huang, X., Tan, C. & Zhang, H. Thin metal nanostructures: Synthesis, properties and applications. *Chemical Science* **6**, 95–111 (2015).

[4] Sayle, D. C., Maicaneanu, S. A. & Watson, G. W. Atomistic models for CeO_2(111), (110), and (100) nanoparticles, supported on yttrium-stabilized zirconia. *Journal of the American Chemical Society* **124**, 11429–11439 (2002).

[5] Nan, C. *et al.* Size and shape control of $LiFePO_4$ nanocrystals for better lithium ion battery cathode materials. *Nano Research* **6**, 469–477 (2013).

[6] Valden, M. Onset of Catalytic Activity of Gold Clusters on Titania with the Appearance of Nonmetallic Properties. *Science* **281**, 1647–1650 (1998).

[7] Yang, F., Deng, D., Pan, X., Fu, Q. & Bao, X. Understanding nano effects in catalysis. *National Science Review* **2**, 183–201 (2015).

[8] Wilcoxon, J. P. & Abrams, B. L. Synthesis, structure and properties of metal nanoclusters. *Chemical Society Reviews* **35**, 1162–1194 (2006).

[9] Zhang, Z.-c., Xu, B. & Wang, X. Engineering nanointerfaces for nanocatalysis. *Chemical Society Reviews* **43**, 7870–7886 (2014).

[10] Wang, D. *et al.* Shape control of CoO and $LiCoO_2$ nanocrystals. *Nano Research* **3**, 1–7 (2010).

[11] Hu, J. *et al.* Engineering stepped edge surface structures of MoS_2 sheet stacks to accelerate the hydrogen evolution reaction. *Energy & Environmental Science* **10**, 593–603 (2017).

[12] Fu, G. *et al.* Synthesis and electrocatalytic activity of Au@Pd core-shell nanothorns for the oxygen reduction reaction. *Nano Research* **7**, 1205–1214 (2014).

[13] Cuddy, M. J. *et al.* Fabrication and atomic structure of size-selected, layered MoS_2 clusters for catalysis. *Nanoscale* **6**, 12463–9 (2014).

[14] Bertuccioli, L. *et al.* Development of Water Electrolysis in the European Union. Tech. Rep., Fuel Cells and Hydrogen Joint Undertaking (2014).

[15] Mayyas, A., Ruth, M., Pivovar, B., Bender, G. & Wipke, K. Manufacturing Cost Analysis for Proton Exchange Membrane Water Electrolyzers. Tech. Rep., National Renewable Energy Laboratory (2019).

[16] European Commission. Report on Critical Raw Materials for the EU, Ad hoc Working Group on defining critical raw materials. Tech. Rep., European Commission (2014).

[17] Deloitte Sustainability, British Geological Survey, Bureau de Recherches Géologiques et Minières & Netherlands Organisation for Applied Scientific Research. Study on the review of the list of critical raw materials : Final report. Tech. Rep. (2017).

[18] Walter, M. G. *et al.* Solar Water Splitting Cells. *Chemical Reviews* **110**, 6446–6473 (2010).

[19] Lewis, N. S. & Nocera, D. G. Powering the planet: Chemical challenges in solar energy utilization. *Proceedings of the National Academy of Sciences* **103**, 15729–15735 (2006).

[20] Roger, I., Shipman, M. A. & Symes, M. D. Earth-abundant catalysts for electrochemical and photoelectrochemical water splitting. *Nature Reviews Chemistry* **1**, 0003 (2017).

[21] anonymous. The Haber Process. *Nature* **111**, 101–102 (1923).

[22] Chorkendorff, I. & Niemantsverdriet, J. W. *Concepts of Modern Catalysis and Kinetics* (Wiley-VCH, 2003).

[23] Bell, R. P. & Hinshelwood, C. N. The theory of reactions involving proton transfers. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* **154**, 414–429 (1936).

[24] Brønsted, J. N. Acid and Basic Catalysis. *Chemical Reviews* **5**, 231–338 (1928).

[25] Evans, M. G. & Polanyi, M. Inertia and driving force of chemical reactions. *Transactions of the Faraday Society* **34**, 11–24 (1938).

[26] van Santen, R. A., Neurock, M. & Shetty, S. G. Reactivity Theory of Transition-Metal Surfaces: A Brønsted-Evans-Polanyi Linear Activation Energy-Free-Energy Analysis. *Chemical Reviews* **110**, 2005–2048 (2010).

