CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS GRADUAÇÃO MULTICÊNTRICO EM QUÍMICA

Lucas Bernardes Pena

Unraveling Molecular Descriptors for Reliable Adsorption Energy Prediction

Belo Horizonte
2025

Lucas Bernardes Pena

Unraveling Molecular Descriptors for Reliable Adsorption Energy Prediction

Dissertação apresentada ao Programa de Pós-graduação Multicêntrico em Química do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito para a obtenção do título de Mestre em Química.

Orientador: Prof. Dr. Breno Rodrigues Lamaghere Galvão

Belo Horizonte
2025

Pena, Lucas Bernardes.
P397u Unraveling molecular descriptors for reliable adsorption energy prediction / Lucas Bernardes Pena. – 2025.
81 f. : il.
Orientador: Breno Rodrigues Lamaghere Galvão.

Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais, Programa de Pós-Graduação Multicêntrico em Química de Minas Gerais, Belo Horizonte, 2025.
Bibliografia.

1. Nanoestruturas. 2. Cluster. 3. Adsorção. 4. Funcionais de densidade. 5. Aprendizado de máquina. I. Galvão, Breno Rodrigues Lamghere. II. Título.

CDD: 541.335

Ficha catalográfica elaborada pela Biblioteca - *campus* Nova Suíça - CEFET-MG
Bibliotecário: Wagner Oliveira Braga - CRB6-3261

Lucas Bernardes Pena

Unraveling Molecular Descriptors for Reliable Adsorption Energy Prediction

Dissertação apresentada ao Programa de Pós-Graduação Multicêntrico em Química, do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito para a obtenção do título de Mestre em Química.

Belo Horizonte, 1 de agosto de 2025.

Aprovado pela Banca Examinadora:

gov.br Documento assinado digitalmente
BRENO RODRIGUES LAMAGHERE GALVAO
Data: 08/09/2025 13:29:51-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Breno Rodrigues Lamaghere Galvão
(orientador – CEFET-MG)

gov.br Documento assinado digitalmente
EMERSON FERNANDES PEDROSO
Data: 04/09/2025 20:11:16-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Emerson Fernandes Pedroso
(avaliador – CEFET-MG)

gov.br Documento assinado digitalmente
MAICON PIERRE LOURENCO
Data: 06/09/2025 11:28:56-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Maicon Pierre Lourenço
(avaliador – UFES)

Resumo

A acumulação de CO_2 na atmosfera terrestre é um problema crescente, com consequências devastadoras em todo o mundo. Uma abordagem promissora para mitigar esse problema é a conversão de CO_2 em combustíveis e produtos químicos de valor econômico utilizando fontes de energia renováveis. Essa estratégia não apenas reduz os níveis atmosféricos de CO_2, mas também contribui para o desenvolvimento de soluções energéticas sustentáveis. Cálculos *ab initio* possibilitam a descoberta de novos catalisadores através de análises exploratórias de sistemas *nanocluster*-adsorbato de alto custo computacional, uma vez que diversas configurações de adsorção precisam ser avaliadas em diferentes estruturas de *clusters*. Dado o vasto volume de dados computacionais já disponíveis para esses sistemas, métodos de aprendizado de máquina são uma alternativa de ferramenta para acelerar o processo de triagem de catalisadores. Entretanto, os *frameworks* atuais são limitados pela falta de diversidade nos dados e pela falta de transferência de conhecimento dos modelos, comprometendo sua confiabilidade especialmente perante novos dados. Neste trabalho, os descritores moleculares *Coulomb matrix* e *many-body tensor representation* foram otimizados para a regressão da energia de adsorção em *nanoclusters* utilizando-se uma nova base de dados diversa. Utilizando o algoritmo de regressão *random forest*, atingiu-se um erro médio absoluto de 0.05 eV na predição da energia de adsorção para ambos os descritores. Para avaliar a generalizabilidade do modelo, um novo conjunto de dados foi gerado para o sistema mais representativo do *dataset* utilizado na etapa de desenvolvimento. Ao testar o modelo neste conjunto externo, nota-se a incapacidade de generalização do modelo, com o aumento do erro médio absoluto para 0.30 eV. Para a melhora da performance, propõe-se uma *feature* eletrônica, a qual corrigiu as predições para um dos novos sistemas adsorvidos inéditos.

Palavras-chave: *nanocluster*, adsorção, descritor molecular, DFT, aprendizado de máquina.

Abstract

The accumulation of CO_2 in Earth's atmosphere is a growing problem with devastating consequences worldwide. A promising approach to mitigating this issue is the conversion of CO_2 into fuels and economically valuable chemicals using renewable energy sources. This strategy not only reduces atmospheric CO_2 levels but also contributes to the development of sustainable energy solutions. *Ab initio* calculations enable the discovery of new catalysts through exploratory analyses of nanocluster adsorbed systems, which are computationally expensive, as multiple adsorption configurations must be evaluated across different cluster structures. Given the vast volume of computational data already available for these systems, machine learning methods are an option to accelerate the catalyst screening process. However, current frameworks are limited by a lack of data diversity and lack of knowledge transferability, compromising their reliability, especially when applied to new data. In this work, the molecular descriptors Coulomb matrix and many-body tensor representation were optimized for adsorption energy regression in nanoclusters using a newly developed diverse dataset. Using the random forest regression algorithm, an absolute mean error of 0.05 eV was achieved in adsorption energy prediction for both descriptors. To assess the model's generalizability, a new dataset was generated with the most representative cluster from the dataset used in the development phase. When tested on this external dataset, the model demonstrated a lack of generalization, with an increase of the absolute mean error to 0.30 eV. For improving the model, an additional electronic feature was proposed, enabling better results for one of the unprecedented adsorbed systems.

Keywords: nanocluster, adsorption, molecular descriptor, DFT, machine learning.

List of Figures

List of abbreviations and acronyms

ML	Machine learning
LCAO	Linear combination of atomic orbitals
DFT	Density functional theory
SCF	Self consistent field
CM	Coulomb matrix
BoB	Bag of bonds
MBTR	Many-body tensor representation
LMBTR	Local many-body tensor representation
PCA	Principal component analysis
PC	Principal component
RSS	Residual sum of squares
RFR	Random forest regression
CART	Classification and regression trees
FHI-aims	Fritz Haber Institute ab initio materials simulations
MAE	Mean absolute error

Contents

1 Introduction

The steady increase in atmospheric CO_2 levels (Fig. 1) has been a critical environmental challenge for decades. During the 1960s, CO_2 concentrations rose by approximately 1 ppm annually, but this rate doubled to 2 ppm by 2020, primarily due to the combustion of carbonaceous fossil fuels, including oil, diesel, petrol, and natural gas (KANDY, 2020). These emissions stem from various industrial processes such as the manufacturing of cement, which corresponds to 8 percent of all worldwide CO_2 emissions (RODGERS, 2018). Natural gas processing and fuel combustion are also the main sources of direct CO_2 emissions, but indirect emissions, such as land use – which is the main contributor of Brazil's carbon footprint – have also been concerning (EVANS, 2021). This trend has pushed atmospheric CO_2 concentrations to 414.11 ppm, significantly exceeding the safe threshold of 350 ppm, and, persisting, atmospheric concentrations could reach 590 ppm by 2100, correlating with a 1.9 °C increase in global temperatures (KHALIL *et al.*, 2019).

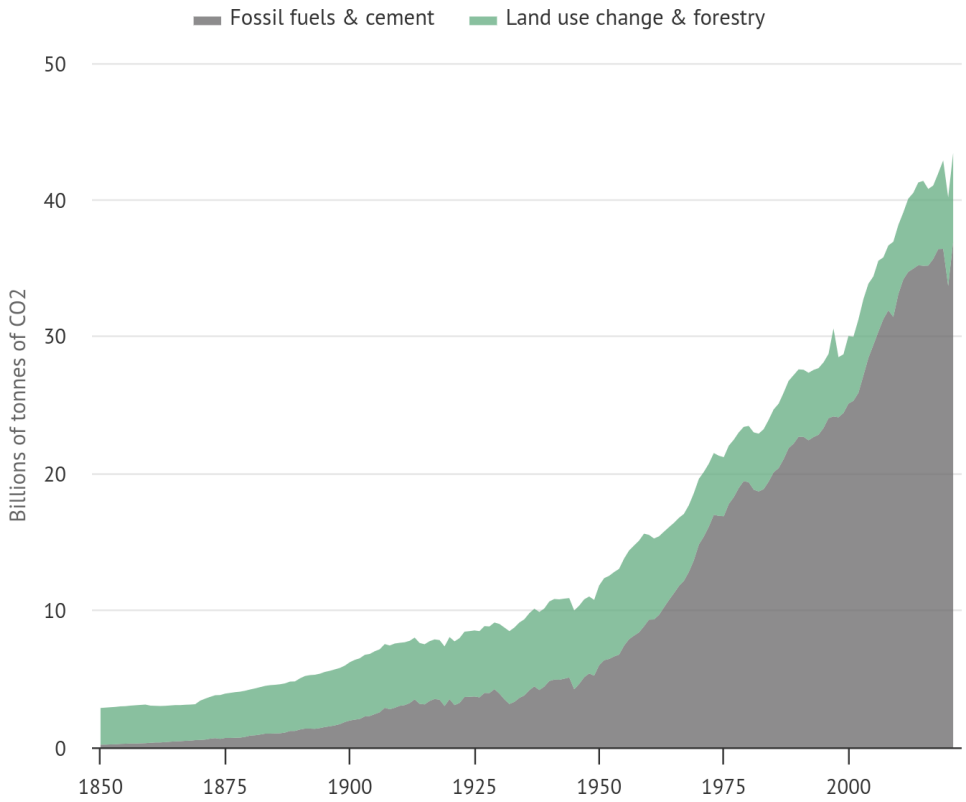

Figure 1 – CO_2 emissions from cement production and land use. Reproduced with permission from Carbon Brief (EVANS, 2021).

To mitigate CO_2 emissions, the CO_2 reduction reaction, combined with renewable energy sources such as wind, solar, and biomass, offers a sustainable alternative to fossil fuels while addressing the problem of atmospheric CO_2 saturation. However, the efficient industrial-scale

conversion of CO_2 remains unfeasible due to its high energy requirements, currently supplied by the combustion of fossil fuels. While CO_2 reduction reactions are far from spontaneous and require significant energy input, they can be facilitated through systems such as photocatalysis or electrocatalysis by incorporating catalysts with enhanced selectivity, activity, and long-term stability. Catalysts act by interacting with reactants and intermediate species in the reaction, allowing for reactions with lower energy barriers, reducing costs and increasing turnover rates for the reaction. The integration of innovative catalytic strategies into these systems could enable efficient and stable industrial-scale implementation, potentially transforming the energy sector (SADASIVUNI *et al.*, 2022).

Metallic nanoclusters have emerged as promising catalysts for this process due to their unique structural and electronic properties. While experimental studies have demonstrated their efficiency in product selectivity and tunability, an optimal catalyst capable of enabling an economically viable CO_2 conversion has yet to be identified. *Ab initio* calculations provide a reliable option for the exploratory search of potential nanocluster catalysts by calculating adsorption energies over these structures for different adsorbates. However, due to the vast configurational space of cluster-adsorbate systems, systematically exploring all possible adsorption configurations is computationally prohibitive, specially considering the great versatility of nanoclusters structures (HU *et al.*, 2023).

The evolution of materials science, illustrated in Fig. 2, reflects a shift from empirical approaches to data-driven methodologies. In this context, machine learning (ML) represents a transformative tool, capable of accelerating material discovery through statistical learning, data mining, and pattern recognition techniques. Models trained on computed datasets can provide rapid and cost-effective predictions of adsorption energies, enabling high-throughput screening of catalytic materials. However, the reliability and transferability of chemical models remain largely unknown due to limited datasets and insufficient model validation, hindering the explainability of various molecular descriptors and their limitations (SCHLEDER *et al.*, 2019).

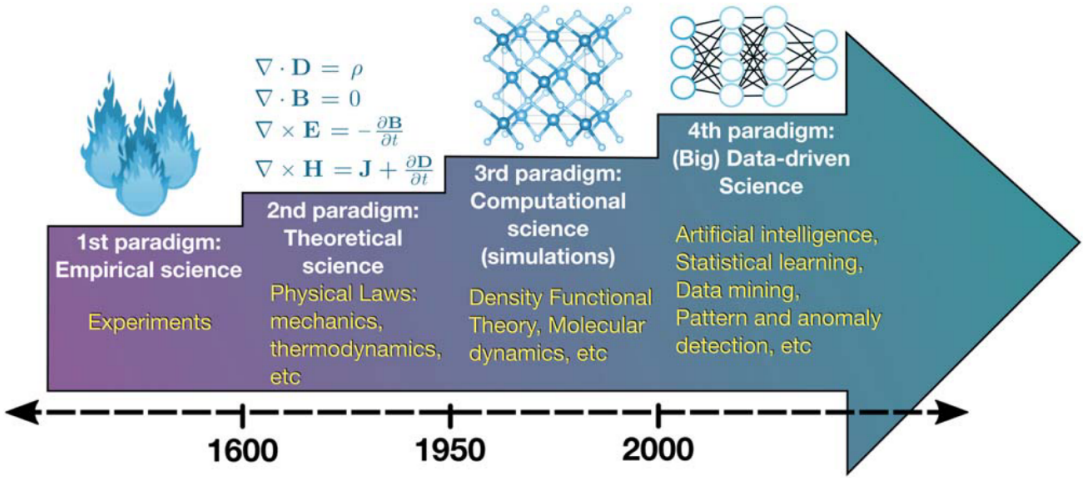

Figure 2 – The four science paradigms that describe the historical evolution of science. Reproduced with permission from IOP publishing (SCHLEDER *et al.*, 2019).

1.1 Motivation

There is a lack of comprehensive databases for nanoclusters, particularly those that integrate structural and adsorption energy data. This dissertation aims to address this gap by developing a machine learning model and evaluating its performance on a diverse nanocluster-adsorbate dataset constructed from previous electronic structure calculations performed for research on a variety of system by the QTNano group (SILVA, 2025). By employing commonly used molecular descriptors, we propose a new perspective for assessing their accuracy in predicting adsorption energies and their ability to generalize to unseen data, an aspect that is not commonly explored in the literature.

1.2 Objectives

The objectives of this dissertation are centered on exploring the application of ML in the prediction of adsorption energies and addressing the associated challenges and opportunities. Specifically, our objectives are:

1. Optimize the Coulomb matrix and many-body tensor representation descriptors for predicting adsorption energies from the utilized database;
2. Generate a new set of nanocluster adsorbed configurations and calculate adsorption energies through DFT single-point calculations for validating the model on unseen data;
3. Assess the performance of the random forest regression algorithm with the optimized descriptors for predicting adsorption energies on the dataset and the constructed external test set;
4. Compare model predictions with DFT results, highlighting discrepancies and areas for improvement;
5. Provide insights into the applicability of ML-driven approaches for efficient screening of adsorption energies on nanocluster materials.

2 Literature Review

In the following sections, the key topics of this work are explored from a bibliographical perspective. The discussion begins with metallic nanoclusters, covering their origin, state-of-the-art experimental techniques, and potential applications. Following, the principles of the CO_2 reduction reaction are addressed, outlining metallic nanoclusters as novel electrocatalysts.

2.1 Metallic Nanoclusters

The fabrication of small metallic particles can be traced back to the Roman Empire, as observed in the Lycurgus Cup (Fig. 3), an artifact dated to 400 B.C. whose color-changing properties under different lighting conditions are explained by the presence of gold, silver, and copper nanocrystals dispersed within its glass matrix (BARBER; FREESTONE, 1990). A similar phenomenon is found in the stained-glass of medieval cathedrals, where production techniques, already widespread in ancient Egypt, enabled the incorporation of different transition metal oxides and compounds into molten glass (CIONTI; STUCCHI; MERONI, 2022). Although ancient methods inadvertently produced metallic nanoparticles, the first formal study of these

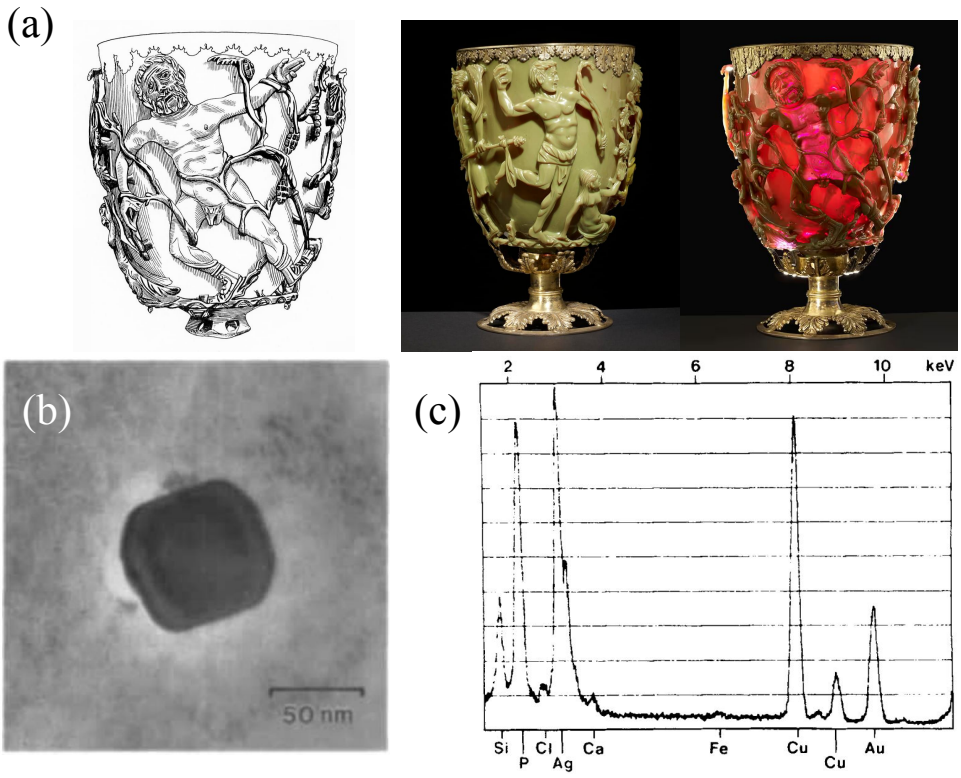

Figure 3 – Characterization of the (a) Lycurgus cup from (b) Transmission Electron Microscopy, and (c) Energy Dispersive X-Ray, which identifies the CuAgAu nanocrystals in its glass matrix. Reproduced with permission from John Wiley and Sons (BARBER; FREESTONE, 1990).

materials is attributed to Faraday (1857), who, while preparing gold sheets for light experiments, observed the formation of colloidal gold, which exhibited distinct optical properties compared to the thin gold leaves (TWENEY, 2006). From these rudimentary processes, metallic nanoclusters and nanoalloys have evolved into a cutting-edge field of research, driving the development of novel technological solutions, as advances in synthesis and characterization techniques have enabled the precise control and understanding of complex nanocluster structures.

2.1.1 Synthesis and Characterization Methods

Synthesis methods for nanoalloys are divided into two broad categories: chemical methods or physical methods. Chemical methods rely on key reactions that directs the nanoalloy formation process, generally conducted in the liquid phase. Physical methods use controlled environments that offer improved atomic precision for the nucleation of the cluster. While chemicals allow for the production of numerous and varied particles at a lower cost with some risk for contamination, physical methods are more precise but costly, time-consuming, and limited to providing lower quantities of these particles (FERRANDO, 2016).

Starting from 1960-1970, colloidal gold nanoclusters began being studied and characterized (MCPARTLIN; MASON; MALATESTA, 1969). Initially, similar to the Faraday (1857) experiment, the first studied nanoclusters were prepared through metal complexes synthesis methods, which comprehends mixing a metal salt (*e.g.*, gold chloride) to a solvent, mixing in a ligand (*e.g.* thiol) and adding the reducing agent (*e.g.* sodium borohydride) (KAWAWAKI; NEGISHI, 2023). This synthesis produces protected metal nanoclusters within a size distribution that is steered by reactant concentrations, and be chemically noted as TM_nL_m, where TM note the metallic species in the particle core and L the ligand molecules in the cluster surface (QIAN *et al.*, 2024).

Chemical methods are mainly related to colloidal formations, which are commonly used for production of ligand-protected clusters. Many factors influence cluster formation in solution, and by controlling reaction conditions, cluster growth may be directed towards different size regimes and alloy orderings (KAWAWAKI; NEGISHI, 2023). The chemical environment provides physico-chemical effects from solvents, ligands, and metals that relate to nanoparticle size and interparticle interaction, guiding a rational design of alloy ordering (QIAN *et al.*, 2024). Gold ligand-protected nanoclusters are the most widely studied, mainly due to the facile synthesis in ambient conditions, non-toxicity and thermodynamic stability of these particles. From colloidal preparation and X-ray characterization it was understood the finite nature of these particles and how their properties are dependent on size and geometry (NEGISHI, 2022).

During the 1980s, the development of molecular beams and laser vaporization enabled the production of free clusters in gas phase, which varied in size from few to thousands of atoms. While colloidal synthesis methods have poor control over particle size – and the particle environment may cause contamination or interference to characterization measurements – gas

phase research together with mass spectrometry and different types of spectroscopy laid the foundation for the understanding of electronic, optical, and magnetic properties of isolated clusters as a function of cluster structure (ANTOINE; BROYER; DUGOURD, 2023).

Gas phase synthesis is an example of a physical method that allows the study of nanoalloys and their electronic properties in an inert environment. From the cluster source (a bulk metal), an inert gas at high pressures is focalized into a vacuum chamber creating a molecular beam that cools upon the adiabatic expansion, creating a saturated solution where cluster growth occurs. The alloy formation process in gas phase can be described in four main steps: (i) vaporization of the cluster source, such as by heating or bombarding, (ii) nucleation of the metal gas (or plasma), forming a well-defined core, (iii) growth of the nuclei by addition of single atoms, and finally (iv) coalescence, which is the process of formation of bigger clusters from possible aggregation (FERRANDO, 2016).

Other physical procedures include the growth of clusters on a solid substrate by the deposition of atoms from an atomic beam that may be produced from a hot-oven, laser, or pulsed arc vaporization of the cluster source. Upon deposition, metallic atoms disperse and begin the clusterization process, forming structures whose shape depends on the interactions with the support, which effectively guides the formation kinetics. Too strong interactions with the substrate, for instance, direct the growth to two-dimensional aggregates, *i.e.* interactions between the deposited atoms should be stronger than with the support for growing clusters. Metallic ions may also be incorporated in insulating matrices, that when subjected to high temperatures and a reducing/oxidizing environment, activates the diffusion process (FERRANDO, 2016).

Most of the current metallic nanoclusters are based on noble metals (Cu, Ag, Pd, Au) due to their intrinsic stability, and more studies on lower cost metals are desired for conceiving their potential industrial applications (PEI; ZHANG; SUN, 2024). Preparations of bi and multi-metallic nanoalloys, rely on cluster templates from these stable compositions, which are employed in techniques such as co-reduction, (anti)galvanic reduction, ligand exchange, reactions with a bulk metal, and intercluster reactions for achieving more complex, multimetallic structures, which may be incapable of being directly grown from the aforementioned methods (KHATUN; PRADEEP, 2020).

Characterization techniques and instrumentation are fundamental for understanding nanoalloy structures and their physico-chemical properties. Specially when dealing with structures compromised from a few atoms only, atomic precision is crucial for accurately understanding the relationship with observed electronic properties and cluster size and geometry. Nanoalloy characterization procedures include a wide array of techniques, such as: mass spectrometry, voltammetry, X-ray crystallography, X-ray photoelectron spectroscopy, X-ray absorption spectroscopy, transmission electron microscopy, scanning tunneling microscopy, atomic force microscopy, UV and IR spectroscopy, among others (SHANG; XU; NIENHAUS, 2019; KANG *et al.*, 2020; YANG *et al.*, 2024). Generally, many characterizations techniques are employed, as

each one provides different information from the synthesized nanomaterial.

2.1.2 Structural and Electronic Properties

Metallic nanoclusters emerge as a class of intermediate materials between the bulk, crystalline phase, and single atoms or metal complexes, presenting unique properties that may be explained from quantum effects on the nanoscale, finite particle regime. As we will see, many factors affect the overall cluster electronic structure, such as the composition, ordering, size, and geometry of the particle. While some metallic cluster theories have been proposed, the great diversity of structures, specially when considering alloying, poses a challenge for a general understanding of these materials, as a single atom may drastically change their properties (FERRANDO, 2016).

Nanoclusters structural motifs are mainly classified in crystalline and noncrystalline. Crystalline motifs are fragments of bulk-phase crystals and may be obtained from mechanical *top-down* methods, with common lattices being the face-centered cubic (fcc), body-centered cubic (bcc), and hexagonal close-packed (hcp). Due to the finite nature of nanoclusters, non-crystalline, symmetrical structures are also possible, being the most common the icosahedral (I_h) and decahedron (D_h) geometries (FERRANDO, 2016). Many stable clusters also assume amorphous, non-symmetrical structures, even in the case of a geometric number of atoms, however symmetrical structures tend to present increased stability which is related to the increased coordination of the composing atoms. Well-defined structures can appear in different size regimes and may be explained, but not restrict to, *magical number* of atoms, which explain concentric shell formations over a well-defined cluster core. In this work, the used dataset was composed of finite particles with symmetrical and amorphouys geometries, and not periodic crystals (KAATZ; BULTHEEL, 2019).

The increasingly theoretical and experimental studies on nanoalloys allowed for the understanding of some factors that affect formation and mixing. Alloying is favored by the relative bond strength between metals, with ordering being dependent on the composing metals radii and surface energies, which are directly related to packing effects that increase coordination, favoring formation (ASHRAF *et al.*, 2023). The electronic structure of metals contribute to the charge transfer from the inner (core) to the outer (shell) metal, creating a positively charged core with an electronic charged surface. This also allows for the functionalization of clusters by ligand-etching over the surface, allowing for different interactions with the chemical environment, which are crucial for biomedical applications where clusters are employed for specific interactions in biochemical systems (ZHENG *et al.*, 2022).

Furthermore, the size and geometry of cluster structures are also dependent on the resulting electronic occupations, as from the displacement of charges, the overall electronic structure is highly changed from the original unary metals, contributing to the stabilization or destabilization of the particle. The jellium model (KNIGHT *et al.*, 1984) is one of the most

famous electronic structure theories that aim for explaining symmetrical clusters by assuming the homogeneous distribution of charges over the entire cluster structure that when related to the cluster valence electrons allows for predicting its electronic structure. This idea was recently expanded to encompass not only symmetrical but different cluster geometries by introducing the concept of the *super-atom* and the symmetry-adapted orbital model (Fig. 4), that correlates compositions and geometries to the resulting electronic discretization (TSUKAMOTO *et al.*, 2019).

