Master Thesis

Design of a Machine Learning Algorithm for Adsorption Energy Prediction on Transition Metal Oxide Surfaces

carried out for the purpose of obtaining the degree of Diplom-Ingenieur_in (Dipl.-Ing.)

submitted at TU Wien

Faculty of Mechanical and Industrial Engineering

by

Alexander POLYAKOV

Mat.No.: 11705337

under the supervision of

Asst. Prof. Dr. Aleix Comas Vives

Institute of Materials Chemistry

Place, Date

_____

Signature

I confirm, that the printing of this thesis requires the approval of the examination board.

Affidavit

I declare, in lieu of oath, that I wrote this thesis and carried out the associated research myself, using only the literature cited in this volume. If text passages from sources are used literally, they are marked as such.

I confirm that this work is original and has not been submitted for examination elsewhere, nor is it currently under consideration for a thesis elsewhere.

I acknowledge that the submitted work will be checked electronically-technically using suitable and state-of-the-art means (plagiarism detection software). On the one hand, this ensures that the submitted work was prepared according to the high-quality standards within the applicable rules to ensure good scientific practice "Code of Conduct" at the TU Wien. On the other hand, a comparison with other student theses avoids violations of my personal copyright.

Place and Date	*Signature*

Acknowledgments

I would like to express my sincerest gratitude to my supervisor, Asst. Prof. Dr. Aleix Comas Vives for providing me the opportunity to engage in this exciting project and for his exceptional guidance throughout my research. As a novice in computational materials science, I deeply appreciate how Aleix patiently elucidated complex quantum chemistry concepts and encouraged me during challenging times. My gratitude also extends to Andres Felipe Usuga from the Department of Chemistry at UAB for his instrumental role in teaching me how to operate with the Vienna Ab-initio Simulation Package and guiding me through the practical aspects of machine learning model implementation. He was also very supportive with troubleshooting software issues, especially when I faced challenging situations. I am also appreciative of my fellow colleague and dear friend Daniel Djambakiev for helping me develop software solutions for challenging tasks and encouraging me along the whole journey.

Furthermore, I am deeply thankful to my family and friends for nurturing my aspirations and providing support throughout the development of this thesis. This thesis is a culmination of my current academic journey and a result of the invaluable contributions of each person mentioned, and I am sincerely grateful for their integral roles in my academic path.

Table of Contents

Abstract

This thesis delves into the capabilities of machine learning (ML) methods for predicting adsorption energies in transition metal oxides (TMOs), highlighting their role as a viable and efficient alternative to the traditional, computationally demanding Density Functional Theory (DFT) simulations in the discovery of novel catalysts. On the basis of 'easy-to-calculate' properties of clean substrates and adsorbates, ML models aim to produce accurate predictions for 'hard-to-calculate' properties, such as adsorption energy, a critical factor in determining catalytic activity. In pursuit of this objective, the study constructs a dataset comprising 102 data points. This dataset, a carefully selected subset of the Open Catalyst 2022 database, focuses on substrates including Ti-O, Al-O, and Zr-O, each paired with an adsorbate (H2O, CO, OH, O, H, C, N) placed in random configuration. Relaxed structures are augmented with additional single-point DFT calculations, enriching them with electronic structure data. This enriched dataset enabled the identification of atoms actively involved in adsorption phenomena and allowed the extraction of their unique geometrical and electronic characteristics. These characteristics, or 'descriptors', were then utilized for training advanced ML algorithms following the 'descriptor-based' approach. Employing Kernel Ridge Regression (KRR), XGBoost, and CatBoost models, the study showcases the capability of these models to capture complex, non-linear relationships between material properties and adsorption energies. However, it is noted that while the models exhibit significant promise, they have yet to achieve further accuracy. The findings indicate that enhancements in data quality and quantity could further refine the accuracy of these models. Such improvements hold the potential to transform these ML models into valuable tools for material screening and catalyst design, offering substantial contributions to the field of catalysis.

Kurzfassung

In dieser Arbeit wird das Potenzial von Methoden des Maschinellen Lernens (ML) zur Prognose von Adsorptionsenergien in Übergangsmetall-Oxiden ausgelotet. Diese Methoden stellen eine vielversprechende und effiziente Alternative zu den herkömmlichen, rechenintensiven Dichtefunktionaltheorie (DFT)-Simulationen für die Entdeckung neuer Katalysatoren dar. Auf der Basis von einfach zu berechnenden Eigenschaften reiner Substrate und Adsorbate, soll ein ML-Model präzise Prognosen für komplexere Eigenschaften wie die Adsorptionsenergie zu erzeugen. Letztere spielt eine entscheidende Rolle für die Bestimmung der katalytischen Aktivität. Die Studie entwickelt einen sorgfältig zusammengestellten Datensatz mit 102 Datenpunkten, der aus einer gezielten Auswahl der Open Catalyst 2022-Datenbank stammt. Dieser konzentriert sich auf Substrate wie Ti-O, Al-O und Zr-O, kombiniert mit einem zufällig gewählten Adsorbaten (H2O, CO, OH, O, H, C, N). Um die Aussagekraft des Datensatzes zu erhöhen, werden zusätzliche DFT-Berechnungen durchgeführt, die wichtige Daten über elektronische Eigenschaften liefern. Mit diesen ergänzten Informationen erfolgt die Identifikation der bei der Adsorption beteiligten Atome und die Extraktion ihrer unikalen geometrischen und elektronischen Eigenschaften. Diese Eigenschaften, auch als 'Deskriptoren' bezeichnet, dienen als Grundlage für das Training fortschrittlicher ML-Algorithmen, die dem 'deskriptorbasierten' Ansatz folgen. Die Anwendung von Kernel Ridge Regression (KRR), XGBoost und CatBoost Modellen in dieser Forschungsarbeit demonstriert deren Fähigkeit, die komplexen, nichtlinearen Zusammenhänge zwischen Materialeigenschaften und Adsorptionsenergien zu erfassen. Die Ergebnisse deuten darauf hin, dass durch Verbesserungen in der Qualität und Quantität der Daten die Präzision dieser Modelle weiter gesteigert werden könnte. Solche Fortschritte können diese ML-Modelle zu wertvollen Instrumenten für das Screening von Materialien und die Gestaltung von Katalysatoren machen und damit einen wesentlichen Beitrag zum Bereich der Katalyse leisten.

List of Figures

List of Tables

List of Abbreviations

DFT..... Density Functional Theory

ML..... Machine Learning

OC20.....Open Catalyst 2020

OC22.....Open Catalyst 2022

GNN .....Graph Neural Network

TM ..... Transition Metal

TMO.....Transition Metal Oxide

MAE..... Mean Average Error

MSE ..... Mean Squared Error

QM .....Quantum Mechanics

HK .....Hohenberg and Kohn

KS ..... Kohn and Sham

XC.....Exchange-Correlation (functional)

LDA.....Local Density Approximation

GGA.....Generalized Gradient Approximation

PBE .....Perdew, Burke, Ernzerhof (functional)

VASP ..... Vienna Ab-initio Simulation Package

PAW.....Projector-Augmented-Wave

RMM-DIIS ..... Residual Minimization Method with Direct Inversion of the Iterative Subspace

OER..... Oxygen Evolution Reaction

DOS ..... Density of States

ORR .....Oxygen Reduction Reaction

KRR .....Kernel Ridge Regression

GBDT .....Gradient Boosted Decision Tree

SHAP ..... SHapley Additive exPlanations

1. Introduction and motivation

Heterogeneous catalysis is an indispensable cornerstone in the modern chemical industry governing a broad spectrum of socioeconomically essential processes such as thermal and electrocatalysis, biofuel production, waste treatment, petroleum refining, etc. The traditional approaches of catalyst design have predominantly been empirical, relying heavily on trial-and-error methodologies. This empirical nature is attributed to the inherent complexity of catalytic processes that involve a multitude of possible adsorption sites as well as a high diversity of geometric and electronic structures [1]. The continuous increase of accessible computational power over the past few decades has enabled researchers to deploy quantum mechanical calculations based on density functional theory (DFT), which have nowadays become widely adopted by researchers in materials science. Yet, despite the notable accuracy offered by DFT calculations, this method presents its challenges, which impose several limitations on screening superior catalysts. Specifically, DFT calculations of catalytic systems are still rather computationally intensive, particularly for large low-symmetry structures, leading to substantial energy consumption, extended calculation times, and, consequently, high costs. As a result, it becomes technically infeasible to thoroughly explore the vast search space of materials and geometrical structures defined at the detail and extent required to perform computational catalytic screening.

Machine learning (ML) emerges as a viable solution to address these challenges as the availability of relevant data steadily increases. ML algorithms have gained tremendous popularity after the public release of large language models [2]. Concurrently, in fields such as materials science, researchers have been exploring meaningful implementations of these algorithms for over a decade [3]. One of the most promising avenues is using machine learning techniques to predict “hard-to-calculate” catalyst properties and adsorption behaviors utilizing the available data to simplify the material screening by replacing the demanding DFT calculations or narrowing the search space to only a few candidates [4], [5], [6], [7]. In 2016, Takigawa et al. [7] demonstrated the applicability of machine learning techniques in the understanding of catalytic materials. Their work involved developing and training a simple predictive model to determine the *d*-band center in metallic and bimetallic surfaces. Subsequently, the unveiling of the Open Catalyst 2020

database (OC20), along with the provision of pre-trained Graph Neural Networks (GNN) models [8] that focused on the prediction of adsorption energy, underscored the usefulness and emerging prominence of deep learning methodologies within the scientific community. While deep learning methods offer advantages in learning complex and generalizable representations, they also face challenges in model efficiency, consistency, and energy conservation [6]. Opposed to them are descriptor-based methods, which are older, simpler, and more computationally efficient, but the representation of the system is limited to one or several specific active sites.

Nevertheless, ML models can efficiently capture complex non-linear relationships, extending the applicability of the descriptor-based approach. C. S. Praveen and A. Comas-Vives applied machine learning to predict adsorption energy for several C-, N-, and O- based adsorbates on 11 transition metals (TM). The descriptor-based approach utilizes the scaling relationships between specific materials' properties and adsorption energy to make predictions of the adsorption energy for unexplored systems. They achieved a mean average error (MAE) of 0.174 eV on the test set, a relatively small error regarding the reference calculations performed with DFT [9].

This thesis aims to identify and evaluate descriptors for transition metal oxide (TMO) surfaces, which exhibit significantly more complicated atomic interactions than metallic surfaces, and develop a generalizable machine learning model for predicting the adsorption energy. TMOs play a crucial role in many industrial applications and demonstrate high potential in renewable technologies (Fischer-Tropsch-Process, energy storage), resulting in the demand for novel catalyst discovery [10], [11]. In this study, 134 DFT calculated trajectories of Zr-O, Al-O, and Ti-O systems, including 9 adsorbates, released in the Open Catalyst 2022 (OC 22) database (https://github.com/Open-Catalyst-Project/ocp/blob/main/DATASET.md), were selected to build a new dataset. This selection forms the basis for the development of a new dataset. The utilization of the OC 22 database facilitates the bypassing of numerous labor-intensive and computationally demanding steps, such as the construction of the adsorbate + substrate systems and the computation of their relaxation trajectories. For the chosen structures, additional single-point DFT simulations are conducted to acquire detailed insights into the electronic properties of the systems. These electronic properties are paramount as they are the descriptors used in our study. A comprehensive dataset will be constructed using the gathered data, featuring both geometrical and electronic descriptors, focusing on specific active sites. Finally, a range of machine learning algorithms will be employed, and their performance will be

evaluated to gain valuable insights into the potential of machine learning methods for adsorption phenomena involving TMOs.

The first part of the thesis presents an in-depth overview of critical concepts and tools essential for this research. That includes the theoretical background for the density functional theory and the Vienna Ab initio Simulation Package (VASP), which is needed to perform the DFT calculations. Subsequently, the OC 22 database and the general concept of the machine learning algorithms and specific regression models applied during this work are presented. The following chapter illustrates the methodology and the details of the data selection, collection, and analysis at each step of the conducted research and its implications on the overall research outcomes. The final part is dedicated to discussing and interpreting the results within the existing scientific literature and offers an outlook for this research area.

2. Theory and Methodology

2.1. Density functional theory

Density Functional Theory (DFT) is the most utilized computational method in materials science and quantum mechanics for electronic structure properties. As it is the primal source of training data for machine learning algorithms in this study, this chapter briefly overviews its fundamental concepts to understand the analyzed data better.

In 1964, P. Hohenberg and W. Kohn formulated and proved a new approach that employed ground state density $r_0(\boldsymbol{r})$ instead of many-body wave functions as the critical variable for evaluating the total energy and other properties. This work was done within the context of the static Schrödinger equation in the Born-Oppenheimer (BO) approximation, thereby avoiding the mathematical complexities associated with many-body wave functions [12]. The Schrödinger equation for static nuclei at R can be written as follows:

$$H\Psi(\boldsymbol{r},\boldsymbol{R}) = E\Psi(\boldsymbol{r},\boldsymbol{R}) \tag{1}$$

with the many-body Hamiltonian expressed as

$$H = -\frac{\hbar^2}{2m_e}\nabla_{\boldsymbol{r}}^2 + V_{\text{ext}}(\boldsymbol{r},\boldsymbol{R}) + V_{\text{ee}}(\boldsymbol{r}) \tag{2}$$

2.1.1. The Thomas-Fermi Model

The DFT approach is fundamentally built upon the principles outlined in the Thomas-Fermi (TF) model. Thomas and Fermi introduced the concept that electrons form a gas, where the classical Coulomb potential determines the electron-electron interaction energy. They were the first to propose using the electron density $r(\mathbf{r})$ as the essential variable for describing the electronic part of the many-body wave function [13]. It allows defining a system of N electrons in a volume element $d\boldsymbol{r}$ with only three independent coordinates as follows:

$$r(\boldsymbol{r}) = \sum_{i=1}^{M} \int |\psi(\boldsymbol{r_1}, s_1, \dots, \boldsymbol{r_M}, s_M)|^2 \delta(\boldsymbol{r_i} - \boldsymbol{r}) d\boldsymbol{r_1}\, s_1 \dots d\boldsymbol{r_M} s_M \tag{3}$$

Within this model, the ground state total energy and other properties can be expressed as functionals of the electron density. The total energy functional was defined as

$$E_{TF}[r(\boldsymbol{r})] = \frac{3}{10}(3\pi^2)^{2/3}\int r(\boldsymbol{r})^{5/3}d\boldsymbol{r} - \sum_{v=1}^{N}\int \frac{Z_v r(\boldsymbol{r})}{|\boldsymbol{r}-\boldsymbol{R}_v|}d\boldsymbol{r} + \frac{1}{2}\int\int \frac{r(\boldsymbol{r})r(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}d\boldsymbol{r}\,d\boldsymbol{r}' \quad (4)$$

The first term refers to the kinetic energy of non-interacting electrons, and the second and the third term denote the classical electrostatic electron-nucleus attraction and electron-electron repulsion. Although the Thomas Fermi model does not provide high accuracy for solid-body physics and quantum chemistry applications, it was the first step in using electron density for solving multielectron systems [12].

2.1.2. The Hohenberg-Kohn Theorems

DFT is based on two fundamental theorems presented by Hohenberg and Kohn (HK) in their publication of 1964 [14]. The first theorem establishes a unique relationship between the external potential $V_{ext}(r)$ (the potential of the ions or nuclei) and the ground state density $r_0(r)$ of an N-electron system via the nondegenerate ground state wave function Ψ. This relationship constrains $V_{ext}(r)$ to within an additive constant.

$$V'_{ext}(\boldsymbol{r}) = V_{ext}(\boldsymbol{r}) + C \quad (5)$$

Hohenberg and Kohn defined the "universal" functional $F[r(\boldsymbol{r})]$ that can be applied to any V_{ext}.

$$F[r(\boldsymbol{r})] = \langle\Psi|T + V_{ee}|\Psi\rangle \quad (6)$$

$T[r]$ represents the kinetic energy functional expressed in terms of the electron density, and $V_{ee}[r]$ is electron-electron interaction energy.