[27] Wang, S., Vorotnikov, V., Sutton, J. E. & Vlachos, D. G. Brønsted-Evans-Polanyi and Transition State Scaling Relations of Furan Derivatives on Pd(111) and Their Relation to Those of Small Molecules. *ACS Catalysis* **4**, 604–612 (2014).

[28] Pallassana, V. & Neurock, M. Electronic Factors Governing Ethylene Hydrogenation and Dehydrogenation Activity of Pseudomorphic Pd_{ML}/Re(0001), Pd_{ML}/Ru(0001), Pd(111), and Pd_{ML}/Au(111) Surfaces. *Journal of Catalysis* **191**, 301–317 (2000).

[29] Sabatier, P. & Senderens, J. Direct hydrogenation of oxides of carbon in presence of various finely divided metals. *Comptes rendus de l'Académie des Sciences* **134**, 689–691 (1902).

[30] Nørskov, J. K. *et al.* Trends in the Exchange Current for Hydrogen Evolution. *Journal of The Electrochemical Society* **152**, J23–J26 (2005).

[31] Parsons, R. The rate of electrolytic hydrogen evolution and the heat of adsorption of hydrogen. *Transactions of the Faraday Society* **54**, 1053 (1958).

[32] Brüssel, M., di Dio, P. J., Muñiz, K. & Kirchner, B. Comparison of Free Energy Surfaces Calculations from Ab Initio Molecular Dynamic Simulations at the Example of Two Transition Metal Catalyzed Reactions. *International Journal of Molecular Sciences* **12**, 1389–1409 (2011).

[33] Chipot, C., Mark, A. E., Pande, V. S. & Simonson, T. Applications of Free Energy Calculations to Chemistry and Biology. In Chipot, C. & Pohorille, A. (eds.) *Free Energy Calculations: Theory and Applications in Chemistry and Biology*, Springer Series in Chemical Physics, 463–501 (Springer, Berlin, Heidelberg, 2007).

[34] Hinnemann, B. *et al.* Biomimetic Hydrogen Evolution: MoS_2 Nanoparticles as Catalyst for Hydrogen Evolution. *Journal of the American Chemical Society* **127**, 5308–5309 (2005).

[35] Seh, Z. W. *et al.* Combining theory and experiment in electrocatalysis: Insights into materials design. *Science* **355**, 4998 (2017).

[36] Ojha, K., Saha, S., Dagar, P. & Ganguli, A. K. Nanocatalysts for hydrogen evolution reactions. *Physical Chemistry Chemical Physics* **20**, 6777–6799 (2018).

[37] Jiao, Y., Zheng, Y., Jaroniec, M. & Qiao, S. Z. Design of electrocatalysts for oxygen- and hydrogen-involving energy conversion reactions. *Chemical Society Reviews* **44**, 2060–2086 (2015).

[38] Zheng, Y., Jiao, Y., Jaroniec, M. & Qiao, S. Z. Advancing the Electrochemistry of the Hydrogen-Evolution Reaction through Combining Experiment and Theory. *Angewandte Chemie International Edition* **54**, 52–65 (2015).

[39] Suryanto, B. H. R., Wang, Y., Hocking, R. K., Adamson, W. & Zhao, C. Overall electrochemical splitting of water at the heterogeneous interface of nickel and iron oxide. *Nature Communications* **10**, 5599 (2019).

[40] Hammer, B. & Nørskov, J. K. Why gold is the noblest of all the metals. *Nature* **376**, 238–240 (1995).

[41] Hammer, B. & Nørskov, J. Theoretical Surface Science and Catalysis — Calculations and Concepts. *Advances in Catalysis* **45**, 71–129 (2000).

[42] Groß, A. Adsorption on Surfaces. In Groß, A. (ed.) *Theoretical Surface Science: A Microscopic Perspective*, 101–163 (Springer, Berlin, Heidelberg, 2009).

[43] Hammer, B. & Nørskov, J. Electronic factors determining the reactivity of metal surfaces. *Surface Science* **343**, 211–220 (1995).

[44] Hammer, B., Nielsen, O. & Nørskov, J. Structure sensitivity in adsorption: CO interaction with stepped and reconstructed Pt surfaces. *Catalysis Letters* **46**, 31–35 (1997).