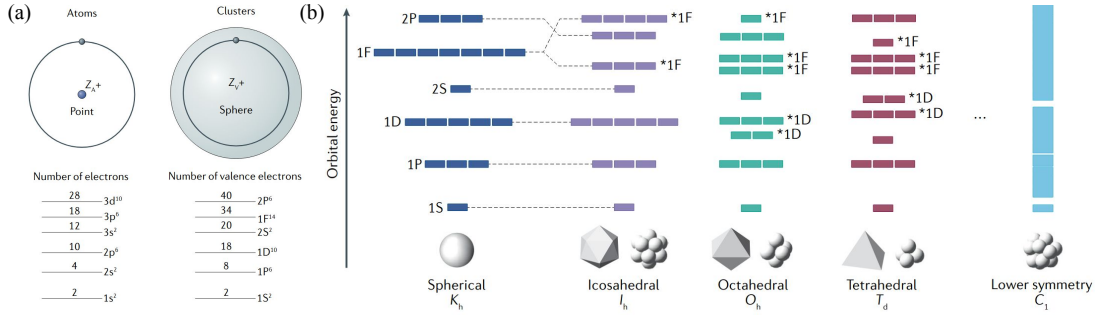

Figure 4 – Comparison of (a) electronic energy levels between a single atom and a jellium model spherical cluster and (b) the symmetry-adapted orbital model with different superatomic orbitals for each cluster symmetry. Reproduced with permission from Springer Nature (TSUKAMOTO *et al.*, 2021).

The symmetry-adapted orbital model arises from the limitation of the jellium model, which is only reasonable for spherical approximations of the nanocluster geometry. To understand this, we refer to Fig. 4, where an atom with atomic number Z_A has a nuclear charge Z_A+ and discretized energy levels corresponding to its atomic orbitals. Analogously, a cluster – referred to as a superatom – possesses a total valence electron count of Z_V and net positive charge Z_V+ within its interior, with the valence electrons occupying superatomic orbitals. The filling of the superatomic orbitals follows distinct Aufbau series as shown in Fig. 4, which may be generalized from (a) the jellium orbitals for spherical clusters to (b) the symmetry-adapted orbital model, which defines a different set of superatomic orbitals (Aufbau series) to each cluster symmetry. Consequently, the jellium model's magic numbers — representing electron counts that stabilize spherical clusters — can be extended to closed-shell electronic configurations that contribute to the stability of clusters with various sizes and geometries (TSUKAMOTO *et al.*, 2019).

The effect of cluster size on the electronic structure has also been previously assessed through voltammetry measurements. By controlling voltage and measuring current, the electronic structure is probed and associated eletronic transitions observed. From voltammetric measurements for different unary Au_n nanocluster sizes, current peaks enable identifying different electronic transitions, showing that for each size regime there are different transitions in number and energy, being related to the absorption in IR and UV wavelengths (KAWAWAKI; NEGISHI, 2023). By tuning structures such by hetero-atom doping, one can regulate the wavelength absorption, with increasingly quantum yields (QIAN *et al.*, 2024), which is of interest for different photoluminescence applications, such as bioimaging or environmental monitoring

(ZHANG *et al.*, 2019).

2.2 CO_2 Electrochemical Reduction

The CO_2 electrochemical reduction is one of the main candidates for sustainable energy and carbon management. It enables the conversion of CO_2 into commercially desirable chemicals and fuels, offering an alternative to close the carbon cycle and reduce greenhouse gas emissions. Hence, optimizing catalysts and reaction conditions is essential to improving efficiency, selectivity, and economic viability of this procedure.

2.2.1 Mechanisms of CO_2 Electrocatalysis

From some reactant to product there is a free-energy landscape that is dictated by thermodynamics. In many cases, this landscape is defined by various steps and includes many intermediates, parallel reactions, and byproducts that may be undesirable for the process of interest. Additionally, several energetic barriers may compose the landscape, requiring high amounts of energy that make the conversion process economically unfeasible. Hence, a role of a catalyst is to tune this energy landscape while attending to the Sabatier principle (Fig. 5), which states that there is an optimal adsorption energy between the surface and adsorbate that enables the catalytic process (OOKA; HUANG; EXNER, 2021).

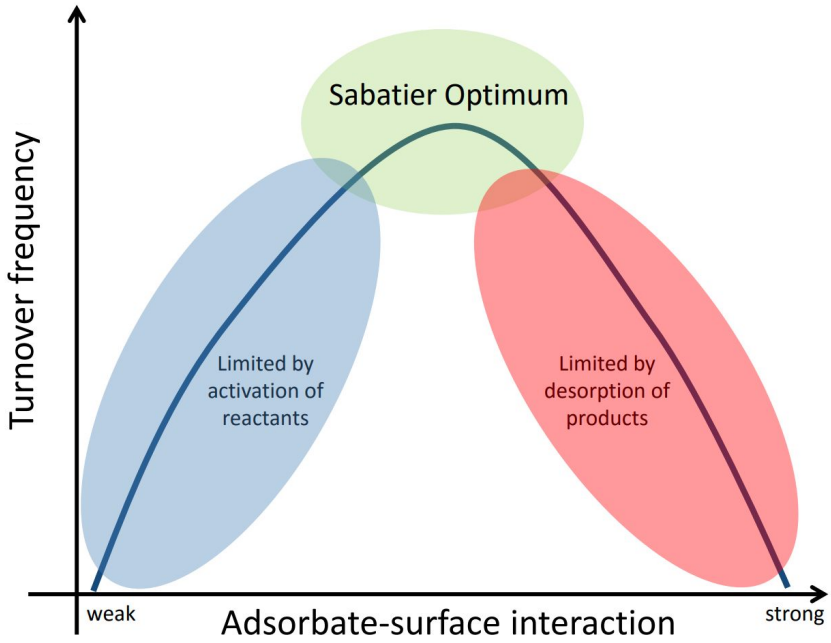

Figure 5 – Sabatier's principle: the catalyst turnover frequency depends on an optimal interaction between adsorbate and surface, not too strong, poisoning the catalyst, neither too weak that it does not activate the reactants. Source: KAKEKHANI (2016)

The CO_2 reduction proceeds through different catalyzed reactions (Fig. 6) that involve the transfer of 2 to 18 electrons. By controlling the number of transferred protons and electrons,

one can direct the conversion of CO_2 to different species which can be classified into C_1 (CO, HCOOH, CH_3OH, HCHO, CH_4) and C_2 (C_2H_4, C_2H_5OH, CH_3COOH) products. C_2 products are more commercially desirable due to higher energy density than C_1 products however their large-scale production is limited due to the required number of protons and higher energy barriers associated with the C=C bond formation (SADASIVUNI *et al.*, 2022).

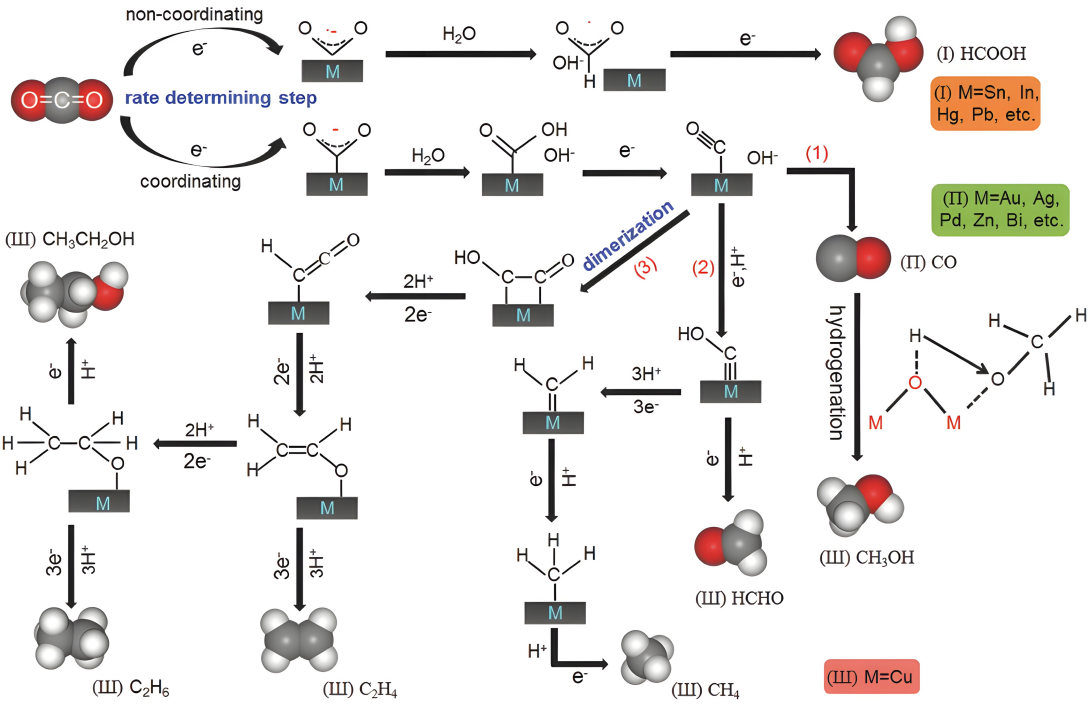

Figure 6 – Different catalytic pathways for the CO_2 electrochemical reduction. Each pathway depends on adsorption nature, substrates and reaction conditions. Reproduced with permission from John Wiley and Sons (ZHANG *et al.*, 2017).

The C–O bond energy (750 kJ/mol) is much greater than, for instance, the C–H (430 kJ/mol) or C–C (336 kJ/mol) bonds, being the initial bottleneck for conducting the CO_2 reduction reaction. Upon adsorbed, the linear, thermodynamically stable and chemically inert O=C=O molecule becomes vulnerable to nucleophilic attacks on the carbon center. Subsequently, a single-electron transfer from an electrochemical cell results in the generation of the radical anion CO_2^- (Eq. 2.1), which initiates the electrochemical reduction (KAWAWAKI *et al.*, 2024).

$$CO_2(g) + e^- \longrightarrow CO_2^- \qquad E^0_{redox} = -1.900\,V \tag{2.1}$$

From the adsorption of CO_2, the electrochemical reduction may be directed to different reaction pathways (Fig. 6) that are dependent on catalyst selectivity and activity, thermodynamic conditions, and voltaic potentials. Directing the reaction to a specific product is a challenge due to close equilibrium pontentials for the different reactions shown in Tab. 1, requiring specific catalysts that are able to thermodinamically prioritize different molecules across the catalytic

pathway. The hydrogen evolution reaction, for instance, is the main competing reaction that consumes available protons (Eq. 2.2), hindering the overall process (QIN *et al.*, 2021).

$$2\,H^{+}(aq) + 2\,e^{-} \longrightarrow H_2(g) \qquad E^{0}_{redox} = 0.000\,V \tag{2.2}$$

Table 1 – Reduction potentials under standard conditions for some of half-reactions in the CO_2 electrochemical reduction, measured against the standard hydrogen electrode (SHE) at pH 7. Source: SADASIVUNI *et al.* (2022).

Reaction	E^0 ([V] *vs.* SHE pH 7)
$2\,H^{+}(aq) + 2\,e^{-} \longrightarrow H_2(g)$	0.0
$O_2(g) + 4\,H^{+}(aq) + 4\,e^{-} \longrightarrow 2\,H_2O(l)$	1.23
$CO_2(aq) + e^{-} \longrightarrow CO_2{}^{-}(g)$	-1.9
$CO_2(g) + 2\,H^{+}(aq) + 2\,e^{-} \longrightarrow CO(g) + H_2O(l)$	-0.52
$CO_2(g) + 2\,H^{+}(aq) + 2\,e^{-} \longrightarrow HCOOH\,(l)$	-0.61
$CO_2(g) + 4\,H^{+}(aq) + 4\,e^{-} \longrightarrow HCHO(aq) + H_2O(l)$	-0.51
$CO_2(g) + 6\,H^{+}(aq) + 6\,e^{-} \longrightarrow CH_3OH(l) + H_2O(l)$	-0.38
$CO_2(g) + 8\,H^{+}(aq) + 8\,e^{-} \longrightarrow CH_4(aq) + 2\,H_2O(l)$	-0.24
$2\,CO_2(g) + 12\,H^{+}(aq) + 12\,e^{-} \longrightarrow C_2H_4(aq) + 4\,H_2O(l)$	-0.35
$2\,CO_2(g) + 12\,H^{+}(aq) + 12\,e^{-} \longrightarrow C_2H_5OH(aq) + 3\,H_2O(l)$	0.33
$2\,CO_2(g) + 14\,H^{+}(aq) + 14\,e^{-} \longrightarrow C_2H_6(aq) + 4\,H_2O(l)$	-0.27
$3\,CO_2(g) + 18\,H^{+}(aq) + 18\,e^{-} \longrightarrow C_3H_7OH(aq) + H_2O(l)$	-0.31

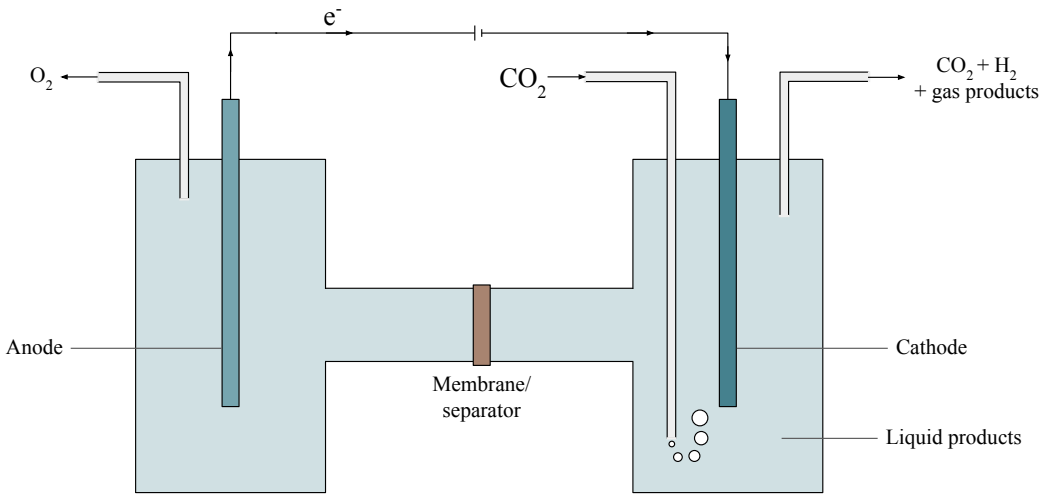

Figure 7 – H-cell scheme for the CO_2 electrochemical reduction. Reproduced with permission from the Royal Society of Chemistry (FERMIN; MARKEN, 2018).

Fig. 7 presents the H-cell assembly for performing the CO_2 electrochemical reduction, which is composed of two compartments separated by an ion-exchange membrane that prevents the mixing of the oxidation-reduction products. The cell is composed of a cathodic and anodic compartments, each with an electrode immerse in an electrolyte medium. The electrodes are connected to an external power supply so that energy may be provided for the reaction. The process of begins with the CO_2 mass transfer from the gaseous phase to the bulk electrolyte. The

oxygen evolution reaction acts the proton source in the anode, through the reaction presented in Eq. 2.3 (EWIS *et al.*, 2023)

$$2\,H_2O\,(l) \longrightarrow O_2(g) + 4\,H^+(aq) + 4\,e^-, \tag{2.3}$$

with protons dispersing in the electrolyte onto the electrolyte-cathode interface, where adsorption occurs. Among the cathode surface, adsorbed protons (H^+) initiate the hydrogenation of adsorbed CO_2 molecules, which transforms into different intermediates such as $\cdot$CO, $\cdot$COH, $\cdot$CHO, and $\cdot$COOH. Electrons are then transferred from the catalyst to the intermediates, reducing them to products which either migrate to bulk liquid phase or evolve as bubbles in the gas phase (EWIS *et al.*, 2023).

A voltage grater than reported for the standard hydrogen electrode is necessary to overcome different energy barriers that occur in electrolysis. These include activation barriers at both cathode and anode, loss of electrode surface due to bubble formation, ohmic loss caused by ion transport across the membrane, and ion conduction in the overall bulk electrolyte. Different components of the circuit can also pose additional resistance, contributing to this over-potential. In the development of novel catalysts, the reduction of over-potential is one of its goals (SADASIVUNI *et al.*, 2022).

2.2.2 Challenges and Perspectives

Substantial research has been done on the CO_2 reduction reaction, analyzing different electrodes, electrolytes and catalyst and their effects on selectivity, turnover rate and Faradaic efficiency. For instance, it is well-known that different metal nanoclusters yield distinct products (KAWAWAKI *et al.*, 2024):

- Au and Ag favor CO production;
- Pb, Sn, and In favor HCOOH;
- Cu directs the catalysis to a variety of organic products, such as hydrocarbons and alcohols, making it particularly versatile for fuel synthesis;
- Pt, Ni, and Pd are highly active in the hydrogen evolution reaction, but not so useful for the CO_2 catalysis due to $\cdot$H adsorption, which prevents the formation of carbon products.

Nanoclusters present a high surface to volume ratio, relating to a high number of low-coordinated metallic atoms among its surface that present high catalytic activity. As cluster geometry and size relate to the nanocluster stability, consistent particle size and site types are exhibited by this material, effectively directing the reduction reaction. From a solved nanocluster structure, through the calculation of its electronic structure and d-band distribution, the adsorption

strength of different adsorbates may be determined (QIAN *et al.*, 2024). Size, geometry, doping, and ligands are all affecting factors on catalytic selectivity and activity as well as Faradaic efficiency for electrocatalysis, enabling tuning structures for different catalytic pathways that may favor specific products (PEI; ZHANG; SUN, 2024).

While Au and Ag clusters have been proven to effectively catalyze CO_2 to CO formation, Cu clusters favors different products due to increased intermediate stability over this composition. While, size, geometric effects may direct selectivity, expanding the structural diversity of nanoclusters beyond Cu-based architectures could further enhance possibilities. Investigating alternative compositions, including dopings with different transition metals, may improve reactivity and stability. Exploring nanoclusters composed of elements from different periodic table groups may unlock new pathways that may be pivotal for advancing CO_2 catalysis, specially for industrial-scale applications (BISWAS *et al.*, 2024).

3 Theoretical Background

The subsequent sections address the mathematical foundations relevant to the study, including key concepts from linear algebra, fundamental concepts of quantum chemistry that are required for performing computational calculations on chemical systems, and cheminformatics modeling, encompassing featurization and regression methods, crucial to the development of this project.

3.1 Linear Algebra

Linear algebra consists of methodologies for manipulating structured data which are crucial for modeling and gathering knowledge, specially in the case of complex problems. These techniques are fundamental mathematic concepts that allowed for the development of many fields in science. The following sections present important concepts that lay the main principles for this work's methodology.

3.1.1 Vectors, Matrices, and Tensors

Vectors and matrices are mathematical objects used to represent and manipulate structured data, often composed of multidimensional spaces. Commonly, in fields such as physics, these objects are used to represent points, directions or quantities in three-dimensional space. Through organizing measurements, we may describe some space across various dimensions from sequences of scalars. Usually, space represented by more than three dimensions is referred to as hyperspace, or hyperdimensional. A n-space vector consists of n dimensions that can be chosen to represent whatever measurement unit that is relevant for the description of some object of interest. Vectors can then be organized into a matrix, or data table, whose rows and columns can be interpreted as features and samples, respectively. All these objects are classified as tensors, which generalize for any array dimensionality, encompassing more complex data structures (Fig. 8) (DEISENROTH; FAISAL; ONG, 2020).

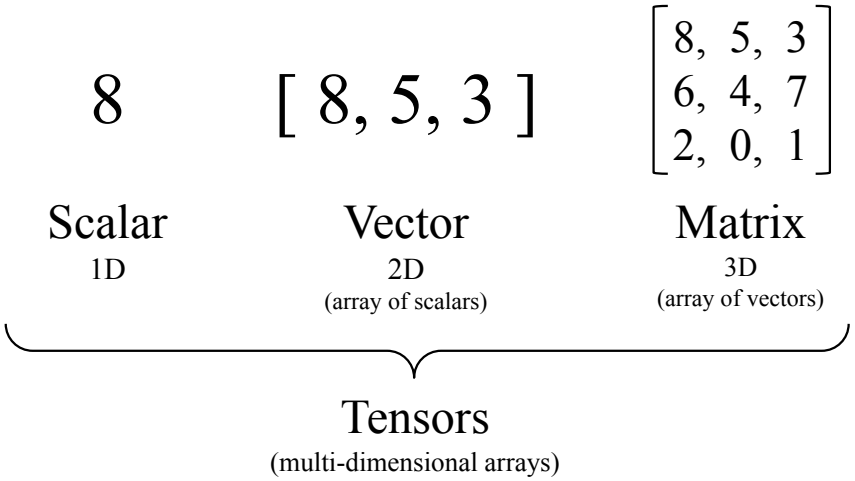

Figure 8 – Tensors as multidimensional arrays.

The geometric vectors are usually denoted by a small arrow above the letter (e.g. $\vec{x}$ or $\vec{y}$), however, here we chose a more general notation for describing an abstract $\mathbb{R}^n$ space. A data table consists of a rectangular matrix $\mathbf{A}$ with I rows ($i = 1, ..., I$, samples) and J columns ($j = 1, ..., J$, features), hence of size $I \times J$. The individual elements of $\mathbf{A}$ are denoted by a_{ij} and compose a given vector $\mathbf{a}_i$ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

3.1.2 Basis Set

Representing some real space $\mathbb{R}^n$ may be achieved through different vector and matrix representations, and, in this work, the change of basis was specifically used in the principal component analysis (Sec. 3.3.2) for representing the featurized dataset on a lower dimensionality. In a vector space $\mathbb{R}^n$, there is a set of vectors $\mathscr{B}$ that posses the property of describing any vector $\mathbf{v} \in \mathbb{R}^n$ from linear combinations. $\mathscr{B}$ consists of the basis set of vectors that can describe any possible vector in $\mathbb{R}^n$ through linear combinations. Vectors from $\mathscr{B}$ are linearly independent, and any other vector added to this set will make it linearly dependent. In $\mathbb{R}^3$, the canonical or standard basis is (Eq. 3.1)

$$\mathscr{B} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}. \tag{3.1}$$

Every vector space $\mathbf{V}$ possesses infinite possible bases, as there is no unique basis. All the basis sets, however, have the same number of elements or basis vectors, which describe each one a space dimension. Decomposing a matrix in its basis vectors allows for simplifying calculations and linear transformations, which is useful for describing and manipulating the vector space coordinate system (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

As any basis can be chosen to represent the coordinate system, we can define a change of basis (Eq. 3.2)

$$\mathbf{PX} = \mathbf{Z}, \tag{3.2}$$

where the original data $\mathbf{X}$ is transformed from a new basis set $\mathbf{P}$ onto a new representation $\mathbf{Z}$. Any resulting vector $\mathbf{z}_i$ is considered to be a projection onto the basis $\{\mathbf{p}_1, \ldots \mathbf{p}_j\}$. Choosing a new basis set to represent data may be useful for efficient computational methods and re-casting data in a more suitable way (SHLENS, 2014).

3.1.3 Determinants

The determinant is a scalar-valued function that maps a square matrix onto a real number, possessing many properties of interest for different linear methods. In this work, Slater determinants (Sec. 3.2.1.2) are utilized as a mathematical tool for inserting the antisymmetry principle in theoretical chemistry calculations. Geometrically, determinants may be interpreted as the area described by a 2×2 matrix or the volume of a 3×3 matrix of spatial vectors. It provides important information about the matrix, such as whether it is invertible. The determinant of an $n \times n$ matrix $\mathbf{A} = [a_{ij}]$ is denoted as $\det(\mathbf{A})$ or $|\mathbf{A}|$ and can be recursively defined using cofactor expansion (DEISENROTH; FAISAL; ONG, 2020)

$$\det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{A}_{ij}), \tag{3.3}$$

where $\mathbf{A}_{ij}$ is the $(n-1) \times (n-1)$ submatrix obtained by removing the i-th row and j-th column from $\mathbf{A}$.

For a 2×2 matrix:

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \det(\mathbf{A}) = ad - bc. \tag{3.4}$$

For a 3×3 matrix:

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}, \quad \det(\mathbf{A}) = aei + bfg + cdh - ceg - bdi - afh. \tag{3.5}$$

A matrix is singular if $\det(\mathbf{A}) = 0$, meaning it is not invertible.

Determinants may also be defined in terms of permutations. A permutation σ is a function that operates over an ordered set providing the same elements in a different order. The permutation can be written as matrix that multiplies the rows on a column vector (Eq. 3.6)

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} = \begin{bmatrix} a_2 \\ a_4 \\ a_1 \\ a_5 \\ a_3 \end{bmatrix}. \tag{3.6}$$

For a set of size n there are a total of $n!$ possible permutations. For achieving each permutation, we may consider a conjunction of simple permutations called transpositions. Transpositions consist of swapping exactly two indices of a set, leaving the rest unaltered. An even or odd number of transpositions indicates the parity of the permutation, denoting the sign of σ as $\text{sgn}\sigma$ as $+1$ if σ is an even permutation and -1 if it is an odd permutation. The determinant of a square $n \times n$ matrix $\mathbf{A}$ is the sum of $n!$ terms, one for each permutation σ of the set $1,\ldots,n$, which may be defined in terms of permutations as (Eq. 3.7). While computing determinants from the permutations perspective is not computationally efficient, it provides another intuitive way of thinking about the determinant. (JOYCE, 2015)

$$|\mathbf{A}| = \sum_{\sigma} \text{sgn}\sigma A_{1\sigma 1},\ldots,A_{n\sigma_n}. \tag{3.7}$$

3.1.4 Eigenvectors and Eigenvalues

Eigen is a German word meaning "characteristic," "self," or "own". In linear algebra, an *eigen* analysis studies special vectors of a matrix known as *eigenvectors* and their associated scalar values called *eigenvalues*. In this work, eigenvectors and eigenvalues are calculated for the principal component analysis (Sec. 3.3.2), allowing for estimating the main dimensions of covariance for a featurized dataset of high dimensionality. Given a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, the eigenvalue equation is:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \tag{3.8}$$

where $\lambda \in \mathbb{R}$ is an eigenvalue of $\mathbf{A}$, and $\mathbf{x} \neq \mathbf{0}$ is an associated eigenvector. To find the eigenvalues, we solve the characteristic equation:

$$\det(\mathbf{A} - \Lambda\mathbf{I}) = 0, \tag{3.9}$$

where $\mathbf{I}$ is the identity matrix of the same dimension as $\mathbf{A}$. The roots of this characteristic polynomial give the eigenvalues λ_i. Once the eigenvalues are found, we determine the eigenvectors by solving the linear system:

$$(\mathbf{A} - \Lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \tag{3.10}$$

Eigenvectors and eigenvalues have widespread applications in differential equations, quantum mechanics, principal component analysis, and many other fields of science and engineering. They also play a fundamental role in diagonalizing matrices and understanding transformations in vector spaces (DEISENROTH; FAISAL; ONG, 2020).