The second theorem states that the ground state energy of the system can be obtained by minimizing the energy functional $E[r(\boldsymbol{r})]$ with respect to $r(\boldsymbol{r})$, under the constraint that $r(\boldsymbol{r})$ integrates to the total number of electron N. HK showed that the energy functional satisfies this statement, called the energy variational principle.

$$E_0 = \min_{r(\boldsymbol{r})}\{E[r(\boldsymbol{r})]\} \quad \text{s.t.} \quad \int r(\boldsymbol{r})\,d\boldsymbol{r} = N \quad (7)$$

Hence, the expression of the system energy in the context of HK theorems

$$E[r(\boldsymbol{r})] = \int r(\boldsymbol{r}) V_{ext}(\boldsymbol{r}) d\boldsymbol{r} + F[r(\boldsymbol{r})] \tag{8}$$

The solution of this Equation is unknown; however, a multitude of successful approximation strategies for the *F[r(**r**)]* term [13] have been proposed.

2.1.3. The Kohn-Sham Equations

A year later, in 1965, Kohn and Sham (KS) introduced equations as a practical approach to implementing DFT. These equations provide a way to approximate the ground state electron density and, consequently, the ground state properties of a many-electron system. The core idea of this approach lies in transforming the interacting many-body problem into a set of non-interacting single-particle problems. A KS orbital describes each electron in this non-interacting system $\phi_i(\boldsymbol{r})$, which satisfies the single particle KS equation.

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{eff}}(\boldsymbol{r})\right]\phi_i(\boldsymbol{r}) = \epsilon_i \phi_i(\boldsymbol{r}) \tag{9}$$

In this Equation, ϵ_i is the eigenvalue associated with the orbital ϕ, which is the energy of the i-th non-interacting electron. The key term in the KS-equation is the effective potential $V_{\text{eff}}(\mathbf{r})$ that can be decomposed into 3 terms:

$$V_{\text{eff}}(\boldsymbol{r}) = V_{\text{ext}}(\boldsymbol{r}) + \int \frac{\mathrm{r}(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r}' + V_{\text{xc}}(\boldsymbol{r}) \tag{10}$$

The first term describes the aforementioned external potential $V_{\text{ext}}(\boldsymbol{r})$. The second term is the Hartree potential, accounting for the classical electron-electron Coulomb repulsion. The third term $V_{\text{xc}}(\boldsymbol{r})$ is the exchange-correlation potential that contains both the quantum mechanical exchange and correlation effects between electrons. The relationship between the exchange-correlation energy functional and potential is given by:

$$V_{\text{xc}}(\boldsymbol{r}) = \frac{\delta E_{\text{xc}}[r_0]}{\delta r_0(\boldsymbol{r})} \tag{11}$$

This term is usually approximated by various methods like LDA, GGA, meta-GGA, etc, an aspect that will be briefly discussed in the next section. The ground state density $r_0(\boldsymbol{r})$ is defined by the KS orbitals:

$$r_0(\boldsymbol{r}) = \sum_{i=1}^{N} \left|\phi_i(\boldsymbol{r})\right|^2 \tag{12}$$

if the normalization condition $\int \mathrm{r}_0(\mathbf{r})\, d\mathbf{r} = N$ is satisfied. An iterative procedure is employed to achieve total convergence when solving a specific problem within the Kohn-Sham framework. Initially, an approximate ground-state electron density $r_0(\boldsymbol{r})$ is posited. This initial guess calculates the effective potential $V_{\text{eff}}(\boldsymbol{r})$. Subsequently, the Kohn-Sham equations are solved to obtain a new $r_0(\boldsymbol{r})$. This iterative process is repeated until the electron density converges within a predefined tolerance, ensuring that the solution accurately represents the system's ground state [13]. These findings provide a computationally efficient yet sufficiently accurate framework for approximating quantum systems.

2.1.4. Exchange-Correlation Functionals

When formulating the Kohn-Sham equations, incorporating an exchange-correlation (XC) functional is indispensable for capturing the missing electron-electron interactions not accounted for by other terms. Given the current state of knowledge, the exact form of the XC functional remains unknown, requiring approximations. This has led to a distinct specialized field of research dedicated to exploring and refining XC functional approximations. Within the scope of this thesis, particular focus will be placed on 2 functionals—Local Density Approximation (LDA) and Generalized Gradient Approximation (GGA). LDA and GGA play an essential role in elucidating the nature of the dataset under investigation and will be the subject of subsequent clarification.

Local Density Approximation

LDA was proposed by Kohn and Sham in the original paper [15] published 1965.

The method employs the ground-state electron density to determine the exchange-correlation energy functional, expressed as:

$$E_{\text{xc}}[r_0] = \int r_0(\boldsymbol{r})\epsilon_{\text{xc}}(r)\, d\boldsymbol{r} \tag{13}$$

where $\epsilon_{\text{xc}}(r)$ is the exchange-correlation energy per electron and is a function of the ground state density function $r_0(\boldsymbol{r})$. This formulation approximates local and kinetic exchange densities as components of a homogeneous electron gas. The exchange-correlation term $\epsilon_{\text{xc}}(r)$ can be decomposed to exchange and correlation contributions $\epsilon_{\text{xc}}(r) = \epsilon_{\text{x}}(r) + \epsilon_{\text{c}}(r)$. The exchange part can be characterized by the Dirac expression $\epsilon_x\big(r(\boldsymbol{r})\big) = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} r(\boldsymbol{r})^{1/3}$. The correlation part is computed through interpolating quantum Monte

Carlo results for a uniform gas with respect to various densities. While this approach provides satisfactory accuracy for a wide range of problems in materials science, it needs to improve in describing strongly correlated systems, such as surface phenomena and chemical reactions, due to the rapid changes in density values.

Generalized Gradient Approximation

GGA serves as an extension to the LDA to address its limitations, particularly in systems where the electron density varies rapidly. In GGA, the exchange-correlation functional is modified to include not only the electron density $\mathrm{r_0(r)}$ but also its gradient $\nabla \mathrm{r_0}$. This inclusion allows GGA to capture more details of the electron distribution, resulting in improved accuracy for some systems.

$$E_{\mathrm{xc}}[r_0] = \int r_0(\boldsymbol{r}) \epsilon_{\mathrm{xc}}(r_0, \nabla r_0)\, d\boldsymbol{r} \tag{14}$$

One particular form of GGA that is widely applied in problems related to surface physics and was used for calculations performed within this thesis was introduced by Perdew, Burke, and Ernzerhof (PBE) [16]. In the PBE formalism, the correlation energy is expressed as follows:

$$E_c^{GGA}[r_0] = \int r_0(\boldsymbol{r}) \epsilon_{\mathrm{unif,c}}(r_0) + F_c(r_0, \nabla r_0)\, dr \tag{15}$$

where $\epsilon_{\mathrm{unif,c}}(r_0)$ is the correlation energy per electron of a uniform gas and the special $F_c(r_0, \nabla r_0)$ function accounts for the dimensionless density gradient. The exchange energy term can be written as

$$E_x^{GGA}[r_0] = \int r_0(\boldsymbol{r}) \epsilon_{\mathrm{unif,x}}(r_0)\, F_x(\nabla r_0)\, d\boldsymbol{r} \tag{16}$$

Here, $\epsilon_{\mathrm{unif,\,x}}(r_0)$ denotes the exchange energy per electron for a uniform gas, similarly to correlation component. The term $F_x(\nabla r_0)$ represents a special enhancement function characterized by a dimensionless density gradient. The GGA is the most employed functional in DFT calculations since it is superior to LDA in many aspects, such as describing molecular geometries and vibrational frequencies. It is worth noting that GGA, particularly in its PBE variant, still has problems, such as the predominantly underestimation of bond energies and the over-delocalization of the electronic structure [17].

2.2. Vienna Ab-initio Simulation Package

The Vienna Ab initio Simulation Package (VASP) is a robust computational code for solving the many-body Schrödinger equation approximately. VASP incorporates many approximation techniques, including both Hartree-Fock and DFT methodologies. Core quantities in VASP, such as one-electron orbitals, electronic charge density, and the local potential, are formulated via plane wave basis sets. The electron-ion interactions are characterized through either norm-conserving pseudopotentials or the projector-augmented-wave (PAW) method. To obtain the electronic ground state, VASP employs efficient iterative matrix diagonalization techniques, such as the Residual Minimization Method with Direct Inversion of the Iterative Subspace (RMM-DIIS) and blocked Davidson algorithms or a combination of them [18], [19].

Essentially, VASP requires the configuration of four critical input files to execute calculations:

INCAR

The INCAR file is the primary control file for VASP simulations, containing a set of key-value pairs that specify the parameters and settings for a given calculation. It controls the type of calculation to be conducted (e.g., geometry optimization, electronic structure calculation), the exchange-correlation functional to be used, and various other computational settings [20].

POSCAR

The POSCAR file provides the atomic positions and types, as well as the lattice vectors of the unit cell. This file is crucial for defining the system's geometry [21].

POTCAR

The POTCAR file contains the pseudopotentials and projector functions necessary for the plane-wave basis set expansion. It provides the information required for the electron-ion interaction within the DFT framework [22].

KPOINTS

The KPOINTS file specifies the k-point grid used for Brillouin zone sampling. The granularity of this grid can significantly affect the accuracy and computational cost of the simulation. It is configured manually or by employing automated mesh generation schemes, such as the Monkhorst-Pack and Γ-centered method [21].

In the Open Catalyst 2022 database used in this study, relaxation trajectories were generated using VASP configured for DFT calculations. Additionally, single-point calculations for the clean surfaces and the adsorbates in the gas phase aimed at acquiring their electronic structure were conducted within the VASP framework during this work. Details concerning the parameterization will be elaborated upon in the following chapters.

2.3. Open Catalyst 2022

The Open Catalyst 2022 (OC22) database results from a collaborative project between Meta AI and the Department of Chemical Engineering at Carnegie Mellon University. This publicly available database was developed to address the challenges of discovering new catalysts using machine learning methods. A detailed elaboration of the dataset and the ML approach is available in a publication [14] accessible via the project's webpage. Hence, only the main aspects of the data generation will be presented further.

The OC22 database comprises 62,331 Density Functional Theory (DFT) relaxation trajectories, equating to approximately 9,854,504 single-point calculations. These trajectories characterize atomic structures as adsorbate + slab (adslab) systems, thereby highlighting the configurational complexity inherent in oxides. Regarding material diversity, the database encompasses 51 metals configured in unary forms (A_xO_y) or binary combinations ($A_xB_yO_z$). The design philosophy behind the database centers on constructing a comprehensive and unbiased dataset, an approach that is remarkably reasonable for developing generalizable machine learning models. As a result of this approach, the atomic structures in OC22 occasionally feature atypical stoichiometry or exhibit moderate suitability for the reaction of interest, which is the oxygen evolution reaction (OER). For surface selection, the process involved using varying surface terminations derived from a randomly sampled bulk structure, constrained to a maximum Miller index of 3. Additionally, the surface may incorporate non-stoichiometric substitutions and vacancies embedded randomly.

The range of adsorbates included in the database—O, OH, H_2O, OOH, O_2, C, CO, H, N—correlates with species commonly involved in OER and associated intermediate reaction pathways. For each generated surface, several adsorbates (0-3) are randomly chosen to bind one of the three distinct types of adsorption sites: 1) undercoordinated surface metal, 2) surface oxygen atom, and 3) oxygen vacancies. Figure 1 provides a visualization of these adsorbate placement strategies.

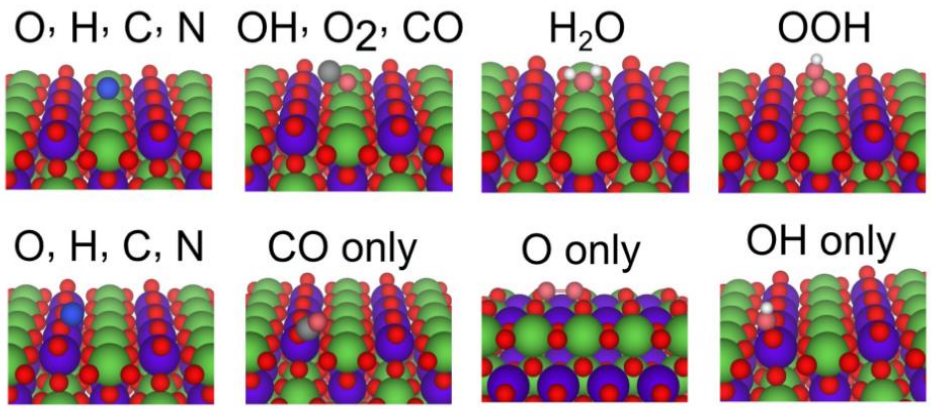


Figure 1 OC22 adsorbate placement strategies. The top row depicts the binding of different adsorbate types to a surface metal. The bottom row shows adsorption through a surface oxygen atom [10].

All DFT calculations for the OC22 database were performed using the VASP code via the Projector Augmented Wave (PAW) methodology. Distinctively, the primary attribute assigned to each entry in OC22 is the total DFT energy, setting it apart from its predecessor, OC20, which focuses on adsorption energy values. The Generalized Gradient Approximation (GGA) was used to model exchange-correlation effects according to the Perdew-Burke-Ernzerhof (PBE) formulation. Calculations incorporated spin-polarization to address the relevance of significant spin states in metal oxides. The Hubbard U correction was also applied in calculations involving certain transition metals, providing a more accurate representation of strong electron-electron interactions. A hybrid algorithm combining the blocked Davidson iteration and Residual Minimization Method-Direct Inversion in the Iterative Subspace (RMM-DIIS) was employed as the default scheme for the energy minimization process. An additional electrostatic potential was introduced as a corrective measure to compensate for the non-physical dipole moment generated by uneven charge distribution between two surfaces— a consequence of adsorbate placement.

The authors explored an approach that utilizes Graph Neural Networks (GNN) for examining the geometrical optimization of atomic systems, with the ultimate aim of determining the system's total energy. In contrast, this thesis focuses on a distinct methodology that analyzes specific adsorption sites and their correlation with adsorption energy. Within the framework of this thesis, the OC22 database serves as an essential resource, as it is used as the data source for the relaxed atomic structures, which are then used to compute the adsorption energy of selected adsorbate/oxides during this master thesis.

2.4. Machine learning

Machine learning algorithms offer a powerful framework for extracting generalizable insights from data, enabling systems to improve their performance based on experience. These algorithms encompass various techniques, ranging from supervised and unsupervised learning to reinforcement learning [23].

2.4.1. Scaling relationships and descriptors

Before delving into machine learning algorithms, it is crucial to present the concept of descriptors and scaling relationships and their role in predicting adsorption energies. The advent of advanced computational tools, such as VASP, has enabled detailed modeling of reactions and the approximation of solid bodies' electronic and geometric properties. Focused investigations of the electronic structure of catalysts, the binding strength of intermediates, and various geometric properties have resulted in the discovery of scaling relationships. They allow the selection of a few specific material properties, named "reactivity descriptors," that correlate with the surface reaction energies and the activation barriers. The significance of reactivity descriptors lies in their ability to simplify complex kinetic models and effectively predict trends of the catalytic performance. A comprehensive review paper [1] about the scaling relationships and up-to-date discovered descriptors was published by Zhao et al. in 2019.

Descriptors can be classified into:

- Electronic descriptors

Electronic descriptors are associated with the electronic properties of a material, such as the density of states (DOS), Bader charges, the position of the *d*-band center, etc. They help capture the electronic interactions between atoms of the system.

- Structural descriptors

Structural descriptors encompass materials' geometric and topological properties, such as coordination number, surface facet orientation, bond lengths, and bond angles. These descriptors are crucial for understanding how the structure of a catalyst affects its reactivity and stability.

- Combined descriptors

Combined descriptors unite both electronic and structural properties through a mathematical function. An example of such a descriptor is the bond-energy-integrated orbital-wise coordination number [24]. By encapsulating both properties, they offer a more nuanced representation of the correlations between material properties and catalytic performance.

It is important to note that descriptors are specific to a single adsorption site, usually accounting for atoms involved in the adsorption process or their neighboring atoms. Descriptors are not limited to the catalyst surface but can represent the adsorbate. It is also recognized that different materials exhibit different quantum mechanical mechanisms. Hence, descriptors are generally valid within one material class or a single reaction (ORR or OER).

2.4.2. Descriptor-based ML approach in heterogeneous catalysis

After understanding the concept of descriptors, a generalized approach for creating a machine learning model for an application in heterogeneous catalysis will be discussed. Adsorption energy is a pivotal parameter in evaluating the effectiveness of a catalyst. It provides critical insights into the catalytic activity since a correlation between the two has been established [25]. Within the scope of heterogeneous catalysis, our focus is mainly on designing supervised learning models to predict the adsorption energies of selected adsorbates on oxides. In this approach, each data entry is annotated or "labeled" with a specific value representing the property of interest - the adsorption energy. The general procedure for developing a descriptor-based machine-learning algorithm can be divided into the following steps [23]:

Step 1: Data Construction

The first step entails the construction of a comprehensive database. In heterogeneous catalysis, both the catalyst surface (specifically, the adsorption site) and the adsorbate are characterized by a feature vector. This vector aggregates a range of electronic and geometric descriptors of the clean material, whether the slab or the adsorbate. The data for these descriptors can be sourced from experimental measurements or computational methods, such as Density Functional Theory (DFT), and then systematically organized into a dataset.