[45] Pallassana, V., Neurock, M., Hansen, L. B., Hammer, B. & Nørskov, J. K. Theoretical analysis of hydrogen chemisorption on Pd(111), Re(0001) and Pd_{ML}/ Re(0001), Re_{ML}/Pd(111) pseudomorphic overlayers. *Physical Review B* **60**, 6146–6154 (1999).

[46] Xin, H. & Linic, S. Communications: Exceptions to the d-band model of chemisorption on metal surfaces: The dominant role of repulsion between adsorbate states and metal d-states. *The Journal of Chemical Physics* **132**, 221101 (2010).

[47] Vojvodic, A., Nørskov, J. K. & Abild-Pedersen, F. Electronic Structure Effects in Transition Metal Surface Chemistry. *Topics in Catalysis* **57**, 25–32 (2014).

[48] Xin, H., Vojvodic, A., Voss, J., Nørskov, J. K. & Abild-Pedersen, F. Effects of d-band shape on the surface reactivity of transition-metal alloys. *Physical Review B* **89**, 115114 (2014).

[49] Anderson, P. W. Localized Magnetic States in Metals. *Physical Review* **124**, 41–53 (1961).

[50] Greiner, M. T. *et al.* Free-atom-like d states in single-atom alloy catalysts. *Nature Chemistry* **10**, 1008–1015 (2018).

[51] Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials* **4**, 053208 (2016).

[52] Vitos, L., Ruban, A. V., Skriver, H. L. & Kollár, J. The surface energy of metals. *Surface Science* **411**, 186–202 (1998).

[53] Kollár, J., Vitos, L. & Skriver, H. L. Surface energy and work function of the light actinides. *Physical Review B* **49**, 11288–11292 (1994).

[54] Skriver, H. L. & Rosengaard, N. M. Surface energy and work function of elemental metals. *Physical Review B* **46**, 7157–7168 (1992).

[55] Methfessel, M., Hennig, D. & Scheffler, M. Trends of the surface relaxations, surface energies, and work functions of the 4d transition metals. *Physical Review B* **46**, 4816–4829 (1992).

[56] Chizallet, C. & Raybaud, P. Density functional theory simulations of complex catalytic materials in reactive environments: Beyond the ideal surface at low coverage. *Catalysis Science & Technology* **4**, 2797 (2014).

[57] Tran, K. & Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nature Catalysis* **1**, 696–703 (2018).

[58] Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. & Nørskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nature Materials* **5**, 909–913 (2006).

[59] Nørskov, J. K., Bligaard, T., Rossmeisl, J. & Christensen, C. H. Towards the computational design of solid catalysts. *Nature Chemistry* **1**, 37–46 (2009).

[60] Greeley, J., Nørskov, J. K., Kibler, L. A., El-Aziz, A. M. & Kolb, D. M. Hydrogen Evolution Over Bimetallic Systems: Understanding the Trends. *ChemPhysChem* **7**, 1032–1035 (2006).

[61] Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nature Materials* **12**, 191–201 (2013).

[62] Vojvodic, A. & Nørskov, J. K. New design paradigm for heterogeneous catalysts. *National Science Review* **2**, 140–143 (2015).

[63] Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials* **2**, 083802 (2018).

[64] Isayev, O. *et al.* Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications* **8**, 15679 (2017).

[65] Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Physical Review Letters* **114**, 105503 (2015).

[66] Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2**, 1–7 (2016).

[67] Ward, L. *et al.* Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B* **96**, 024104 (2017).

[68] Ghiringhelli, L. M. *et al.* Learning physical descriptors for materials science by compressed sensing. *New Journal of Physics* **19**, 023017 (2017).

[69] Ma, X. Orbitalwise Coordination Number for Predicting Adsorption Properties of Metal Nanocatalysts. *Physical Review Letters* **118**, 036101 (2017).

[70] Wexler, R. B., Martirez, J. M. P. & Rappe, A. M. Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni_2P from Nonmetal Surface Doping Interpreted via Machine Learning. *Journal of the American Chemical Society* **140**, 4678–4683 (2018).

[71] Gasper, R., Shi, H. & Ramasubramaniam, A. Adsorption of CO on Low-Energy, Low-Symmetry Pt Nanoparticles: Energy Decomposition Analysis and Prediction via Machine-Learning Models. *The Journal of Physical Chemistry C* **121**, 5612–5619 (2017).