3.2 Quantum Chemistry

The field of quantum chemistry emerges in the end of the XIX century, when different experiments put the nature of matter into question. The double slit experiment is one of the most well known to show the duality of light, but other rationalizations such as the black-body radiation also cast doubt on classical physics (ATKINS, 2018). The following sections introduce quantum chemistry theory, specifically the foundations for the DFT calculations, which are nowadays one of the main methods for computational material sciences.

3.2.1 Time-independent Schrödinger Equation

Differently from classical mechanics, where objects travel along defined trajectories, in quantum mechanics particle states are described by a wave function (or state function) Ψ, which is distributed across space, in opposition to being localized. The wave function contains all possible information about a quantum system and can be used for obtaining properties of interest. This idea was first conceptualized by SCHRÖDINGER in 1926, who proposed the equations governing its change with time. For many applications in quantum chemistry, however, the simpler time-independent Schrödinger equation is used as it suffices calculating molecular electronic structures of stationary states (LEVINE, 2014).

Describing an atomic system state requires considering all energies and positions of this system particles. Eq. 3.11 presents the compact form of the time-independent Schrödinger equation, where this information is included in the form of operators

$$\hat{H}\Psi = E\Psi. \tag{3.11}$$

The wave function Ψ is an eigenfunction of the Hamiltonian operator $\hat{H}$, whose eigenvalues are the system's total energy E. From solving for the eigenvalues and eigenfunctions, possible energy levels and their associated wave functions (states) are achieved. Eq. 3.12 presents the expanded Hamiltonian of a molecule, constructed from operators that are associated to different energy contributions (ATKINS, 2018):

$$\hat{H} = \hat{K}_e + \hat{K}_N + \hat{V}_{ee} + \hat{V}_{NN} + \hat{V}_{eN}. \tag{3.12}$$

The composing operators of the Hamiltonian are the electrons and nuclei kinetic energies, $\hat{K}_e$ and $\hat{K}_N$, and potential energies from electron-electron repulsions, nucleus-nucleus

Table 2 – Atomic units up to five significant digits for the four independent units required for a mechanical-electromagnetic system. Source: ECHENIQUE; ALONSO (2007).

Unit of mass	Electron rest mass: $m_e = 9.1094 \times 10^{-31}$ kg
Unit of charge	Charge of the proton: $e = 1.6022 \times 10^{-19}$ C
Unit of length	1 bohr: $a_0 = 4\pi\varepsilon\hbar^2/m_e e^2 = 0.529\,18$ Å $= 5.2918 \times 10^{=11}$m
Unit of energy	1 hartree: $\hbar^2/m_e a_0^2 = 6627.51$kcal mol$^{-1} = 4.3597$ J

repulsions and electron-nucleus attractions, $\hat{V}_{ee}$, $\hat{V}_{NN}$, and $\hat{V}_{eN}$, respectively. Eq. 3.13 exemplifies the Hamiltonian operator for the helium atom using SI units:

$$\hat{H} = -\frac{\hbar^2}{2m_e}\nabla_1^2 - \frac{\hbar^2}{2m_e}\nabla_2^2 - \frac{2e^2}{4\pi\varepsilon_0 r_1} - \frac{2e^2}{4\pi\varepsilon_0 r_2} + \frac{e^2}{4\pi\varepsilon_0 r_{12}}, \tag{3.13}$$

where $\hbar$ is the reduced Planck constant, e the elementary charge, m_e the electron rest mass, and ε_0 the vacuum permittivity. Here, $\hat{H}$ is expressed in terms of electron positions (r_1 and r_2) and ∇_1, ∇_2 refer to each electron's position differential operator. For simplifying writing the Schrödinger equation, Tab. 2 presents the Atomic Units system, which introduces four fundamental units for describing quantum systems. This convention avoids numerical work from high powers of 10, simplifying the notation in terms of units (ECHENIQUE; ALONSO, 2007). Adopting the Atomic Units convention, Eq. 3.13 can be rewritten as 3.14:

$$\hat{H} = -\frac{1}{2}\nabla_1^2 - \frac{1}{2}\nabla_1^2 - \frac{2}{r_1} - \frac{2}{r_2} + \frac{2}{r_{12}}. \tag{3.14}$$

Hence, since all relevant expressions in quantum chemistry are derived from the molecular Hamiltonian, the simplifications by using atomic units were adopted for all subsequent formalism. Most quantum chemistry software packages also express physical quantities using these units.

A basic postulate of quantum mechanics is how to construct an operator corresponding to a given observable. The position and linear momentum operators presented in Eq. 3.15 are utilized for building operators for different observables, as every physical property has a corresponding quantum-mechanical operator. The position operator is simply represented by multiplication with a spatial variable (x, y, z), whereas the linear momentum operator incorporates the reduced Planck constant $\hbar = \frac{h}{2\pi}$ and the imaginary unit i which guarantees the operator's hermiticity (LEVINE, 2014).

$$\hat{x} = x\times \quad , \quad \hat{p}_x = \frac{-\hbar}{i}\frac{\partial}{\partial x} \tag{3.15}$$

The Born interpretation of the wave function proposes $|\Psi|^2 = \Psi^*\Psi$ to be considered as the probability density, which is one way of conceptualizing the wave function physical significance. From this idea of probability distribution, the expectation value, or the average value of a significantly large number of measurements, for any given observable operator $\hat{\Omega}$ can

be calculated from $\langle\Psi|\hat{\Omega}|\Psi\rangle$, which is the Dirac's notation for $\int \Psi^* \hat{\Omega} \Psi d\tau$. From this concept, the standard deviation (σ) and other statistical quantities may be also calculated as shown in Eq. 3.16 (ATKINS, 2018)

$$\sigma_{\hat{\Omega}} = \sqrt{\langle\hat{\Omega}^2\rangle - \langle\hat{\Omega}\rangle^2}. \tag{3.16}$$

3.2.1.1 Born-Oppenheimer Approximation

The atomic nucleus is much heavier than its orbiting electrons and consequently moves relatively slowly. Hence, the Born-Oppenheimer approximation states that the nuclei may be treated as stationary when treating the electronic motion, allowing the separation of the wave function into the electronic and nuclear part (Eq. 3.17)

$$\Psi(r,R) = \psi_e(r;R)\psi_N(R). \tag{3.17}$$

The electronic wave function is described by ψ_e, which is explicitly dependent on the electron coordinates r and parametrically dependent on nuclei coordinates R. Assuming fixed nuclei in space, a resulting potential is stablished that allows the separation of the full wave function into the electronic and nuclei portions ψ_e and ψ_N respectively (3.17). This formulation allows for fixing different positions for the nuclei, calculating their effective potential, and solving the electronic portion of the Schrödinger equation independently (ATKINS, 2018).

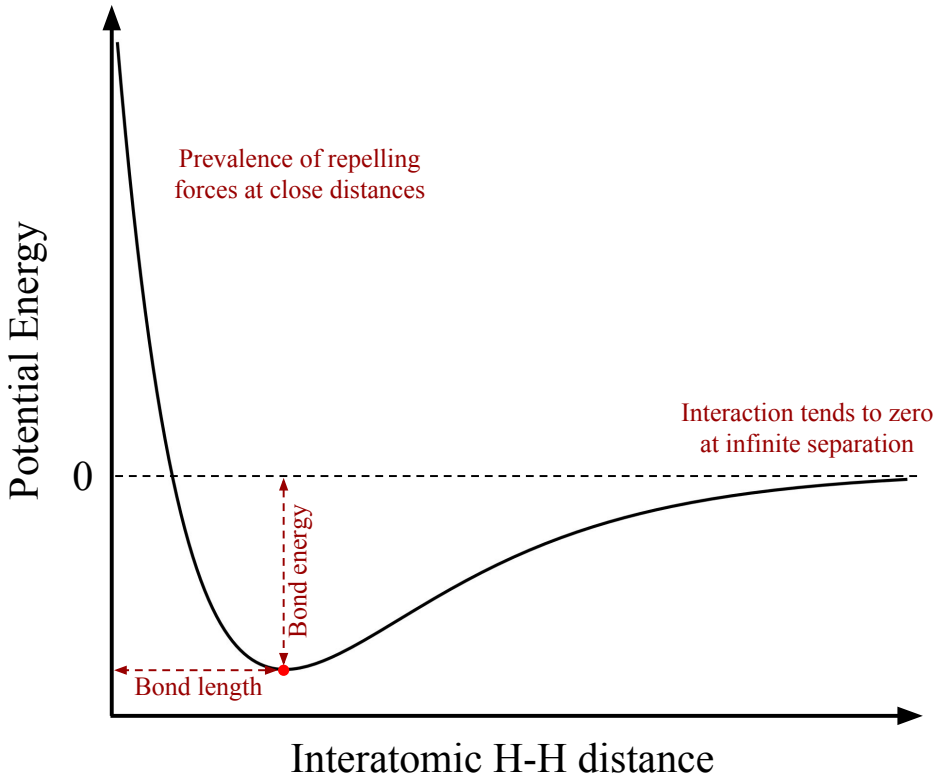

Figure 9 – Illustration for the potential energy surface for different interatomic distances of the H_2 molecule.

Fixing the nuclear coordinates allows for the calculation of single-point energies that compose the potential energy surface of the chemical system. In the case of the H_2 molecule (Fig. 9), for instance, the four-body problem of two pairs of nuclei and electrons is transformed into a two-body problem that may be analytically solved utilizing the fixed potential and the electronic wave function. This approximation has negligible errors as electron contributions have a larger importance in the overall system energy than nuclei kinetics. The Born-Oppenheimer approximation is fundamental for the development of further computational chemistry methods by providing a way to *ab initio* calculations programmatically exploring the potential energy landscape in search of an optimal geometry for the system (JENSEN, 2007).

3.2.1.2 Many-electrons Wave Functions

While the Schrödinger equation can be exactly solved for hydrogen-like atoms or one-electron systems, the wave function cannot be calculated exactly for molecules with $n > 2$ electrons due to the existence of inseparable terms with coupled electron-coordinates, whose energetic contributions are not mathematically known. That being said, computational chemistry methods still make use of the wave function for iteratively approximating energy values through operators. Thus, there should be a way for approximating an unknown wave function so that an initial guess to the iterative process can be provided (JENSEN, 2007).

For approximating a many-electron wave function, one could start assuming that the electrons do not interact with each other. This idea completely neglects effects from electron-electron interactions (exchange and correlation), but allows for approximating the wave function through the Hartree product, shown in Eq. 3.18 (LEVINE, 2014):

$$\psi_{\mathrm{HP}}(\vec{r}_1, \vec{r}_2, ..., \vec{r}_n) = \prod_{i=1}^{n} \phi_i, \tag{3.18}$$

where the Hartree product HP describes the approximated total wave function of the system and is calculated from the product of spatial orbitals ϕ_i. This formulation does not satisfy the antisymmetry requirement, or Pauli's exclusion principle, but can be further improved by adding spin functions leading to the Slater determinant (Eq. 3.19) ((SZABO, 1996))

$$\psi(\vec{r}_1, \omega_1, ..., \vec{r}_n, \omega_n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_1(\vec{r}_1, \omega_1) & \phi_2(\vec{r}_1, \omega_1) & \dots & \phi_{(n-1)}(\vec{r}_1, \omega_1) & \phi_n(\vec{r}_1, \omega_1) \\ \phi_1(\vec{r}_2, \omega_2) & \phi_2(\vec{r}_2, \omega_2) & \dots & \phi_{(n-1)}(\vec{r}_2, \omega_2) & \phi_n(\vec{r}_2, \omega_2) \\ \dots & \dots & \dots & \dots & \dots \\ \phi_1(\vec{r}_n, \omega_n) & \phi_1(\vec{r}_n, \omega_n) & \dots & \phi_{(n-1)}(\vec{r}_n, \omega_n) & \phi_n(\vec{r}_n, \omega_n) \end{vmatrix}, \tag{3.19}$$

where each spin-orbital is a product of a spatial orbital ϕ_i with a spin function ω_i that may assume two distinct values, α or β, representing possible electrons spins that are orthogonal, including the antisymmetry requirement of the wave function. The determinant is also multiplied by the factor $\frac{1}{\sqrt{N}}$, which normalizes the resulting wave function (LEWARS, 2016).

While this formulation allows for calculating the approximated total wave function, the spatial orbitals ϕ_i still need to be defined so that calculations may be performed. For that, each spatial orbital ϕ_i is expressed through a linear combination of atomic orbitals (LCAO), which are defined from basis functions $\mathscr{X}_r$ with coefficients c_{ri}. Basis sets are pre-defined and of great importance for the calculations as they will describe the overall wave function shape (Eq. 3.20) (PIELA, 2007).

$$\phi_i = \sum_i c_{ri} \mathscr{X}_r. \tag{3.20}$$

The basis set is always limited, as there is an associated computational cost for storing all information for larger basis sets, rapidly increasing with the total number of atoms in the system. Different basis sets may be better suited for different systems, and generally computational chemistry research employs many basis sets for assessing accuracy. For improving computations, Gaussian-type functions are the most used basis functions, as their mathematical properties allows for an analytical integration of LCAOs, facilitating optimization algorithms (LEVINE, 2014).

3.2.2 Density Functional Theory

While the wave function is a useful tool for calculating n-electron systems through the use of approximations such as the Hartree-Fock method (ECHENIQUE; ALONSO, 2007), it still depends on $3n$ spatial and n spin coordinates. Consequently, the computational complexity of said methods scales as $O(n^4)$, making them unsuitable for larger systems such as many-atoms clusters. This motivated the search for different functions with fewer variables, such as the electronic density (LEVINE, 2014).

The electronic density $\rho(r)$ is a physical quantity of a system which may be interpreted as the electron probability density. It can be obtained from the summation of one-electron probabilities as shown in Eq. 3.21

$$\rho(r) = \sum_{i=1}^{N} P_i(r), \tag{3.21}$$

where the indistinguishable uni-electronic probability densities P_i are the integration of all *other* electrons probability densities (Eq. 3.22)

$$P_1(r_1) = \sum_{m_s} \int_{-\infty}^{+\infty} |\Psi(r_1, r_2, ..., r_N)|^2 d\tau_2, d\tau_3, ...d\tau_N, \tag{3.22}$$

with summation of all possible spin states m_s for all electrons in the system (LEVINE, 2014). The one-electron probabilities compose each a part of the electronic density, with the total number of electrons in system (N) obatined as (Eq. 3.23)

$$\rho(r) = N \sum_{m_s} \int_{-\infty}^{+\infty} |\Psi(r_1, r_2, ..., r_N)|^2 d\tau_2, d\tau_3, ...d\tau_N \tag{3.23}$$

$$N = \int_{-\infty}^{+\infty} \rho(r) d\tau. \tag{3.24}$$

Differently from the wave function, the electronic density depends only on the three spatial coordinates (x, y, z) independently to the system size. It is also an experimental observable, measurable by X-ray diffraction, electron diffraction, Transition Electron Microscopy (TEM), Atomic Force Microscopy (AFM), and Scanning Transmission Micrsocopy (STM) (LEWARS, 2016).

3.2.2.1 Hohenberg-Khon Theorems

In a molecular system, electrons interact not only with each other but also with the nuclei, which exert an electrostatic influence on them. This interaction may be described by the external potential $v_{ext}(\vec{r})$, which arises from the Coulomb attraction between the electrons and the nuclei. Since $v_{ext}(\vec{r})$ depends only on the nuclear charges and positions, it may be utilized for calculating the electronic density. From this idea, two theorems on the electron density proposed by Hohenberg & Kohn (1964) allowed for an analytical treatment of the electronic density (LEVINE, 2014).

The first Hohenberg & Kohn theorem proves by contradiction that the external potential $v_{\text{ext}}(\vec{r})$, is a unique functional of the ground state electron density $\rho_0(r)$. Since $v_{\text{ext}}(\vec{r})$ fixes the Hamiltonian operator, any ground state property of the molecule is a functional of $\rho_0(r)$, *i.e.* there is a one-to-one correspondence between the property and the electron density, as expressed in Eq. 3.25 for the ground state energy E_0. This ensures that there is, in principle, some way to calculate molecular properties from the electronic density, even if the exact functional is unkown (LEWARS, 2016).

$$E_0 = E[\rho_0(x, y, z)]. \tag{3.25}$$

The proof of existence guarantees that no two external potentials could give the same ρ_0, but lacks a practical way for approximating the ground state density. For establishing this procedure, the second Hohenberg & Kohn theorem is analogous to the Variation Principle, stating that an approximate density function must yield an energy higher or equal to the exact ground state energy E_0, as shown in Eq. 3.26 (LOWE; PETERSON, 2006).

$$E_{\text{approx}}[\rho_{\text{approx}}] \geq E_0[\rho_0]. \tag{3.26}$$

This proposition assumes a fixed $v_{\text{ext}}(\vec{r})$, restraining the optimization of ρ to the exchange and correlation electronic effects. These theorems guarantee that there is a way of calculating

optimal values for the ground state energy from the electronic density through the refining of the electron density. However, the electron-electron interactions part of the energy functional and the real electronic density are unknown and need to be approximated somehow (LEWARS, 2016).

3.2.2.2 Kohn-Sham Self-Consistent Method

For estimating the electron density, Kohn & Sham proposed in 1965 considering a fictitious reference system $\rho_s(\vec{r})$ with N non-interacting electrons. This electron density may be regarded as an ideal gas of electrons, subjected to an external potential $v_s(\vec{r})$. We can choose any $v_s(\vec{r})$ such that the resulting $\rho_s(\vec{r})$ matches the true ground-state electronic density $\rho_0(\vec{r})$. This is justified by the first Hohenberg-Kohn theorem, which states that $\rho_s(\vec{r})$ uniquely determines the external potential $v_s(\vec{r})$. Eq. 3.27 presents the Kohn-Sham (KS) Hamiltonian for the reference system (LEVINE, 2014)

$$\hat{H}_{\mathrm{KS}} = \sum_{i=1}^{N}[-\frac{1}{2}\nabla_i^2 + v_s(r_i)] \equiv \sum_{i=1}^{N}\hat{h}_i^{\mathrm{KS}}, \tag{3.27}$$

where the first and second terms are the N electron kinetic energies and electron-nucleus attractions to the M nucleus that compose, respectively. The electron-electron contributions are absent due to the supposition of non-interactingelectrons. Each uni-electronic Hamiltonian $\hat{h}_i^{\mathrm{KS}}$ has its corresponding uni-electronic KS orbital ϕ_i^{KS}, that in conjunction may be used for approximating the electronic density (Eq. 3.28) (LEWARS, 2016)

$$\rho_s(\vec{r}) = \sum_{i=1}^{N}|\phi_i^{\mathrm{KS}}(\vec{r})|^2. \tag{3.28}$$

With a way for estimating the electron density, we may try to write the energy functional, but firstly, the errors from the reference system approximation should be considered. Electron repulsions are expected to affect electron kinetics and the associated error to the electron kinetic energy (ΔT) may be defined as in Eq. 3.29 from the unknown, exact kinetic energy functional $T[\rho]$ and the approximated functional to the reference system, which may be written as the integration of the electron density and the reference potential

$$\langle \Delta T[\rho] \rangle \equiv \langle \Delta T[\rho] \rangle - \frac{1}{2}\sum_{i=1}^{N}\langle \phi_i^{KS}|\,\nabla_i^2\,|\phi_i^{KS}\rangle \tag{3.29}$$

Considering the non-interactive nature of the reference system, the inter-electron repulsions may be approximated in a classical way from the integration of the homogeneous charge density of the electronic density in the Coulomb term. Analogously, the error for inter-electron repulsions is expressed as in Eq. 3.30 (LEVINE, 2014)

$$\langle \Delta V_{ee}[\rho] \rangle \equiv \langle V_{ee}[\rho] \rangle - \frac{1}{2}\int\int\frac{\rho(r_1)\rho(r_2)}{r_{12}}dr_1dr_2. \tag{3.30}$$

Hence, the ground-state energy functional may be written as (Eq. 3.31)

$$E_0 = \int \rho(r)v(r)dr - \frac{1}{2}\sum_{i=1}^{N} \langle \phi_i^{KS} \nabla_i^2 | \rangle \phi_i^{KS}| + \frac{1}{2}\int\int \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 + \Delta T[\rho] + \Delta V_{ee}[\rho], \tag{3.31}$$

and differences $\Delta T[\rho]$ and $\Delta V_{ee}[\rho]$, that describe electronic exchange and correlation effects and may be included in the same functional (Eq. 3.32) (LEWARS, 2016)

$$E_{\mathrm{XC}}[\rho_0] \equiv \Delta T[\rho] + \Delta V_{ee}[\rho]. \tag{3.32}$$

Consequently, the main factor affecting the accuracy of DFT calculations is how the exchange-correlation functional is approximated. Exchange-correlation functionals range from the simplest Local Density Approximation (LDA) and Generalized Gradient Approximation (GGA) to complex hybrid functionals that included Hartree-Fock corrections that increase the overall computational cost and may be unfeasible for bigger systems. While functionals may be sufficiently accurate for system's total energies, other properties such as band-gaps may be poorly approximated, requiring further treatments (JENSEN, 2007).

As the electronic density is achieved from the KS orbitals, for including the exchange-correlation effect in the optimization of said orbitals the exchange-correlation potential v_{XC} is included in the Hamiltonian. The relation in Eq. 3.33 is utilized for obtaining this potential (LEVINE, 2014)

$$v_{\mathrm{XC}}(r)\frac{\partial E_{\mathrm{XC}[\rho(r)]}}{\partial \rho(r)} \tag{3.33}$$

Hence, all tools necessary for calculating energies from the electron density are attained. For minimizing energies, the Self-Consistent Field (SCF) process is utilized, as schematized in Fig. 10. The KS equations must be solved iteratively using a SCF method from a guessed electronic density, minimizing the energy upon some convergence criteria, such as the maximum energy difference from the previous and current attained electron densities. Upon convergence, the calculated energy is utilized in the geometric optimization of the system, directing to another SCF cycle (JENSEN, 2007).

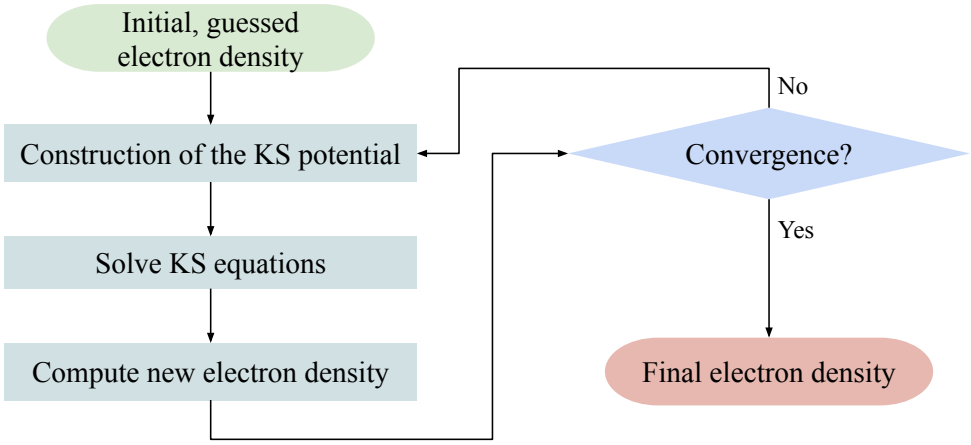

Figure 10 – SCF procedure from the Kohn-Sham formalism.

3.3 Cheminformatics Modeling

Cheminformatics is the field of using data-driven methodologies for solving chemical problems, which are commonly characterized as exploratory. This section introduces molecular descriptors, which are fundamental in the featurization of chemical systems into ML-able vectors that describe the data structural information. These features can be analyzed and improved through unsupervised feature analysis methods, essential for treating or exploring the sample space to be used with the regression algorithms which will operate for predicting properties of interest.

3.3.1 Molecular Descriptors

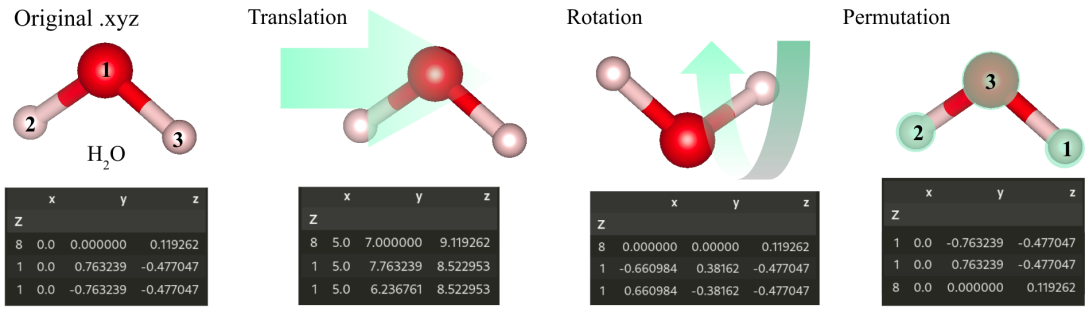

Figure 11 – Translations, rotations, or permutations changes data contained in the `.xyz` file format even if the molecule remains unaltered.

While the use of `.xyz` files is sufficient for describing molecular systems for *ab initio* calculations, this file format lacks properties to uniquely and numerically represent different systems. As shown in Fig. 11, any translation, rotation, or simply changing the ordering of atoms in the file creates a different set of coordinates, that while still describing the same molecule, posses different numerical values, hence making `.xyz` files not an ideal system fingerprint. Ideally, a molecular descriptor should be (LANGER; GOEßMANN; RUPP, 2022):

1. Invariant: should be unchanged by any transformation that preserves the predicted property;
2. Unique: two systems with different values for one property should be mapped to different representations;
3. Continuous and differentiable: enables the use of differentiable ML models, requiring less training data;
4. Computationally efficient: the computational cost by employing the model should be significantly lower than of DFT calculations, justifying its use;
5. General: capable of enconding any atomistic system regardless of size or composition with consistent feature vectors.

The numerical representation of molecules is an area that consistently receives new contributions and ideas for more efficient ways to describe chemical systems. The following sections details three structural descriptors that have been utilized in this study for the development of the ML model.

3.3.1.1 Coulomb Matrix

The Coulomb matrix (CM) descriptor, as introduced by Rupp *et al.* (2012), is one of the most well-known molecular descriptors. It has been widely used in numerous studies due to its straightforward implementation, low computational cost, and effective representation (BATISTA *et al.*, 2016; MORAIS; ANDRIANI; SILVA, 2021; SILVA *et al.*, 2022; MENDONCA *et al.*, 2022). By its formulation, the CM inherently guarantees rotational and translational invariance, that is, it uniquely identifies some chemical system regardless of its position or orientation in space. For achieving this, the CM is constructed from the Coulombic repulsions between all atoms in the system as shown in Eq. 3.34:

$$M_{ij} = \begin{cases} 0.5Z^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{|R_i - R_j|} & \text{for } i \neq j \end{cases}. \tag{3.34}$$

The off-diagonal elements of the CM correspond to the electrostatic repulsion between nuclei, which depend on their charges Z and positions R. The diagonal elements, encode a fit of the atom's potential energy based on its nuclear charge Z. This results in a matrix that effectively describes the system, independent of its position or orientation, since its construction relies solely on the relative position of atoms within the system. However, the CM is not invariant to permutations, as it can be constructed from any declared order of atoms. For instance, for the water molecule the matrix can be declared with orderings such as (H,H,O), (H,O,H), and (O,H,H), where each hydrogen atom is unique as it corresponds to a different atom in the molecule. This limitation is mainly mitigated through the use of the eigenspectrum approximation, a vector

composed by the CM eigenvalues in descending order (HIMANEN *et al.*, 2020). Fig. 12 elucidates this featurization process by using the water molecule as an example.

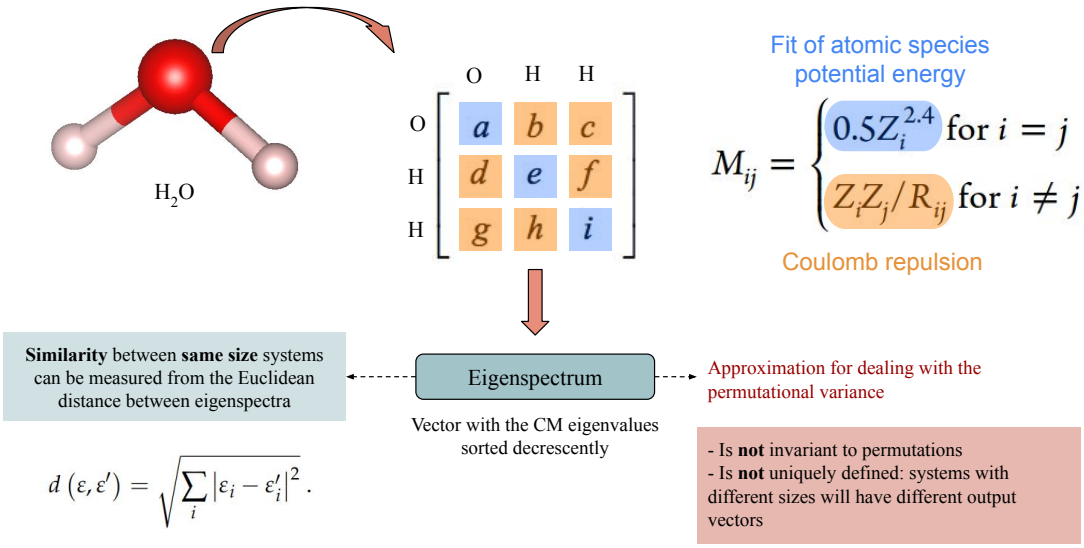

Figure 12 – Coulomb matrix eigenspectrum featurization process for the water molecule.

The CM is widely used in similarity assessments, serving as a metric for comparing the similarity between two chemical systems, thereby accelerating configurational searches in systems with numerous composition possibilities.(PENA *et al.*, 2023; SOUZA *et al.*, 2024) The similarity measure is done by taking the Euclidean distance between two configurations eigenspectra, being restricted to systems with the same composition and size. For modeling systems of different sizes (in number of atoms), it is necessary to append zeros (dummy features) to the end of the eigenspectrum for guaranteeing that all CM resulting eigenspectra have consistent sizes. Additionally, there are formulations of the CM for periodic systems, such as the Sine Matrix or Ewald Sum Matrix, which include terms for replicating the unit cell coulomb interactions peridically and a damping term (HIMANEN *et al.*, 2020).

3.3.1.2 Bag of Bonds

In the development of invariant molecular descriptors, it is ideal that the descriptor is capable of describing systems of different sizes and species. This is not achieved in the CM descriptor, as the output feature vector size depends on the number of atoms contained in the molecular system. Additionally, this feature vector can be constructed in any order, and, even with the eigenspectrum approximation, vector components (features) may differ in significance for each molecule, either due to chemical species or atom position correspondence.

Due to this problem, the Bag of Bonds (BoB) representation, first proposed by Hansen *et al.* (2015), tries to include permutation invariance by building upon the CM descriptor. While this descriptor was not used in our study, it is fundamental for understanding the (L)MBTR descriptors. This descriptor aims also for computing the pairwise Coulomb potentials, but sorting them differently into "bags" which describe different types of bonds between chemical

species. Through the sorting of atomic numbers, the permutational invariance is achieved, guaranteeing that all molecular systems have corresponding features that refer to the same combinations of nuclear charges. For molecular systems with different sizes, zeros-padding (dummy features) is utilized for normalizing the feature vectors. Note that, as the CM descriptor, the BoB representation cannot distinguish between homometric molecules (PATTERSON, 1939), *i.e.* molecules with different geometries but the same set of pairwise distances. Hence, the BoB can be viewed as an $N_e \times N_e \times d$ tensor where N_e is the number of elements and d is sufficiently large to encompass all existing bags for one bond. For using this descriptor in regression tasks, the tensor is flattened maintaining a consistent ordering of the bags across different systems. By including the permutational invariance, an improvement in total energy prediction is observed in relation to the CM (HUO; RUPP, 2022).

3.3.1.3 Many-Body Tensor Representation

The many-body tensor representation (MBTR), proposed by Huo & Rupp (2022), expand on the concept of modeling chemical systems in terms of groups of atoms, as introduced in the BoB representation. In addition to the two-body terms (element pairs), the MBTR proposes to include other many-body terms such as one-body, and three-body terms. By including higher order terms, additional information on the chemical system is included and a better fitting for properties expected (LANGER; GOEßMANN; RUPP, 2022).

Alike the BoB representation, the MBTR can be viewed as a tensor that stores contributions from different atom groupings. Each many-body term has a geometry function g_k that encodes information on atomic species groupings for a system:

- $g_2(R_l, R_m)$: distance in angstroms between two atoms;
- $g_3(R_l, R_m, R_n)$: cosine of the angle from three atoms.

The MBTR retains the stratification by elements, but avoids the permutation and vector length problems (dummy variables) for different systems by arranging the g_k outputs on a real-space axis (x) through a kernel density estimation process. Each geometry function transforms a group of k atoms into a single scalar, which is smoothed using Gaussian kernels (Eq. 3.35) to ensure the continuity of the descriptor in the real space (HUO; RUPP, 2022)

$$D_2^{l,m}(x) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x-g_2(R_l,R_m))^2}{2\sigma_2^2}}, \tag{3.35}$$

$$D_3^{l,m,n}(x) = \frac{1}{\sigma_3\sqrt{2\pi}} e^{-\frac{(x-g_3(R_l,R_m,R_n))^2}{2\sigma_3^2}}. \tag{3.36}$$

For each k term, an associated smearing factor σ_k can be selected for adequately separating distributions in the real axis. The Gaussians are normalized to a unit area by the factor $\frac{1}{\sigma_k\sqrt{2\pi}}$

and the exponent denominator $2\sigma_k^2$. The resulting scalar from each geometry function g_k will shift the Gaussian the kernel along a real line whose limits are pre-defined by the user. From this formulation, the MBTR is computed for every possible combination of k atoms in the system (Eq. 3.37)

$$\mathrm{MBTR}_2^{Z_1,Z_2}(x) = \sum_{l}^{|Z_1|}\sum_{m}^{|Z_2|} w_2^{l,m} D_2^{l,m}(x), \tag{3.37}$$

$$\mathrm{MBTR}_3^{Z_1,Z_2,Z_3}(x) = \sum_{l}^{|Z_1|}\sum_{m}^{|Z_2|}\sum_{n}^{|Z_3|} w_3^{l,m,n} D_3^{l,m,n}(x), \tag{3.38}$$

where the sums for l, m, and n run over all system atoms with atomic number Z_1, Z_2, or Z_3, respectively. w_k is a weighting function that is used to control the importance of different terms, giving less importance to distant pairs of atoms in the case of g_2, for instance. The weighting function also enables this descriptor to be utilized in periodic systems, where periodic copies of atoms in the unit cell are taken into account until being significantly small. This formulation enables for great flexibility in representing chemical systems as the user can arbitrarily choose any combination of geometry functions, smearing factors, and weightings. Additionally, the featurization process can be restricted to specific ranges of interatomic distances or angles by defining the start and end values of the real axis, for which the resolution (discretization) can also be set (HIMANEN *et al.*, 2020).

Differently from other molecular descriptors, this representation can also be visualized as shown in Fig. 13 for the example of the water molecule. In these graphs, the y axes represent the frequency for the two (distance) and three-body (cosine) terms, and the x axes are the flattened MBTR feature vectors. As observed, each species group – hereafter referred as a MBTR class – has a kernelization grid with a discretization of n, that is, n features (scalars) corresponding to the kernelization results. As the MBTR iterates over all atoms for a given MBTR class (such as H – H or H – H – H), the kernels are the occurrences of each distance or cosine for that class. The user defines a minimum and maximum values for the kernelization grid for each term, used for the kernelization of the term classes, and the kernels may be cutoff or simply not appear if outside of this range (HUO; RUPP, 2022).

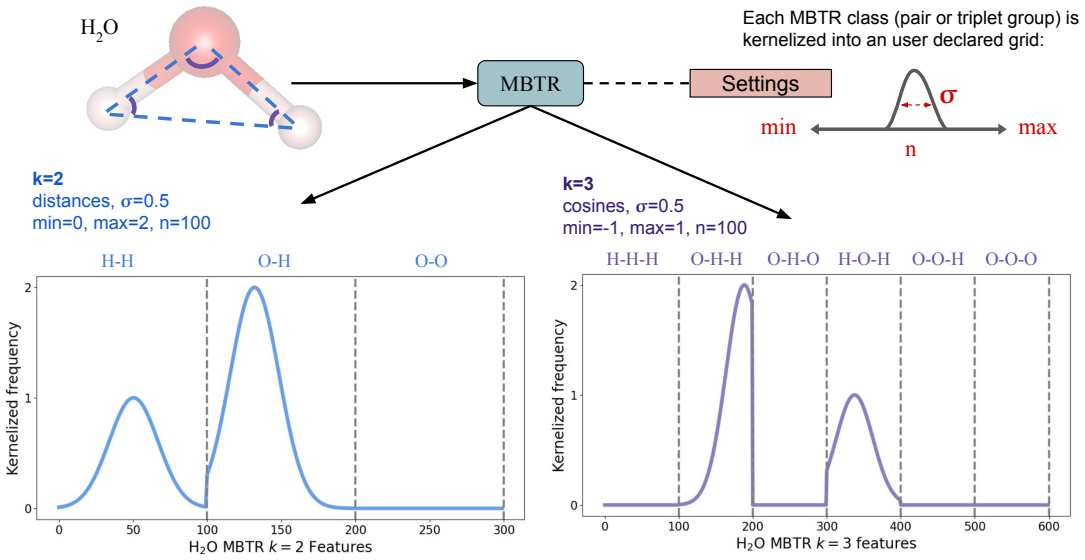

Figure 13 – Illustration of the MBTR kernelization for the distances ($k = 2$) and cosines ($k = 3$) of the H_2O molecule. Kernelization is controlled by settings such as the gaussian width σ, minimum, maximum values and discretization n for the MBTR classes kernelization grid. Bellow, the flattened tensor for each k term is plotted, showing the kernelization results for each MBTR class.

The MBTR, however, has limitations due to its class-oriented nature. For instance, in the example above we could not featurize the CO_2 molecule or any other chemical system with atomic species other than H and O, as the feature vectors in Fig. 13 do not have classes including groupings with this new atomic species. Hence, the MBTR should be declared already with all possible atomic species to be encountered in the dataset or any other data employed in testing the model, as new species are not applicable in the same feature vectors. Additionally, the MBTR vector size grows quickly with the combinatorial nature of two-body and three-body groups for different species, requiring lower resolutions (n) for the same amount of memory as with a few classes only.

Table 3 presents the MBTR parameters that may be set for defining and improving the featurization procedure. Essentially, we aim for choosing the parameters that capture the most important information and do not account for less crucial data for the property of interest. For instance, we need to define a range for distances and angles that are relevant for determining adsorption energies, which is done with the kernelization grid `min` and `max` parameters. This is crucial, as with many different atomic species a lower resolution (n) should be used for making the descriptor computationally affordable, thus limiting the grid makes lower resolutions more efficient than in larger grids. Additionally, there is also the `sigma` (σ) parameter, which effectively defines the overlapping of different kernels for a same class. Kernel overlapping may cause some loss of information, as the individual kernels cannot properly be identified.

Table 3 – MBTR parameters for the featurization process.

Parameter	Description
`species`	Set of all possible species encountered in data
`geometry`	Geometry function
`gridMax`	Kernelization grid maximum
`gridMin`	Kernelization grid minimum
`n`	Kernelization grid discretization (resolution)
`sigma`	Gaussian width for kernelization
`scale`	Scale of the exponential weighting

While the MBTR is calculated for all possible combinations of two-body or three-body atomic species classes, its local version, the local many-body representation (LMBTR), restrains the classes to declared system coordinates (centers). In many cases, such as in adsorption systems, not all atoms in the system are necessarily relevant for the property of interest, and disconsidering irrelevant regions is desired. As shown in Fig. 14, the LMBTR utilizes the same kernelization procedure but includes the pseudo-species X_n which corresponds to declared centers. Any number of centers may be declared, with the resulting feature vector being the concatenated flattened feature vectors for each center. Hence, LMBTR maps the systems in terms of k groupings where one of the species is X, making it possible to encode spatial information in a localized fashion with the combination of specific kernelization settings (HIMANEN *et al.*, 2020).

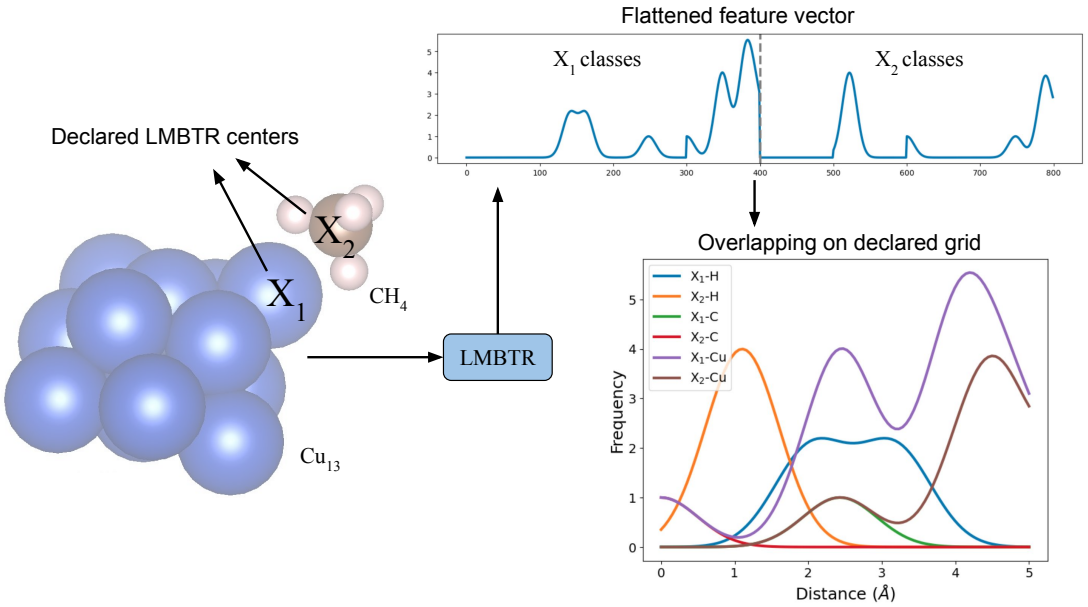

Figure 14 – LMBTR featurization of the $Cu_{13} + CH_4$ adsorbed system with two declared centers X_1 and X_2. The flattened feature vector is shown above, corresponding to the concatenation of each center flattened vector. Bellow, the feature vector is viewed on the declared distance grid from 0 to 5 angstroms, with colors identifying each LMBTR class.

If the number of centers X_n is much lower than the number of atoms, the resulting

classes for both $k = 2$ and $k = 3$ are drastically reduced in comparision to the MBTR, as each class now refers to a combination with pseudo-species X_n. This means that the total number of many-body classes is not anymore the product of combinations of all possible species in the dataset (excluding symmetrical terms), but only two-body or three-body classes with the center species X_n are accounted. Hence, the LMBTR feature vector size grows with the number of X_n centers declared, as an equal number of classes for all possible encountered species in data is created per center, as shown in Fig. 14.

3.3.2 Principal Component Analysis

High dimensionality datasets contain measurements on many variables, which can pose a high computational cost, specially when dealing with a large sample size. Many variables also tend to be correlated, hence containing redundant information. Due to this problem, the principal component analysis (PCA) method is one of the most widely used for dimensionality reduction, being effective for extracting relevant information from data (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). In this work, the PCA was used for reducing the final concatenated $k = 2$, $k = 3$ LMBTR feature vector, which has a very high dimensionality and sparsity (many zeros), making it computationally costly.

PCA concentrates on variances and aims for reducing the dataset dimensionality while maintaining the most possible original variance. This is achieved by transforming the dataset variables into another set of variables called principal components (PCs), removing correlated features, thus leaving orthogonal ones. While the dataset features are generally interpretable measurements and quantities, the PCs do not possess any significance other than being a linear combination of variables that explains how data is spread out (DEISENROTH; FAISAL; ONG, 2020).

One of the requirements for conducting PCA is the preprocessing of the data table through scaling and centering of each feature. Imagine, for instance, variables measured in degree Celsius, seconds, and mol L^{-1}. Since these variables have different units and magnitudes, their variances will differ significantly, leading to biased PCs. To ensure that all features contribute equally, a standardization preprocessing (Eq. 3.39) is applied by transforming each variable to have a mean of zero and a variance of one

$$x_{\text{scaled}} = \frac{x - \bar{x}}{\sigma_x}, \tag{3.39}$$

where x is the original feature, $\bar{x}$ this feature's mean, and σ_x the standard deviation of this feature over all samples. It is worth noting that, from this preprocessing, zeros no longer represent an absence of signal, but indicate an average value. By performing this normalization, all variables will have an equal opportunity of expressing its variance and hence being accounted for the PCA (SHLENS, 2014).

The standardization is not included in the PCA equations but is a crucial preprocessing. This has been observed in performing the PCA for the final LMBTR feature vector, which had a different number of resulting PCs if standardized or not. Having a fewer number of PCs indicates that some of the dimensions were not accounted in the overall variance of the featurized dataset, and was the case for performing the PCA on the not standardized LMBTR. In this case, the PCs were biased to whichever geometric term and class had the most occurrences in the final LMBTR output vector, as the larger the values the greater the variance (variances are not to scale).

Since the data table $\mathbf{X}$ is centered, its values represent the deviation from the mean for each feature. For understanding correlations in $\mathbf{X}$, a covariance matrix $\mathbf{C_X}$ is utilized, being defined by (Eq. 3.40)

$$\mathbf{C_X} \equiv \frac{1}{n-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}. \tag{3.40}$$

Diagonal elements from $\mathbf{C_X}$ represent variances for each measurement type while non-diagonal elements covariances between measurement types. The covariance measures the joint variability from two random variables, with its sign showing the tendency of linear relationship between the variables. The ij^{th} element of $\mathbf{C_x}$ is the dot product between vectors from the i^{th} measurement with the j^{th} measurement, hence composing the square matrix of all possible correlations. If the covariance of two measurements is zero, they are orthogonal, and the larger the covariance, the more redundancy between measurements, that is, both variables tend to variate together indicating a strong relationship or dependence (SHLENS, 2014).

To determine the PCs, the eigenvalue decomposition of $\mathbf{C_X}$ is performed (Eq. 3.41)

$$\mathbf{C_X}\mathbf{V} = \mathbf{V}\Lambda, \tag{3.41}$$

where $\mathbf{V}$ is the matrix of eigenvectors (PC directions) and Λ is the diagonal matrix containing the eigenvalues, which represent the variance explained by each corresponding eigenvector. The eigenvectors are sorted in descending order of their eigenvalues, ensuring that the first PC captures the highest variance, the second PC captures the second highest variance, and so on. A crucial step in PCA analysis is determining how many components should be retained. This is often done by examining the explained variance ratio (Eq. 3.42) (DEISENROTH; FAISAL; ONG, 2020)

$$\text{Explained Variance} = \frac{\lambda_{\text{PC}}}{\sum\limits_{i=1} \lambda_i} \times 100, \tag{3.42}$$

where λ_{PC} is the eigenvalue of one of the PCs, representing a percentage of all calculated eigenvalues λ_i from the data table. The cumulative explained variance from a different number of PCs – the more PCs, the more explained variance in data – is typically plotted to identify the

"elbow point" where adding more components yields diminishing returns in variance preservation. Hence, we can choose any number of the calculated PCs (from the larger to the smaller eigenvalue) to project our data onto, for instance, choosing fewer PCs decreases the resulting dimensionality of the projection, however losing some of explained variance in data (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

3.3.3 K-Means Clusterization Algorithm

The k-means algorithm is an exploratory clustering algorithm that is used to automatically group data based on a similarity score. This method is utilized for uncovering possible data clusters, which otherwise could not be properly identified. From this arranging, different data clusters with similar intrinsic characteristics are achieved, and meaning may be extracted from the clusterization results (DEISENROTH; FAISAL; ONG, 2020).

k-means is an iterative clustering algorithm that groups samples based on a similarity metric, typically the Euclidean distance between feature vectors. This metric quantifies how (dis)similar two samples are, guiding their assignment to clusters. Clustering is performed relative to anchor points called centroids, which are vectors of the same dimensionality as the data samples. These centroids are randomly initialized within the range of the data features at the beginning of each algorithm run. The number of centroids (k) must be pre-specified and determines the number of clusters into which the data will be partitionated. k-means operates through an iterative process consisting of two main steps, repeated until convergence (IKOTUN *et al.*, 2023):

- Expectation: each sample is assigned to the closest (most similar) centroid.
- Maximization: centroid positions are updated to the mean of their corresponding clusters.

k-means clusterization results are only dependent on the randomly initialized centroids, being the algorithm deterministic upon initialization and its convergence – centroids stop updating their positions – guaranteed. As the Euclidean distance is utilized for calculating similarity, the k-means algorithm is most suitable for spherical group distributions. The optimal number for k that provides the best clusterization is unknown, and it is necessary to test different numbers of clusters for assessing the grouping quality throughout metrics such as the silhouette score or visualization of the data partitioning (IKOTUN *et al.*, 2023).

3.3.4 Regression Algorithms

Regression is a statistical method that consists of finding a function f that maps inputs $\mathbf{x}$ (independent variables or features) to outputs $\hat{y}$ (dependent/target variable). The goal of regression is to fit a model capable of estimating, with minimal errors, the target variable on a test set based

on analyzed relations from a training set. From the modeled relations, the regression algorithm is tested on a test-set with new samples, aiming for assessing its generalizability upon new data. The following sections introduce two linear methods and the non-linear random forest algorithm.

3.3.4.1 Linear Regression

The linear regression algorithm is the simplest regression algorithm as it estimates an output variable y from linear combinations of input variables x. It was first conceptualized by GALTON in 1894, whose first insights about regression arose studying pea sizes. In his studies, Galton observed different generations of self-fertilizing sweet peas, noting how a straight line could explain the constant variability in sizes across different lineages (STANTON, 2001). In this work, the linear regression was utilized as a baseline regression model for directing descriptors optimization, as improvements in performance could be attributed to the system modeling instead of the algorithm parameters, which in the linear regression case are none.

For an input vector $\mathbf{x}^T = (x_1, x_2, \ldots, x_p)$, the linear regression model has the form (Eq. 3.43)

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j, \tag{3.43}$$

where $f(\mathbf{x})$ is the prediction output ($\hat{y}$) for the feature vector $\mathbf{x}$. Typically, a set of training data $(\mathbf{x}_1, y_1 \ldots \mathbf{x}_N, y_N)$ of N samples is utilized for estimating the linear parameters or coefficients β_j that will provide the best fit (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). As any regression algorithm, the linear regression aims for optimizing a loss function which quantifies the quality of the resulting fit to the target variable y. The most common estimation method is the least squares, where the coefficients $\beta^T = \beta_0, \beta_1, \ldots, \beta_p$ are optimized for minimizing the residual sum of squares (RSS) (Eq. 3.44)

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - \hat{y})^2 \tag{3.44}$$

$$= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \tag{3.45}$$

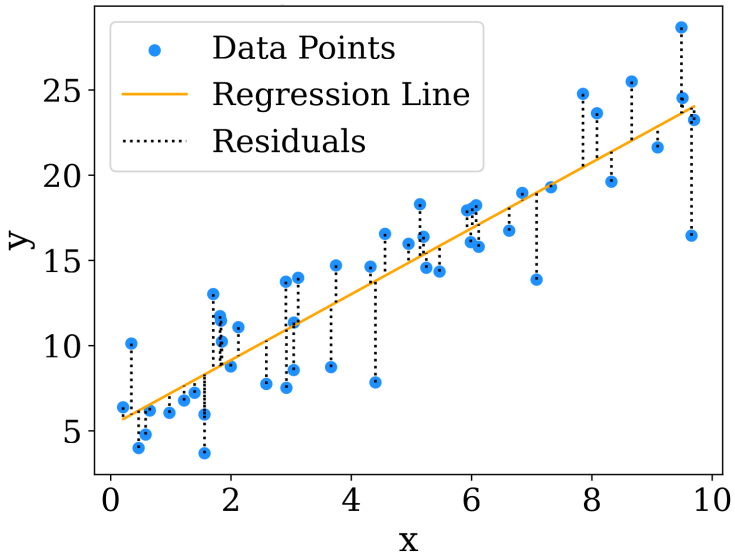

Figure 15 – Optimal linear fit and its residuals for a one-dimensional feature space.

Squaring the residuals facilitates treating negative and positive errors, avoiding cancelation, and also ensures that larger errors have a greater impact than smaller ones. Most importantly, the squared function is differentiable, which allows for analytical solutions. Fig. 15 illustrates the optimal line and its residuals in relation to data points, which may be generalized for any dimension. The linear regression will rotate this line (or hyperplane) for finding the optimal coefficients that minimize the RSS for the overall training set. For that, let $\mathbf{X}$ be the $N \times (p+1)$ matrix of input vectors, with the additional column corresponding to the intercept β_0, and $\mathbf{y}$ the N-vector of the target variable in the training set. RSS may be written as (Eq. 3.46) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta), \tag{3.46}$$

which differentiating with respect to β results in (Eq. 3.47)

$$\frac{\partial \text{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta). \tag{3.47}$$

Since the RSS is a quadratic function, we can equate this first derivative to 0 so that the optimal linear coefficients β that minimize the function may be encountered (Eq. 3.48)

$$\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta) = 0. \tag{3.48}$$

3.3.4.2 Ridge Regression

Overfitting refers to when a model learns not only the underlying patterns in the training set, but also noise and possible outliers. As a result, the model performs exceptionally well on the training set, but when employing it in new data (test set) predictions are poor in accuracy,

indicating a lack of generalization capability due to being overly complex. In the case of the linear regression, this may be observed in high-dimensional feature spaces, such as the MBTR and LMBTR descriptors, which have thousands of dimensions (features). As the model will irrestrictly choose the optimal linear coefficients (which may explode) that minimizes the RSS, the linear regression model may be overfitted to the training data and perform poorly in test samples. Due to this problem, the ridge regression, restricts the coefficients optimization by a regularization term added to the loss function (DEISENROTH; FAISAL; ONG, 2020).

The ridge regression shrinks the regression coefficients by employing a penalized RSS (Eq. 3.49)

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left(\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \tag{3.49}$$

which is achieved by the regularization term λ. If $\lambda = 0$, there is no regularization and the RSS is optimized as in a linear regression. λ has to be positive and the greater its value the higher the penalization constraint. By limiting the magnitude of the linear coefficients through the length of β, overfitting problems may be avoided (HOERL, 2020).

3.3.4.3 Random Forest Regression

The random forest regression (RFR) is one of the most versatile, accurate and general techniques for ML. First introduced by BREIMAN in 2001, the RFR is an ensemble method that builds upon decision trees for dealing with high-dimensional feature spaces and diverse datasets. Due to its scalability, simplicity, and resilience to overfitting, the random forest became one of the most popular algorithms, being employed in a wide array of classification and regression problems (BIAU; SCORNET, 2016).

Classification and regression trees (CART) is the most common algorithm for decision trees due to its robustness and ability to handle both classifications and regression problems (BREIMAN *et al.*, 1984). As illustrated in Fig. 16 for some random data, a decision tree utilizes features x_1 and x_2 to recursively partition the feature space into distinct leaf regions R_m according to binary rules, autonomously defined by the algorithm. A decision tree starts to grown from a root node which contains its fitting data, from where the algorithm computes splits ($s = 1, 2, 3, \ldots$) that partitions data based on a set of features (x_j) and a split points (t_s), which define the binary rules $x_j \leq t_s$. Hence, from the root node, the algorithm recursively splits the fitting data into two decision nodes until some convergence criteria is achieved, characterizing the terminal regions as leaf nodes. Once the tree is fitted, new input data can be fed into the model, traversing it according to the learned decision rules. The prediction is then computed at the leaf node, typically as the mean response of samples in the leaf region ($\bar{y}_m$) for regression tasks (BREIMAN *et al.*, 2017).

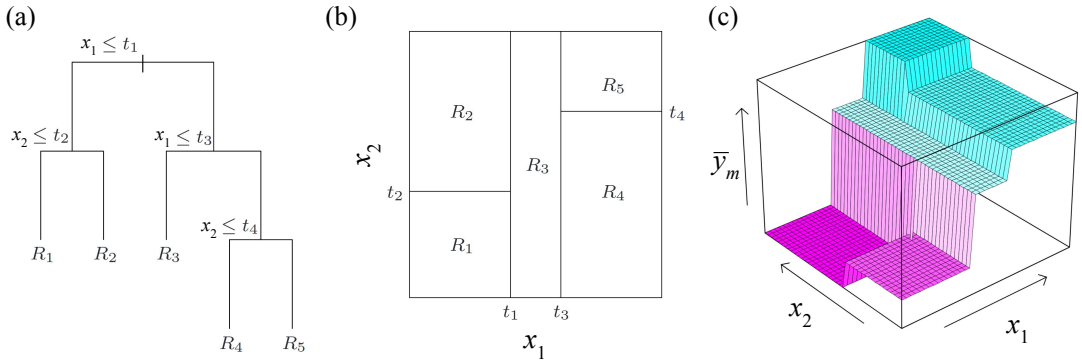

Figure 16 – Scheme for (a) a fitted decision tree with (b) its corresponding partitionated two-dimensional feature space and (c) prediction surface. Reproduced with permission from Springer Nature (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

The goal of a decision tree is to establish logical pathways based on a selected set of input features, ensuring that the leaf nodes are as homogeneous as possible with respect to the target variable y. In other words, each leaf node should contain samples with similar y values, maximizing the purity of the leaf region. This structure enables the tree to generalize well, ensuring that new inputs are assigned to leaves that provide accurate predictions based on the learned binary rules. For deciding upon these logical pathways, an impurity measure is utilized, that is, an evaluation to how well the logical split s separates y values based on some feature x_j and split point t_s (RUTKOWSKI *et al.*, 2014). In the case of regression trees, a common impurity measure is the y variance for the resulting split regions, which is defined by (Eq. 3.50)

$$\text{var}(R_m) = \frac{1}{N_m} \sum_{i \in R_m}^{N_m} (y_i - \bar{y}_m)^2, \tag{3.50}$$

where $\text{var}(R_m)$ is the variance of y in the region R_m, with N_m samples (y_i) and mean $\bar{y}_m$. The region R_m could be any node of the tree with a set of samples, such as the child decision nodes *during* a splitting procedure. Observe that upon fitting, only the leaf nodes of the tree will have a set of samples and could be considered a region (as noted in Fig. 16), and the decision nodes only direct the prediction of the tree (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). For any split, the pair of child nodes is (Eq. 3.51)

$$R_l(j, t_s) = \mathbf{P} \mid x_j \leq t_s \text{ and } R_r(j, t_s) = \mathbf{P} \mid x_j > t_s, \tag{3.51}$$

where R_l and R_r are the resulting left and right regions from split s over the parent set of samples $\mathbf{P}$. In the case of the root node, $\mathbf{P}$ is the complete fitting set of the tree, which is partitionated into different branches until finally reaching terminal regions R_m (leaf nodes), which are defined by different convergence criteria. For data splitting, CART selects a feature index j and a split point t_s within the range of feature x_j to create the binary inequalities that determine the assignment of samples from $\mathbf{P}$ to each child region (R_l or R_r). CART searches for t_s and j that solves (Eq. 3.52)

$$(j^*, t_s^*) = \arg\min_{j,t_s} \left[\frac{1}{N_l} \sum_{i \in R_l} (i - \bar{y}_l)^2 + \frac{1}{N_r} \sum_{x_i \in R_r} (y_i - \bar{y}_r)^2 \right], \tag{3.52}$$

which reads as finding the optimal values j^* (feature index) and t_s^* (split point t for split s) that minimize the sum of variances from the child regions R_l and R_r. Encountering a minimum for this expression indicates a reduction in impurity, as the overall variances for y in each child region are minimized, *i.e.* each region target variables are the most homogeneous as possible from the split. This procedure is conducted iteratively until all resulting regions are transformed into leaf nodes, which may be determined by different convergence criteria based on the number of samples or impurity metrics (`min_samples_leaf`, `min_samples_split`, `min_impurity_decrease`), as well as the overall tree structure (`max_depth`, `max_leaf_nodes`). The decision tree parameters are equally important in defining the growth of the tree based on the characteristics of the training set and features, effectively directing the tree to avoiding under or overfitting data. Hence, choosing the optimal parameters for the decision tree depends ultimately on the underlying characteristics and distribution of features and samples. By controling the decision tree hyperparameters (Tab. 4), the tree complexity is tuned for achieving the best possible fit from training to test sets (BREIMAN *et al.*, 2017).

Table 4 – Decision tree hyperparameters.

Parameter	**Description**
`max_features`	Fraction of sampled features for fitting
`max_depth`	Max. depth of the tree (subsequent nodes) is achieved
`max_leaf_nodes`	Max. n° of leaf nodes is achieved
`min_samples_leaf`	A split at any depth is only considered if it leaves `min_samples_leaf` on R_l and R_r
`min_samples_split`	Min. n° of samples required to split an internal node
`min_impurity_decrease`	Min. impurity decrease from the split is not achieved

Finding the best sequence of binary partitions of the overall tree for reducing some loss function over the training set is computationally infeasible. Hence, CART is a greedy algorithm that will opt for the local minima in each split, without worrying about further consequences. However, trees may be fitted in different conditions, such as with different subsets of the training set or with a different set of features. Additionally, the random selection of split points, data samples, and features adds versatility to the fitting. Still, while many trees may be fitted and the one with the lowest errors chosen as the best, a single tree may not have the predictive power for generalizing upon different data as it offers a single, fixed logical structure fitted on the training data. For solving this problem, we may utilize a forest of trees in our regression model for capturing more complex patterns in data by means of majority voting and/or averaging (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

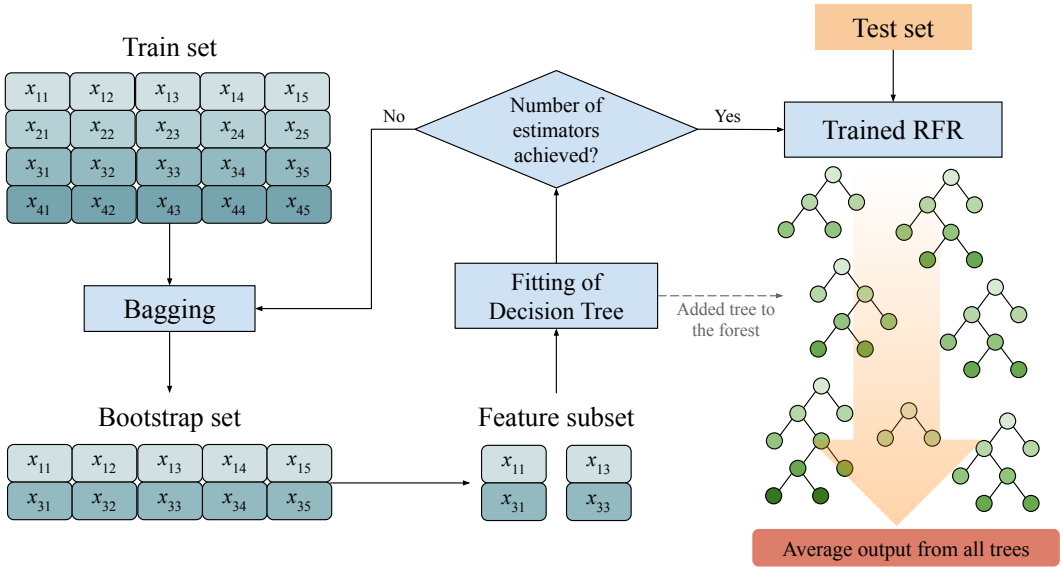

Figure 17 – Fluxogram describing the construction of a RFR. Starting from the training set with samples i and features x_{ij}, the process of bagging randomly selects samples for constructing a bootstrap set for fitting a decision tree. A subset of features is randomly sampled for further increasing the randomness in fitting the decision trees. Upon convergence, the tree is added to the forest and this process is conducted iteratively until the number of estimators (trees) in the forest is achieved. With the trained random forest, a test set is feed for evaluating the model with predictions computed as the average response from all trees.

For constructing a forest of trees we fix the decision tree parameters presented in Tab. 4. As these parameters are convergence criteria and conditions for the fitting of the decision trees, for constructing a random forest different methods may be utilized for further improving tree versatility. A forest of trees is constructed randomly, further enhancing the predictive power of differently fitted trees. For that, different sampling techniques from the training data may be utilized, either with replacement or not, with different subsets or not. In this work, we used the RFR algorithm with *bagging* (bootstrap aggregating), a technique that generates multiple training subsets by randomly sampling data with replacement, as presented in Fig. 17 (BREIMAN, 2001). Each tree is trained on a different subset from the training data, improving model stability. Additionally, for each tree, the algorithm selects a random subset of features for fitting, further increasing diversity among the trees. By averaging the predictions of all trees, the forest achieves a more robust and generalizable model, with the number of trees (`n_estimators`) being its main hyperparameter. Furthermore, controlling the decision tree hyperparameters such as the convergence criteria of leaf nodes allows for fine-tuning for avoiding overfitting problems (DEISENROTH; FAISAL; ONG, 2020).

4 Methodology

This chapter outlines the steps taken to develop a ML approach for predicting adsorption energies of adsorbate species on metallic and oxide nanoclusters based on structural information alone. By leveraging structural data and ML techniques, the proposed framework offers a pathway to streamline the screening of adsorption possibilities while maintaining reasonable accuracy and computational cost in relation to DFT calculations. The following sections describe the dataset preparation, molecular system delimitation, and optimization processes involved in constructing the predictive model.

4.1 Dataset

The dataset used in this work was constructed from various works on cluster-adsorbate systems from the QTNano group (ANDRIANI; MUCELINI; SILVA, 2020; FELíCIO-SOUSA; ANDRIANI; SILVA, 2021; ANDRIANI *et al.*, 2022; COLLACIQUE; OCAMPO-RESTREPO; SILVA, 2022; SOUSA *et al.*, 2022; GOMES, 2022; OCAMPO-RESTREPO; VERGA; SILVA, 2023; FELíCIO-SOUSA, 2024), which are detailed in Tab. 5. These references are investigations on catalytic intermediates and adsorption activations, being the great majority related to the CO_2 electrochemical reduction. All dataset entries correspond to DFT optimizations within the PBE functional (PERDEW; BURKE; ERNZERHOF, 1996) and performed in the FHI-aims software (BLUM *et al.*, 2009; HAVU *et al.*, 2009) with the `light-tier1` or `light-tier2` basis set from the simulation package, and van der Waals (vdW) corrections using the Tkatchenko & Scheffler (2009) method. Each work explores different metallic and oxide clusters which vary in size, geometry, and composition with different sets of adsorbate species related to catalytic pathways.

For the construction of the dataset, each calculation had their geometric optimization frames – single-point energies calculated during the geometry optimization – extracted from the `aims.out` file. By including all frames from the geometric optimization, the dataset reflects not only the final optimized geometry but also intermediate, less-stable configurations. This adds richness to the dataset, enabling the exploration of energy landscapes and structural evolution during optimization. Hence, the dataset is composed of structural information of the adsorption system (input variables) and the associated adsorption energy (target variable), which is calculated as (Eq. 4.1)

$$\Delta E_{\text{tot}} = E_{\text{tot}}^{\text{system}} - (E_{\text{tot}}^{\text{cluster}} + E_{\text{tot}}^{\text{adsorbate}}), \tag{4.1}$$

where $E_{\text{tot}}^{\text{system}}$, $E_{\text{tot}}^{\text{cluster}}$, and $E_{\text{tot}}^{\text{adsorbate}}$ are the DFT total energy of the adsorption system, isolated cluster, and isolated adsorbate, respectively. The original data contained a total of 451,535

Table 5 – Dataset sources and their substrates and adsorbates.

Source	Substrates	Adsorbates	Samples
Andriani 2020[a]	13 atoms unary clusters: Fe, Co, Ni, Cu	H, C, CH, CH_2, CH_3, CH_4	8,551
Felício-Sousa 2021[b]	13 atoms unary clusters: Fe, Co, Ni, Cu	H_2, CH_4, CO, CH_3OH	38,313
Andriani 2022[c]	4 to 15 atoms unary clusters: Fe, Co, Ni, Cu	H, CH_3, CH_4	180,605
Collacique 2022[d]	8 atoms unary clusters: Fe, Co, Ni, Cu, Ru, Pd	CO_2	86,911
De Sousa 2022[e]	Ni_8, Ga_8, Ni_5Ga_3	H, H_2, CO, HCO, CO_2, HCOO, CH_4	37,617
Gomes 2022[f]	Zr oxide cluster: $Zr_{16}O_{32}$	H, H_2, H_2O, H_2O_2, C, CH, CH_2, CH_3, CH_4, CO, CO_2, O, O_2, OH, N, NH_3, S, SO_2, CH_3CH_2OH, CH_3OH, CH_3OOH	5,369
Ocampo Restrepo 2023[g]	$CuCu_{55}$, $Cu_{42}Zn_{13}$	CH_3CO, CHCHO, CHCOH, CH_2CO, CH_3CHO, CH_3COH, CH_2CHO, CH_2COH, CH_3CH_2O, CH_3CH_2O, $CHCH_2O$, CHCHOH	5,859
Felício-Sousa 2024[h]	Oxide and doped Zr clusters: $Zr_{16}O_{32}$, $RhZr_{16}O_{31}$, $LaZr_{15}O_{31}$, $La_2Zr_{14}O_{31}$, $RhLa_2Zr_{14}O_{31}$	H, CH_3, CH_4	34,277
Andriani et al. (in progress)	13 atoms unary clusters: Fe, Co, Ni, Cu	CO, COH, CH_2O, CH_3O, CH_2OH, CH_3OH	100,238

[a] Andriani, Mucelini & Silva (2020)
[b] Felício-Sousa, Andriani & Silva (2021)
[c] Andriani *et al.* (2022)
[d] Collacique, Ocampo-Restrepo & Silva (2022)
[e] Sousa *et al.* (2022)
[f] Gomes (2022)
[g] Ocampo-Restrepo, Verga & Silva (2023)
[h] Felício-Sousa (2024)

samples, however, for limiting the model scope to only adsorption cases, all samples with $\Delta E_{\text{tot}} > 0$ were removed, which amounted to (19,897 samples) 4.6%, resulting in 431,638 samples for the final, utilized dataset. Hence, each sample corresponds to a single point calculation.

Fig. 18 presents a visualization of which clusters and adsorbates compose the dataset. Each bar chart is a perspective of the total number of samples in the dataset, showing either the count of those samples (adsorbed configurations) in relation to cluster sizes, cluster compositions, and adsorbate species. As noted, there is a great disparity in sample representation, whether it be due to a difference in the distribution of cluster sizes, compositions, or adsorbate species.

Figs. 19 and 20 present the distribution for the target property ΔE_{tot} in relation to cluster compositions and adsorbate species, respectively. Most of the cluster compositions have similar distributions for ΔE_{tot}, with some exceptions such as Pd_n, Au_n, and Ru_n, that present in their great majority physiosorbed cases, which present lower values for ΔE_{tot}. Adsorbate species present much more expressive distributions for ΔE_{tot}, showing that they may be the guiding factor in modeling this property.

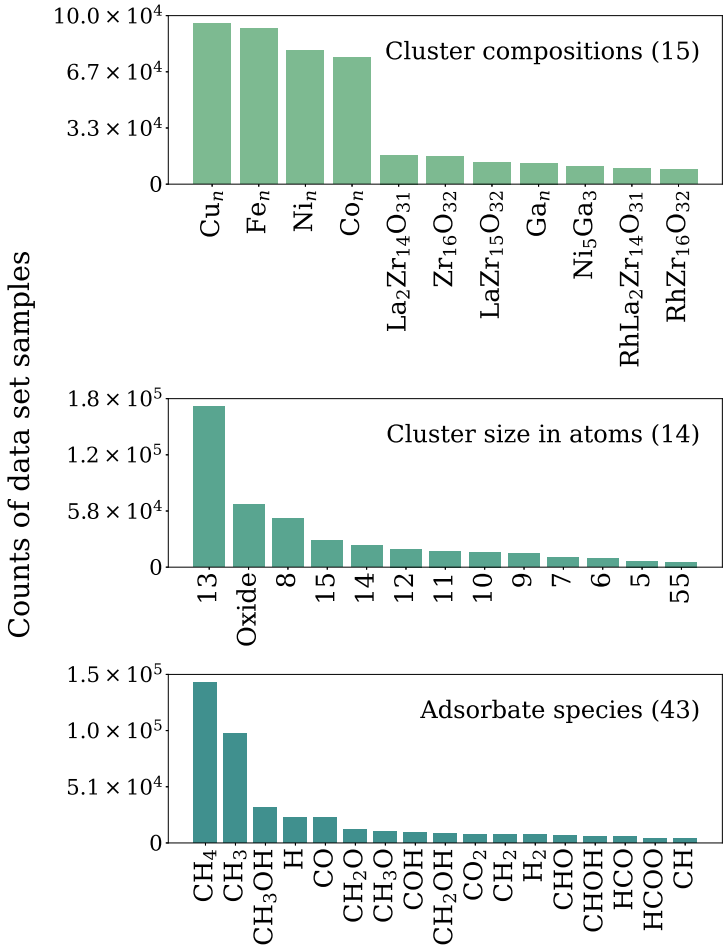

Figure 18 – Dataset breakdown by cluster size, composition, and adsorbate species. Categories that include less than 1 % are omitted for clarity. The total number of categories present in the dataset are provided in parentheses for each graph.

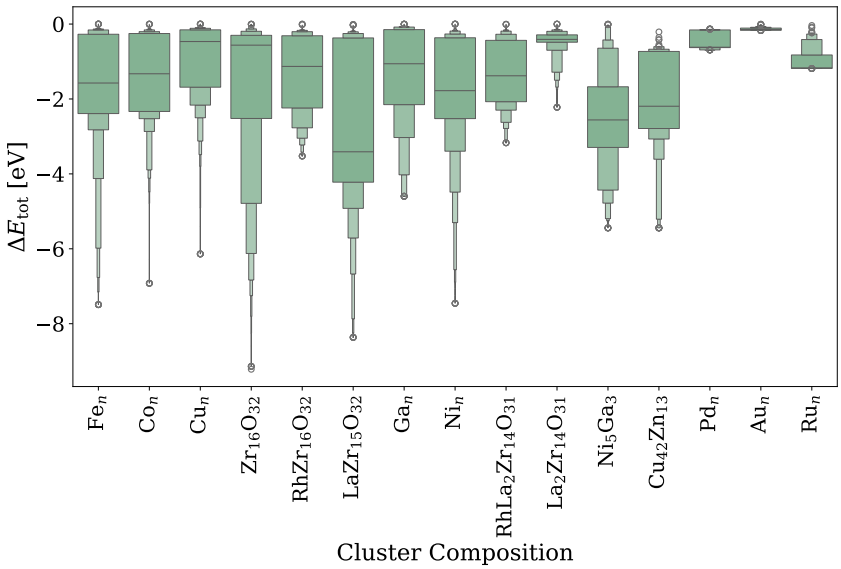

Figure 19 – Distribution of ΔE_{tot} values for each cluster composition. For each composition, the distribution includes all adsorbates in the dataset that share the same underlying cluster.

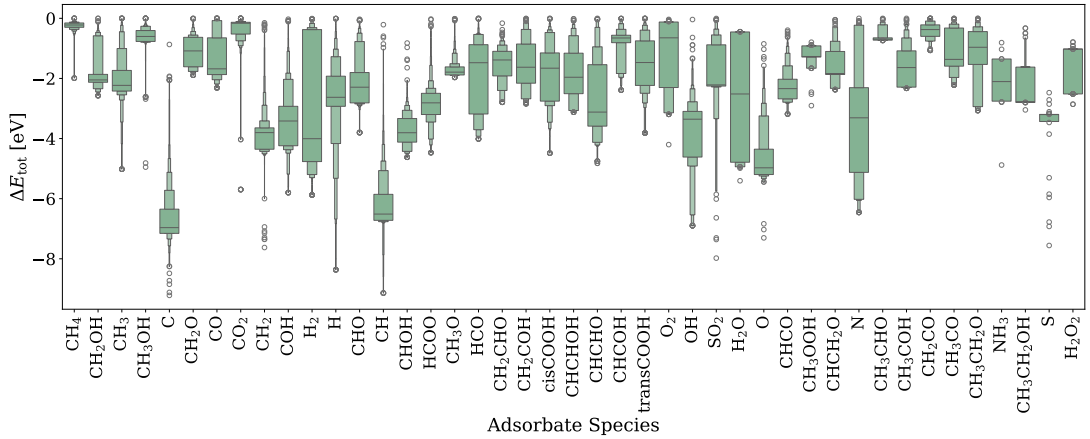

Figure 20 – Distribution of ΔE_{tot} values for each adsorbate species. For each species, the distribution includes all adsorbed systems in the dataset where that adsorbate is present, regardless of the underlying cluster.

4.2 Descriptor Optimization

Considering the various cluster sizes present in the dataset and the objective of a general model capable of screening adsorption energies in any substrate size, such as surfaces or clusters with hundreds of atoms, each system structure information was scoped down to only the adsorption site region when the descriptor is constructed. This delimitation aligns the model development towards a more general and cost-effective predictive framework by discarding less relevant information from the cluster structure for modeling ΔE_{tot}. The implemented function `find_closest_atoms_to_mol` takes as input the Cartesian coordinates for the adsorbate species and cluster structure, calculating the Euclidean distances for all possible molecule-cluster atom pairs. After calculating these pairwise distances, the function identifies the s closest cluster atoms to the molecule, where s is a user-defined parameter. Finally, it returns the resulting adsorbed system, consisting of the selected s cluster atoms and the adsorbate species, as shown in Fig. 21. Consequently, the descriptor is constructed in the region composed of the adsorbate species and the cluster site atoms, which could be the whole nanocluster if its size in atoms is smaller or equal to s.

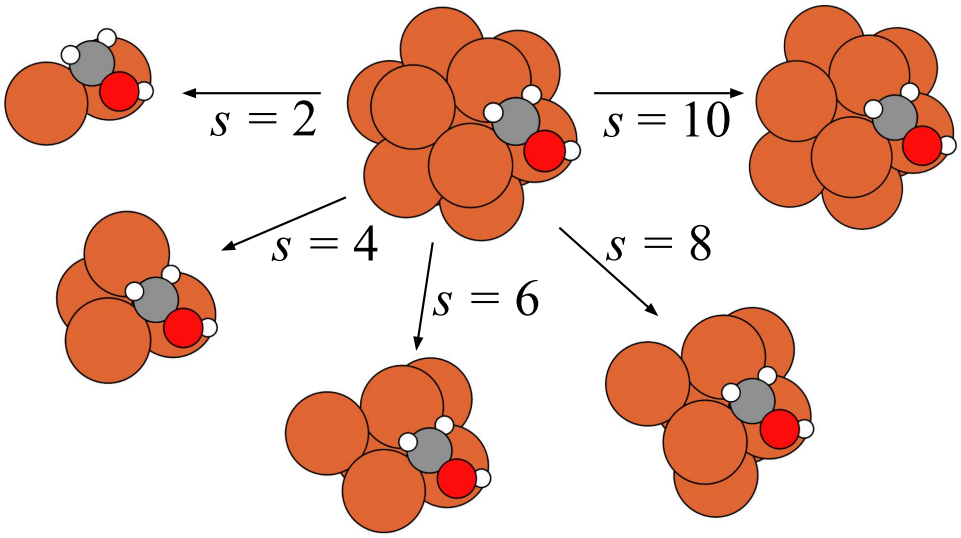

Figure 21 – Constraining of the $Fe_{13} + CH_2OH$ adsorbed system structural information for different `site_size` (s) values. The output, composed of the cluster site atoms and adsorbate species, is the sample structural information fed to the featurization process.

For understanding the influence of the descriptors parameters in modeling data, separately, linear regression algorithms were utilized, as they are the simplest in form and require few or no hyperparameters. Therefore, improvements in regression capability may be solely attributed to variations on descriptors parameters, instead of the regression algorithm. For all experiments, the mean absolute error (MAE) was utilized for assessing the regression results and directing descriptor optimization (Eq. 4.2)

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|. \tag{4.2}$$

Two levels of descriptors were chosen, the CM with the eigenspectrum approximation and the (L)MBTR, both implemented in the DScribe python package (HIMANEN *et al.*, 2020). For the CM, which does not possess any parameters for adjusting feature extraction, only the effect of the `site_size` parameter was analyzed by varying this parameter from $s = 1$ to $s = 30$. The linear regression (`LinearRegressor`) algorithm from the Scikit-learn ML python package (PEDREGOSA *et al.*, 2012) was used for performing predictions. In the case of (L)MBTR, the ridge regressor (`Ridge`) model from the same package with the default regularization term was utilized. All tests were performed with a randomized split without replacement (per execution) of 90% training data (388,474 samples) and 10% test data (43,164 samples) from the final treated dataset. Due to the large sample size, smaller splits for the training data resulted in comparable results, and the 90%–10% split was chosen as the definitive dataset split as it provided a sufficient number of samples in the test set for assessing the model.

Unlike the CM, the (L)MBTR descriptor involves multiple tunable parameters that significantly impact featurization, leading to a vast combinatorial space of possible configurations.

Given the impracticality of an exhaustive search, an adaptive approach was adopted. Initial parameter grids were constructed based on insights from previous studies (JäGER *et al.*, 2018) on the descriptor. Subsequent refinements were guided by observing the performance of experiments conducted, resulting in a total of 657 tests for this descriptor. For the (L)MBTR, the `site_size` parameter varied from $s = 2$ to $s = 10$ cluster atoms in steps of 2. As the local version requires declaring the centers for the featurization, in our implementation, we include the `points` argument as a setting for the LMBTR initialization. The user can select to initialize any number and combinations of centers, such as

- `molC`: geometric center of the adsorbate species.
- `siteC`: geometric center of the cluster site atoms.
- `siteA`: each atom in the cluster site is initialized as a center for the LMBTR.
- `interface`: geometric mean (midpoint) between `molC` and `siteC`.

4.3 Regression Algorithm Optimization

With the putative optimal parameters for both CM and LMBTR from the linear optimization, both descriptors were employed with the RFR (`RandomForestRegressor`) model from the Scikit-learn ML python package (PEDREGOSA *et al.*, 2012). Due to the complexity of the RFR and the high number of achieved LMBTR features, a PCA was conducted for reducing the LMBTR dimensionality, making it computationally affordable to be optimized in many executions of the RFR. The Optuna hyperparameter optimization framework (AKIBA *et al.*, 2019) was utilized for autonomously optimizing the RFR hyperparameters within the ranges specified in Tab. 6.

Table 6 – RFR hyperparameter optimization ranges for each descriptor.

RFR Parameter	CM	LMBTR
`n_estimators`	20-200	20-200
`max_depth`	10-150	10-50
`max_features`	1.0	0.1-1.0
`min_samples_split`	8-20	8-20
`min_samples_leaf`	1-20	1-10

To perform the RFR hyperparameter optimization, a different dataset split than in the structural descriptors optimization was utilized. Since Optuna performs hyperparameter tuning through iterative model evaluations, a validation set was required during optimization. To accommodate this, 90% of the full dataset was allocated to training and further split into 90% for training and 10% for validation. The remaining 10% of the dataset was held out as a fixed test

set, used exclusively for final model evaluation after tuning. This separation ensures an unbiased assessment, as the test data is not involved in training or parameter selection.

4.4 Generation of External Test Set

Upon finishing the optimization of descriptors and algorithms, the ML model was tested on new generated DFT calculations not included in the dataset, an external test set. The Cu_{13} cluster is the most well represented in the dataset and was selected as the substrate for the external test set, with adsorbates CO, CH_3, CH_4, CH_3OH, which range from $4,000$ to $10,000$ examples with the Cu_{13} cluster in the dataset. Additionally, we include the water molecule (H_2O) which is not present in the dataset as an adsorbate for this cluster, aiming for validating the model over unprecedented adsorption cases. It is worth noting that the water molecule *is* included in the dataset, but as an adsorbate on oxide clusters. In the case of the CH_3 molecule, which is planar at the gaseous ground-state but assumes a pyramidal geometry when adsorbed, both geometries were included with a 50%-50% ratio, with the pyramidal geometry being denoted by py-CH_3.

The generation of adsorbed configurations was conducted via the `cluster_adsorption` software (ZIBORDI-BESSE *et al.*, 2016), which randomly places adsorbate species along a cluster surface on a specified distance, with variations of 0.1 Å. For generating adsorption configurations, 4 evenly-spaced cluster-adsorbate distances were selected from the first and third interquartile ranges for each selected adsorbed system in the dataset, whose distributions are presented in Fig. 22. For each selected distance, $2,000$ configurations were generated, totaling $8,000$ total configurations for each adsorbed system. However, we select a small subset for performing the extra DFT calculations, and thus use the k-means algorithm for clustering the $8,000$ generated configurations into 50 representative adsorption systems using the CM Eigenvector as the similarity metric (BATISTA *et al.*, 2021). This ensures that the external test set retains the underlying characteristics of the dataset while providing an independent benchmark for evaluating the model performance.

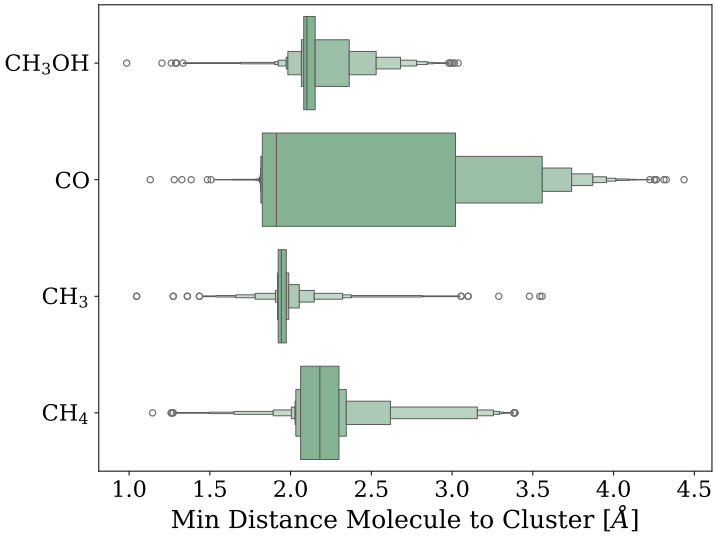

Figure 22 – Minimum Cu_{13}-adsorbate distance distributions in dataset.

4.4.1 Total Energy Calculations

Calculating ΔE_{tot} for the generated external test set was required for comparing the model predictions with the real DFT calculated, single-point relative energies. The Kohn–Sham (KS) equations were solved using The Fritz-Haber Institute *ab initio* materials simulations (FHI-aims) package,(BLUM *et al.*, 2009; HAVU *et al.*, 2009) where the KS orbitals were expanded to numerical atom-centered orbitals (NAO),(HAVU *et al.*, 2009) expanded from the minimum basis set (free atoms) up to the second improvement of the basis set (*light-tier2*). Total energy calculations were performed with the spin-polarized DFT method with the semi-local PBE exchange-correlation energy functional (PERDEW; BURKE; ERNZERHOF, 1996). For accounting for dispersion interactions which are important for weakly bonded systems, the van der Waals (vdW) correction using the Tkatchenko & Scheffler (2009) method was utilized for correcting the DFT-PBE total energy accounting for molecular vdW interactions. The Kohn–Sham (KS) equations were solved using the FHI-aims package (BLUM *et al.*, 2009; HAVU *et al.*, 2009), with KS orbitals expanded in the FHI-aims `light-tier2` basis set. For all calculations, a Gaussian broadening parameter of 10 meV was used to achieve the correct occupation of the electronic states, which is critical for systems with a small energy separation for band gaps. All calculations for the external test set examples were single point (no geometry optimization) for direct comparison with the ML-model outputs.

5 Results and Discussion

5.1 Structural Descriptors Linear Optimization

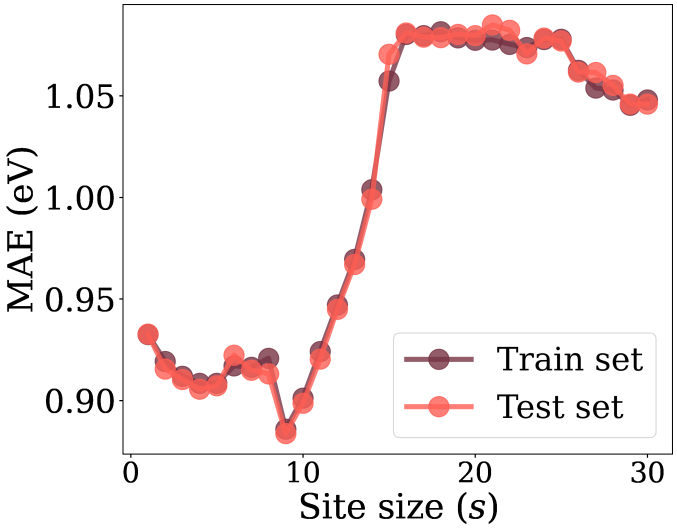

Figure 23 – Linear optimization results for the CM.

The optimization of the `site_size` parameter (s) for the CM is shown in Fig. 23. From the observed MAE metric, it is noted that there is not a significant difference from varying the number of site atoms in the CM, with MAEs ranging from 0.875 eV to 1.075 eV. Including more atoms in the representation allows for additional information for identifying larger clusters, such as oxides and 55 atom cluster in the dataset (Cu_{55}, $Cu_{42}Zn_{13}$), however, smaller clusters (< 13 atoms), which are the great majority, will have many dummy features. This relation may be observed in Fig. 24, where for $s = 18$, the linear predictions concentrate on values from -2.2 to -0.7 eV, between the two main calculated ΔE_{tot} regions, while for $s = 9$ the calculated regions are better approximated, even though there are some outliers with vastly overestimated ΔE_{tot}. Due to the initial proposition of locally modeling adsorption sites, the lower MAE, and a few number of outliers, $s = 9$ atoms were selected as the optimal `site_size` for the CM.

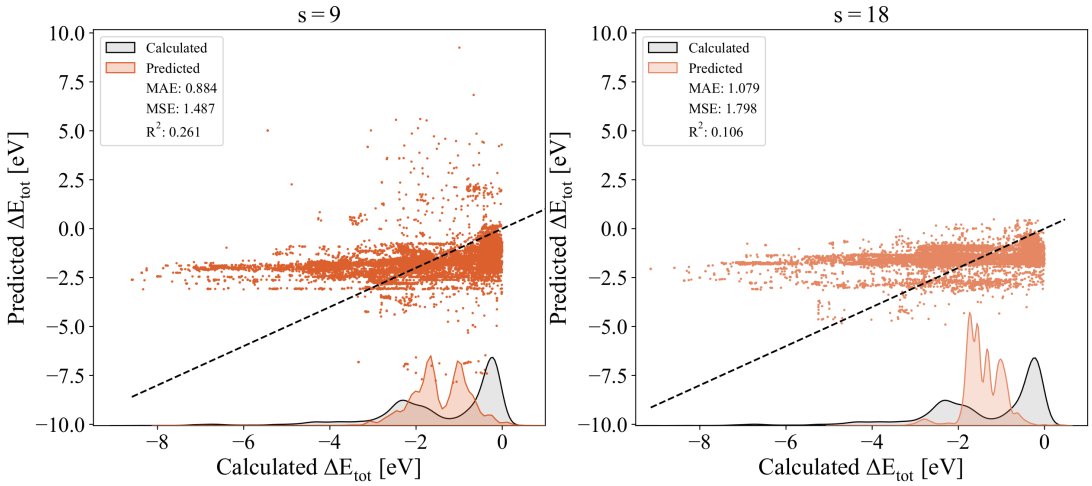

Figure 24 – Test set linear regression prediction distribution for the CM with $s = 9$ and $s = 18$.

Optimization results of (L)MBTR parameters for the two ($k = 2$) and three-body ($k = 3$) terms are presented in Figs. 25 and 26, respectively. As observed in the MAE, the MBTR and LMBTR descriptors perform much better than the CM using the linear regression. There is a significant improvement in the MAE from the CM MAE of 0.85 eV to 0.23 eV for the $k = 2$ and 0.21 eV for the $k = 3$ LMBTR terms with the putative best parameters achieved. Due to limits on computational resources, the optimization of the (L)MBTR parameters had some constraints which are addressed in the following paragraphs. However, some trends are observable from the total of 657 tests conducted for this descriptor.

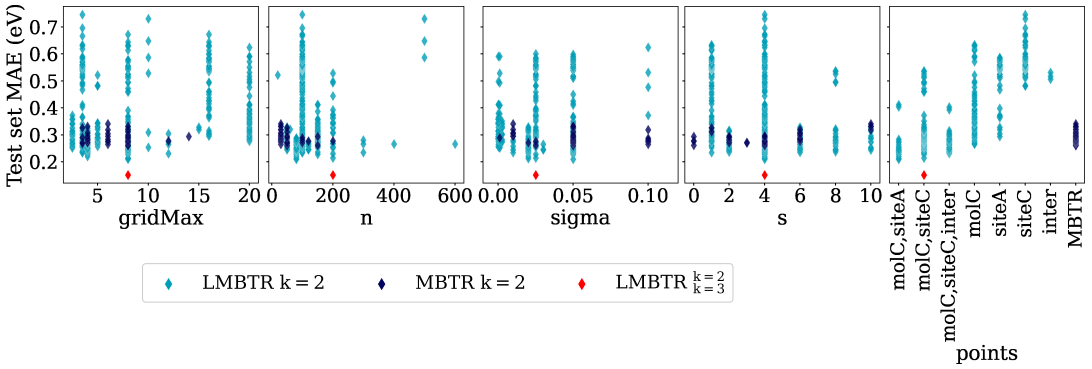

Figure 25 – Scatter plot showing the test set MAE from all 572 executions using the ridge regressor and the (L)MBTR $k = 2$ term. Each graph displays all executions, highlighting one of the varied parameters on the x-axis while the others varied. The result of combining the final selected $k = 2$ and $k = 3$ parameters is marked in red.

The diversity of chemical species in the dataset – which is directly related to this descriptor's feature vector size – did not allow for numerous experiments with the MBTR, which initializes the product of combinations of two and three body terms for all atomic species in the dataset, further increasing the computational cost. Specifically, the $k = 3$ term poses an even higher cost than the $k = 2$ term due to all possible combinations of three-body classes, being not

feasible for the MBTR, and, in the case of the LMBTR, only with a small grid discretization (n), such as $n = 10$ or $n = 20$.

For the $k = 2$ term (Fig. 25), the localized LMBTR performed better than MBTR, which presented consistent values around 0.3 eV for all combinations of varied parameters. As observed, most parameters do not present clear correlations or convergences towards lower MAES, except for the `points` argument, which delimitates where LMBTR centers are declared. Declaring a center in the molecule's geometric mean (`molC`) showed great importance, with tests without this center being notably worse. Declaring centers on all site atoms (`siteA`) was in its majority unefficient, and the intermediate point between cluster and molecule (`inter`) also showed poor results. Combining `molC` with the site atoms geometric center (`siteC`), however, presented the best results, and in this case `siteA` had a lower MAE than `siteC`. However, declaring only two centers `molC` and `siteC` showed most efficiency, as the computational cost for three additional centers in the site atoms (for $s = 4$) – from `molC` + `siteC` to `molC` + `siteA` – would not justify the small reduction in the order of 10^{-2} in the MAE metric.

For the `site_atoms` parameter (s), differently from the CM with 9 site atoms ($s = 9$), for the LMBTR $s = 4$ atoms showed the lowest values for the MAE in initial optimizations, being fixed for most of the experiments as adsorptions are commonly described in terms this number of substrate atoms. Additionally, $s = 4$ atoms is a smaller number of atoms for computing the LMBTR, being computationally cheaper. For the grid resolution (n), a flattening of the MAE is observed at $n = 100$, indicating that higher resolutions do not led to better results. The grid size and also did not seem important for improving prediction capability with the ridge regression, however, the kernel width parameter (`sigma`) showed a clear convergence at around `sigma`= 0.05, which is related to a better separation of the kernel density estimation, *i.e.* no overlapping Gaussians which may lead to information loss.

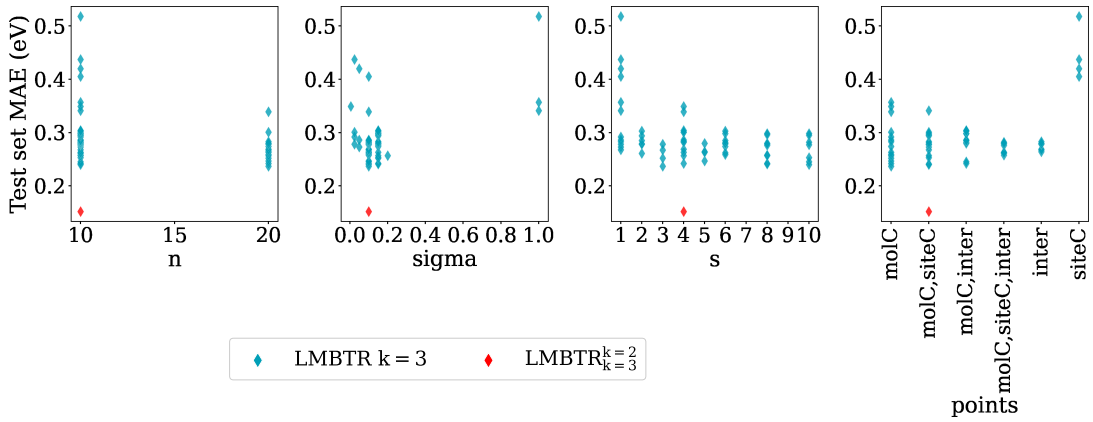

Figure 26 – Scatter plot showing the test set MAE from all 85 executions using the ridge regressor and the (L)MBTR $k = 3$ term. Each graph displays all executions, highlighting one of the varied parameters on the x-axis while the others also varied. The result of combining the final selected $k = 3$ and $k = 2$ parameters is marked in red.

The $k = 3$ term for the LMBTR (Fig. 26) showed some differences to the $k = 2$ term,

with the `site_size` (s) and `points` parameters not being as relevant for the minimization of the ridge regression MAE. There is a clear improvement from increasing the grid discretization from $n = 10$ to $n = 20$, which allows for smoothing the kernelization output resolution, and further improvements are expected from increasing this parameter, which was unfortunately not possible due to our sample size and chemical diversity. The `sigma` parameter showed more importance, which may be explained by the lower resolution and the need for an optimal kernelization for the output cosines. For maintaining a similar modeling for both terms, the final `site_size` (s) and `points` parameters were chosen to be the same as for the $k = 2$ term, namely $s = 4$ and `molC,siteC`, respectively. The best parameters for the LMBTR $k = 2$ and $k = 3$ terms are presented in Table 7. The $k = 3$ showed slightly better results for the MAE (0.209 eV) in relation to the $k = 2$ term (0.235 eV), and by concatenating the feature vectors of $k = 2$ and $k = 3$ an improvement in modeling with the ridge regression is observed with the reduction of the MAE to 0.170 eV (Fig. 27). Lower errors than 0.15 were to be expected if the dataset had not such diversity of chemical species, which greatly increases the feature vector size, demanding substantial computational resources throughout the project pipeline

Table 7 – Best LMBTR parameters for $k = 2$ and $k = 3$ terms from linear optimization.

MAE [eV]	`geometry`	`site_size`	`min`	`max`	`n`	`sigma`	`scale`	`points`
0.235	distance	4	0	3.5	100	0.05	0.25	molC, siteC
0.209	cosine	4	−1	1	10	0.1	0	molC, siteC

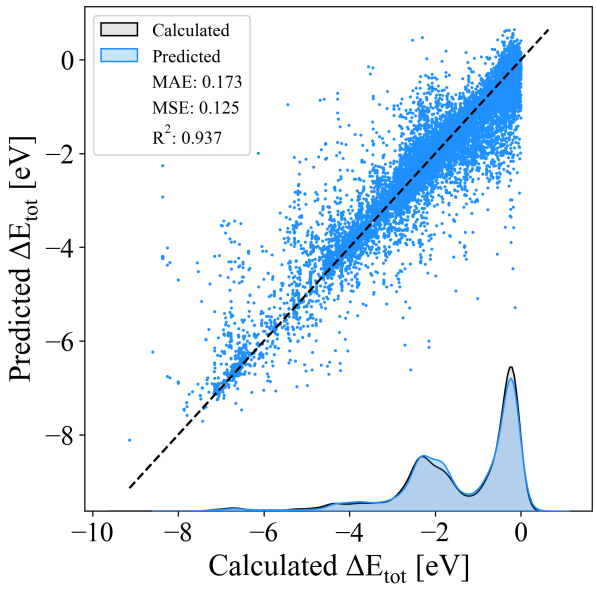

Figure 27 – ridge regression results for the concatenated $k = 2$, $k = 3$ LMBTR.

The concatenated LMBTR $k = 2$ and $k = 3$ output vector has a total of 18,540 features, requiring a great amount of working memory for being used. For treating these features for reducing the required amount of computational resources, the PCA method was employed specifically for reducing the input data dimensionality (thus computational cost), with results

present in Fig. 28. For performing the PCA, the LMBTR concatenated feature vector was fitted for two ranges of percentages of explained of variance in the training data: from 10% to 90% and 91% to 99%. The explained variance denotes the overall variance to be kept in the featurized train set for the LMBTR concatenated vector. While the PCA reduces the feature vector dimensionality (based on the train set distribution) by transforming correlated features in orthogonal ones through linear combinations, the explained variance metric defines the magnitude of this reduction. Hence, the higher the explained variance the more PCs (dimensions) to be kept, and the lower the explained variance the fewer PCs to be included in the PCA transformed vector, however with a loss of information for discerning between systems, *i.e.* some of variance is lost (unexplained).

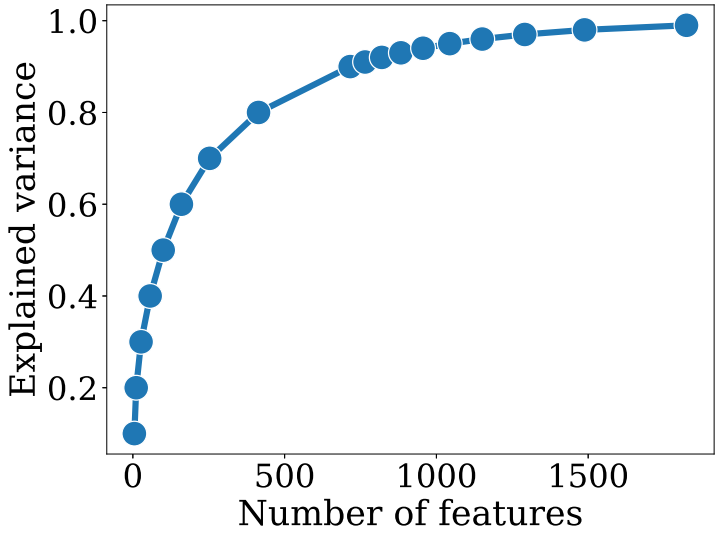

Figure 28 – PCA results for the cocatenated LMBTR $k = 2$ and $k = 3$ feature vectors.

For 99% of explained variance in the featurized train set, the $18,540$ features were transformed to $1,824$, which were sufficiently small considering the available working memory. The reduction to just 9.8% of the original number of features highlights the high sparsity of the LMBTR descriptor, as the majority of features contributed minimally to the variance. This is due to the (L)MBTR class-nature featurization, which organize atomic groups by species from the existing chemical species in data (HIMANEN *et al.*, 2020). Therefore, for any given sample, a great percentage of features will be zeros as they are related to other chemical species also present in the dataset. Additionally, many species combinations do not exist in conjunction in any sample from the dataset (*e.g.* Cu, Fe) but are initialized by default by the LMBTR $k = 3$ term – for instance X–Cu–Fe or Cu–X–Fe.

5.2 Random Forest Model

Table 8 – Optimal RFR hyperparameters for each descriptor with MAEs.

RFR Parameter	CM	LMBTR
`n_estimators`	100	136
`max_depth`	78	31
`max_features`	0.64	0.67
`min_samples_split`	8	8
`min_samples_leaf`	1	4
MAE	0.048 eV	0.054 eV

The optimal hyperparameters obtained for the RFR with each descriptor are summarized in Tab. 8. Both models achieved a very similar performance on the test set, with MAE values around 0.05 eV. Notably, the fraction of features used for each tree (`max_features`) remained close for both descriptors, at approximately 60%. However, key differences emerged in other hyperparameters, particularly tree depth (`max_depth`), which appears to be influenced by the number of features in each descriptor. The CM descriptor, which consists of fewer than 20 features, required a much deeper tree ((`max_depth`)= 78) compared to the PCA transformed LMBTR, which has 1,800 features and reached optimal performance with a shallower depth (`max_depth`= 31). This contrast suggests that as the CM has much less feature thant he LMBTR, it benefits from deep trees to fully capture feature interactions,while the LMBTR relies more on feature selection and sampling strategies to improve results. Additionally, the number of estimators varied, with the LMBTR requiring more trees (`n_estimators`= 136) than the CM (`n_estimators`= 100), presumably to compensate for the increased complexity of its feature space. Furthermore, we see that the `min_samples_split` and `min_samples_leaf` remained close, possibly compensating for overfitting based on the sample distribution, regardless of the descriptor. These differences highlight the impact of descriptor dimensionality on the RFR optimization and the importance for a hyperparameter optimization.

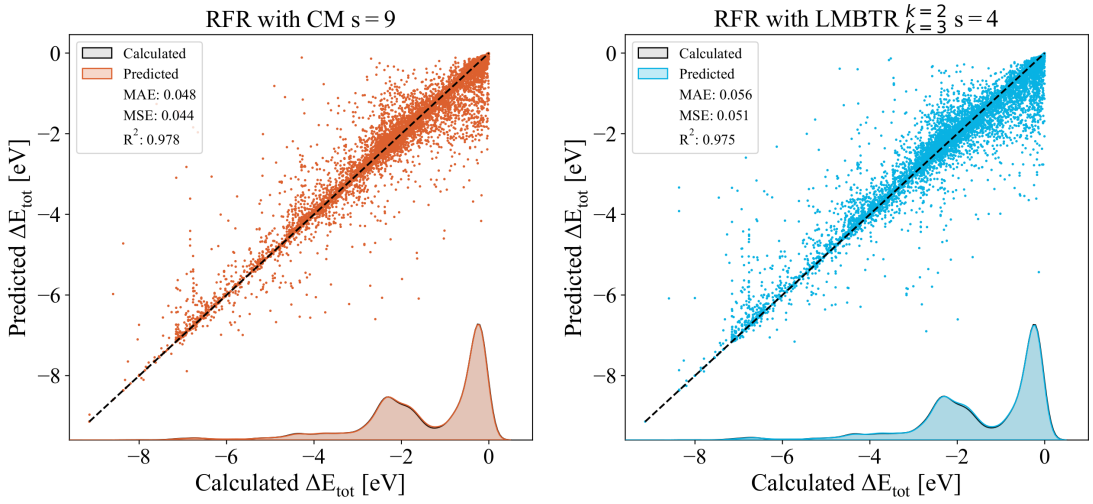

Figure 29 – Prediction results over the test set for the RFR algorithm with both optimized descriptors.

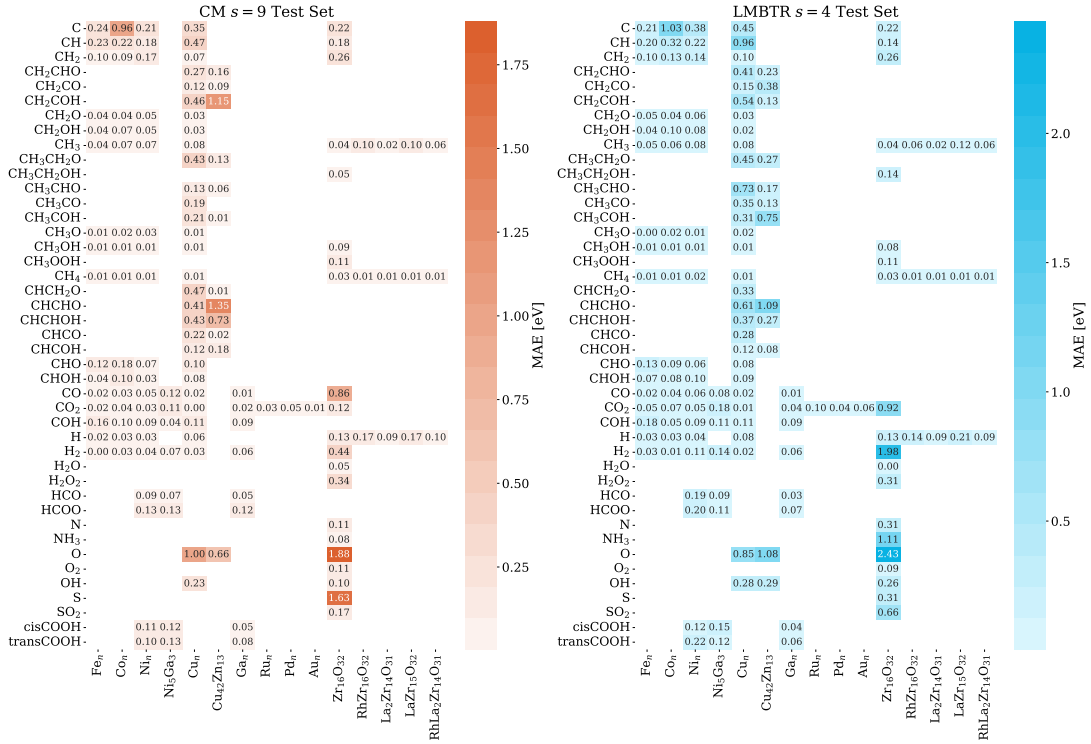

Figure 30 – Categorical MAEs for combinations of cluster compositions and adsorbate species.

Fig. 29 presents the prediction distribution for the RFR with the CM and the LMBTR. Both descriptors present MAEs of 0.05 eV with no observable biases from the prediction distribution towards under or overestimating any ΔE_{tot} region and there is a small number of outliers that are observable as single points in both test sets. It is noted that, by employing the RFR, both descriptors had their MAEs minimized to 0.048 and 0.055 eV for the CM and the LMBTR respectively. While the simpler CM descriptor performed much worse than the robust LMBTR in the linear regression, the RFR is a complex enough algorithm that nivelates both descriptors in their performances on the dataset used in this work.

A small number of outliers appear as isolated points in both test sets. These high-error cases are associated with specific adosrbate systems, as highlighted in Fig. 30. Notably, the $Zr_{16}O_{32}$ and $Cu_{42}Zn_{13}$ clusters exhibit the largest MAEs, ranging from approximately 1.0 eV to 2.0 eV. Higher MAEs at around 0.4 eV are also observed for the Cu_n cluster composition with some of the adsorbate species, which are related to the Cu_{55} cluster. Other than these cases, only the outlier MAE of 0.96 eV for the Co_n+C adsorbed systems is observed.

The great majority of predictions with either the CM or LMBTR with the RFR fall below or close to the expected error margin of 0.1 eV typically associated with DFT calculations on adsorbed systems compared to experimental results (HENSLEY *et al.*, 2017; SHARADA *et al.*, 2019; ARAUJO *et al.*, 2022). This indicates that the model can achieve an accuracy comparable to that of DFT, regardless of the structural descriptor employed. Therefore, other factors must be contributing to variations in model performance. To explore what might be increasing the MAEs for specific adsorbed systems, we turn to Fig. 31, which presents the MAEs per category

in relation to (A) cluster compositions, (B) cluster sizes, and (C) adsorbate species.

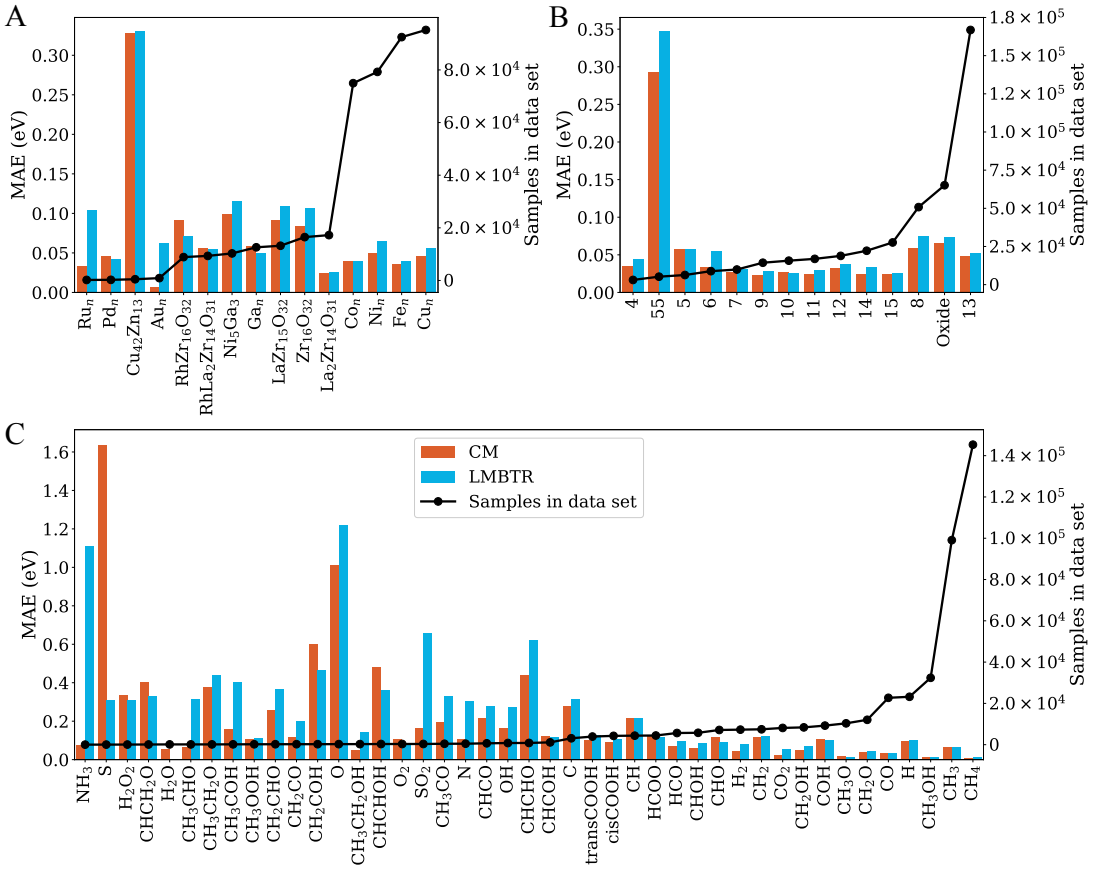

Figure 31 – MAEs per category and total sample count in dataset for (A) cluster compositions, (B) cluster sizes, and (C) adsorbate species. The MAEs are given as color coded bar plots for each descriptor with the RFR, with the black line denotating the total sample count of each category in the dataset.

As observed in Fig. 31, with the exception of 55 atom clusters, which greatly differ from the smaller clusters that compose most of the dataset, the total sample count of different cluster compositions and cluster sizes do not correlate to lower MAEs. However, it is observed that the MAE correlates in some proportion to the total count of samples of adsorbate species. This effect is especially evident in species such as NH_3, S, and O that presented MAEs exceeding 1.0 eV, well above the typical range. In contrast, adsorbates with larger sample sizes, starting at around 4.000 total samples in the dataset, consistently achieve MAEs below 0.2 eV, underscoring a strong link between adsorbate species sample representativity and model accuracy. This is in accordance to modeling different adsorption configurations, either from adsorption distances, adsorbate position and orientation to the cluster site, which effectively directs adsorption regime to either physi or chemisorption, the main factor affecting adsorption energies. Again, we note no substantial difference between the CM and LMBTR structural descriptors, with the observed differences mainly being related to sample count in the training sets, as executions were performed with different samplings for the dataset split. These results suggest that improving the balance of samples across adsorbate types could substantially enhance the performance and generalizability of ML models.

5.3 Final Model Assessment

For evaluating the model generalizability, that is, its capability for accurately predicting ΔE_{tot} values in new data, an external test set was constructed through the autonomous generation and selection of representative adsorbed systems through the k-means algorithm. The most prevalent cluster in data (Cu_{13}) was selected as the adsorbent for generation of the adsorbed systems with adsorbates CH_3, CH_4, CO, and CH_3OH. We also included the H_2O molecule, which is not present in the dataset with this cluster for assessing the knowledge transferability of the model from seen to unprecedented cases.

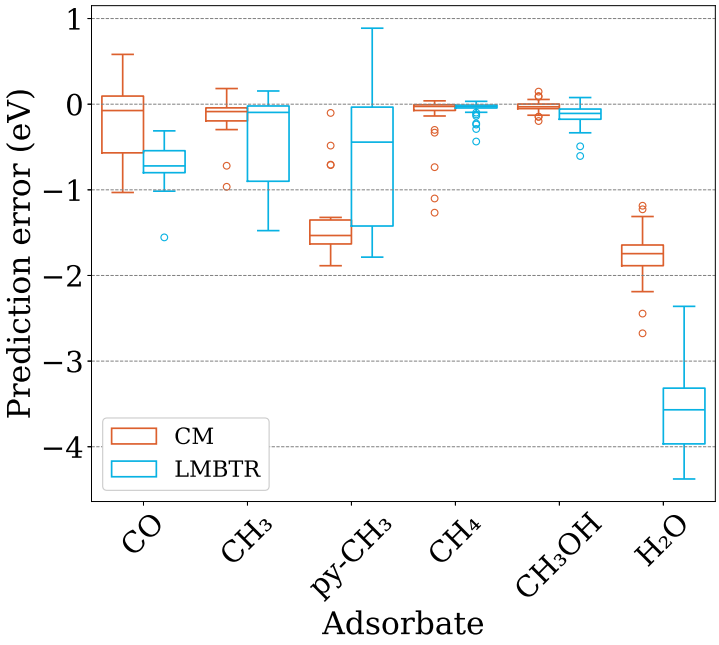

Figure 32 – Boxplot showing errors for assessed descriptors with the RFR on the external test set.

Fig. 32 presents the errors boxplots for both descriptors across the selected adsorbates on the Cu_{13} cluster. In general, both descriptors with the RFR perform similarly, ranging from lower prediction errors around -0.5 eV to higher errors around -1.5 eV for the known adsorbates in the dataset. The CH_4 and CH_3OH molecules yielded the best results, with the majority of predictions presenting errors smaller than 0.5 eV, however with some underestimated outliers for the CH_4 molecule with the CM and for the CH_3OH molecule with the LMBTR. The CO molecule exhibited markedly different prediction distributions for each descriptor, with the CM outperforming the LMBTR descriptor, however still presenting some uncertainty with errors down to -1.0 eV. The greatest discrepancies among the employed descriptors, however, appears for the CH_3 molecule, whose adsorbed configurations were generated with both planar and pyramidal geometries.

While for the LMBTR descriptor the prediction distribution spreads over a wide range of values for both geometries, the CM descriptor provides consistently accurate predictions for the planar CH_3 molecule, while poorly predicting the pyramidal py-CH_3 molecule with

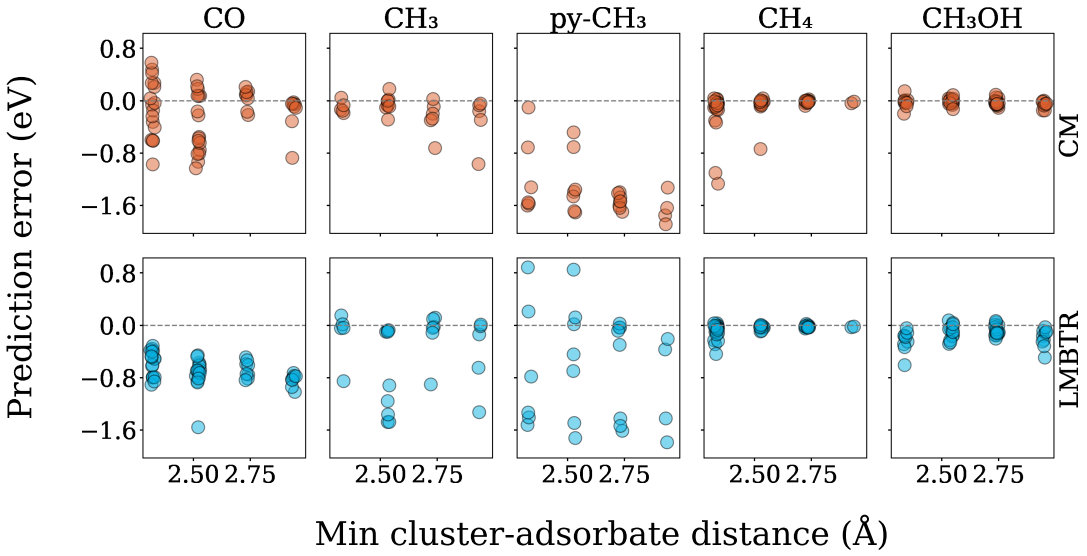

Figure 33 – Prediction distribution for both structural descriptors with the RFR in the external test set in relation to the minimum cluster-adsorbate distance.

only a few outliers next to the true DFT values. The pyramidal py-CH_3 molecule showed significant differences between descriptors: the LMBTR achieved 50 % of predictions with low errors (ranging from -0.5 eV to 0 eV), whereas CM exhibited higher errors around -1.5 eV, consistently underestimating the true values. For the planar CH_3, the CM performed better, producing fewer outliers compared to LMBTR, which underestimated energies by up to -1.5 eV. These results show a significant variation from the CM in estimating the same molecule with different geometries (planar or pyramidal CH_3) which is not as accentuated for the LMBTR. As observed in the boxplot, H_2O predictions for the Cu_{13} cluster exhibited consistently high errors, with neither structural descriptor achieving an error bellow 1.0 eV. This suggests that the structural descriptors are not effectively learning the physicochemical aspects of adsorbed systems, as it fails to transfer knowledge from previously seen H_2O adsorbates on oxide clusters in dataset to the Cu_{13} adsorbed configurations in the external test set.

We may also analyze the results in terms of cluster-adsorbate distance, which are presented in Fig. 33. As observed, the distance is not a major factor affecting prediction capability, however some takeaways may be noticed considering some of the observed predictions. For the CO molecule, the CM descriptor performed significantly better at larger distances compared to LMBTR, which had some increment in errors for larger distances. For the planar CH_3 molecule, no clear error correlation with distance is observed for the LMBTR descriptor, while outliers are observed in larger distances for the CM. In the case of the py-CH_3 molecule (pyramidal, adsorbed geometry), the LMBTR is capable of accurately predicting samples regardless of distance while the had CM only close predictions at lower distances, even though most of the configurations had high errors regardless of the descriptor. For the CH_4 molecule, LMBTR is more robust not showing the same outliers as the CM for closer distances, and for the CH_3OH molecule, while the most accurately predicted adsorbate for both descriptors, LMBTR had higher errors for the

closest and furthest distances.

5.3.1 Unpaired Electrons as Feature

Having evaluated the developed ML models with the CM and LMBTR on the external test set, we propose an additional electronic feature to the structural descriptors for improving model performance. The proposed feature is the total number of unpaired electrons in the adsorbate molecule, defined as a single digit for each dataset sample. This electronic feature is defined for all adsorbate species in the dataset as a `.json` file, being simply concatenated to the optimized CM eigenspectrum feature vector as a new, separate descriptor. Finally, a new RFR optimization is conducted with the same hyperparameter ranges as for the standard CM descriptor as defined in Tab. 4, and the optimized model for this new descriptor evaluated in the external test set. This adddition is motivated due to known adsorption phenomena: the presence of unpaired electrons is a key factor that influences adsorption strength, with closed-shell molecules generally weakly interacting, whereas radicals tend to form stronger interactions with surfaces.

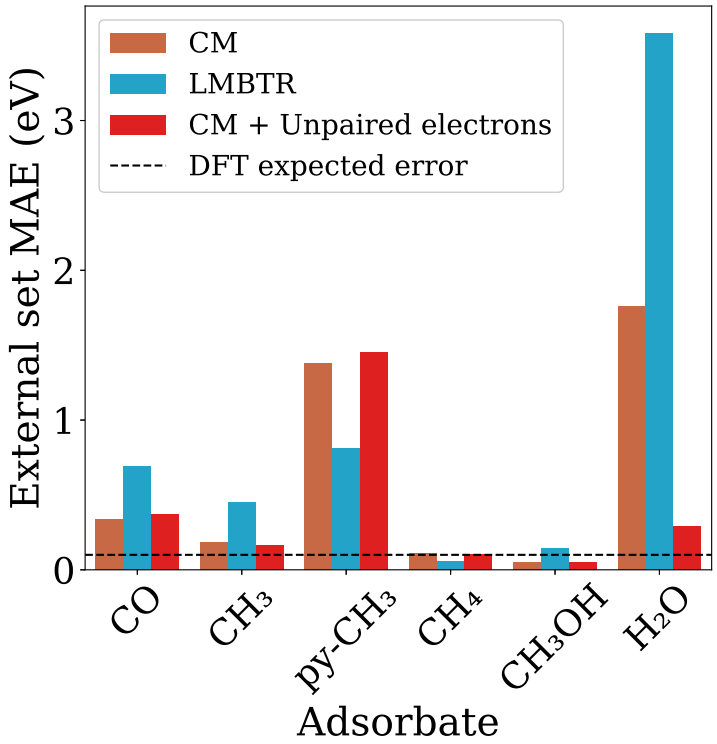

Figure 34 – MAE comparison for both descriptors with the RFR for each adsorbed system from the external test set and the corresponding train-test MAEs from the dataset performance.

The overall MAEs for each adsorbate and descriptor assessed in the external test set is given in Fig. 34, with the expected DFT error in relation to experimental results on adsorbed systems outlined (HENSLEY *et al.*, 2017; SHARADA *et al.*, 2019; ARAUJO *et al.*, 2022). Overall, the model only achieved MAEs under the expected DFT error for the CH_4 and CH_3OH adsorbates with both descriptors, while for adsorbates CO and CH_3 only the CM came close to

the expected DFT error. For the py$-CH_3$ molecule the LMBTR outperformed the CM, which may indicate its superiority in accurately modeling different molecule geometries, however still relying on sample representativity as observed by the high errors for this adsorbate. In general, the better performance of the CM may be explained as its resilience to overfitting the training data due to a much smaller number of features than the LMBTR. For the H_2O molecule, the inclusion of the adsorbate unpaired electrons as an additional electronic feature to the CM feature vector substantially improved prediction MAEs, which indicates that electronic information on the adsorbed system is crucial for the model generalizability. By including the electronic feature of the number of adsorbate unpaired electrons, it is observed that the errors for the H_2O molecule are reduced down to around 0.5 eV. Nevertheless, prediction errors were significantly higher in the external test set than in the dataset samples, as discussed previously, even though this external test set was generated with the most represented nanocluster (Cu_{13}) in the dataset. This outlines that the model cannot deal with unprecendented adsorbed configurations, such as the pyramidal py-CH_3 molecule in not chemisorption distances, effectively indicating that sample representativity – besides adsorbate species and cluster size or composition – is critical for model performance.

6 Conclusion

While many works in the literature have achieved optimal results for ML models using geometric featurization, they are often limited to few adsorbed system cases. Several studies demonstrated that molecular descriptors provide a clear but small improvement when employing more advanced formulations, with the CM commonly used as a baseline for comparison (JäGER *et al.*, 2018). In this work, we observe that even when utilizing two levels of molecular descriptors, similar predictive accuracy can be achieved by employing a sufficiently complex regression algorithm such as the RFR over a chemically diverse dataset of nanocluster adsorbed systems.

However, molecular descriptors lack generalizability and are insufficient for developing a robust and transferable ML model for adsorption energy prediction across diverse cluster structures, compositions, and adsorbates. While low errors bellow the expected DFT error of 0.1 eV (HENSLEY *et al.*, 2017; SHARADA *et al.*, 2019; ARAUJO *et al.*, 2022) are obtained for well-represented systems, the model fails to generalize to underrepresented cases. This is mainly observed when employing the developed model on new generated data of adsorbed configurations using the most represented cluster in the dataset. In this external test set, the model presented higher prediction errors and required an additional electronic feature to the structural descriptors for accurately predicting the novel adsorption case of the H_2O molecule with the Cu_{13} cluster, which was not present in the dataset. This suggests that the quantity and diversity of training data is far more critical than currently available featurization techniques or regression algorithms.

Recently, many works focused on adsorption energy prediciton have employed electronic features such as cluster d-orbital energy and occupation, which showed great importance for accurately predicting adsorption energies (SHARMA *et al.*, 2024; BHARADWAJ *et al.*, 2025; LIU *et al.*, 2025). This indicates a new pathway for the development of reliable and explainable ML models in this field of research. Furthermore, the model's strong dependence on sample representativity for accurately determined adsorbed configurations severely limits its practical utility, as the effort required to generate a sufficiently large training dataset demands computational cost, time, and human effort. To address these limitations, future research should explore active learning strategies or hybrid ML-DFT approaches that benefits from the strengths of both methodologies, while simultaneously addressing their limitations. From this framework, research on nanocluster adsorbed system could be accelerated, effectively reducing the dependence on exhaustive DFT calculations while maintaining accuracy and reliability, bringing us closer to the discovery of breakthroughs in renewable energy catalysis.

Bibliography

AKIBA, T. *et al.* Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2019. (KDD19), p. 2623–2631. Disponível em: <http://dx.doi.org/10.1145/3292500.3330701>. Citado na página 57.

ANDRIANI, K. F. *et al.* Role of quantum-size effects in the dehydrogenation of ch4 on 3d tmn clusters: Dft calculations combined with data mining. *Catalysis Science & Technology*, Royal Society of Chemistry (RSC), v. 12, n. 3, p. 916–926, 2022. ISSN 2044-4761. Disponível em: <http://dx.doi.org/10.1039/D1CY01785C>. Citado 2 vezes nas páginas 52 and 53.

ANDRIANI, K. F.; MUCELINI, J.; SILVA, J. L. F. D. Methane dehydrogenation on 3d 13-atom transition-metal clusters: A density functional theory investigation combined with spearman rank correlation analysis. *Fuel*, Elsevier BV, v. 275, p. 117790, set. 2020. ISSN 0016-2361. Disponível em: <http://dx.doi.org/10.1016/j.fuel.2020.117790>. Citado 2 vezes nas páginas 52 and 53.

ANTOINE, R.; BROYER, M.; DUGOURD, P. Metal nanoclusters: from fundamental aspects to electronic properties and optical applications. *Science and Technology of Advanced Materials*, Informa UK Limited, v. 24, n. 1, jun. 2023. ISSN 1878-5514. Disponível em: <http://dx.doi.org/10.1080/14686996.2023.2222546>. Citado na página 16.

ARAUJO, R. B. *et al.* Adsorption energies on transition metal surfaces: towards an accurate and balanced description. *Nature Communications*, Springer Science and Business Media LLC, v. 13, n. 1, nov. 2022. ISSN 2041-1723. Disponível em: <http://dx.doi.org/10.1038/s41467-022-34507-y>. Citado 3 vezes nas páginas 66, 70, and 72.

ASHRAF, S. *et al.* Bimetallic nanoalloy catalysts for green energy production: Advances in synthesis routes and characterization techniques. *Small*, Wiley, v. 19, n. 43, jun. 2023. ISSN 1613-6829. Disponível em: <http://dx.doi.org/10.1002/smll.202303031>. Citado na página 17.

ATKINS, P. *Atkins's Physical Chemistry: International Eleventh Edition*. [S.l.: s.n.], 2018. Citado 2 vezes nas páginas 28 and 30.

BARBER, D. J.; FREESTONE, I. C. An investigation of the origin of the colour of the lycurgus cup by analytical transmission electron microscopy. *Archaeometry*, Wiley, v. 32, n. 1, p. 33–45, fev. 1990. ISSN 1475-4754. Disponível em: <http://dx.doi.org/10.1111/j.1475-4754.1990.tb01079.x>. Citado 2 vezes nas páginas 6 and 14.

BATISTA, K. E. A. *et al.* A theoretical investigation of the structural and electronic properties of 55-atom nanoclusters: The examples of Y-Tc and Pt. *J. Chem. Phys.*, American Institute of Physics publishing, v. 144, n. 5, p. 054310, Feb. 2016. ISSN 0021-9606, 1089-7690. Disponível em: <https://doi.org/10.1063/1.4941295>. Citado na página 37.

BATISTA, K. E. A. *et al.* Energy decomposition to access the stability changes induced by CO adsorption on transition-metal 13-atom clusters. *J. Chem. Inf. Model.*, American Chemical Society, v. 61, n. 5, p. 2294–2301, May 2021. ISSN 1549-9596, 1549-960X. Disponível em: <https://doi.org/10.1021/acs.jcim.1c00097>. Citado na página 58.

BHARADWAJ, N. *et al.* Machine learning-driven screening of atomically precise pure metal nanoclusters for oxygen reduction. *ACS Materials Letters*, American Chemical Society (ACS), v. 7, n. 2, p. 500–507, jan. 2025. ISSN 2639-4979. Disponível em: <http://dx.doi.org/10.1021/acsmaterialslett.4c01737>. Citado na página 72.

BIAU, G.; SCORNET, E. A random forest guided tour. *TEST*, v. 25, n. 2, p. 197–227, jun. 2016. ISSN 1133-0686, 1863-8260. Citado na página 48.