Step 2: Descriptor Identification

Given the high dimensionality of the input feature space, it is essential to pinpoint the descriptors most strongly correlated with the adsorption energy. This involves a multi-faceted approach ranging from reviewing existing scientific literature to employing advanced statistical techniques like Sure Independence Screening and Sparsifying Operator (SISSO) [26]. This step aims to refine the existing feature set, making it relevant and manageable for subsequent model training. It is important to note that the choice of the descriptors is strongly dependent on the catalyst and adsorbates under investigation.

Step 3: Modeling and Validation

The final step involves training the predictive machine learning model using the refined dataset. Various algorithms, such as linear regression, support vector machines, or ensemble methods like Random Forest, can be explored and compared at this stage. Model fine-tuning also encompasses the optimization of hyperparameters, and validation is usually performed using either experimental data, computationally derived data, or statistical techniques like cross-validation. The latter partitions the initial dataset into training and validation subsets to assess the model's generalizability.

This workflow is depicted in Figure 2 and is designed to be iterative, allowing for continual improvement of the models by revising and/or expanding the dataset under investigation.

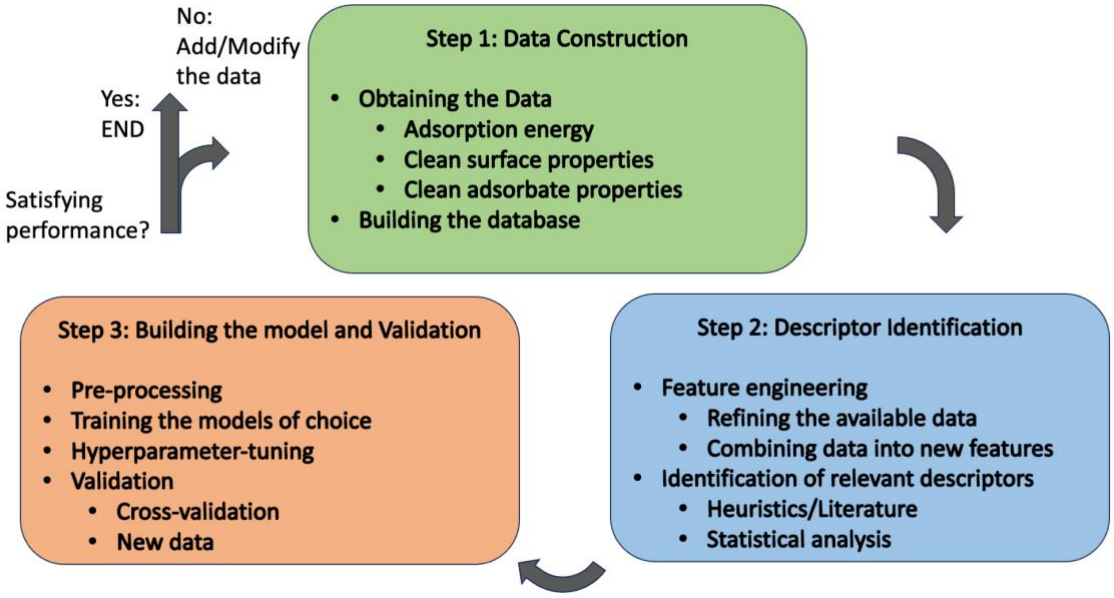


Figure 2 Workflow for the descriptor-based machine learning model training in heterogeneous catalysis

2.4.3. Machine learning algorithms

This section briefly overviews the machine-learning techniques relevant to this thesis. As discussed earlier, the approach employed here relies on supervised learning methods, which are characterized by assigning specific 'labels' to each data point. These labels represent the property of interest — called Y in data science. In the context of this project, the 'label' is the computationally demanding property: adsorption energy. Moreover, each data point comprises a feature vector X, representing the selected descriptors — properties that are easy to calculate and define the adsorbate + slab system, mainly the active site. Machine learning models are developed to explore the correlation between the feature vector (X) and the label (Y), enabling the prediction of Y for a given X. Before the training process, the data is split into the training set (used for shaping the model) and the test set (used for the performance evaluation). The model's performance is characterized by the difference between the predicted and actual values, usually computed for several given systems (test set). The commonly used performance metrics are the Mean Average Error (MAE) and Mean Squared Error (MSE) for training and the test splits. Generally, selecting an appropriate machine learning model for a given task is inherently heuristic. This process typically involves a trial-and-error approach, primarily due

to the “black box” nature of these models, which can yield unpredictable performance metrics [27]. We will now proceed to the discussion of the ML models used in the present thesis, i.e., the CatBoost [28], the XGBoost [20], and the Kernel Ridge Regression (KRR) [30].

CatBoost and XGBoost – Decision Tree ensembles and Gradient Boosting

CatBoost and XGBoost models are based on Decision Tree (DT) ensembles using the Gradient Boosting (GB) technique, i.e., DTGB, which can tackle regression and classification problems. A DT is a flowchart-like model in which an internal node represents a feature, the branch represents a decision rule, and each leaf node represents the outcome. The Gradient Boosting workflow (Figure 3) within decision tree ensembles includes the following steps [31]:

1. Base model

GB starts with a base model (e.g., a single decision tree) representing the initial prediction, for example, the mean of the target values.

2. Residual calculation

The model calculates the residuals (the difference between the initial predictions and actual values) and builds a new decision tree to predict the residuals. Therefore, the loss function at the n-th tree is generally given by:

$$L_n = \sum_i \frac{y^i - \hat{y}^i}{2} \tag{17}$$

where L_n is the loss function, y^i and $\hat{y}^i$ the true and predicted values, respectively.

The process is then repeated, with each new tree being fitted to the residuals of the ensemble of the cumulative previous trees. The loss function minimization occurs along the following gradient:

$$y^i - \hat{y}^i = -\frac{\partial L_n}{\partial p_n(X)} \tag{18}$$

with $p_n(X)$ as the predicted values dependent on the training data X.

3. Final model

The final model consists of hundreds to thousands of trees, all contributing to the predicted output value.

Additionally, DTGB includes three regularization parameters: the learning rate (scaling factor for a new iteration), the depth of the tree, and the number of trees. Gradient Boosting Decision Trees perform well for small and noisy datasets with non-linear feature-label correlations [31], [32]. The algorithm specifics for the open-sourced CatBoost and XGBoost models can be found in [33] and [25], respectively.

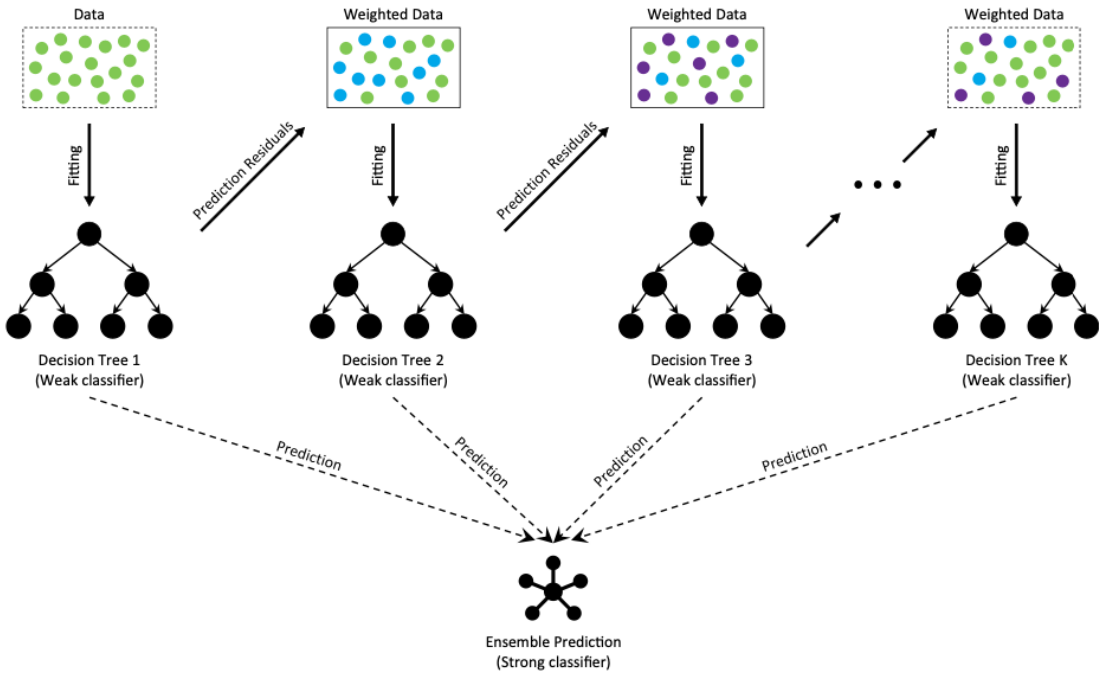


Figure 3 The general workflow of the Gradient Boosting Decision Tree algorithm [32]

Kernel Ridge Regression

Kernel Ridge Regression is a supervised learning technique that combines Ridge Regression (linear least squares with an L2-norm regularization) with the kernel trick. Ridge Regression modifies simple linear regression by adding an L2 penalty on the size of the coefficients, which is the sum of the squares of the coefficients multiplied by a parameter λ. The kernel trick is a mathematical technique that enables mapping non-linear input data into a high dimensional space, where a linear model is applied. The kernel function then calculates the inner products between the images of all data pairs in the feature space. In KRR, this principle is applied to Ridge Regression, allowing it to perform linear regression on non-linear data in the higher-dimensional feature space without ever computing the data coordinates but instead of calculating inner products between the images of the data in this space. The loss function of the KRR is given by:

$$J(\boldsymbol{w}) = \sum_{i=1}^{n} (y^i - \sum_{j=1}^{n} \alpha_j \, K(x_i, x_j))^2 + \lambda \sum_{j=1}^{n} \alpha_i \alpha_j \, K(x_i, x_j) \tag{19}$$

where α_i and α_j represent the dual coefficients and $K(x_i, x_j)$ is the kernel function [35].

The performance of the KRR algorithm for a specific dataset is highly dependent on the hyperparameters and the choice of the kernel function that accounts for the non-linear relationships. The commonly available kernel functions in the scikit-learn KRR implementation [30] are

- Linear function

$$\gamma \langle x_i, x_j \rangle + r \tag{20}$$

This kernel function is the linear dot product between two points, being equivalent to no kernel function implemented. γ and r represent kernel coefficient terms:

- Polynomial function

$$(\gamma \langle x_i, x_j \rangle + r)^d \tag{21}$$

In this kernel function d is the degree of the polynomial.

- Radial basis function

$$\exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \tag{22}$$

- Sigmoidal function

$$tanh(\gamma \langle x_i, x_j \rangle + r) \tag{23}$$

The KRR algorithm is known to be robust against overfitting, particularly in cases with a high-dimensional feature vector and a small number of dataset entries. It offers high flexibility via parameterization of the kernel function.

3. Results and Discussion

This chapter provides a discussion of the results achieved and a comprehensive description of the steps taken in the research. Each step, from the data gathering to the model training, was performed in a distinctly configured Python 3.9.16 environment. The software specifics will be provided to ensure transparency and to illustrate the path leading to the presented results.

The geometrical structure data used in this research arises from the OC22 dataset. This choice was made because the OC22 dataset provides essential information on relaxed structures' atomic positions and system energy. By utilizing this database, I eliminated the need for several time-consuming and energy-demanding steps:

- Modelling of the oxide bulk structures
- Identifying the most stable surface and surface termination per oxide (oxide surfaces tend to exert unstable behaviors [10])
- Screening the most stable adsorption sites (either Metal and Oxygen atoms can provide adsorption sites)
- Computing the relaxation trajectory, which is connected with high energy demand

This approach, with its limitations, enabled me to build my database for investigating the hypothesis of the descriptor-based ML – approach on oxide surfaces. However, OC22 lacked data regarding the electronic properties and the system energy of clean slabs and adsorbates. Thus, during this thesis, additional DFT Single Point calculations under the settings provided by the OC22 research group [10] were performed to fill this gap. The results were analyzed and subsequently integrated into a database, encompassing geometrical and electronic data about the structures under study. The following step entailed identifying and examining adsorption sites and formulating a comprehensive set of descriptors. The refined data was ultimately employed to train and evaluate various machine learning models, including Kernel Ridge Regression (KRR), XGBoost, and CatBoost algorithms, to predict adsorption energies.

3.1. Dataset construction

3.1.1. Data selection

The core information, including the geometrical data and total system energy, was retrieved from the OC22 GitHub webpage that provides 56 367 raw trajectories (~86 GB) stored as .traj [36] files. The dedicated Python environment included the following libraries required for operating with atomic structure data:

-) Atomic Simulation Environment (ASE) [37]

-) Pymatgen [38]

Given the diverse characteristics of the OC22 dataset, specific criteria were established to ensure a uniform description of the active sites. These criteria included:

- Only one adsorbate per surface

- Exclusion of combined oxides

- Absences of substitution defects

- Focus on chemisorption phenomena

The initial scope of evaluated materials included 11 systems: Al-O, Co-O, Cr-O, Cu-O, Fe-O, Mn-O, Ni-O, V-O, W-O, Zn-O, and Zr-O. The configuration for the VASP input files, while only provided in plain text format, required validation according to the database results. The validation metric was the total system energy. Consequently, a single-point VASP calculation was performed for a randomly chosen structure from each material category.

Upon completing the validation step, a decision was made to narrow the focus to Al-O, Zr-O, and Ti-O systems. This decision was driven by the following considerations, ensuring a balance between generalizability and resource constraints:

- **Computational Cost**

During validation, computational times varied significantly, ranging from 4 to 18 hours. This variation was attributable to the high atom count and challenging convergence behaviors in specific systems, leading to a considerable computational cost. Given the limited resources available for this project, these costs restricted the number of feasible calculations.

- **Space-Intensive Output Data**

Each calculation generated substantial data, occupying between 1.7 GB and 3.8 GB of storage space. This requirement necessitated careful consideration of the available memory capacity of the workstation.

- **Hubbard U Correction**

For oxides involving elements Co, Cr, Fe, Mn, Mo, Ni, V, and Wo, adding the Hubbard U correction was necessary to accurately represent the interactions of strongly correlated electrons. Additionally, these oxides often exhibit magnetic polymorphism, which necessitated specifying unique magnetic moments for each structure, further adding to the complexity of the analysis. Hence, oxides involving the aforementioned materials were not included in the dataset.

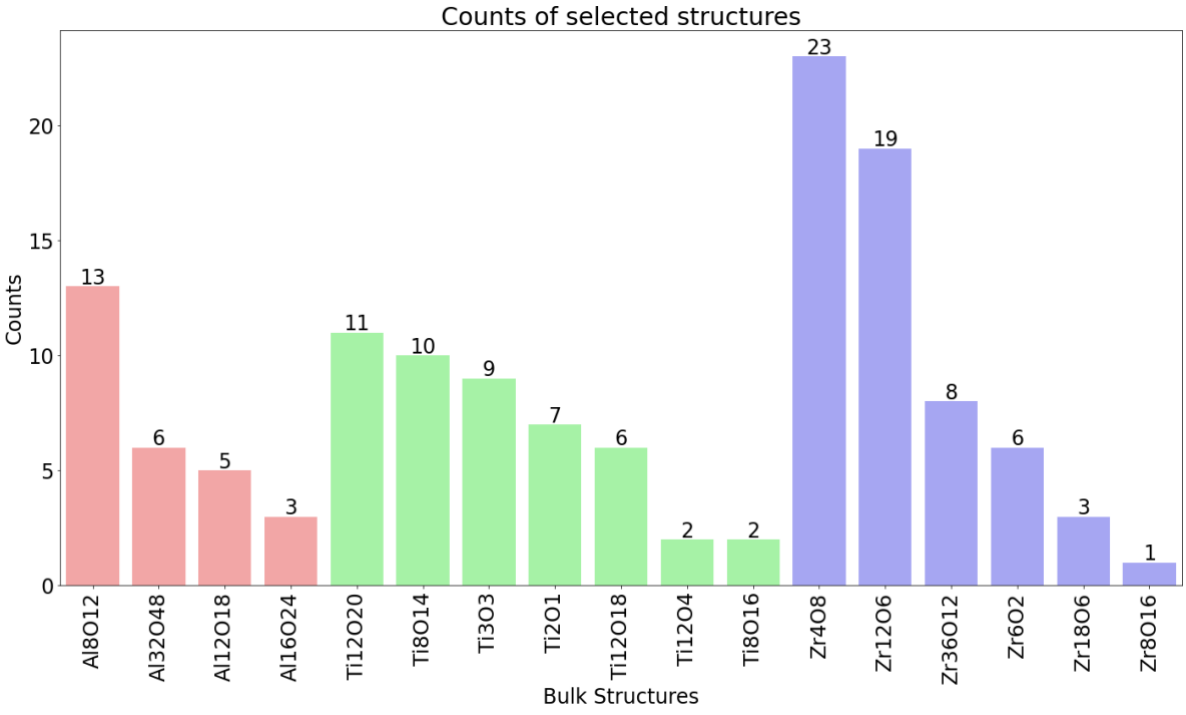


Figure 4 Distribution of the selected bulk structures. Al-O in red; Ti-O in green; Zr-O in blue

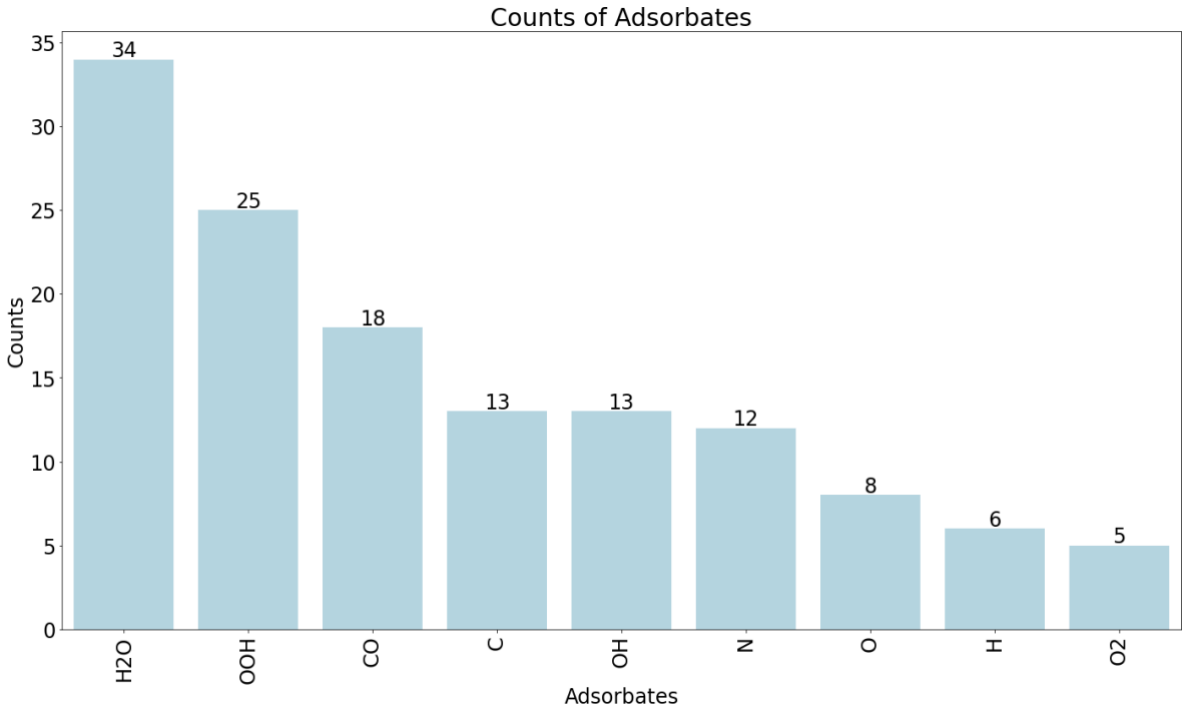


Figure 5 Distribution of the involved adsorbates in the selected subset

Subsequently, 134 systems from the OC22 database were considered for further analysis. Figure 4 and Figure 5 depict the distribution of the new dataset entries regarding the slab's chemical formula and the adsorbate. Firstly, a significant diversity of bulk structures within each material class is evident. This diversity originates from the specific settings of the **bulk selection process** [10], which prioritized structure diversity over its electrochemical stability, resulting in a wide range of bulk structures within each material class. Secondly, the dataset exhibits a disproportionate distribution of both adsorbates and bulk structures. The OC22 publication indeed reports preferences towards certain adsorbates and structure types. However, they diverge from the patterns observed in this dataset, most likely as a result of statistical variances. Such occurrences might introduce a bias into the model and subsequently affect the model's performance on unseen data [6].

3.1.2. Obtaining the missing data

The next critical step involved creating our dataset by calculating essential yet missing details like the electronic structure of the clean slab, its total energy, and the energy of the adsorbate. This required performing an extra DFT calculation for each entry, covering 134 clean slab structures and 9 adsorbate ones.

To streamline this process, I developed a program to automate the setup of VASP input files. This program primarily extracted atomic positions from trajectory files, removed the adsorbate or the slab as needed and then generated the POSCAR file with these changes.

One challenge was that the initial trajectory data did not differentiate between atoms in the bulk, on the surface, or in the adsorbate. However, the OC22 project provided an additional .lmdb file (pre-processed database for deep learning models) containing a specific tensor that helped identify whether an atom was part of the bulk, surface, or an adsorbate. Using a metadata file that linked specific trajectory identifiers with OC22 identifiers (sid), the pre-processed data was connected with the raw trajectories, and the "roles" were assigned to the atoms. Following this approach, other input files were automatically generated, specifically KPOINTS, INCAR, POTCAR, and vasp_run.slm. The VASP configuration matched the settings validated in the previous step. The k-point values in the KPOINTS file were tripled for better accuracy in the electronic properties. Details regarding the VASP configuration may be found in the appendix of the OC22 publication [10]. Additionally, the program identified the materials involved and automatically created a POTCAR file, which included the necessary pseudopotentials. Figure 6 depicts the workflow iteratively applied over each data point to generate the configurational files for VASP.

All DFT calculations ran on the CSUC cluster [39] using the vasp/5.4.4 module. 132 out of 134 successfully converged, and the successful output files were then analyzed with the help of the following tools:

- Code: Bader Charge Analysis [40]

This program, provided by the University of Texas at Austin, is designed to perform the Bader charge analysis on the raw VASP output files. Bader charges, representing an artificial method to distinguish atoms within a system based solely on the electron density, demonstrated potential as a descriptor [9].

- VASPKIT [41]

A powerful post-processing tool for VASP output files was used to calculate and extract the electronic properties, such as occupation number, spin density, and electronic density of states (DOS).

- pymatgen.io.vasp.outputs [38]

This tool is an integral part of the Python pymatgen package that allows reading VASP output files and directly extracting the information of interest.

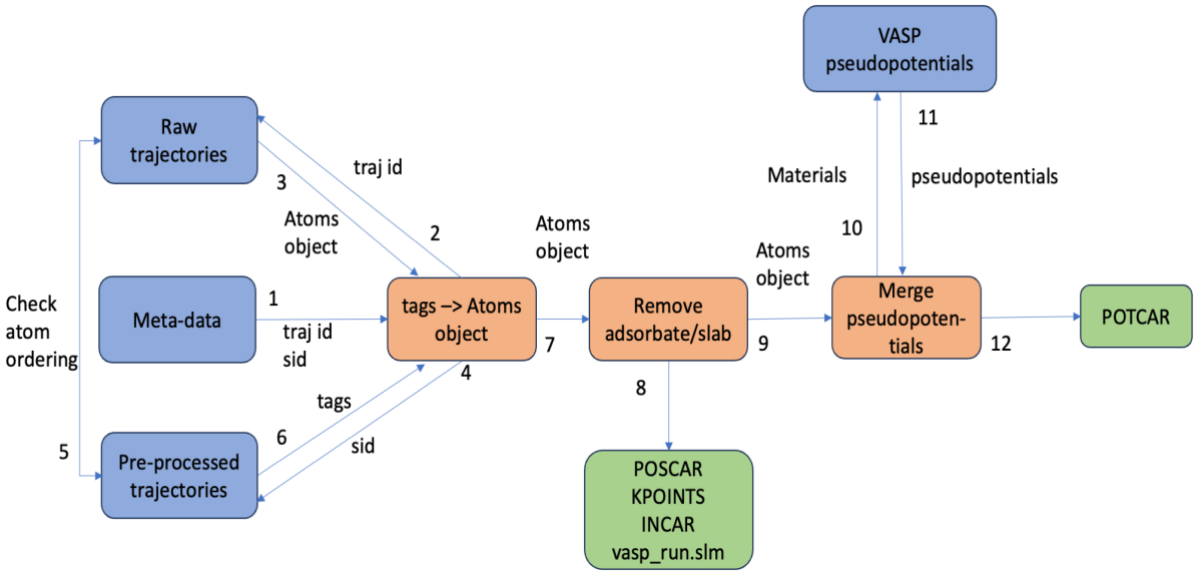


Figure 6 Automatic VASP configuration files generation workflow

3.2. Dataset details

3.2.1. Collected data

In this section, I'll delve into the details of the data collected, which formed the basis for our machine learning dataset, as shown in Figure 7. Firstly, I collected data on the total energy of three different setups: the combined adsorbate and slab system, the clean slab, and the adsorbate. This data is crucial for calculating adsorption energy (as detailed in Section 3.2.4) and serves as our machine-learning models' primary output variable. Data on atomic positions, cell parameters, and the specific roles of atoms in the system (slab, surface, or adsorbate) was gathered regarding geometric information. Finally, the electronic data consists of the electronic structure and additional information on Pauling electronegativity and electron affinity.

Analytical and processing steps are essential to transform the collected data into a dataset suitable for machine learning applications. The initial phase involved accurately identifying the active site within each system. After this step, the selected features are extracted/calculated for each data point and merged into a comprehensive machine-learning dataset.

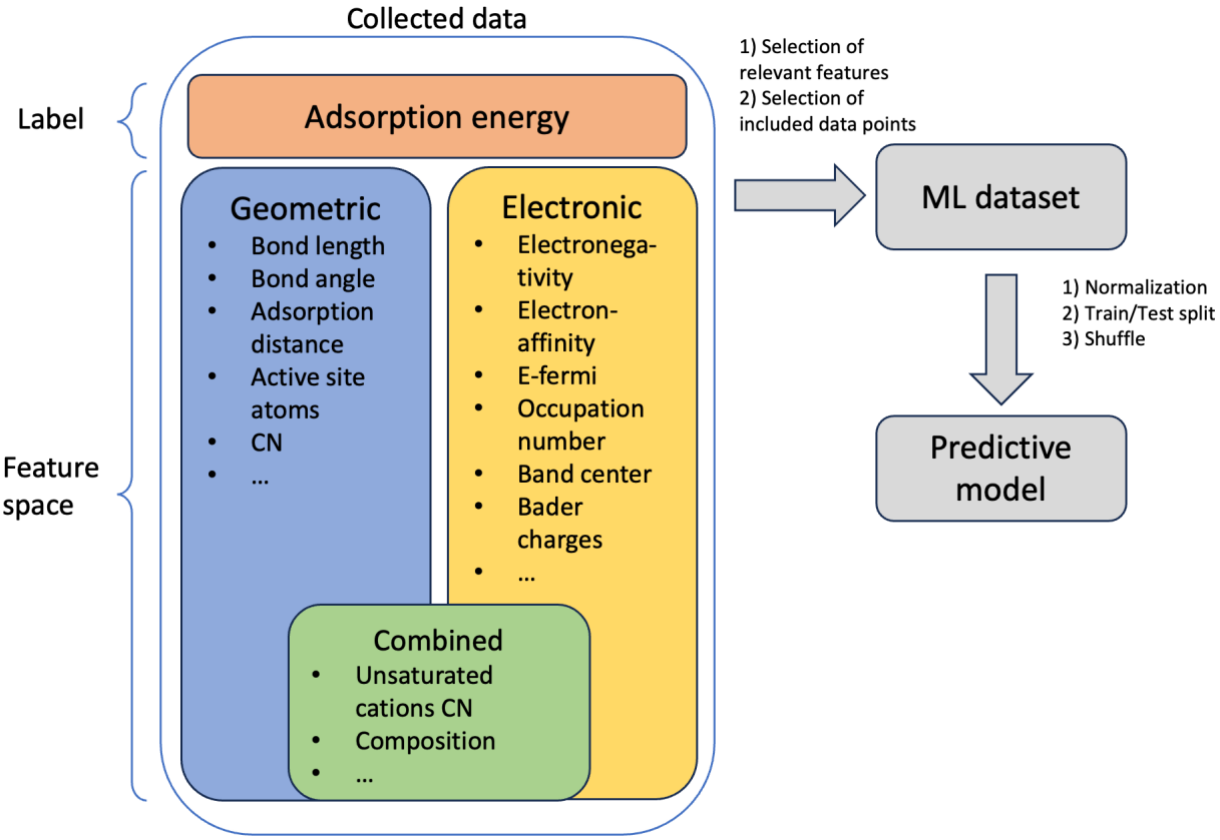


Figure 7 Illustration of the database contents and the workflow developed for predictive model training

3.2.2. Representation of the active site

Descriptors are always associated with a specific active site, and they characterize unique properties of the atoms involved in adsorption or their neighboring atoms. My initial task was to create a script that could automatically and accurately identify the active site and its kind or type to account for this important aspect. Moreover, further complexity arises when including adsorbates like H_2O and OOH; often, an atom (H or O) from the molecule detaches and autonomously adsorbs on a different site. This different site can correspond to another metal (Me) atom or an oxygen (O) atom, particularly in cases where H splits off. Additionally, it's important to note that each active site can be composed of 1 to 4 atoms, as shown in Figure 8.

These factors result in a wide variation in the number of surface atoms involved in adsorption, typically ranging from 1 to 8. Given that a model requires a consistent feature vector format (consisting of numerical values) across the entire dataset, it became clear that an efficient method for describing active sites was needed. To address this, I

incorporated two active sites into the feature vector. For each site, I assigned a descriptor that indicates the number of participating surface atoms. Other parameters, such as bond length, Bader charge, occupation number, etc., were averaged across all involved atoms and attributed to each active site. In cases where a system has only one active site, all values related to the second site were set to zero. This strategy significantly reduced the feature vector's size, lowering the risk of overfitting [42].

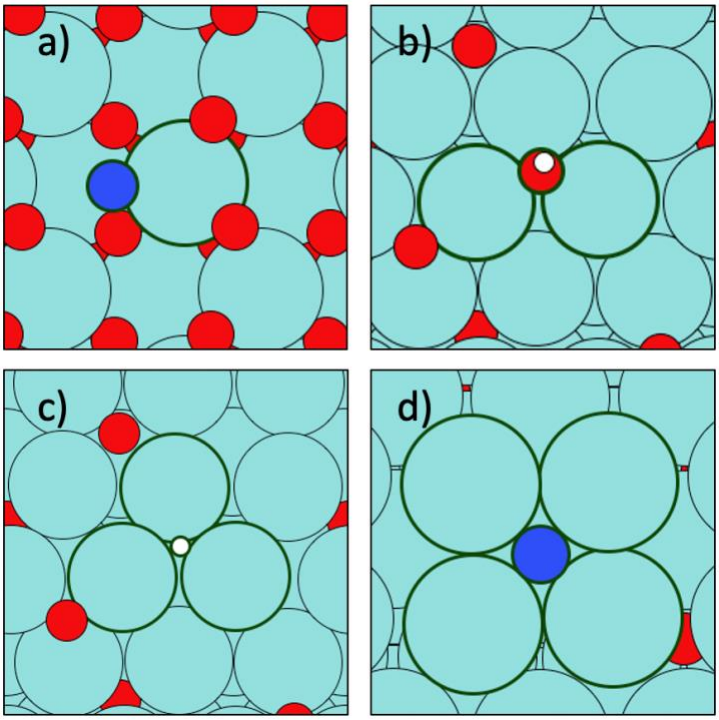


Figure 8 Different types of active sites a) N on a Zr-O surface. The active site consists of one Zr surface atom. b) OH on a Zr-O surface. Two surface atoms characterize the active site. c) H on a Zr-O surface with 3 participating surface atoms characterizing the active site. d) N on a Zr-O surface. The active site involves 4 surface atoms.

3.2.3. Feature vector

Figure 7 illustrates that the feature space comprises geometric and electronic data relevant to the chosen structures. In a technically rigorous interpretation, the feature space includes all potential descriptors derived from the available geometric and electronic data. While primary geometric and electronic data offer a finite set of possible descriptors, more complex descriptors that integrate various properties using mathematical functions create a vastly large (theoretically infinite) feature space. In this research, my focus was solely on descriptors that have been documented or suggested in the scientific literature [1], [9],

[23], [26] due to the novelty of the overall task and the resource constraints. Findings of 1) quality assessment of the underlying data, which included scrutinizing its scientific coherence and comparing it to the available reference values from similar systems, and 2) performance evaluation of multiple predictive models, the properties outlined in Table 1 were integrated into the feature vector. Atom-specific Adsorbate and Surface properties characterized each active site according to Section 3.2.2. For systems comprising a single active site, the feature vectors assigned a value of 0 to parameters related to a hypothetical second active site.

Surface		Adsorbate
Number of participating atoms	*Electronegativity*	*Fermi energy*
Coordination number	*Electron affinity*	*p-occupancy (up/down)*
Unsaturated cation number	*p-occupancy (up/down)*	*Bader charge*
Fermi energy	*d-occupancy (up/down)*	*Electronegativity*
Distance to neighbors	*Bader charge*	*Electron affinity*

Table 1 Descriptors implemented into the feature vector

The exclusion of the Band Gap as a descriptor in this study is due to the specific limitations of the PBE GGA functional used in the work, which does not accurately compute band gaps [43]. Therefore, the evaluation of band gaps with a more accurate DFT method was not carried out due to resource constraints.

3.2.4. Adsorption energy

Adsorption energy, which is the target quantity of this study, was not directly provided by the OC22 dataset. Additional DFT computations were required to obtain data on clean slabs and adsorbates. The formula outlined in Equation (24) was employed to obtain the adsorption energy:

$$E_{adsorption} = E_{adslab} - E_{slab} - E_{gas} \tag{24}$$

Adsorbate	Standard deviation σ [eV]	Range [eV]
CO	1.20	3.61
H_2O	1.55	7.08
OOH	3.76	11.70
C	1.78	5.56
O	2.35	7.17
OH	1.34	4.53
O_2	3.11	9.21
N	3.16	8.69
H	1.12	3.23

Table 2 Standard deviation of adsorption energies across adsorbates

Statistical examination of the derived values in Table 2 indicates that three adsorbate types—OOH, O_2, and N—display a notably high standard deviation (> 3 eV) and the widest value range. This phenomenon could be attributed to the evidenced tendency of the GGA functional to overestimate the O-O bond strength [10], explaining these high deviations for OOH and O_2. Further supported by the observation of significant enhancement in ML-model performance upon their exclusion, it was decided to remove all entries related to these two adsorbates from the dataset, resulting in a refined dataset comprising 103 data points. Data points classified as outliers in systems involving N, as shown in Figure 9, were closely inspected. This examination failed to reveal anomalies other than the adsorption energy value, indicating the erroneous computation.

The plot in Figure 9 illustrates the distribution and the range of adsorption energies across the adsorbates.

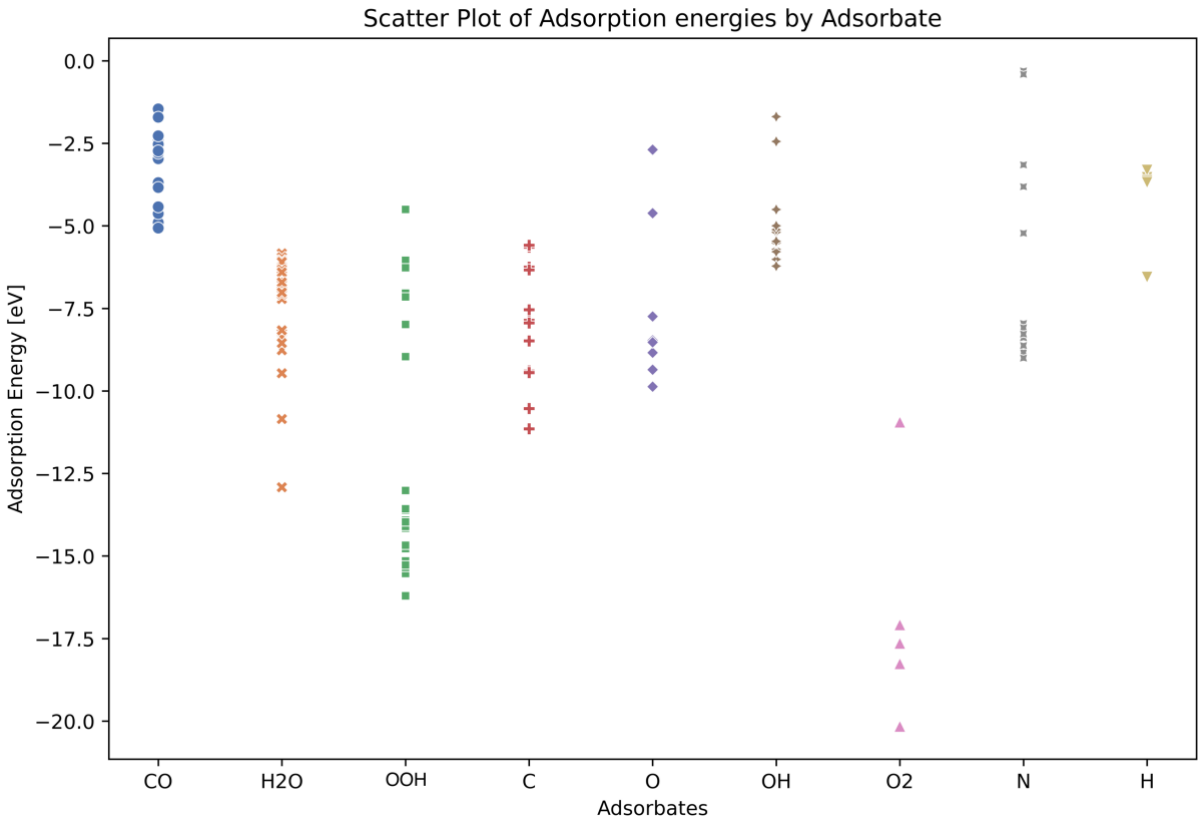


Figure 9 Adsorption energies distribution grouped by Adsorbate type

3.3. Model training

The final dataset comprised 103 data points, which is assumed to be insufficient considering the complexity of the investigated catalytic systems. The presence of "outliers" in Figure 9 might indicate potential computational errors and the high diversity of QM interaction mechanisms, which the current dataset does not cover adequately. The ability of these models to generalize to these catalytic systems could be deprecated, hence the overall model performance. In addition, the models are only trained on a limited number of active sites, which are not always consistent in oxide surfaces [10].

3.3.1. Additional model information

The efficacy of the machine learning model is determined not merely by the selection of an appropriate algorithm or the integrity of the employed data. Still, it is also influenced by the strategies of data pre-processing, hyperparameter optimization, and evaluation methodologies. This chapter outlines the approaches to refine and enhance the model's performance through the work,

Train/Test split

The Test set size was configured to 20 % of the whole dataset, resulting in 21 data points in the Test set and 82 data points in the Train set.

RobustScaler [44]

This scaler is instrumental in standardizing and normalizing the feature values to ensure a balanced contribution of each feature to the model's internal weights, thereby enhancing the model's predictive performance. It is particularly adept at managing outliers by considering the median value over the mean, making it a robust data normalization tool. Additionally, the features are randomly shuffled to reduce a potential bias arising from the selected ordering of the features.

K-Fold [45]

The K-Fold cross-validation technique is crucial for assessing machine learning models, especially when dealing with limited datasets. It involves partitioning the training data into 'k' subsets (with k=5 in this study). During each round of validation, a single subset is retained as the validation data for testing the model, while the remaining subsets are used for training. This approach thoroughly assesses the model's performance and enhances its generalizability and reliability.

Hyperparameter Tuning

In my research, I employed a brute-force parameter search methodology. This technique involves an exhaustive iteration over a vast parameter space to identify the optimal set of parameters. The evaluation metrics for hyperparameter tuning include MAE and MSE of the train and test sets

3.3.2. Performance of ML models

	MAE Train [eV]	MAE Test [eV]	MSE Train [eV]	MSE Test [eV]	Accuracy	Standard deviation
CatBoost	10^{-5}	0.883	10^{-9}	0.946	0.478	0.181
XGBoost	0.0005	0.931	10^{-6}	1.690	0.330	0.246
KRR	0.0922	1.146	0.021	2.388	0.290	0.237

Table 3 Performance metrics summary of the investigated ML algorithms: Mean Average Error (MAE) and Mean Squared Error (MSE) of the Train and Test splits, Kfold (k=5) accuracy, and K-fold deviation.

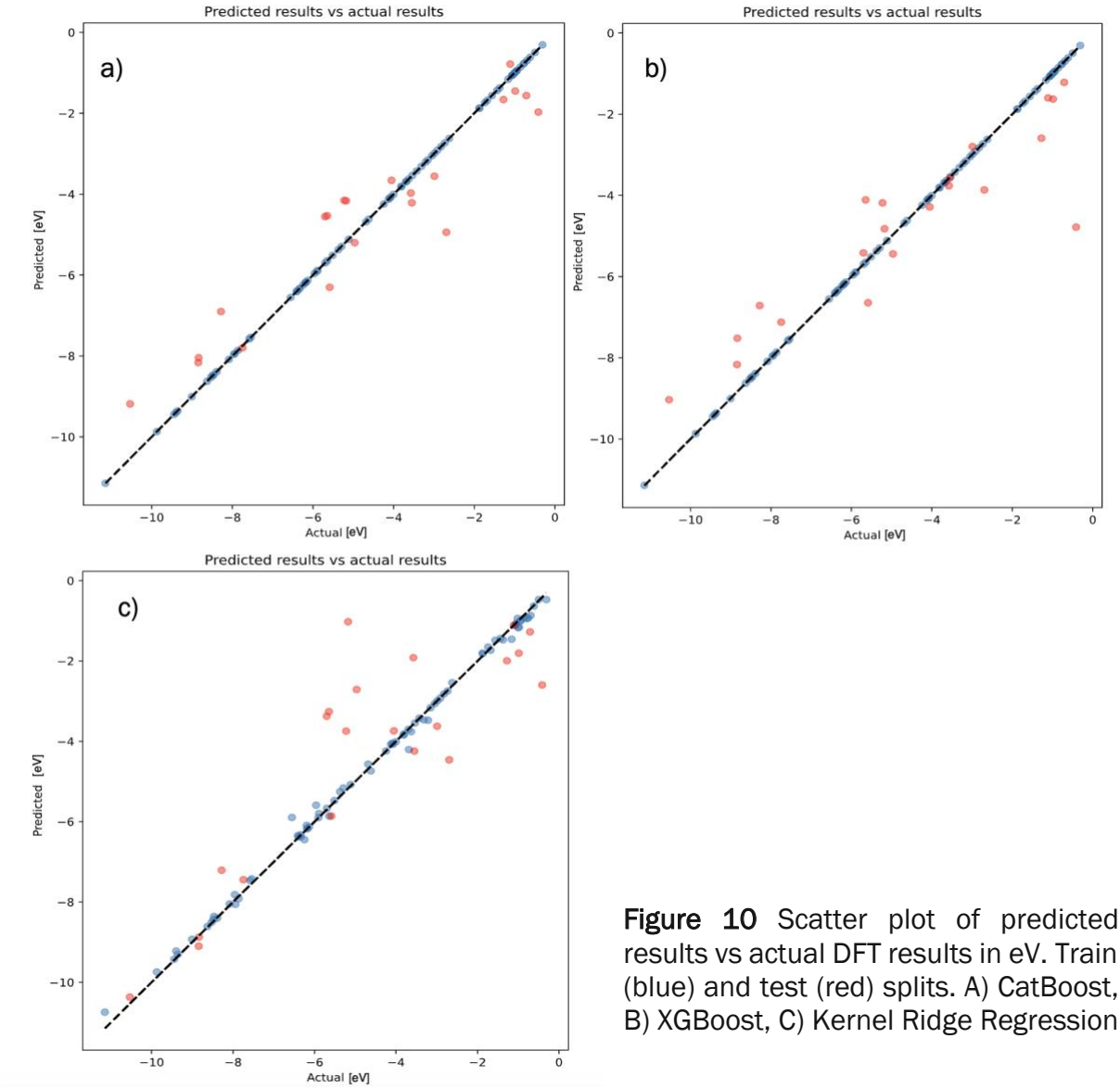


Figure 10 Scatter plot of predicted results vs actual DFT results in eV. Train (blue) and test (red) splits. A) CatBoost, B) XGBoost, C) Kernel Ridge Regression

Table 3 provides a detailed summary of the performance metrics for the investigated machine learning algorithms after extensive hyperparameter tuning. The metrics include Mean Average Error (MAE) and Mean Squared Error (MSE) for both the training and testing datasets, along with a K-fold (k=5) cross-validation score and standard deviation of errors across folds. The primary evaluation metric is the predictive model capacity based on the MAE of the Test set. Figure 10 shows scatter plots of the predicted results against each model's actual DFT calculated adsorption energy values.

CatBoost

CatBoost demonstrates the MAE on the training set of practically 0 eV. Still, a significantly higher MAE on the test set (0.883 eV) points at an overfitting scenario where the model

has learned specific details of the training set, including noise and outliers from the training data. Despite this, its performance on unseen data is superior to other models, producing the lowest MAE and MSE on the test set and the highest k-fold accuracy of 0.478.

XGBoost

XGBoost exhibits higher error rates (MAE of 0.931 eV on the test set) and inferior robustness to different datasets compared to CatBoost. Notably, the substantial difference between the MAE and MSE indicates the presence of outliers with significant error margins. The difference in the error magnitude between the Train / Test splits also suggests overfitting.

Kernel Ridge Regressor

KRR shows the highest MAE on the test set (1.107 eV) and the lowest cross-validation accuracy (0.291). The gap in error rates between Train and Test splits implies overfitting, as was the case with the previous models.

3.3.3. Discussion of the results

All models exhibit a significant error difference between training and testing errors, indicating overfitting. This result suggests that while the models learn the training data well, they struggle to generalize to new data effectively. Similarly, the cross-validation metrics indicate relatively poor stability and robustness across multiple datasets. The overall performance suggests the patterns in the underlying dataset exist but could not be fully captured. Several potential factors could result in this behavior:

Insufficient training data

For the task at hand, the dataset comprised 103 data points after excluding OOH and O_2 structures. Given the intrinsic complexity of heterogeneous catalysis processes involving metal oxides, this dataset is limited. The implication is that the dataset does not adequately represent the wide variety of possible systems and certainly does not reflect the whole variety of facets originally conceived in OC22. The related study by Praveen and Comas-Vives [9] on metallic surfaces, which included 623 data points, was less susceptible to overfitting. That study investigated metallic surfaces that were geometrically somewhat similar, modeling phenomena of lesser complexity. Here, the geometric differences between the considered oxides are much more significant. Moreover, that study evidenced an enhancement in the test set performance when increasing the set size to a threshold

of approximately 400 data points. It is assumed that the limited size of the dataset in the present study is a significant contributor to the overfitting observed in the models.

Underrepresentation of certain structures

The dataset exhibits a shifted distribution favoring specific structures and adsorbates, as highlighted in Section 3.1.1, where particular structures are only represented by one or a few data points and others by double-digit numbers. This overrepresentation can inherently introduce bias into the algorithm's performance. The prevalence of specific structures and adsorbates might influence the model's weights and limit its ability to generalize to less common but valid scenarios. The substantial difference further supports this assumption, as observed in the MSE and MAE results of the test sets. Higher MSE, which is, by definition, more sensitive to outliers, implies that some predictions had a significant margin of error.

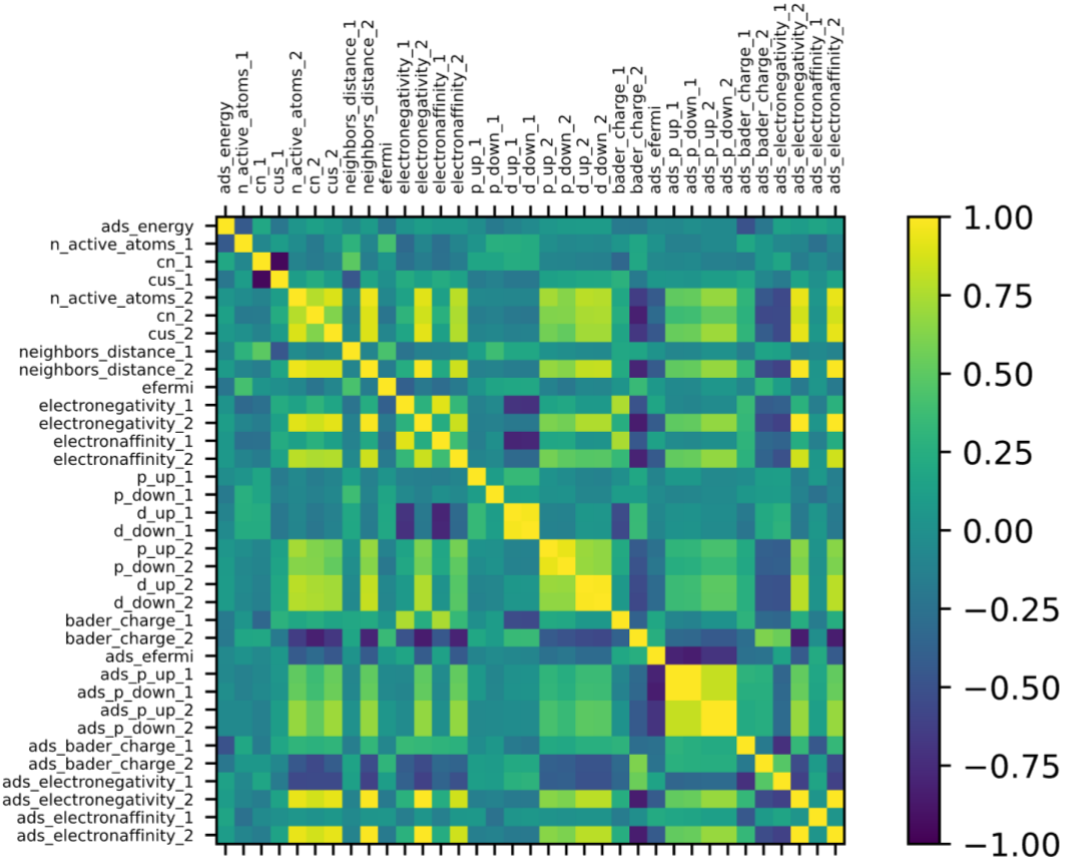


Figure 11 Correlations between all features and the adsorption energy

Features

To elucidate the relationship between features and adsorption energy (denoted as “ads_energy”), I have illustrated the correlations in Figure 11. This figure delineates the correlation range between -1 and 1, where adsorption energy shows the highest absolute correlation with several key features: the number of atoms participating at the surface, the Bader charge of the adsorbate atom, and the coordination number. Notably, features related to secondary adsorption sites generally exhibit a lower correlation with adsorption energy. This is primarily because structures with only one adsorption site fill these values with zeros, diluting their overall impact.

One conventional strategy to combat overfitting involves reducing the dimensionality of the feature vector. To discern which features could be eliminated without sacrificing predictive performance, I employed SHapley Additive exPlanations (SHAP) [46]. This advanced tool, rooted in cooperative game theory, assigns a SHAP value to each feature, reflecting its contribution to the final prediction. Generally, a higher absolute SHAP value indicates greater feature importance.

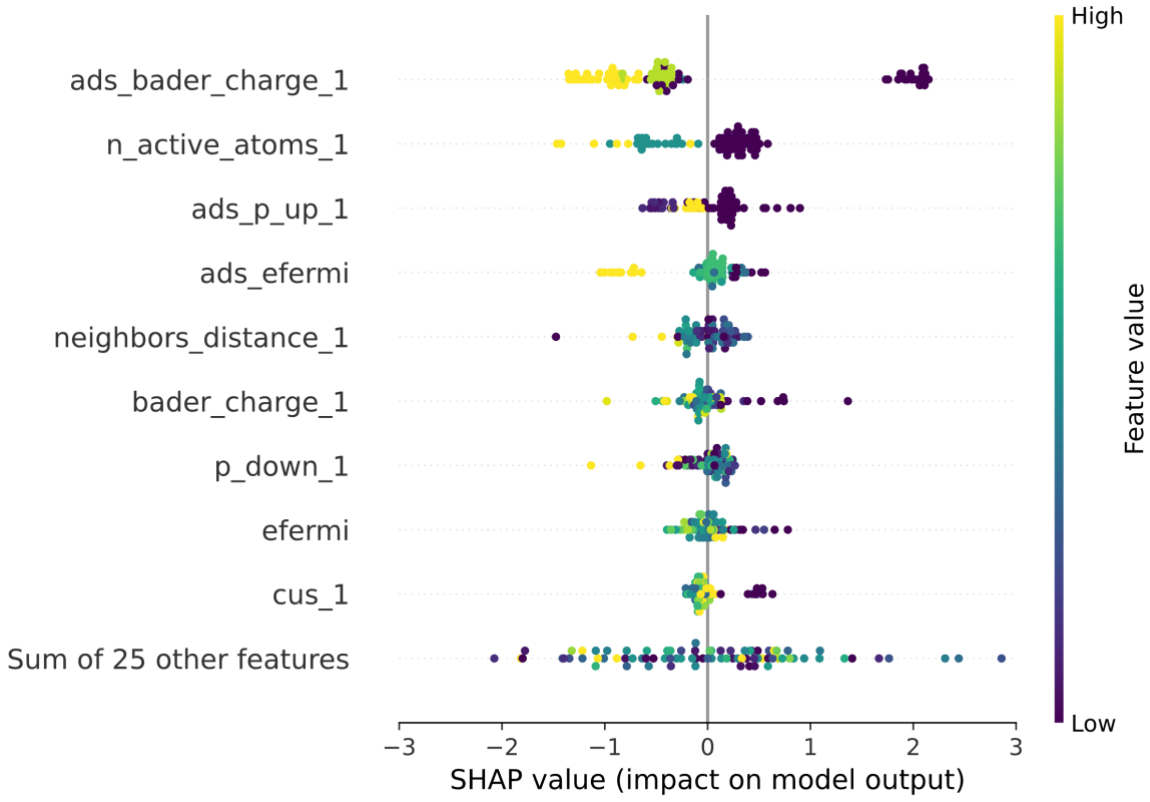


Figure 12 SHAP values of 9 most essential features for CatBoost

Figure 12 presents the SHAP analysis results for the CatBoost model, which performed best among the evaluated algorithms. Features are listed top-down based on their overall impact on the model's prediction. Figures containing all features for CatBoost, XGBoost, and KRR are presented in the Supporting Information. As anticipated, electronic features typically show higher importance due to the intricate electronic interactions within oxide structures [1]. Features about the secondary adsorption site are generally of lesser importance, which can be again explained with many data points possessing only 1 adsorption site.

To evaluate the impact of eliminating less critical features on the model's performance, I constructed a "reduced" feature vector containing only 21 entries. In the "reduced" feature vector, all the features describing the second active site were dropped, which creates significant gaps in the distinction of the systems (occurs only with CO and H2O) where two atoms bonded to the surface. The results are presented in Table 4.

	MAE Train [eV]	MAE Test [eV]	MSE Train [eV]	MSE Test [eV]	Accuracy	Standard deviation
CatBoost	0.07	1.028	0.007	1.485	0.483	0.152
XGBoost	0.0005	0.995	10^{-6}	1.869	0.311	0.315
KRR	0.05	1.308	0.007	2.897	0.418	0.205

Table 4 Performance metrics of ML algorithms with reduced feature vector

All three models demonstrated inferior performance in predictive capabilities based on MAE/MSE metrics. CatBoost experienced a moderate rise of MAE (1.028 eV now vs. 0.883 eV previously) and a substantial increase in MSE (1.485 eV now vs 0.946 eV previously) of the test set, indicating the rising significance of "big" outliers. The marginal improvement of the k-fold accuracy (0.483 now vs 0.478 previously) does not suggest any significant difference in robustness. XGBoost and KRR produced similar results, only slightly underperforming the models trained with the "full" feature vector in MAE/MSE. Unlike XGBoost, KRR improved its k-fold accuracy (0.418 now vs 0.290 previously). The issue of overfitting persists in all models since the discrepancy between MAE and MSE of the Train and Test is substantial. These results highlight that the high dimensionality of the feature vector is not the primary cause for overfitting.

However, the intricate nature of the investigated systems necessitates a feature vector of 35 dimensions. These features were needed to account for multiple adsorption sites, appropriately represent metal and oxygen atoms as active sites, and obtain a better overall

performance. In practice, constructing a feature vector requires a delicate balance. It's essential to adequately describe the adsorption site while avoiding excessive specificity that might negatively affect the model's predictive capabilities.

3.3.4. Limitations

Some limitations and constraints have been identified during this work, which significantly influenced its development and goals.

Data Accuracy and Reliability:

The structures of the OC22 dataset are highly heterogeneous, presenting a challenge in ensuring reliability across the data. The absence of reference values to benchmark against complicates further identifying irregularities or anomalies. While efforts were made to replicate the system energy values using the same functional as the OC22 researchers, it is acknowledged that the GGA functional, despite its increased k-point density, may not always yield the most accurate electronic structures compared to more complex functionals, i.e. meta-GGA or hybrid functionals. This limitation is significant, as inaccuracies in the foundational data can develop through the model, affecting the overall model performance.

Potential Inherited Errors:

Throughout the research process, instances of potential errors within the OC22 dataset were identified, such as the misclassification of physisorption as chemisorption. Given the dataset's size and the fact that another team generated it, it is plausible to assume that some errors, possibly during the VASP calculations, might have remained undetected.

Limited Scope of Investigated Materials:

This study investigated a limited selection of three materials. Given the complex and unique nature of atomic interactions in oxide structures, predicting whether the algorithm would demonstrate enhanced or reduced performance on a broader range of materials is challenging. This limitation underscores the need for cautious interpretation of the results and their applicability beyond the studied materials.

Surface defects:

The OC22 database includes surface defects, specifically oxygen vacancies, introduced randomly across various structures. These oxygen vacancies might be a significant factor due to the strong electronic interactions associated with oxygen. Such defects can notably

alter the adsorption behavior, leading to variations in the adsorption energy values. This factor is crucial in analyzing catalytic activity, as it highlights the potential irregularities in the adsorption energy values in the database.

4. Conclusions

In the scope of this research, the application of machine learning algorithms was systematically explored to predict adsorption energy in systems involving TMOs. The global objective of this research area is to establish a more efficient alternative to the computationally intensive Density Functional Theory simulations. Such an advancement is anticipated to accelerate the discovery of novel catalysts significantly. I have trained several ML models, incorporating features linked explicitly to specific active sites to obtain a physically meaningful description of the systems. This approach, commonly called the 'descriptor-based method', is designed to capture these active sites' unique characteristics accurately. By doing so, the models are expected to provide precise and reliable predictions of adsorption energy.

Among the ML algorithms tested, Decision Tree Gradient Boosting (DTGB) exhibited superior performance in the task defined by this study, i.e., the prediction of adsorption energies on oxides. Specifically, the dataset included Ti-O, Al-O, and Zr-O substrates features with one random gas-phase adsorbate out of CO, H2O, OH, O, C, N, and H. The best-performing DTGB had MAE of 0.883 on the Test set with a cross-validation accuracy of 0.478 via the CatBoost algorithm, closely followed by XGBoost when trained with a full feature vector. Although the accuracy is still considerably distant from the benchmark accuracy needed in computational catalysis (error <0.1 eV with respect to the DFT calculations), the results highlight the potential of advanced predictive models to discern patterns in inherently complex tasks such as heterogeneous catalysis on oxide surfaces, even with relatively small training sets of ca. 80 data points.

A previous comparable study using a similar methodology focused on predicting adsorption energies on metal surfaces from our group [9] reported a MAE of 0.174, with a cross-validation accuracy of 0.987. The model used much more data than the present study, and the systems were significantly more homogeneous, achieving a performance our models did not reach the same extent. This difference also indicates a more complex electronic structure of oxides and higher differences among them than metal-based systems.

Although not entirely comparable, the Open Catalyst Project researchers reported a MAE of 0.678 eV using the deep-learning GemNet-OC model, trained on a subset of around 700

systems from the OC22 database to predict adsorption energies [10]. Notably, the CatBoost and XGBoost baseline models demonstrate a comparable performance despite using a significantly smaller training set.

Overall, despite the efforts of the present work, overfitting remains a challenge to be solved when predicting the adsorption energies on the oxides, given the significant disparity between the errors of the Train and Test splits. However, there are possible ways to improve the results. Firstly, the training set is relatively small; therefore, increasing it seems a potential and straightforward way of improving the results. Catalytic systems involving TMOs exhibit several surfaces and facets, which are known to influence the material's catalytic activity significantly [46]. As already shown by an analogous work from our group on metal surfaces, increasing the size of the training set improved the predictive performance on unseen data. Secondly, an excessive number of features may also contribute to overfitting, as explained in [47]. The challenges associated with the variety and multitude of active sites, as detailed in Section 3.2.2, necessitated a high-dimensional feature vector for an appropriate description of the active sites. In this study, I sampled a relatively small subset from the OC22 dataset (102 out of 56 367), which did not fully reflect the diversity of the provided data and, more importantly, resulted in an underrepresentation of specific structures.

In a broader context, it is crucial to recognize that the OC22 dataset was initially conceived for a different purpose. It utilizes data points on a scale two orders of magnitude larger than what was employed in this study. Sampling data from such a large dataset introduces a set of risks and challenges, as evidenced throughout this thesis. Given the critical importance of data quality in smaller datasets, the results suggest generating our own dataset, having more control of the data and its quality seems to be the way to go. Moreover, it is critical to select the materials carefully, the evaluated facets, and active sites in terms of a single species bonded to the surface to reduce the variability of the dataset.

Nevertheless, exciting features of the parameter space of the properties of the oxides were gathered during this master thesis. The feature analysis also provided the most relevant properties determining adsorption energies on these materials, corresponding to the Bader Charge and p-occupation of the active adsorbate atom. Interestingly, the number of participating surface atoms consistently substantially impacted the prediction's outcome and the distance to neighbors. Future research on predicting adsorption energies on oxide

surfaces will build upon these results while tackling the challenges and limitations described in this study and the suggested solutions as an outlook.

List of References

1 Zhao Z.-J., Liu S., Zha S., Cheng D., Studt F., Henkelman G., Gong J., “Theory-guided design of catalytic materials using scaling relationships and reactivity descriptors,” *Nat Rev Mater*, vol. 4, no. 12, Art. no. 12, Dec. 2019, doi: 10.1038/s41578-019-0152-x.

2 Leswing J. V. Kif., “ChatGPT and generative AI are booming, but the costs can be extraordinary,” CNBC. Accessed: Nov. 13, 2023. [Online]. Available: https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html

3 Behler J., Parrinello M., “Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces,” *PHYSICAL REVIEW LETTERS*, vol. 98, p. 146401, 2007, doi: 10.1103/PhysRevLett.98.146401.

4 Yuan X., Suvarna M., Low S., Dissanayake P. D., Lee K. B., Li J., Wang X., Ok Y. S., “Applied Machine Learning for Prediction of CO_2 Adsorption on Biomass Waste-Derived Porous Carbons,” *Environ. Sci. Technol.*, vol. 55, no. 17, pp. 11925–11936, Sep. 2021, doi: 10.1021/acs.est.1c01849.

5 Ghanekar P. G., Deshpande S., Greeley J., “Adsorbate chemical environment-based machine learning framework for heterogeneous catalysis,” *Nat Commun*, vol. 13, no. 1, p. 5788, Oct. 2022, doi: 10.1038/s41467-022-33256-2.

6 Kolluru A., Shuaibi M., Palizhati A., Shoghi N., Das A., Wood B., Zitnick C. L., Kitchin J. R., Ulissi Z. W., “Open Challenges in Developing Generalizable Large-Scale Machine-Learning Models for Catalyst Discovery,” *ACS Catal.*, vol. 12, no. 14, pp. 8572–8581, Jul. 2022, doi: 10.1021/acscatal.2c02291.

7 Takigawa I., Shimizu K., Tsuda K., Takakusagi S., “Machine-learning prediction of the d-band center for metals and bimetals,” *RSC Adv.*, vol. 6, no. 58, pp. 52587–52595, 2016, doi: 10.1039/C6RA04345C.

8 Chanussot L., Das A., Goyal S., Lavril T., Shuaibi M., Riviere M., Tran K., Heras-Domingo J., Ho C., Hu W., Palizhati A., Sriram A., Wood B., Yoon J., Parikh D., Zitnick C. L., Ulissi Z., “Open Catalyst 2020 (OC20) Dataset and Community Challenges,” *ACS Catal.*, no. 11, pp. 6059–6072, 2021, doi: 10.1021/acscatal.0c04525.

9 Praveen C. S., Comas-Vives A., “Design of an Accurate Machine Learning Algorithm to Predict the Binding Energies of Several Adsorbates on Multiple Sites of Metal Surfaces,” *ChemCatChem*, vol. 12, no. 18, pp. 4611–4617, Sep. 2020, doi: 10.1002/cctc.202000517.

10 Tran R., Lan J., Shuaibi M., Wood B. M., Goyal S., Das A., Heras-Domingo J., Kolluru A., Rizvi A., Shoghi N., Sriram A., Therrien F., Abed J., Voznyy O., Sargent E. H., Ulissi Z., Zitnick C. L., “The Open Catalyst 2022 (OC22) Dataset and Challenges for Oxide Electrocatalysts,” *ACS Catal.*, vol. 13, no. 5, pp. 3066–3084, Mar. 2023, doi: 10.1021/acscatal.2c05426.

11 Takanabe K., “Photocatalytic Water Splitting: Quantitative Approaches toward Photocatalyst by Design,” *ACS Catal.*, vol. 7, no. 11, pp. 8006–8022, Nov. 2017, doi: 10.1021/acscatal.7b02662.

12 Keith J. A., Anton J., Kaghazchi P., Jacob T., “Modeling Catalytic Reactions on Surfaces with Density Functional Theory,” in *Modeling and Simulation of Heterogeneous Catalytic Reactions: From the Molecular Process to the Technical System*, 2012, pp. 1–37.

13 Jones R. O., “Density functional theory: Its origins, rise to prominence, and future,” *Rev. Mod. Phys.*, vol. 87, no. 3, pp. 897–923, Aug. 2015, doi: 10.1103/RevModPhys.87.897.

14 Hohenberg P., Kohn W., “Inhomogeneous Electron Gas,” *Phys. Rev.*, vol. 136, no. 3B, pp. B864–B871, Nov. 1964, doi: 10.1103/PhysRev.136.B864.

15 Kohn W., Sham L. J., “Self-Consistent Equations Including Exchange and Correlation Effects,” *Phys. Rev.*, vol. 140, no. 4A, pp. A1133–A1138, Nov. 1965, doi: 10.1103/PhysRev.140.A1133.

16 Perdew J. P., Burke K., Ernzerhof M., “Generalized Gradient Approximation Made Simple,” *PHYSICAL REVIEW LETTERS*, vol. 77, no. 18, 1996.

17 Perdew J. P., Ruzsinszky A., Csonka G. I., Vydrov O. A., Scuseria G. E., Constantin L. A., Zhou X., Burke K., “Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces,” *PHYSICAL REVIEW LETTERS*, vol. 100, p. 136406, 2008, doi: 10.1103/PhysRevLett.100.136406.

18 Hafner J., “Ab-initio simulations of materials using VASP: Density-functional theory and beyond,” *Journal of Computational Chemistry*, vol. 29, no. 13, pp. 2044–2078, 2008, doi: 10.1002/jcc.21057.

19 “Category:Theory - Vaspwiki.” Accessed: Oct. 19, 2023. [Online]. Available: https://www.vasp.at/wiki/index.php/Category:Theory

20 “INCAR - Vaspwiki.” Accessed: Oct. 19, 2023. [Online]. Available: https://www.vasp.at/wiki/index.php/INCAR

21 “KPOINTS - Vaspwiki.” Accessed: Oct. 19, 2023. [Online]. Available: https://www.vasp.at/wiki/index.php/KPOINTS

22 “POTCAR - Vaspwiki.” Accessed: Oct. 19, 2023. [Online]. Available: https://www.vasp.at/wiki/index.php/POTCAR

23 Guan Y., Chaffart D., Liu G., Tan Z., Zhang D., Wang Y., Li J., Ricardez-Sandoval L., “Machine learning in solid heterogeneous catalysis: Recent developments, challenges and perspectives,” *Chemical Engineering Science*, vol. 248, p. 117224, Feb. 2022, doi: 10.1016/j.ces.2021.117224.

24 Wu D., Dong C., Zhan H., Du X.-W., “Bond-Energy-Integrated Descriptor for Oxygen Electrocatalysis of Transition Metal Oxides,” *J. Phys. Chem. Lett.*, vol. 9, no. 12, pp. 3387–3391, Jun. 2018, doi: 10.1021/acs.jpclett.8b01493.

25 Wang Y., Qiu W., Song E., Gu F., Zheng Z., Zhao X., Zhao Y., Liu J., Zhang W., “Adsorption-energy-based activity descriptors for electrocatalysts in energy storage applications,” *National Science Review*, vol. 5, no. 3, pp. 327–341, May 2018, doi: 10.1093/nsr/nwx119.

26 Andersen M., Levchenko S. V., Scheffler M., Reuter K., “Beyond Scaling Relations for the Description of Catalytic Materials,” *ACS Catal.*, vol. 9, no. 4, pp. 2752–2759, Apr. 2019, doi: 10.1021/acscatal.8b04478.

27 Raschka S., Mirjalili V., *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, 2019.

28 “GitHub - catboost/catboost: A fast, scalable, high performance Gradient Boosting on Decision Trees library, used for ranking, classification, regression and other machine learning tasks for Python, R, Java, C++. Supports computation on CPU and GPU.” Accessed: Nov. 06, 2023. [Online]. Available: https://github.com/catboost/catboost

29 “eXtreme Gradient Boosting.” Distributed (Deep) Machine Learning Community, Nov. 06, 2023. Accessed: Nov. 06, 2023. [Online]. Available: https://github.com/dmlc/xgboost

30 “sklearn.kernel_ridge.KernelRidge,” scikit-learn. Accessed: Nov. 06, 2023. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.kernel_ridge.KernelRidge.html

31 Jiang J., Wang R., Wang M., Gao K., Nguyen D. D., Wei G.-W., “Boosting Tree-Assisted Multitask Deep Learning for Small Scientific Datasets,” *J Chem Inf Model*, vol. 60, no. 3, pp. 1235–1244, Mar. 2020, doi: 10.1021/acs.jcim.9b01184.

32 Deng H., Zhou Y., Wang L., Zhang C., “Ensemble learning for the early prediction of neonatal jaundice with genetic features,” *BMC Med Inform Decis Mak*, vol. 21, no. 1, p. 338, Dec. 2021, doi: 10.1186/s12911-021-01701-9.

33 Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. V., Gulin A., “CatBoost: unbiased boosting with categorical features.” arXiv, Jan. 20, 2019. Accessed: Nov. 07, 2023. [Online]. Available: http://arxiv.org/abs/1706.09516

34 Chen T., Guestrin C., “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

35 Murphy K. P., “Kernels,” in *Machine learning: a probabilistic perspective*, 4th ed., Cambridge, Mass.: MIT Press, 2013, pp. 479–514.

36 “Trajectory files — ASE documentation.” Accessed: Nov. 16, 2023. [Online]. Available: https://wiki.fysik.dtu.dk/ase/ase/io/trajectory.html

37 “About — ASE documentation.” Accessed: Dec. 05, 2023. [Online]. Available: https://wiki.fysik.dtu.dk/ase/about.html

38 “Home,” pymatgen. Accessed: Dec. 05, 2023. [Online]. Available: https://pymatgen.org/

39 “Supercomputers | CSUC.” Accessed: Nov. 16, 2023. [Online]. Available: https://www.csuc.cat/en/serveis/supercomputadors

40 “Code:Bader Charge Analysis.” Accessed: Dec. 11, 2023. [Online]. Available: http://theory.cm.utexas.edu/henkelman/code/bader/

41 Wang V., Xu N., Liu J.-C., Tang G., Geng W.-T., "VASPKIT: A user-friendly interface facilitating high-throughput computing and analysis using VASP code," *Computer Physics Communications*, vol. 267, p. 108033, Oct. 2021, doi: 10.1016/j.cpc.2021.108033.

42 Brownlee J., "Overfitting and Underfitting With Machine Learning Algorithms," MachineLearningMastery.com. Accessed: Dec. 13, 2023. [Online]. Available: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

43 "Band gap of Si with the Perdew-Burke-Ernzerhof (PBE) and PBE0 functionals," VASP. Accessed: Dec. 19, 2023. [Online]. Available: https://www.vasp.at/tutorials/latest/hybrids/part1/

44 "sklearn.preprocessing.RobustScaler," scikit-learn. Accessed: Dec. 28, 2023. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

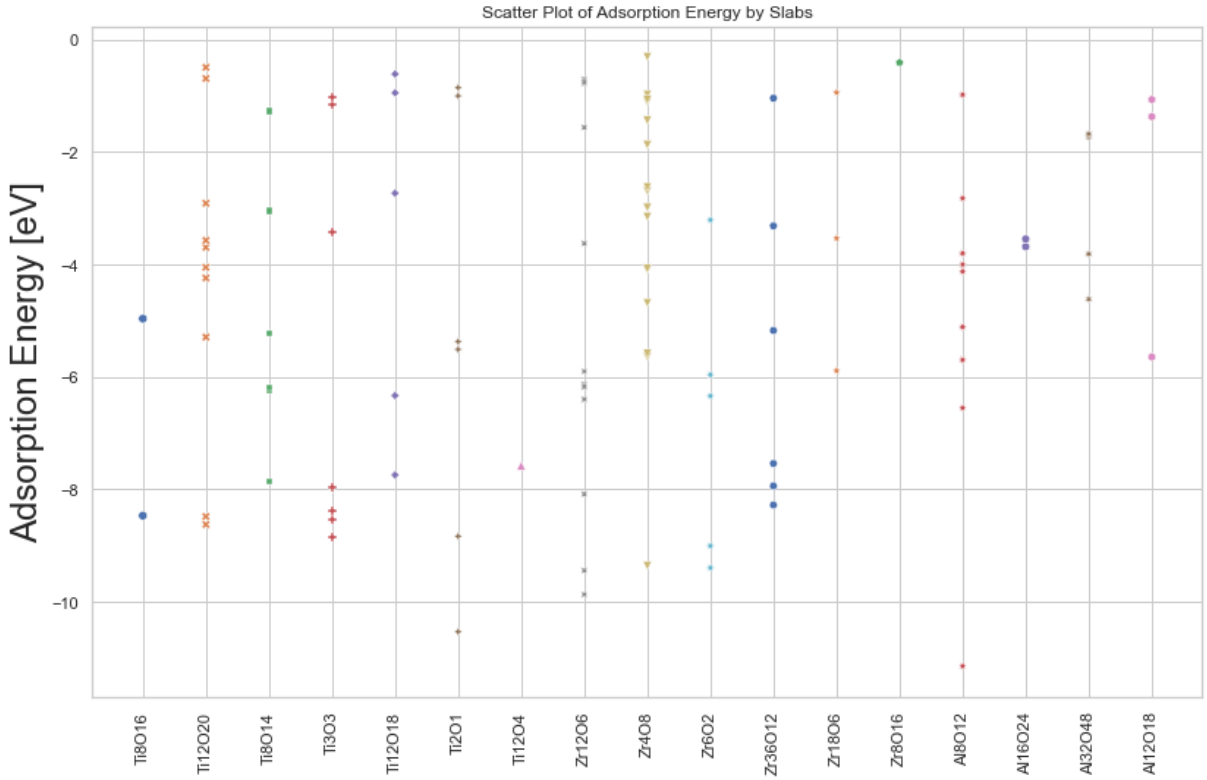
45 "sklearn.model_selection.KFold," scikit-learn. Accessed: Dec. 28, 2023. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.model_selection.KFold.html

46 Lundberg S. M., Lee S.-I., "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Jan. 04, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

47 Tetko I. V., Livingstone D. J., Luik A. I., "Neural network studies. 1. Comparison of overfitting and overtraining," *J. Chem. Inf. Comput. Sci.*, vol. 35, no. 5, pp. 826–833, Sep. 1995, doi: 10.1021/ci00027a006.

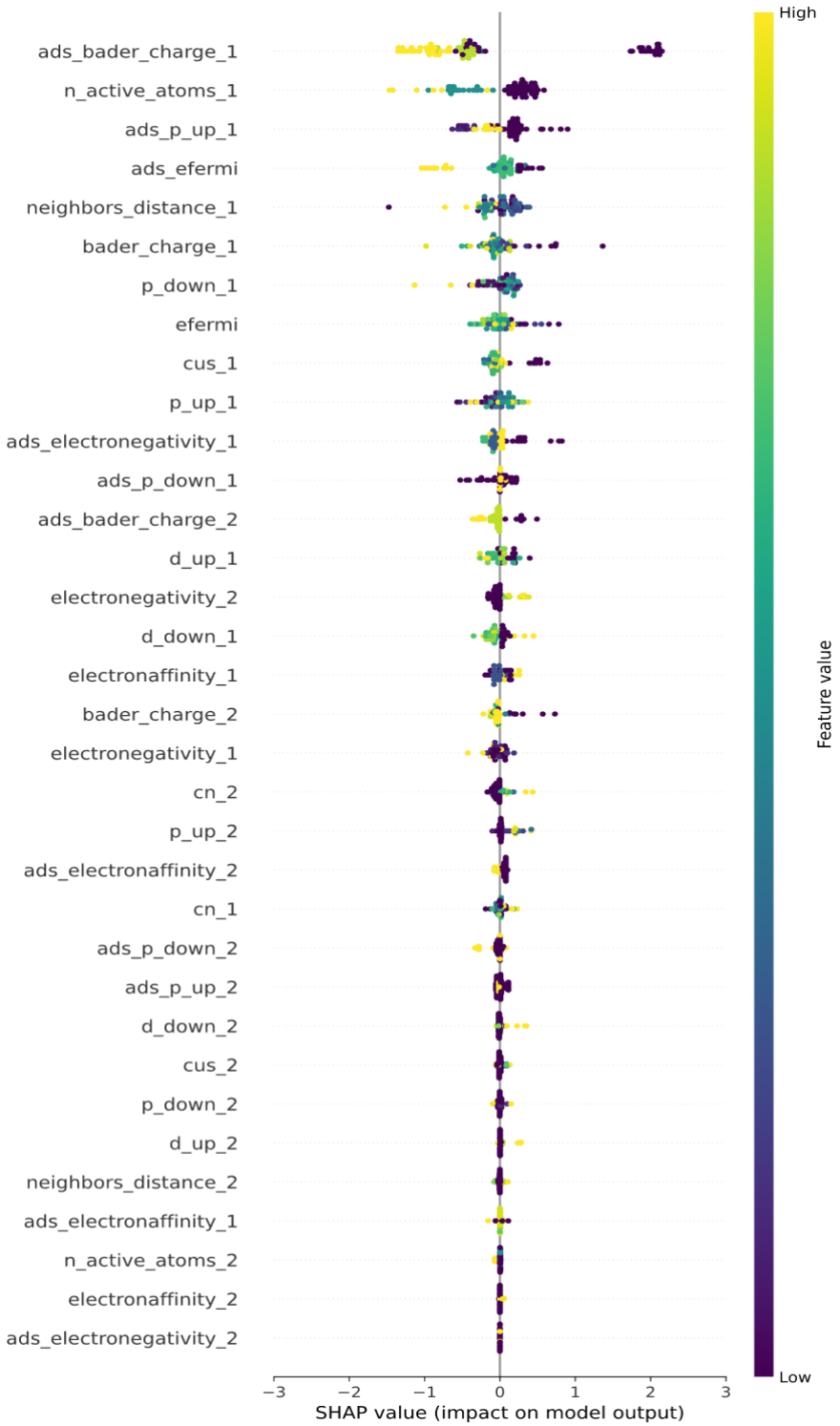
Supporting Information

Adsorption energy distribution grouped by the chemical formula of slabs

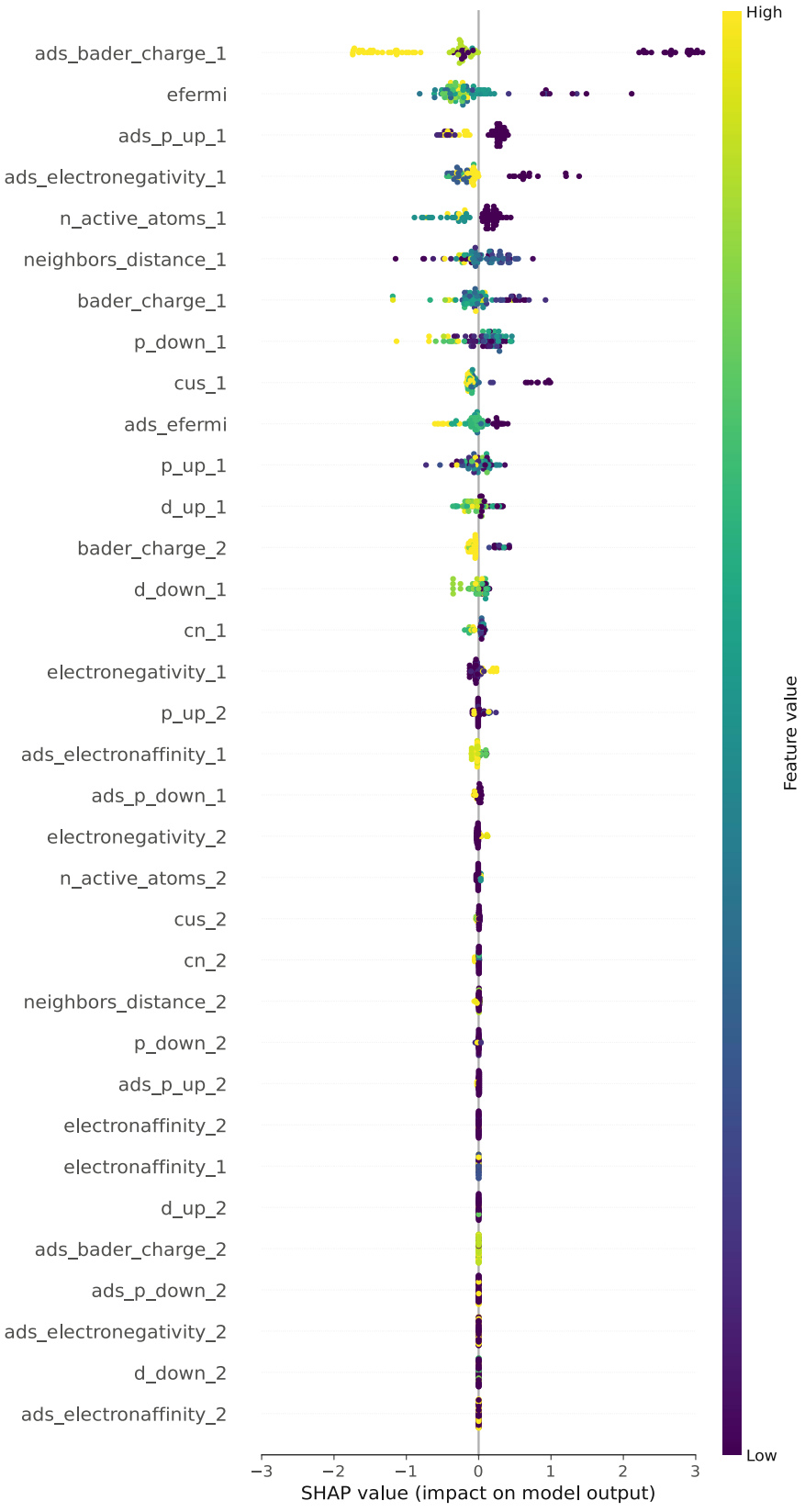


Complete SHAP analysis for models trained with the “full” feature vector

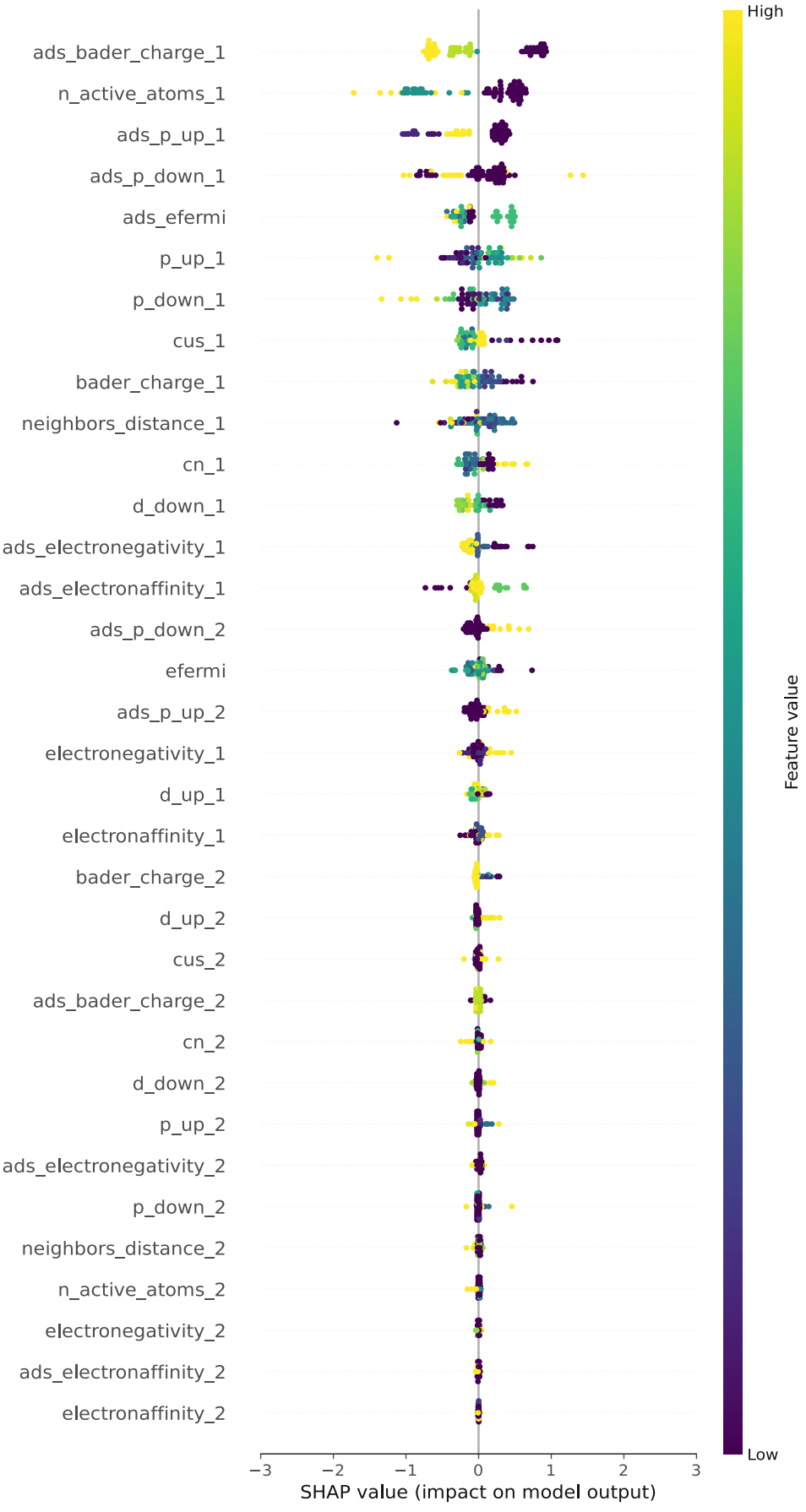
CatBoost



XGBoost

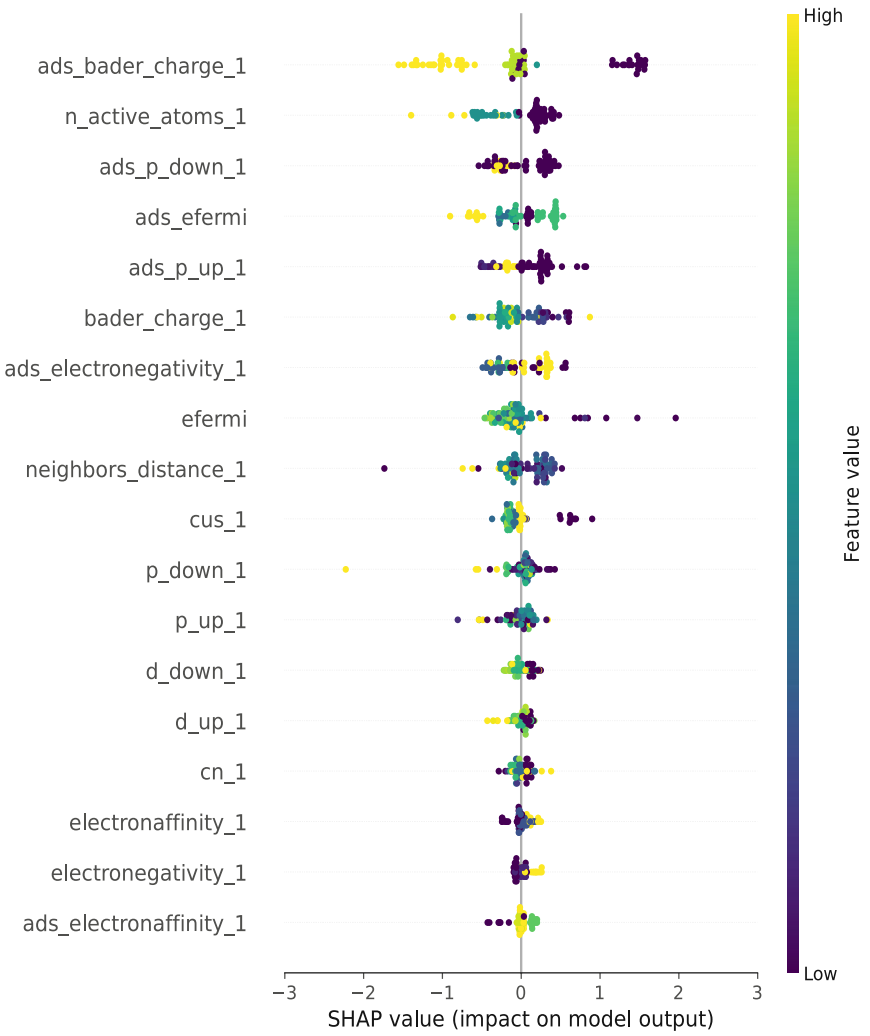


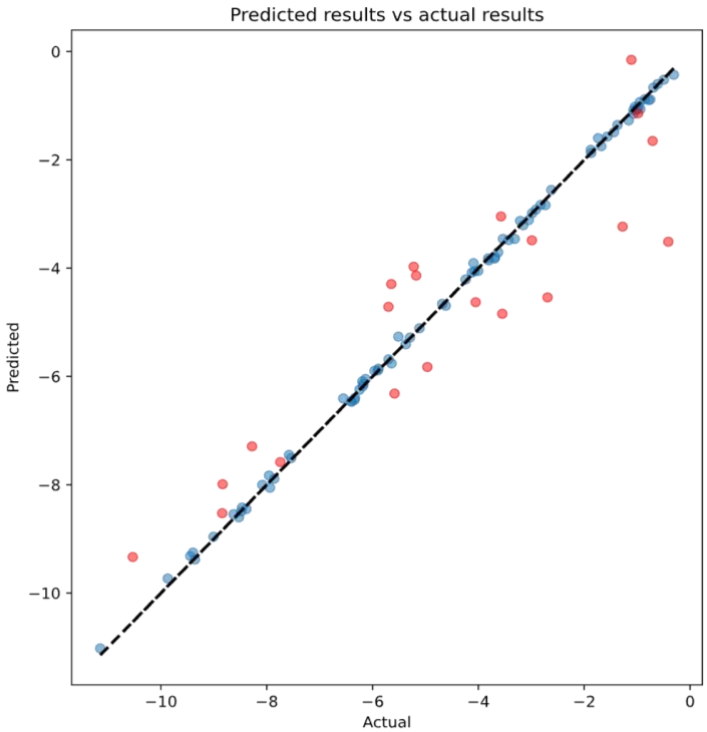
Kernel Ridge Regression



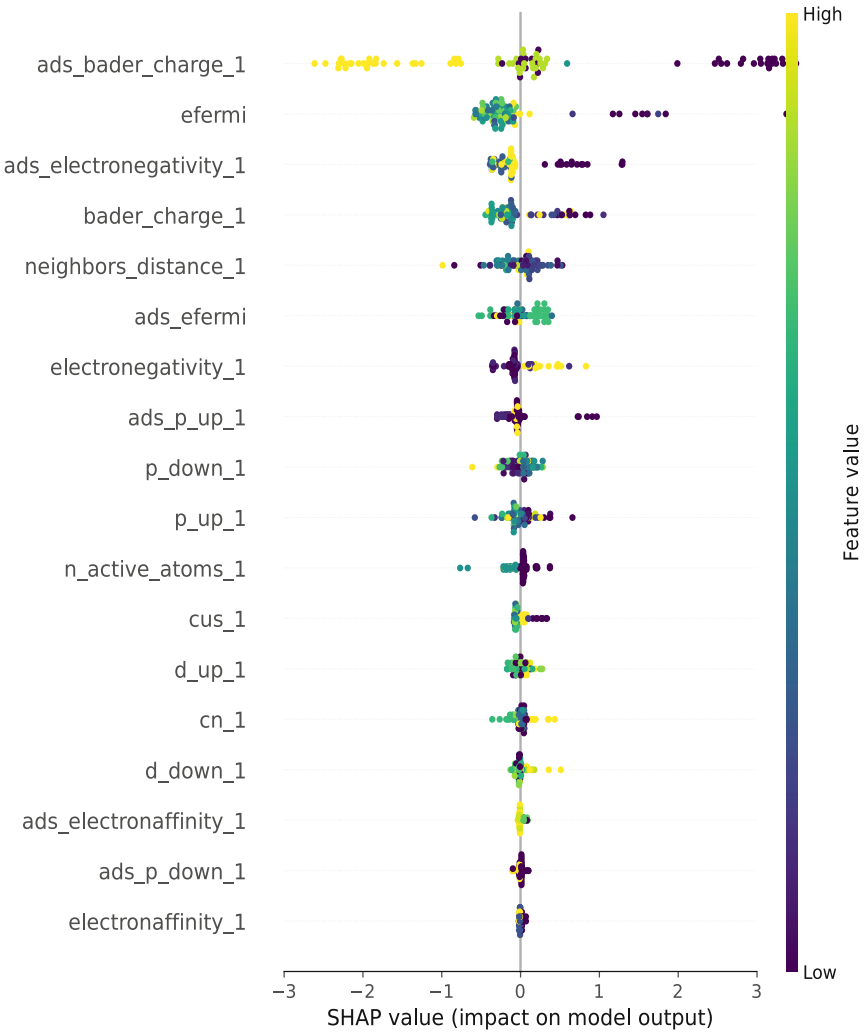
Predictions plot and SHAP analysis for models trained with “reduced” vector

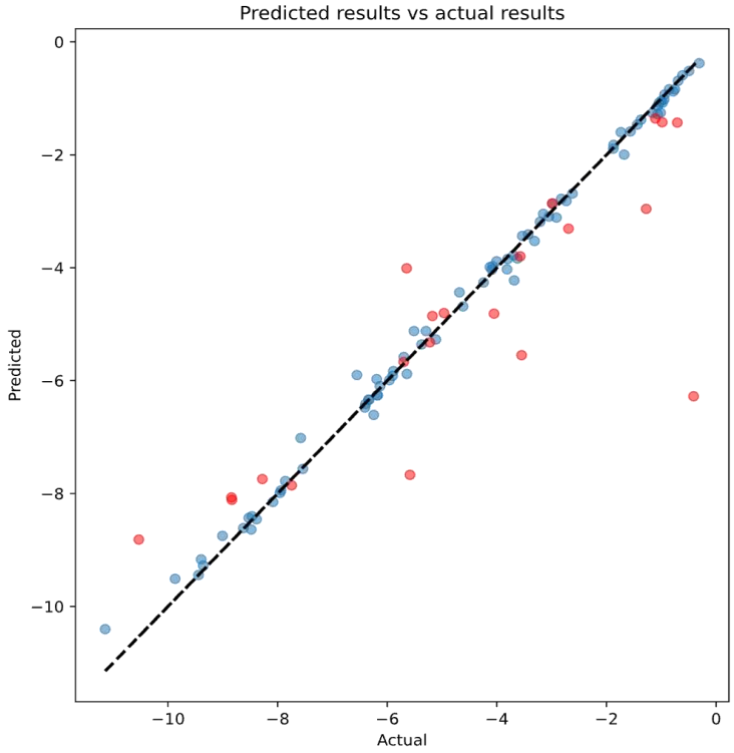
CatBoost





XGBoost





Kernel Ridge Regression