[72] Peterson, K. A., Feller, D. & Dixon, D. A. Chemical accuracy in ab initio thermochemistry and spectroscopy: Current strategies and future challenges. *Theoretical Chemistry Accounts* **131**, 1079 (2012).

[73] Bartlett, R. J. & Musiał, M. Coupled-cluster theory in quantum chemistry. *Reviews of Modern Physics* **79**, 291–352 (2007).

[74] Crawford, T. D. & Schaefer, H. F. An Introduction to Coupled Cluster Theory for Computational Chemists. In *Reviews in Computational Chemistry*, 33–136 (John Wiley & Sons, Ltd, 2007).

[75] Kirklin, S. *et al.* The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Computational Materials* **1**, 1–15 (2015).

[76] Hohenberg, P. & Kohn, W. Inhomogeneous Electron Gas. *Physical Review* **136**, B864–B871 (1964).

[77] Kohn, W. & Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review* **140**, A1133–A1138 (1965).

[78] Parr, R. G. & Yang, W. *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).

[79] Janthon, P., Kozlov, S. M., Viñes, F., Limtrakul, J. & Illas, F. Establishing the Accuracy of Broadly Used Density Functionals in Describing Bulk Properties of Transition Metals. *Journal of Chemical Theory and Computation* **9**, 1631–1640 (2013).

[80] Martin, R. M. *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, 2004).

[81] Hutter, J., Iannuzzi, M., Schiffmann, F. & VandeVondele, J. CP2K: Atomistic simulations of condensed matter systems. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **4**, 15–25 (2014).

[82] Vandevondele, J., Borštnik, U. & Hutter, J. Linear scaling self-consistent field calculations with millions of atoms in the condensed phase. *Journal of Chemical Theory and Computation* **8**, 3565–3573 (2012).

[83] Vandevondele, J. *et al.* Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Computer Physics Communications* **167**, 103–128 (2005).

[84] Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **77**, 3865–3868 (1996).

[85] VandeVondele, J. & Hutter, J. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *The Journal of chemical physics* **127**, 114105 (2007).

[86] Goedecker, S., Teter, M. & Hutter, J. Separable dual-space Gaussian pseudopotentials. *Physical Review B* **54**, 1703–1710 (1996).

[87] Krack, M. Pseudopotentials for H to Kr optimized for gradient-corrected exchange-correlation functionals. *Theoretical Chemistry Accounts* **114**, 145–152 (2005).

[88] Hartwigsen, C., Goedecker, S. & Hutter, J. Relativistic separable dual-space Gaussian Pseudopotentials from H to Rn. *Physical Review B* **58**, 3641–3662 (1998).

[89] Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *Journal of Chemical Physics* **132** (2010).

[90] Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the Damping Funtion in Dispersion Corrected Density Functional Theory. *Journal of Computational Chemistry* **32**, 1456–1465 (2011).

[91] Barber, D. *Bayesian Reasoning and Machine Learning* (Cambridge University Press, 2012).

[92] Huo, H. & Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. *Submitted Results* (2017). 1704.06439.

[93] Faber, F., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry* **115**, 1094–1101 (2015).

[94] Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B* **95**, 144110 (2017).

[95] Hansen, K. *et al.* Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *The Journal of Physical Chemistry Letters* **6**, 2326–2331 (2015).

[96] Choudhary, K., DeCost, B. & Tavazza, F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Physical Review Materials* **2**, 083801 (2018).

[97] Gastegger, M., Schwiedrzik, L., Bittermann, M., Berzsenyi, F. & Marquetand, P. wACSF—Weighted atom-centered symmetry functions as descriptors in machine learning potentials. *The Journal of Chemical Physics* **148**, 241709 (2018).

[98] Pronobis, W., Tkatchenko, A. & Müller, K.-R. Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules. *Journal of Chemical Theory and Computation* **14**, 2991–3003 (2018).

[99] Faber, F. A., Christensen, A. S., Huang, B. & von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of Chemical Physics* **148**, 241717 (2018).

[100] Faber, F. A. *et al.* Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *Journal of Chemical Theory and Computation* **13**, 5255–5264 (2017).

[101] Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters* **108**, 58301 (2012).

[102] Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics* **134**, 074106 (2011).