BISWAS, S. *et al.* Advancing electrochemical co2 reduction with group 11 metal nanoclusters for renewable energy solutions. *EcoEnergy*, Wiley, v. 2, n. 3, p. 400–418, jul. 2024. ISSN 2835-9399. Disponível em: <http://dx.doi.org/10.1002/ece2.56>. Citado na página 23.

BLUM, V. *et al. Ab initio* molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.*, Elsevier BV, v. 180, n. 11, p. 2175–2196, Nov. 2009. ISSN 0010-4655. Disponível em: <https://doi.org/10.1016/j.cpc.2009.06.022>. Citado 2 vezes nas páginas 52 and 59.

BREIMAN, L. *Machine Learning*, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A:1010933404324>. Citado 2 vezes nas páginas 48 and 51.

BREIMAN, L. *et al. Classification And Regression Trees*. Routledge, 1984. ISBN 9781315139470. Disponível em: <http://dx.doi.org/10.1201/9781315139470>. Citado na página 48.

BREIMAN, L. *et al. Classification And Regression Trees*. Routledge, 2017. ISBN 9781315139470. Disponível em: <http://dx.doi.org/10.1201/9781315139470>. Citado 2 vezes nas páginas 48 and 50.

CIONTI, C.; STUCCHI, M.; MERONI, D. Mimicking stained glass: A hands-on activity for the preparation and characterization of silica films colored with noble metal ions and nanoparticles. *Journal of Chemical Education*, American Chemical Society (ACS), v. 99, n. 3, p. 1516–1522, fev. 2022. ISSN 1938-1328. Disponível em: <http://dx.doi.org/10.1021/acs.jchemed.1c01141>. Citado na página 14.

COLLACIQUE, M. N.; OCAMPO-RESTREPO, V. K.; SILVA, J. L. F. D. Ab initio investigation of the role of the d-states on the adsorption and activation properties of co2 on 3d, 4d, and 5d transition-metal clusters. *The Journal of Chemical Physics*, AIP Publishing, v. 156, n. 12, mar. 2022. ISSN 1089-7690. Disponível em: <http://dx.doi.org/10.1063/5.0085364>. Citado 2 vezes nas páginas 52 and 53.

DEISENROTH, M. P.; FAISAL, A. A.; ONG, C. S. *Mathematics for Machine Learning*. [S.l.]: Cambridge University Press, 2020. Citado 8 vezes nas páginas 24, 26, 28, 43, 44, 45, 48, and 51.

ECHENIQUE, P.; ALONSO, J. L. A mathematical and computational review of hartree-fock scf methods in quantum chemistry. *Molecular Physics*, Informa UK Limited, v. 105, n. 23–24, p. 3057–3098, dez. 2007. ISSN 1362-3028. Disponível em: <http://dx.doi.org/10.1080/00268970701757875>. Citado 2 vezes nas páginas 29 and 32.

EVANS, S. *Analysis: Which countries are historically responsible for climate change?* Carbon Brief, 2021. Accessed: 2025-01-14. Disponível em: <https://www.carbonbrief.org/

analysis-which-countries-are-historically-responsible-for-climate-change/>. Citado 2 vezes nas páginas 6 and 11.

EWIS, D. *et al.* Electrochemical reduction of co2 into formate/formic acid: A review of cell design and operation. *Separation and Purification Technology*, Elsevier BV, v. 316, p. 123811, jul. 2023. ISSN 1383-5866. Disponível em: <http://dx.doi.org/10.1016/j.seppur.2023.123811>. Citado na página 22.

FARADAY, M. *Philosophical Transactions of the Royal Society of London*, The Royal Society, v. 147, p. 145–181, dez. 1857. ISSN 2053-9223. Disponível em: <http://dx.doi.org/10.1098/rstl.1857.0011>. Citado na página 15.

FELíCIO-SOUSA, P. *Investigação Ab initio da ativação de metano em clusters metálicos e óxidos: integração com mineração de dados*. 236 p. Tese (Doutorado) — São Carlos Institute of Chemistry, 2024. Citado 2 vezes nas páginas 52 and 53.

FELíCIO-SOUSA, P.; ANDRIANI, K. F.; SILVA, J. L. F. D. Ab initio investigation of the role of the d-states occupation on the adsorption properties of h2, co, ch4 and ch3oh on the fe13, co13, ni13 and cu13 clusters. *Physical Chemistry Chemical Physics*, Royal Society of Chemistry (RSC), v. 23, n. 14, p. 8739–8751, 2021. ISSN 1463-9084. Disponível em: <http://dx.doi.org/10.1039/D0CP06091G>. Citado 2 vezes nas páginas 52 and 53.

FERMIN, D. J.; MARKEN, F. Chapter 1. introduction to the eletrochemical and photo-electrochemical reduction of co2. In: ____. *Electrochemical Reduction of Carbon Dioxide*. Royal Society of Chemistry, 2018. p. 1–16. ISBN 9781782620426. Disponível em: <http://dx.doi.org/10.1039/9781782623809-00001>. Citado 2 vezes nas páginas 6 and 21.

FERRANDO, R. *Structure and properties of nanoalloys*. Amsterdam: Elsevier, 2016. (Frontiers of nanoscience, volume 10). ISBN 978-0-08-100212-4. Citado 3 vezes nas páginas 15, 16, and 17.

GALTON, F. *Natural inheritance, by Francis Galton.* Macmillan and co.,, 1894. Disponível em: <http://dx.doi.org/10.5962/bhl.title.46339>. Citado na página 46.

GOMES, I. L. *Investigação Teórica da Adsorção de Moléculas sobre Nanocluster de Óxido de Zircônia*. Tese (Doutorado) — São Carlos Institute of Chemistry, 2022. Disponível em: <https://bdta.abcd.usp.br/item/003099841>. Citado 2 vezes nas páginas 52 and 53.

HANSEN, K. *et al.* Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The Journal of Physical Chemistry Letters*, American Chemical Society (ACS), v. 6, n. 12, p. 2326–2331, jun. 2015. ISSN 1948-7185. Disponível em: <http://dx.doi.org/10.1021/acs.jpclett.5b00831>. Citado na página 38.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. Springer New York, 2009. ISSN 2197-568X. ISBN 9780387848587. Disponível em: <http://dx.doi.org/10.1007/978-0-387-84858-7>. Citado 8 vezes nas páginas 7, 25, 43, 45, 46, 47, 49, and 50.

HAVU, V. *et al.* Efficient integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.*, Elsevier BV, v. 228, n. 22, p. 8367–8379, Dec. 2009. ISSN 0021-9991. Disponível em: <https://doi.org/10.1016/j.jcp.2009.08.008>. Citado 2 vezes nas páginas 52 and 59.

HENSLEY, A. J. R. *et al.* Dft-based method for more accurate adsorption energies: An adaptive sum of energies from rpbe and vdw density functionals. *The Journal of Physical Chemistry C*, American Chemical Society (ACS), v. 121, n. 9, p. 4937–4945, fev. 2017. ISSN 1932-7455. Disponível em: <http://dx.doi.org/10.1021/acs.jpcc.6b10187>. Citado 3 vezes nas páginas 66, 70, and 72.

HIMANEN, L. *et al.* Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, Elsevier BV, v. 247, p. 106949, fev. 2020. ISSN 0010-4655. Disponível em: <http://dx.doi.org/10.1016/j.cpc.2019.106949>. Citado 5 vezes nas páginas 38, 40, 42, 56, and 64.

HOERL, R. W. Ridge regression: A historical context. *Technometrics*, Informa UK Limited, v. 62, n. 4, p. 420–425, out. 2020. ISSN 1537-2723. Disponível em: <http://dx.doi.org/10.1080/00401706.2020.1742207>. Citado na página 48.

HOHENBERG, P.; KOHN, W. Inhomogeneous Electron Gas. *Physical Review*, v. 136, n. 3B, p. B864–B871, nov. 1964. ISSN 0031-899X. Disponível em: <https://link.aps.org/doi/10.1103/PhysRev.136.B864>. Citado na página 33.

HU, E. *et al.* Machine learning assisted understanding and discovery of co2 reduction reaction electrocatalyst. *The Journal of Physical Chemistry C*, American Chemical Society (ACS), v. 127, n. 2, p. 882–893, jan. 2023. ISSN 1932-7455. Disponível em: <http://dx.doi.org/10.1021/acs.jpcc.2c08343>. Citado na página 12.

HUO, H.; RUPP, M. Unified representation of molecules and crystals for machine learning. *Machine Learning: Science and Technology*, IOP Publishing, v. 3, n. 4, p. 045017, nov. 2022. ISSN 2632-2153. Disponível em: <http://dx.doi.org/10.1088/2632-2153/aca005>. Citado 2 vezes nas páginas 39 and 40.

IKOTUN, A. M. *et al.* K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, Elsevier BV, v. 622, p. 178–210, abr. 2023. ISSN 0020-0255. Disponível em: <http://dx.doi.org/10.1016/j.ins.2022.11.139>. Citado na página 45.

JäGER, M. O. J. *et al.* Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials*, Springer Science and Business Media LLC, v. 4, n. 1, jul. 2018. ISSN 2057-3960. Disponível em: <http://dx.doi.org/10.1038/s41524-018-0096-5>. Citado 2 vezes nas páginas 57 and 72.

JENSEN, F. *Introduction to computational chemistry*. 2nd ed. ed. Chichester, England; Hoboken, NJ: John Wiley and Sons, 2007. ISBN 978-0-470-01186-7. Citado 2 vezes nas páginas 31 and 35.

JOYCE, D. *Permutations and Determinants*. 2015. Math 130 Linear Algebra, Fall 2015. Disponível em: <https://aleph0.clarku.edu/~ma130/permutations.pdf>. Citado na página 27.

KAATZ, F. H.; BULTHEEL, A. Magic mathematical relationships for nanoclusters. *Nanoscale Research Letters*, Springer Science and Business Media LLC, v. 14, n. 1, maio 2019. ISSN 1556-276X. Disponível em: <http://dx.doi.org/10.1186/s11671-019-2939-5>. Citado na página 17.

KAKEKHANI, A. Using ferroelectrics to tackle fundamental challenges in catalysis. Unpublished, 2016. Disponível em: <http://rgdoi.net/10.13140/RG.2.2.26916.94081>. Citado 2 vezes nas páginas 6 and 19.

KANDY, M. M. Carbon-based photocatalysts for enhanced photocatalytic reduction of co2 to solar fuels. *Sustainable Energy and; Fuels*, Royal Society of Chemistry (RSC), v. 4, n. 2, p. 469–484, 2020. ISSN 2398-4902. Disponível em: <http://dx.doi.org/10.1039/c9se00827f>. Citado na página 11.

KANG, X. *et al.* Atomically precise alloy nanoclusters: syntheses, structures, and properties. *Chemical Society Reviews*, Royal Society of Chemistry (RSC), v. 49, n. 17, p. 6443–6514, 2020. ISSN 1460-4744. Disponível em: <http://dx.doi.org/10.1039/C9CS00633H>. Citado na página 16.

KAWAWAKI, T.; NEGISHI, Y. Elucidation of the electronic structures of thiolate-protected gold nanoclusters by electrochemical measurements. *Dalton Transactions*, Royal Society of Chemistry (RSC), v. 52, n. 42, p. 15152–15167, 2023. ISSN 1477-9234. Disponível em: <http://dx.doi.org/10.1039/D3DT02005C>. Citado 2 vezes nas páginas 15 and 18.

KAWAWAKI, T. *et al.* Atomically precise metal nanoclusters as catalysts for electrocatalytic co2 reduction. *Green Chemistry*, Royal Society of Chemistry (RSC), v. 26, n. 1, p. 122–163, 2024. ISSN 1463-9270. Disponível em: <http://dx.doi.org/10.1039/D3GC02281A>. Citado 2 vezes nas páginas 20 and 22.

KHALIL, M. *et al.* Photocatalytic conversion of co2 using earth-abundant catalysts: A review on mechanism and catalytic performance. *Renewable and Sustainable Energy Reviews*, Elsevier BV, v. 113, p. 109246, out. 2019. ISSN 1364-0321. Disponível em: <http://dx.doi.org/10.1016/j.rser.2019.109246>. Citado na página 11.

KHATUN, E.; PRADEEP, T. New routes for multicomponent atomically precise metal nanoclusters. *ACS Omega*, American Chemical Society (ACS), v. 6, n. 1, p. 1–16, dez. 2020. ISSN 2470-1343. Disponível em: <http://dx.doi.org/10.1021/acsomega.0c04832>. Citado na página 16.

KNIGHT, W. D. *et al.* Electronic shell structure and abundances of sodium clusters. *Physical Review Letters*, American Physical Society (APS), v. 52, n. 24, p. 2141–2143, jun. 1984. ISSN 0031-9007. Disponível em: <http://dx.doi.org/10.1103/PhysRevLett.52.2141>. Citado na página 17.

KOHN, W.; SHAM, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, v. 140, n. 4A, p. A1133–A1138, nov. 1965. ISSN 0031-899X. Disponível em: <https://link.aps.org/doi/10.1103/PhysRev.140.A1133>. Citado na página 34.

LANGER, M. F.; GOEßMANN, A.; RUPP, M. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *npj Computational Materials*, Springer Science and Business Media LLC, v. 8, n. 1, mar. 2022. ISSN 2057-3960. Disponível em: <http://dx.doi.org/10.1038/s41524-022-00721-x>. Citado 2 vezes nas páginas 36 and 39.

LEVINE, I. N. *Quantum chemistry*. Seventh edition. Boston: Pearson, 2014. ISBN 978-0-321-80345-0. Citado 7 vezes nas páginas 28, 29, 31, 32, 33, 34, and 35.

LEWARS, E. G. *Computational Chemistry*. Cham: Springer International Publishing, 2016. ISBN 978-3-319-30914-9 978-3-319-30916-3. Disponível em: <http://link.springer.com/10.1007/978-3-319-30916-3>. Citado 4 vezes nas páginas 31, 33, 34, and 35.

LIU, W. *et al.* Advancing the understanding and prediction accuracy of molecular adsorption energy with artificial intelligence. *Physical Review B*, American Physical Society (APS), v. 111, n. 19, maio 2025. ISSN 2469-9969. Disponível em: <http://dx.doi.org/10.1103/PhysRevB.111.195413>. Citado na página 72.

LOWE, J. P.; PETERSON, K. A. *Quantum chemistry*. 3rd ed. ed. Burlington, MA: Elsevier Academic Press, 2006. OCLC: ocm60798495. ISBN 978-0-12-457551-6. Citado na página 33.

MCPARTLIN, M.; MASON, R.; MALATESTA, L. Novel cluster complexes of gold(0)–gold(<scp>i</scp>). *J. Chem. Soc. D*, Royal Society of Chemistry (RSC), v. 0, n. 7, p. 334–334, 1969. ISSN 0577-6171. Disponível em: <http://dx.doi.org/10.1039/C29690000334>. Citado na página 15.

MENDONCA, J. P. A. de *et al.* Theoretical framework based on molecular dynamics and data mining analyses for the study of potential energy surfaces of finite-size particles. *Journal of Chemical Information and Modeling*, American Chemical Society (ACS), v. 62, n. 22, p. 5503–5512, out. 2022. ISSN 1549-960X. Disponível em: <http://dx.doi.org/10.1021/acs.jcim.2c00957>. Citado na página 37.

MORAIS, F. O.; ANDRIANI, K. F.; SILVA, J. L. F. D. Investigation of the stability mechanisms of eight-atom binary metal clusters using dft calculations and k-means clustering algorithm. *Journal of Chemical Information and Modeling*, American Chemical Society (ACS), v. 61, n. 7, p. 3411–3420, jun. 2021. ISSN 1549-960X. Disponível em: <http://dx.doi.org/10.1021/acs.jcim.1c00253>. Citado na página 37.

NEGISHI, Y. Metal-nanocluster science and technology: my personal history and outlook. *Physical Chemistry Chemical Physics*, Royal Society of Chemistry (RSC), v. 24, n. 13, p. 7569–7594, 2022. ISSN 1463-9084. Disponível em: <http://dx.doi.org/10.1039/D1CP05689A>. Citado na página 15.

OCAMPO-RESTREPO, V. K.; VERGA, L. G.; SILVA, J. L. F. D. Ab initio study for late steps of co2 and co electroreduction: from chco* toward c2 products on cu and cuzn nanoclusters. *Physical Chemistry Chemical Physics*, Royal Society of Chemistry (RSC), v. 25, n. 48, p. 32931–32938, 2023. ISSN 1463-9084. Disponível em: <http://dx.doi.org/10.1039/D3CP03315E>. Citado 2 vezes nas páginas 52 and 53.

OOKA, H.; HUANG, J.; EXNER, K. S. The sabatier principle in electrocatalysis: Basics, limitations, and extensions. *Frontiers in Energy Research*, Frontiers Media SA, v. 9, maio 2021. ISSN 2296-598X. Disponível em: <http://dx.doi.org/10.3389/fenrg.2021.654460>. Citado na página 19.

PATTERSON, A. L. Homometric structures. *Nature*, Springer Science and Business Media LLC, v. 143, n. 3631, p. 939–940, jun. 1939. ISSN 1476-4687. Disponível em: <http://dx.doi.org/10.1038/143939b0>. Citado na página 39.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in python. arXiv, 2012. Disponível em: <https://arxiv.org/abs/1201.0490>. Citado 2 vezes nas páginas 56 and 57.

PEI, G. X.; ZHANG, L.; SUN, X. Recent advances of bimetallic nanoclusters with atomic precision for catalytic applications. *Coordination Chemistry Reviews*, Elsevier BV, v. 506, p. 215692, maio 2024. ISSN 0010-8545. Disponível em: <http://dx.doi.org/10.1016/j.ccr.2024.215692>. Citado 2 vezes nas páginas 16 and 23.

PENA, L. B. *et al.* Underlying mechanisms of gold nanoalloys stabilization. *The Journal of Chemical Physics*, AIP Publishing, v. 159, n. 24, dez. 2023. ISSN 1089-7690. Disponível em: <http://dx.doi.org/10.1063/5.0180906>. Citado na página 38.

PERDEW, J. P.; BURKE, K.; ERNZERHOF, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, v. 77, n. 18, p. 3865–3868, out. 1996. ISSN 0031-9007, 1079-7114. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevLett.77.3865>. Citado 2 vezes nas páginas 52 and 59.

PIELA, L. *Ideas of quantum chemistry*. 1st ed. ed. Amsterdam ; Boston: Elsevier, 2007. OCLC: ocm78525103. ISBN 978-0-444-52227-6. Citado na página 32.

QIAN, J. *et al.* Molecular interactions in atomically precise metal nanoclusters. *Precision Chemistry*, American Chemical Society (ACS), v. 2, n. 10, p. 495–517, ago. 2024. ISSN 2771-9316. Disponível em: <http://dx.doi.org/10.1021/prechem.4c00044>. Citado 3 vezes nas páginas 15, 18, and 23.

QIN, L. *et al.* Atomically precise metal nanoclusters for (photo)electroreduction of co2: Recent advances, challenges and opportunities. *Journal of Energy Chemistry*, Elsevier BV, v. 57, p. 359–370, jun. 2021. ISSN 2095-4956. Disponível em: <http://dx.doi.org/10.1016/j.jechem.2020.09.003>. Citado na página 21.

RODGERS, L. *Climate change: Where we are in seven charts and what you can do to help*. BBC News, 2018. Accessed: 2025-01-14. Disponível em: <https://www.bbc.com/news/science-environment-46455844>. Citado na página 11.

RUPP, M. *et al.* Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters*, v. 108, n. 5, p. 058301, jan. 2012. ISSN 0031-9007, 1079-7114. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevLett.108.058301>. Citado na página 37.

RUTKOWSKI, L. *et al.* The cart decision tree for mining data streams. *Information Sciences*, Elsevier BV, v. 266, p. 1–15, maio 2014. ISSN 0020-0255. Disponível em: <http://dx.doi.org/10.1016/j.ins.2013.12.060>. Citado na página 49.

SADASIVUNI, K. K. *et al. 2D Nanomaterials for CO2 Conversion into Chemicals and Fuels*. [S.l.]: The Royal Society of Chemistry, 2022. ISBN 978-1-83916-311-1. Citado 4 vezes nas páginas 12, 20, 21, and 22.

SCHLEDER, G. R. *et al.* From dft to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, IOP Publishing, v. 2, n. 3, p. 032001, maio 2019. ISSN 2515-7639. Disponível em: <http://dx.doi.org/10.1088/2515-7639/ab084b>. Citado 2 vezes nas páginas 6 and 12.

SCHRÖDINGER, E. An undulatory theory of the mechanics of atoms and molecules. *Physical Review*, American Physical Society (APS), v. 28, n. 6, p. 1049–1070, dez. 1926. ISSN 0031-899X. Disponível em: <http://dx.doi.org/10.1103/PhysRev.28.1049>. Citado na página 28.

SHANG, L.; XU, J.; NIENHAUS, G. U. Recent advances in synthesizing metal nanocluster-based nanocomposites for application in sensing, imaging and catalysis. *Nano Today*, Elsevier BV, v. 28, p. 100767, out. 2019. ISSN 1748-0132. Disponível em: <http://dx.doi.org/10.1016/j.nantod.2019.100767>. Citado na página 16.

SHARADA, S. M. *et al.* Adsorption on transition metal surfaces: Transferability and accuracy of dft using the ads41 dataset. *Physical Review B*, American Physical Society (APS), v. 100, n. 3, jul. 2019. ISSN 2469-9969. Disponível em: <http://dx.doi.org/10.1103/PhysRevB.100.035439>. Citado 3 vezes nas páginas 66, 70, and 72.

SHARMA, R. K. *et al.* Machine-learning-assisted screening of nanocluster electrocatalysts: Mapping and reshaping the activity volcano for the oxygen reduction reaction. *ACS Applied Materials and Interfaces*, American Chemical Society (ACS), v. 16, n. 46, p. 63589–63601, nov. 2024. ISSN 1944-8252. Disponível em: <http://dx.doi.org/10.1021/acsami.4c14076>. Citado na página 72.

SHLENS, J. *A Tutorial on Principal Component Analysis*. arXiv, 2014. Disponível em: <https://arxiv.org/abs/1404.1100>. Citado 3 vezes nas páginas 26, 43, and 44.

SILVA, J. L. F. D. *Quantum Theory of Nanomaterials Group - IQSC/USP*. 2025. Accessed: 2025-04-04. Disponível em: <https://qtnano.iqsc.usp.br>. Citado na página 13.

SILVA, L. R. da *et al.* Theoretical investigation of the stability of a55-b nanoalloys (a, b = al, cu, zn, ag). *Computational Materials Science*, Elsevier BV, v. 215, p. 111805, dez. 2022. ISSN 0927-0256. Disponível em: <http://dx.doi.org/10.1016/j.commatsci.2022.111805>. Citado na página 37.

SOUSA, R. A. D. *et al.* Ab initio study of the adsorption properties of co2 reduction intermediates: The effect of ni5ga3 alloy and the ni5ga3/zro2 interface. *The Journal of Chemical Physics*, AIP Publishing, v. 156, n. 21, jun. 2022. ISSN 1089-7690. Disponível em: <http://dx.doi.org/10.1063/5.0091145>. Citado 2 vezes nas páginas 52 and 53.

SOUZA, T. M. *et al.* Data-driven stabilization of nimpdn–m nanoalloys: a study using density functional theory and data mining approaches. *Physical Chemistry Chemical Physics*, Royal Society of Chemistry (RSC), v. 26, n. 22, p. 15877–15890, 2024. ISSN 1463-9084. Disponível em: <http://dx.doi.org/10.1039/D4CP00672K>. Citado na página 38.

STANTON, J. M. Galton, pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, Informa UK Limited, v. 9, n. 3, jan. 2001. ISSN 1069-1898. Disponível em: <http://dx.doi.org/10.1080/10691898.2001.11910537>. Citado na página 46.

SZABO, N. S. O. A. *Modern Quantum Chemistry Introduction to Advanced Eletronic Structure Theory*. [S.l.: s.n.], 1996. Citado na página 31.

TKATCHENKO, A.; SCHEFFLER, M. Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.*, American Physical Society, v. 102, n. 7, p. 073005, Feb. 2009. ISSN 0031-9007, 1079-7114. Disponível em: <https://doi.org/10.1103/physrevlett.102.073005>. Citado 2 vezes nas páginas 52 and 59.

TSUKAMOTO, T. *et al.* Periodicity of molecular clusters based on symmetry-adapted orbital model. *Nature Communications*, Springer Science and Business Media LLC, v. 10, n. 1, ago.

2019. ISSN 2041-1723. Disponível em: <http://dx.doi.org/10.1038/s41467-019-11649-0>. Citado na página 18.

TSUKAMOTO, T. *et al.* Modern cluster design based on experiment and theory. *Nature Reviews Chemistry*, v. 5, n. 5, p. 338–347, maio 2021. ISSN 2397-3358. Disponível em: <http://www.nature.com/articles/s41570-021-00267-4>. Citado 2 vezes nas páginas 6 and 18.

TWENEY, R. D. Discovering discovery: How faraday found the first metallic colloid. *Perspectives on Science*, MIT Press - Journals, v. 14, n. 1, p. 97–121, mar. 2006. ISSN 1530-9274. Disponível em: <http://dx.doi.org/10.1162/posc.2006.14.1.97>. Citado na página 15.

YANG, Z. *et al.* When metal nanoclusters meet smart synthesis. *ACS Nano*, American Chemical Society (ACS), v. 18, n. 40, p. 27138–27166, set. 2024. ISSN 1936-086X. Disponível em: <http://dx.doi.org/10.1021/acsnano.4c09597>. Citado na página 16.

ZHANG, W. *et al.* Progress and perspective of electrocatalytic co2 reduction for renewable carbonaceous fuels and chemicals. *Advanced Science*, Wiley, v. 5, n. 1, set. 2017. ISSN 2198-3844. Disponível em: <http://dx.doi.org/10.1002/advs.201700275>. Citado 2 vezes nas páginas 6 and 20.

ZHANG, Y. *et al.* Ultrasmall au nanoclusters for biomedical and biosensing applications: A mini-review. *Talanta*, Elsevier BV, v. 200, p. 432–442, ago. 2019. ISSN 0039-9140. Disponível em: <http://dx.doi.org/10.1016/j.talanta.2019.03.068>. Citado na página 19.

ZHENG, Y. *et al.* Antibacterial metal nanoclusters. *Journal of Nanobiotechnology*, Springer Science and Business Media LLC, v. 20, n. 1, jul. 2022. ISSN 1477-3155. Disponível em: <http://dx.doi.org/10.1186/s12951-022-01538-y>. Citado na página 17.

ZIBORDI-BESSE, L. *et al.* Ethanol and water adsorption on transition-metal 13-atom clusters: A density functional theory investigation within van der waals corrections. *J. Phys. Chem. A*, American Chemical Society, v. 120, n. 24, p. 4231–4240, June 2016. ISSN 1089-5639, 1520-5215. Disponível em: <https://doi.org/10.1021/acs.jpca.6b03467>. Citado na página 58.