[103] Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Physical Review B* **87**, 184115 (2013).

[104] De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics* **18**, 13754–13769 (2016). 1601.04077.

[105] SIN. DScribe. https://github.com/SINGROUP/cluskit (2020).

[106] SIN. DScribe Documentation. https://singroup.github.io/dscribe/ (2020).

[107] Kermode, J. QUIP - QUantum mechanics and Interatomic Potentials. https://github.com/libAtoms/QUIP (2020).

[108] Khorshidi, A. & Peterson, A. A. Amp: A modular approach to machine learning in atomistic simulations. *Computer Physics Communications* **207**, 310–324 (2016).

[109] Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **152**, 60–69 (2018).

[110] Haghighatlari, M. *et al.* ChemML: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data. *WIREs Computational Molecular Science* **n/a**, e1458 (2020).

[111] Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **5**, 1–36 (2019).

[112] Xie, T., France-Lanord, A., Wang, Y., Shao-Horn, Y. & Grossman, J. C. Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. *Nature Communications* **10**, 2667 (2019).

[113] Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters* **120**, 145301 (2018).

[114] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 1263–1272 (JMLR.org, Sydney, NSW, Australia, 2017).

[115] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[116] Zhou, Y. *et al.* Bonding similarities and differences between Y–Sb–Te and Sc–Sb–Te phase-change memory materials. *Journal of Materials Chemistry C* **8**, 3646–3654 (2020).

[117] Fabrizio, A., Meyer, B. & Corminboeuf, C. Machine learning models of the energy curvature vs particle number for optimal tuning of long-range corrected functionals. *The Journal of Chemical Physics* **152**, 154103 (2020).

[118] Fujii, S., Yokoi, T., Fisher, C. A. J., Moriwake, H. & Yoshiya, M. Quantitative prediction of grain boundary thermal conductivities from local atomic environments. *Nature Communications* **11**, 1854 (2020).

[119] Stuke, A. *et al.* Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Scientific Data* **7** (2020).

[120] Chehaibou, B., Badawi, M., Bučko, T., Bazhirov, T. & Rocca, D. Computing RPA Adsorption Enthalpies by Machine Learning Thermodynamic Perturbation Theory. *Journal of Chemical Theory and Computation* **15**, 6333–6342 (2019).

[121] Singh, A. R., Rohr, B. A., Gauthier, J. A. & Nørskov, J. K. Predicting Chemical Reaction Barriers with a Machine Learning Model. *Catalysis Letters* **149**, 2347–2354 (2019).

[122] Smith, A., Keane, A., Dumesic, J. A., Huber, G. W. & Zavala, V. M. A machine learning framework for the analysis and prediction of catalytic activity from experimental data. *Applied Catalysis B: Environmental* **263**, 118257 (2020).

[123] Schlexer Lamoureux, P. *et al.* Machine Learning for Computational Heterogeneous Catalysis. *ChemCatChem* **11**, 3581–3601 (2019).

[124] Mamun, O., Winther, K. T., Boes, J. R. & Bligaard, T. High-throughput calculations of catalytic properties of bimetallic alloy surfaces. *Scientific Data* **6**, 1–9 (2019).

[125] Li, Z., Wang, S., Chin, W. S., Achenie, L. E. & Xin, H. High-throughput screening of bimetallic catalysts enabled by machine learning. *Journal of Materials Chemistry A* **5**, 24131–24138 (2017).

[126] Ma, X., Li, Z., Achenie, L. E. K. & Xin, H. Machine-Learning-Augmented Chemisorption Model for CO_2 Electroreduction Catalyst Screening. *The Journal of Physical Chemistry Letters* **6**, 3528–3533 (2015).

[127] Ulissi, Z. W. *et al.* Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO_2 Reduction. *ACS Catalysis* **7**, 6600–6608 (2017).

[128] Hoyt, R. A. *et al.* Machine Learning Prediction of H Adsorption Energies on Ag Alloys. *Journal of Chemical Information and Modeling* **59**, 1357–1365 (2019).

[129] Zahrt, A. F. *et al.* Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363** (2019).

[130] Li, Z., Ma, X. & Xin, H. Feature engineering of machine-learning chemisorption models for catalyst design. *Catalysis Today* **280**, 232–238 (2017).

[131] Jinnouchi, R. & Asahi, R. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *The Journal of Physical Chemistry Letters* **8**, 4279–4283 (2017).

[132] Jinnouchi, R., Hirata, H. & Asahi, R. Extrapolating Energetics on Clusters and Single-Crystal Surfaces to Nanoparticles by Machine-Learning Scheme. *The Journal of Physical Chemistry C* **121**, 26397–26405 (2017).

[133] Panapitiya, G. *et al.* Machine-Learning Prediction of CO Adsorption in Thiolated, Ag-Alloyed Au Nanoclusters. *Journal of the American Chemical Society* **140**, 17508–17514 (2018).

[134] Goldsmith, B. R., Esterhuizen, J., Liu, J.-X., Bartel, C. J. & Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AIChE Journal* **64**, 2311–2323 (2018).

[135] Jäger, M. O. J., Morooka, E. V., Canova, F. F., Himanen, L. & Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials* **4**, 1–8 (2018).

[136] McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* **5**, 115–133 (1943).

[137] Sørensen, S. G., Füchtbauer, H. G., Tuxen, A. K., Walton, A. S. & Lauritsen, J. V. Structure and electronic properties of in situ synthesized single-layer MoS_2 on a gold surface. *ACS Nano* **8**, 6788–6796 (2014).

[138] Bruix, A. *et al.* In Situ Detection of Active Edge Sites in Single-Layer MoS_2 Catalysts. *ACS Nano* **9**, 9322–9330 (2015).

[139] Walton, A. S., Lauritsen, J. V., Topsøe, H. & Besenbacher, F. MoS_2 nanoparticle morphologies in hydrodesulfurization catalysis studied by scanning tunneling microscopy. *Journal of Catalysis* **308**, 306–318 (2013).

[140] Bartók, A. P. *et al.* Machine learning unifies the modeling of materials and molecules. *Science Advances* **3**, e1701816 (2017).

[141] Baletto, F. & Ferrando, R. Structural properties of nanoclusters: Energetic, thermodynamic, and kinetic effects. *Reviews of Modern Physics* **77**, 371–423 (2005).

[142] Baletto, F., Ferrando, R., Fortunelli, A., Montalenti, F. & Mottet, C. Crossover among structural motifs in transition and noble-metal clusters. *The Journal of Chemical Physics* **1161** (2002).

[143] Tian, N., Zhou, Z.-Y. & Sun, S.-G. Platinum Metal Catalysts of High-Index Surfaces: From Single-Crystal Planes to Electrochemically Shape-Controlled Nanoparticles. *The Journal of Physical Chemistry C* **112**, 19801–19817 (2008).

[144] Ferrando, R., Jellinek, J. & Johnston, R. L. Nanoalloys: From Theory to Applications of Alloy Clusters and Nanoparticles. *Chemical Reviews* **108**, 845–910 (2008).

[145] Wang, L.-L. & Johnson, D. D. Predicted Trends of Core-Shell Preferences for 132 Late Transition-Metal Binary-Alloy Nanoparticles. *Journal of the American Chemical Society* **131**, 14023–14029 (2009).

[146] Guedes-Sobrinho, D., Nomiyama, R. K., Chaves, A. S., Piotrowski, M. J. & Da Silva, J. L. F. Structure, Electronic, and Magnetic Properties of Binary Pt_nTM_{55-n} (TM = Fe, Co, Ni, Cu, Zn) Nanoclusters: A Density Functional Theory Investigation. *The Journal of Physical Chemistry C* **119**, 15669–15679 (2015).

[147] Corona, B., Howard, M., Zhang, L. & Henkelman, G. Computational screening of core@shell nanoparticles for the hydrogen evolution and oxygen reduction reactions. *The Journal of Chemical Physics* **145**, 244708 (2016).

[148] Xie, P. *et al.* Highly efficient decomposition of ammonia using high-entropy alloy catalysts. *Nature Communications* **10**, 1–12 (2019).

[149] Zhang, G. *et al.* High entropy alloy as a highly active and stable electrocatalyst for hydrogen evolution reaction. *Electrochimica Acta* **279**, 19–23 (2018).

[150] Zeng, M. & Li, Y. Recent advances in heterogeneous electrocatalysts for the hydrogen evolution reaction. *Journal of Materials Chemistry A* **3**, 14942–14962 (2015).

[151] Darby, M. T., Stamatakis, M., Michaelides, A. & Charles Sykes, E. H. Lonely Atoms with Special Gifts: Breaking Linear Scaling Relationships in Heterogeneous Catalysis with Single-Atom Alloys. *Journal of Physical Chemistry Letters* **9**, 5636–5646 (2018).

[152] Piotrowski, M. J. *et al.* Theoretical Study of the Structural, Energetic, and Electronic Properties of 55-Atom Metal Nanoclusters: A DFT Investigation within van der Waals Corrections, Spin–Orbit Coupling, and PBE+ U of 42 Metal Systems. *The Journal of Physical Chemistry C* **120**, 28844–28856 (2016).

[153] Heggen, M., Gocyla, M. & Dunin-Borkowski, R. E. The growth and degradation of binary and ternary octahedral Pt–Ni-based fuel cell catalyst nanoparticles studied using advanced transmission electron microscopy. *Advances in Physics: X* **2**, 281–301 (2017).

[154] Wang, Z. *et al.* An ultrafine platinum–cobalt alloy decorated cobalt nanowire array with superb activity toward alkaline hydrogen evolution. *Nanoscale* **10**, 12302–12307 (2018).

[155] Lu, Q. *et al.* Highly porous non-precious bimetallic electrocatalysts for efficient hydrogen evolution. *Nature Communications* **6**, 6567 (2015).

[156] Mathew, K. *et al.* Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Computational Materials Science* **139**, 140–152 (2017).

[157] Huber, S. P. *et al.* AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *arXiv:2003.12476 [cond-mat]* (2020). 2003.12476.

[158] Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: Automated interactive infrastructure and database for computational science. *Computational Materials Science* **111**, 218–230 (2016).

[159] Jain, A. *et al.* FireWorks: A dynamic workflow system designed for high-throughput applications. *Concurrency and Computation: Practice and Experience* **27**, 5037–5059 (2015).

[160] Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).

[161] SIN. Critcatworks. https://github.com/SINGROUP/critcatworks (2020).

[162] Zhang, R. F. *et al.* An informatics guided classification of miscible and immiscible binary alloy systems. *Scientific Reports* **7**, 1–12 (2017).

[163] Todorović, M., Gutmann, M. U., Corander, J. & Rinke, P. Bayesian inference of atomistic structure in functional materials. *npj Computational Materials* **5**, 1–7 (2019).

[164] SIN. Cluskit. https://github.com/SINGROUP/cluskit (2019).

[165] Caro, M. A., Aarva, A., Deringer, V. L., Csányi, G. & Laurila, T. Reactivity of Amorphous Carbon Surfaces: Rationalizing the Role of Structural Motifs in Functionalization Using Machine Learning. *Chemistry of Materials* **30**, 7446–7455 (2018).

[166] Zhang, Y., Apley, D. W. & Chen, W. Bayesian Optimization for Materials Design with Mixed Quantitative and Qualitative Variables. *Scientific Reports* **10**, 4924 (2020).

[167] Trasatti, S. Work function, electronegativity, and electrochemical behaviour of metals: III. Electrolytic hydrogen evolution in acid solutions. *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry* **39**, 163–184 (1972).

[168] McCrory, C. C. L. *et al.* Benchmarking Hydrogen Evolving Reaction and Oxygen Evolving Reaction Electrocatalysts for Solar Water Splitting Devices. *Journal of the American Chemical Society* **137**, 4347–4357 (2015).

[169] Lu, S. & Zhuang, Z. Electrocatalysts for hydrogen oxidation and evolution reactions. *Science China Materials* **59**, 217–238 (2016).

[170] Greeley, J. & Mavrikakis, M. Alloy catalysts designed from first principles. *Nature Materials* **3**, 810–815 (2004).

[171] Koistinen, O.-P., Dagbjartsdóttir, F. B., Ásgeirsson, V., Vehtari, A. & Jónsson, H. Nudged elastic band calculations accelerated with Gaussian process regression. *The Journal of Chemical Physics* **147**, 152720 (2017).

[172] Dunn, A., Brenneck, J. & Jain, A. Rocketsled: A software library for optimizing high-throughput computational searches. *Journal of Physics: Materials* **2**, 034002 (2019).

Aalto-DD 130/2020

ISBN 978-952-64-0016-7 (printed)
ISBN 978-952-64-0017-4 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Applied Physics
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS