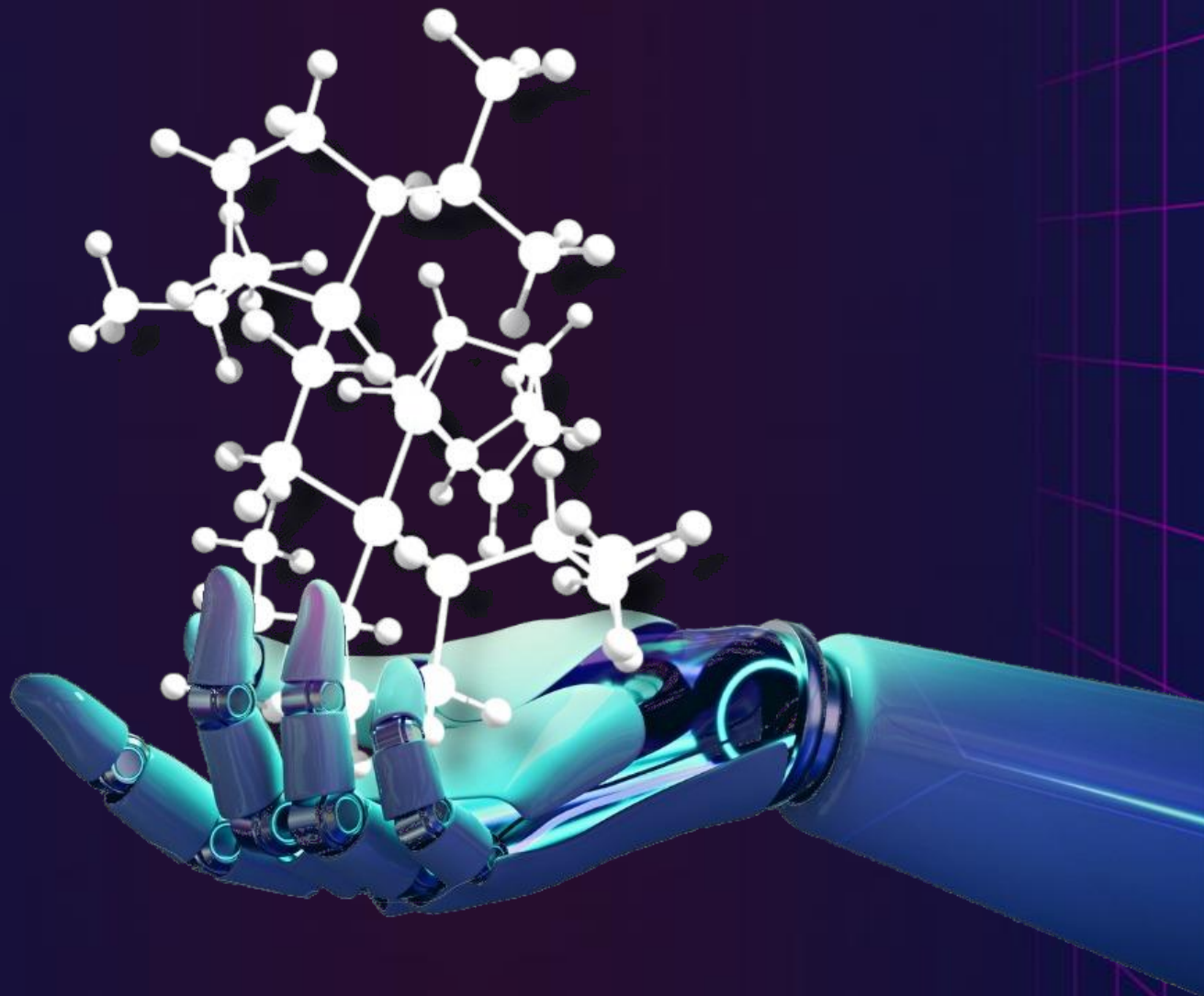
Research in high-throughput conformer search methods for homogeneous catalysis

Sára Finta



TU Delft

Research in high-throughput conformer search methods for homogeneous catalysis

by

Sára Finta

to obtain the degree of Master of Science
at the Delft University of Technology,
to be publicly defended on Thursday, July 11, 2024 at 9:00 AM.

Performed at:
Inorganic Systems Engineering
Faculty of Applied Sciences

Under supervision of:
Prof. Dr. E. A. Pidko
A.V. Kalikadien MSc.

Student number:	5857678	
Project duration:	January 8, 2024 – July 11, 2024	
Thesis committee:	Prof. Dr. E. A. Pidko,	TU Delft, AS
	Dr. A. Bansode,	TU Delft, AS
	Dr. J. M. Weber,	TU Delft, EEMCS

Abstract

Transition metal complexes as homogeneous catalysts enable high enantioselectivity in hydrogenation reactions, making them especially beneficial for the pharmaceutical industry. The development of data-driven prediction models enhances high-throughput catalyst design. However, these models often focus solely on static molecular representation, neglecting the dynamic behavior of the system, such as the formation of conformer ensembles. Currently, no method is available to systematically account for these conformational effects at reasonable costs. In light of this, the study aimed to develop a practical tool that allows predictive models to incorporate the dynamic characteristics of catalysts via conformer ensembles. A dataset of Rh-based precatalysts with mainly bidentate ligands was utilized. Three cheminformatic tools -RDKit, OpenBabel, and CREST- were explored for reliable, automated conformer ensemble generation. Among them, only CREST proved feasible, although it exhibited several limitations and required manual modification. A mapping between the conformer geometries obtained from GFN2-xTB and DFT calculations was achieved based on the relative energies and root mean square deviations. This revealed that many conformers generated by CREST converge into the same DFT local minimum. A classification method was developed to bridge the gap between conformers obtained from the two quantum chemical calculations by selecting a subset of conformers from the CREST ensemble that appear as distinct conformers in the DFT ensemble. This approach allows DFT calculations to be performed only on conformers that would result in different DFT minima on the potential energy surface, thereby eliminating redundant calculations and saving significant costs. This unsupervised DBSCAN clustering algorithm was applied to the GFN2-xTB energy and RMSD of the conformers, reducing the number of redundant conformers by 46% in the original dataset of Rh-based precatalyst structures.

Contents

List of Figures

List of Tables

List of Code Listings

Acronyms

AI	Artificial intelligence
AO	Atomic orbital
CREST	Conformer-rotamer ensemble sampling tool
DFT	Density functional theory
FF	Force field
GFN	Geometries, Frequencies, Noncovalent
GGA	Generalized gradient approximation
GTO	Gaussian type orbital
HF	Hartree-Fock
HOMO	Highest occupied molecular orbital
HTE	High-throughput experimentation
IR	Infrared
KS-DFT	Kohn-Sham density functional theory
LDA	Local density approximation
LUMO	Lowest unoccupied molecular orbital
meta-GGA	Meta generalized gradient approximation
ML	Machine learning
MO	Molecular orbital
NBD	Norbornadiene
NBO	Natural bonding orbital
NMR	Nuclear magnetic resonance

PBE	Perdew–Burke-Ernzerhof
PCA	Principal component analysis
PDE	Partial differential equation
PES	Potential energy surface
QSAR	Quantitative structure-activity relationship
QM	Quantum chemistry
RMSD	Root mean square deviation
RMSE	Root-mean-square error
SMILES	Simplified molecular-input line-entry system
STO	Slater type orbital
TM	Transition metal
UFF	Universal force field
UMAP	Uniform manifold approximation and projection
XC	Exchange correlation

1

Introduction

In modern chemical industrial processes, catalysts undoubtedly play a crucial role [1]. By definition, a catalyst is a chemical substance that accelerates the reaction rate by providing a new reaction mechanism with lower activation energy without being permanently consumed [1–3]. Catalytic steps are incorporated into approximately 90% of chemicals processes, from bulk chemicals to pharmaceutical industries, due to their numerous advantages [4–6].

Catalysts are often categorized into three main groups: homogeneous, heterogeneous and biocatalysis [3]. Homogeneous catalysts are in the same phase as the reactants, whereas the phase of heterogeneous catalysts differ from them [7]. Biocatalysis refers to the process in which chemical reactions are catalysed by cells or cell components [8]. Although heterogeneous catalysis is employed in 85% of the catalytic processes, homogeneous catalysts offer a key advantage of high selectivity, making them particularly relevant in industries such as fine chemicals or pharmaceuticals [9]. In many cases, when chirality is introduced to an atom, only one enantiomer is pharmacologically active while the other one may have undesirable effects [10]. Examples of such drugs include thalidomide or propranolol among many others [11]. The utilization of catalysts capable of selectively producing only the desired enantiomer can reduce the costs of downstream processes and lead to a higher production yield.

The use of transition metal (TM) complexes for homogeneous catalysis has recently gained traction [1]. Their use is widespread in the pharmaceutical industry to produce active pharmaceutical ingredients in processes such as asymmetrical hydrogenation [12]. Rhodium-based catalysts can be used to asymmetrically hydrogenate imines to obtain amines with specific chirality. In this manner, high enantiomer selectivity can be achieved [13, 14]. Chiral amines are key components in the pharmaceutical industry as 40% of pharmaceutically active compounds have a chiral amine in the structure [15]. An example of a reaction scheme, where chirality is introduced to an amine by a transition metal-catalysed asymmetric hydrogenation can be seen in Figure 1.1.

The primary advantage of TM complexes lies in the tunable nature of their ligand properties making them suitable control the catalyst performance [17]. Given the great

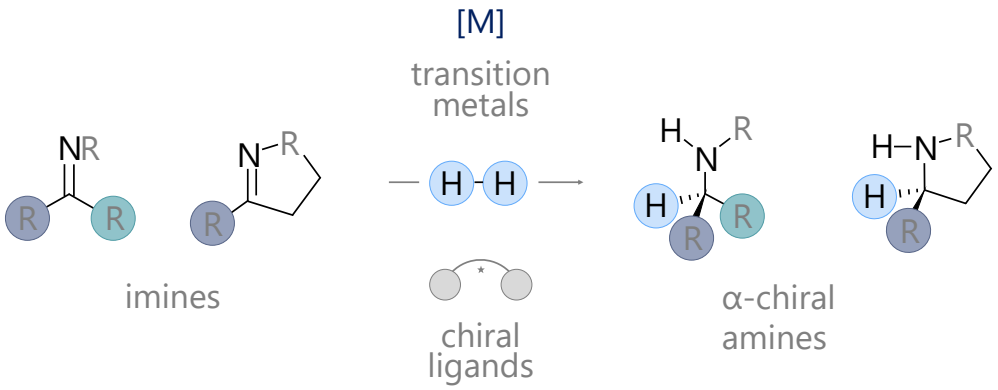


Figure 1.1: Asymmetric hydrogenation of imines using transition metal based catalysts. Image recreated from [16].

amount of available ligands with the absence of straightforward ligand-performance correlation, it is necessary to develop techniques for high-throughput catalyst screening [17]. As a consequence, TM catalysts have become a major research area within novel catalyst design, discovery and optimization [18, 19]. The appearance of high-throughput experimentation (HTE) has allowed for extended experimental catalyst screening [20]. However, as it remains based on trial and error methodology, it is costly, time and resource consuming [6]. To accelerate catalyst design and development several computer based tools emerged to complement experimental methods. Data-driven statistical methods like Machine Learning (ML) aim to find correlations between the catalyst properties and performance such as catalyst activity or selectivity [6, 21]. The predictions obtained by these models enable the reduction in the number of experiments needed to identify a catalyst with the desired properties [21]. An example of an automated approach towards catalyst design is shown in Figure 1.2. The starting point of these models is a database containing

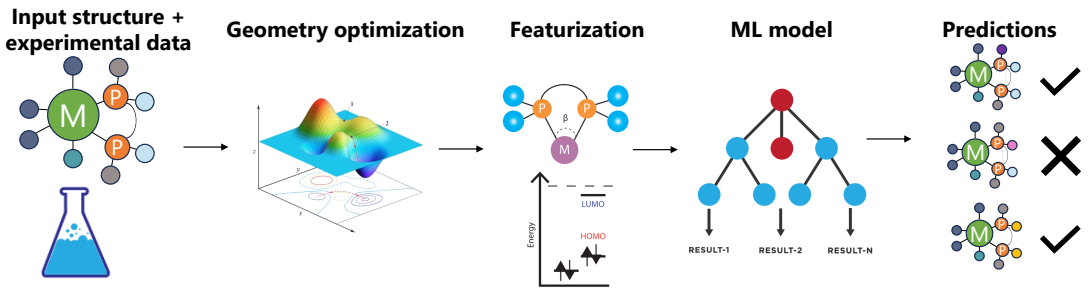


Figure 1.2: General approach to automated catalyst design: The workflow begins with the digital representation of catalyst structures alongside their experimental data. The next step involves optimizing the geometry of these structures, followed by a featurization process. Finally, a machine learning model is trained to predict the experimental data for new structures. Image created using [22].

an initial representation of the input catalyst structures as well as the corresponding experimental data on their performance indicators. These structures are subjected to geometry optimization, followed by the quantification of catalytic properties using descriptors. These input features are then employed by various ML algorithms to capture correlations between catalyst properties and performance [6, 21]. The obtained model can be further utilized to make accurate catalyst performance predictions based on the molecular descriptors.

An appropriate input structure representation is vital to the success of the model, as it impacts every subsequent step [23]. When stereochemistry is introduced to a molecule, conformational dynamics play a crucial role in characterizing the behaviour of the molecular system [24]. As shown in Figure 1.3, a simple 2D representation cannot accurately capture the geometric and electronic differences between conformers. Therefore, descrip-

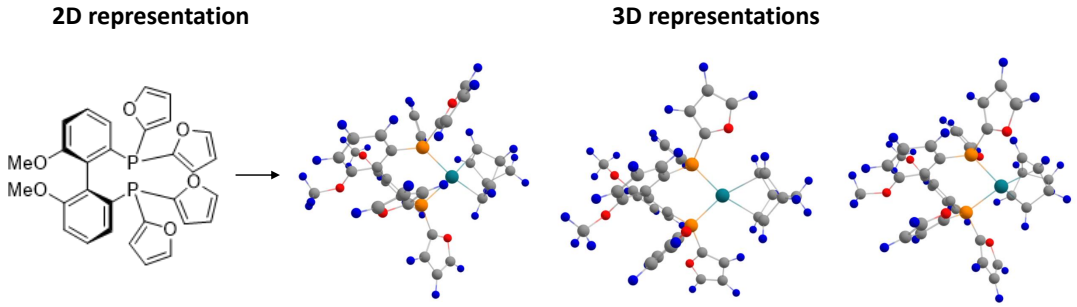


Figure 1.3: Various 3D geometries (right) from a 2D drawing (left) of a molecular structure.

tors that exclusively featurize a static molecular representation may prove inadequate in describing the underlying chemistry of the system [23, 24]. An additional challenge lies in the limitations of analytical methods to observe the dynamic behavior of catalyst structures in situ during reactions [25]. Classical analytical measurements, such as nuclear magnetic resonance (NMR) or infrared (IR) spectroscopy are hindered by the low concentration and short time-frame in which the substrate-catalyst complex exists [1]. Due to the lack of knowledge about the actual conformational behavior during reactions, current computational approaches for catalyst design either consider only the most stable conformer or perform a full computational study at exceptionally high costs [26]. Currently there is no developed method that systematically considers conformational ensembles in a high-throughput manner at reasonable costs.

In the light of current challenges in the field, the research presented in this thesis is aimed at establishing a dynamic representation of catalyst structures via conformer ensembles and benchmarked methods to generate them in a high-throughput, automated manner. During this study, the following questions were investigated:

- Does the use of various conformer searching engines based on distinct methodologies result in varied conformer properties? Which conformer searching tools are

applicable to our TM dataset?

- How does further geometry optimization influence the properties of the ensembles obtained from conformer searching?
- Is it possible to select a subset of conformers that can accurately represent the DFT refined ensemble to reduce the amount of DFT calculations required?

The structure of this thesis is as follows: First, the applicable theoretical background is provided, followed by a description of the computational methods utilized. The main part of the thesis discusses the obtained results and observations. Finally, a section of conclusions and further recommendations is provided.

2

Theoretical background

This chapter aims to provide an overview of the theory behind this research. First, an overview of TM complexes is presented, followed by a discussion of the four pillars of state-of-the-art catalyst design.

2.1 Transition metal complexes

Transition metals are well-suited for utilization as homogeneous catalysts due to their unique properties arising from their partially filled d- (or f-) subshell [27, 28]. They can be present in various oxidation states [28], and can form both σ and π bonds due to their valence atomic orbital (AOs) that can construct hybrid molecular orbitals (MOs) with various kinds of molecules [27]. The selection of the metal center is crucial for successful catalyst design [29]; typically used metals include cobalt, rhodium, platinum, or ruthenium among many others [30]. Ligands (atoms or molecular fragments [31]) typically bond to the metal centre in such way that the coordination number of the metal centre is either 4 or 6 [28]. These ligands can modify the electronic and geometric properties of the catalyst environment and therefore the active site of the catalyst [27]. Hence the choice of the ligand can be used to tune the performance of the catalyst [27]. The most commonly encountered ligands are the phosphorus ligands [32], but ligands containing nitrogen or oxygen atoms have also gained significant attention in various applications [33–35]. Phosphorus atoms can be characterized on the Lewis acid-base scale as soft, ligating atoms that have been demonstrated to effectively enhance the efficiency of the designed catalysts for various reactions [29]. It is often thermodynamically more advantageous for ligands to have more than one atom connected to the metal centre, forming polydentate complexes [28]. Phosphorus ligands are commercially available and utilized as mono- bi- and polydentate ligands [29]. This research mainly focused on biphosphine bidentate ligands, but the catalyst library also included TM complexes with monophosphines, aminophosphines, and phosphoramidites [36].

2.2 Computer-aided catalyst design

The appearance of computer algorithms revolutionized homogeneous catalyst design enabling a more efficient and affordable exploration of the chemical space [6]. Several computational tools have emerged aiming to provide accurate molecular representations and find statistical correlations between molecular properties and catalytic performance [18, 37, 38]. An ML-aided catalyst design workflow including experimental chemistry, quantum chemical methods and ML-based modeling is presented in Figure 2.1 [39]. Current

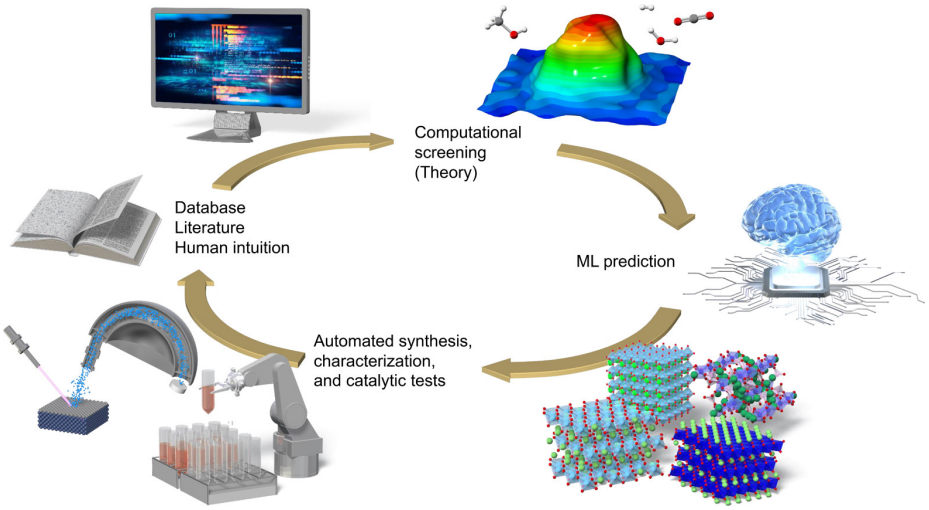


Figure 2.1: Catalyst design using ML tools: Initially, the data acquired through experiments is used to establish a relationship between catalyst properties and performance. The ML model then predicts the performance of new catalyst structures, reducing the number of experiments required. Image taken from [39].

state-of-the-art computer aided catalyst design workflows rely on 4 scientific fields: experimental chemistry, quantum chemical (QC) methods, cheminformatics and statistical (artificial intelligence) models [40].

2.2.1 Chemical databases

While one of the main objectives of future design processes is the independency from experimental data [41], it is not yet achievable [42]. The success of predictive data-driven models lies within the size and accuracy of experimental and computational databases [19]. One of the biggest challenge computer aided catalyst design still faces is the lack of large, robust, representative datasets [19, 43]. Datasets containing quantified properties of catalyst structures and experimental parameters as well as indications of the experimentally obtained catalyst performance can be used for models to identify correlations and make predictions [6, 40]. The desired target performance indicators that are experimentally measured are usually the activity, conversion or enantioselectivity [6], while various molecular descriptors are used as input features.

Descriptors

For accurate data-driven algorithms, applying a computer-readable representation of molecules is essential [44]. Due to the complex nature of computational tasks and circumstances, there is no absolute way of representation [44]. Most cheminformatic programs are actively developed to include and convert between representations with different level of dimensionality [19]. The most common representation in chemistry is using string-like chemical SMILES notation [45]. Due to the complexity of TM complexes, 1D or 2D representations like chemical SMILES fail to reliably represent TM complexes [46]. When modeling a delicate quantity such as enantioselectivity, the model can become highly dependent on structural changes, making at least 3D representation necessary [40, 47, 48]. In the featurization of catalytic structures via descriptors, the chemical information of molecular geometries and the electronic structure of molecules is converted into mathematical (numerical) representation [49, 50]. Descriptors are utilized by quantitative structure-activity relationships (QSAR) models to investigate potential relationships between molecular descriptors and catalyst activity [18, 21]. A comprehensive list of used descriptors can be found in Appendix B.

In this study, the descriptors representing the 3D structure of the molecule are divided into three categories: electronic, steric, and geometric descriptors. Electronic descriptors

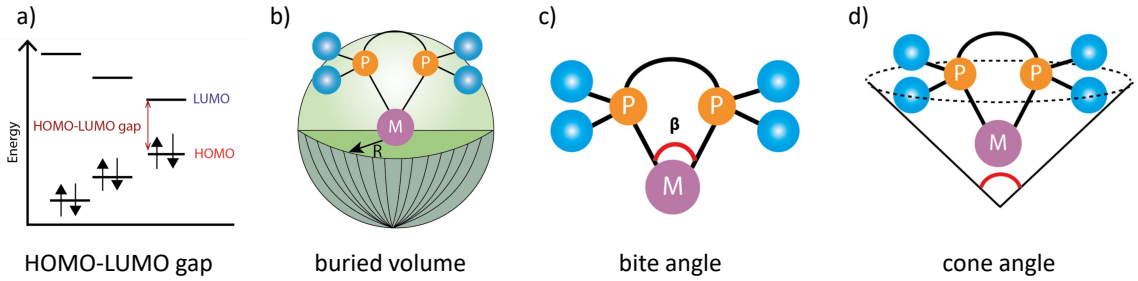


Figure 2.2: Example of key descriptors, (a) HOMO-LUMO gap, (b) buried volume, (c) bite angle, and (d) cone angle.

are designed to capture the main electronic structure of the molecule including the electron densities and the local charge distribution [51, 52]. Some example of electronic descriptors include the natural bond orbitals (NBOs), dispersion, nucleophilicity and the HOMO-LUMO gap. The HOMO-LUMO gap, significantly impacting the reactivity of the molecule [53, 54] is the energy gap between the highest unoccupied and the lowest occupied molecular orbital (illustrated in Figure 2.2 (a)). The second category of the descriptors are steric descriptors, which are used to characterize the steric effects that stems from each atom occupying a specific amount of space [23]. These effects are non-bonding interactions from the overlap of electron clouds [23]. A fundamental steric descriptor for predictive models is the buried volume [40]. The buried volume quantifies the percentage of a sphere with a given radius is taken up by the ligand around a specific atom as a centre [55]. An illustration of the buried volume with the metal as the centre of the sphere is shown in Figure

2.2 (b). The third category is the geometric descriptors, mostly consisting of bond lengths and angles [21]. Two key descriptors to mention are the bite angle and the cone angle, presented in Figure 2.2 (c) and (d) respectively. The bite angle, that has been proven to have be a significant feature of prediction models, describes the angle between the metal centre and two connecting ligand atoms [56]. The cone angle is utilized for symmetrical bidentate ligands to express the angle of a cylindrical cone that takes the metal as centre and the outmost atoms' Van der Waals radii as edges [57, 58]. It is therefore a suitable parameter to indicate the size of the ligand.

2.2.2 Density functional theory

In computational chemistry, density functional theory (DFT) is widely used for various fields of applications [59, 60]. Concerning DFT, numerous methods and techniques have been developed and effectively applied leading to an efficient, robust and widespread approach [60]. In the field of homogeneous catalysis, DFT is commonly utilized for numerous purposes including predicting reactivity [61], determining energy barriers and therefore reaction rate constants [62] and comprehending underlying reaction mechanisms [63]. DFT can serve as a valuable tool for geometry optimization [64] as well as determining the energy of fixed molecular structures [65].

Despite other quantum chemical methods, the primary focus of DFT is on the electron density rather than the total energy of the system [66]. To describe and understand DFT, we must begin with the fundamental equation of quantum chemistry: the time-independent, non-relativistic Schrödinger equation, which can be expressed by the following form [67]:

$$\hat{H}\psi = \hat{E}\psi \tag{2.1}$$

In this equation, $\hat{H}$ denotes the Hamiltonian operator and ψ stands for the wavefunction. The Hamiltonian operator represents the total energy of the system, thus in order to solve the Schrödinger equation, it has to be expressed. Its form can vary depending on the system. In computational chemistry, where multiple nuclei and electron systems are present, the expanded form is as follows for M nuclei and N electrons [67, 68]:

$$\hat{H} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \frac{1}{2}\sum_{A=1}^{M}\frac{1}{M_A}\nabla_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}} \tag{2.2}$$

The equation represents the kinetic energy of electron, kinetic energy of nuclei, attractive Coulomb interaction between nuclei and electrons, repulsion between electrons and repulsion between nuclei respectively [68].

The exact analytical solution of the Schrödinger equation can only be obtained for one-electron systems [69]. For more complex system with multiple electrons and nuclei, approximations are required. The main challenge that increase the complexity of the Schrödinger equation arises from the representation of both electrons and nuclei within the same wavefunctions. The Born-Oppenheimer approximation is based on the princi-

ple that the electron is roughly 1800 times lighter than the nuclei, allowing electrons and nuclei to be treated almost completely separately. This enables them to move around and react more independently, leading to disregard the kinetic energy of the nulcei and approximating the repulsion of nuclei as a constant. Thus the Hamiltonian can be simplified the following way [67]:

$$\hat{\mathrm{H}} = -\frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{N}}\nabla_{\mathrm{i}}^{2} - \sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{A}=1}^{\mathrm{M}}\frac{\mathrm{Z_A}}{\mathrm{r_{iA}}} + \sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}>\mathrm{i}}^{\mathrm{N}}\frac{1}{\mathrm{r_{ij}}} \tag{2.3}$$

To construct the wavefunctions, multiple approximations have been developed and applied, such as the Hartree-Fock (HF) method.

As the primary focus of DFT shifts from wavefunctions to the electron density $\rho(r)$, the dimensionality, and therefore computational costs, are drastically reduced [70]. Electron density allows to only rely on three space coordinates instead of the previous 3N variables for an N electron system [70]. Using the electron density to describe the total energy of the system is based on the two theorems of Hohenberg and Kohn. The first theorem states that the Hamiltonian operator of the system is uniqually determined by the ground state electron density [67]. Relying on the electron density is therefore sufficient to acquire all the system properties [67]. The second theorems states that the lowest (ground state) electronic energy of the system is a functional of the ground state electronic density [67]. Combining these theorems and the Born– Oppenheimer approximation, the energy of the system as a functional of density can be expressed in the following form (Kohn-Sham equation [71]):

$$E[\rho(r)] = V_{NN} + V_{eN}[\rho(r)] + E_J[\rho(r)] + E_T[\rho(r)] + E_Q[\rho(r)] \tag{2.4}$$

Where V_{NN} and V_{eN} are the nuclei-nuclei and nuclei-electron attraction energies, E_J stands for the classical repulsion between 2 electrons and E_T is the functional of the kinetic energy of electrons [72]. The last term (E_Q) represents the quantum (non-classical) interaction of electrons [72]. The two latter terms are unknown to the system, making the main objective of DFT development the determination of appropriate approximations for them [72]. By selecting suitable forms for the unknown functional, based on the two theorems of Hohenberg and Kohn, the ground state energy and therefore all system properties can be approached by the minimization of E_ρ [71].

Exchange-Correlation Functionals

The two unknown terms from the previously presented Kohn-Sham equation can be combined to define an exchange-correlation $E_{XC}[\rho(r)]$ term [72]. This term is a functional of the electron density and represents the sum of electron-electron interaction and kinetic energy corrections [70]. Since finding an accurate, efficient, and robust method to express $E_XC[\rho(r)]$ is critical for successful DFT calculations [60], several approximations emerged, both non-empirical and semi-empirical [72]. Different dependencies are being taken into account by these functional leading to significant differences between their accuracy and

computational expenses. The so-called "Jacobs's ladder" (Figure 2.3) can be used to collect and categorize these functionals based on the above-mentioned criteria [73]. The lad-

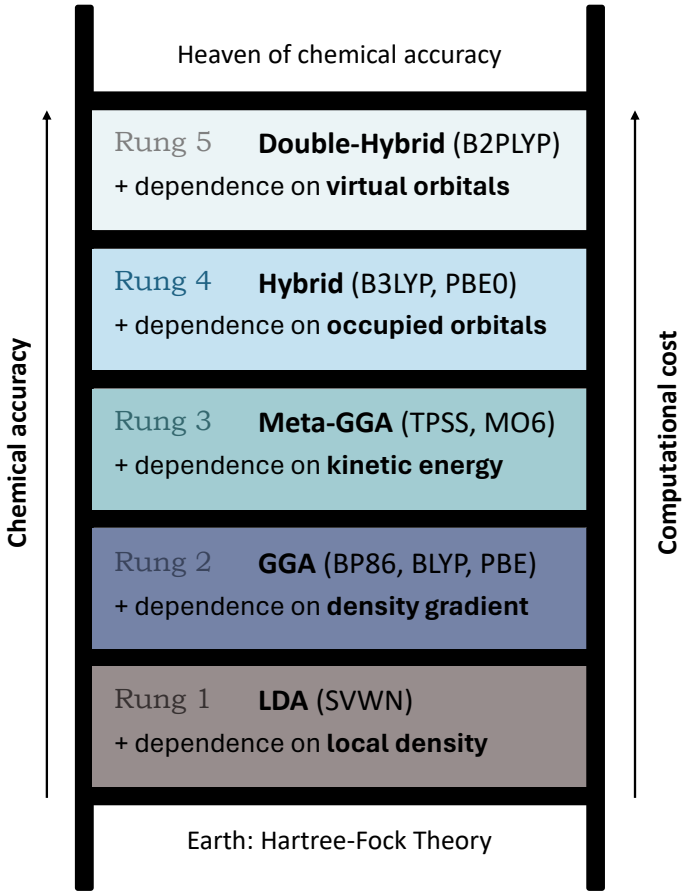


Figure 2.3: The Jacob's ladder of exchange-correlation functionals, starting from the Hartree-Fock theory. Climbing the ladder involves increasing dependencies that enhance both chemical accuracy and computational cost at each step.

der starts with the Hartree approximation ("Earth") assuming zero exchange correlation energy ($E_{XC}[\rho(r)] = 0$) and ends with the highest form ("heaven") of chemical accuracy [73]. The exchange correlation functionals in between are divided to five rungs, each introducing additional corrections to the energy density [73], raising both the chemical accuracy and required computational power. The lowest three rungs are the (semi-) local rungs starting from "local density approximation" (LDA), "generalized gradient approximation" (GGA) and "meta generalized gradient approximation" (meta-GGA) [73, 74]. A significant difference between them and the higher level rungs is that they do not take the Fock exchange into account [60]. They include a self-interaction error (SIE), that is partially replaced by the Fock exchange in the higher level of functionals (rung 4 and 5: hybrid and double-hybrid functionals) [60]. The ratio between the DFT-exchange and the Fock exchange differentiates within the various types of hybrid potentials [60]. The most commonly used exchange correlation functionals for transition metal complexes are the

B3LYP and PBE0 hybrid functionals [75, 76]. The PBE0 hybrid functional that was used for the current study utilizes the Fock and the PBE (GGA) exchange energy in a 1:3 ratio [60, 77].

Basis sets

To approximate the Kohn-Sham equation (Eq. 2.4), Kohn and Sham introduced the use of one-electron orbitals [70, 78]. The wavefunction is parameterized into unknown molecular orbitals that can be determined by the linear combination of known atomic orbitals (LCAO theory) [78]. Basis sets are used to construct these molecular orbitals in such way that computationally solvable algebraic equations are generated [78]:

$$\phi = \sum_{\alpha=1}^{M} c_\alpha \chi_\alpha \tag{2.5}$$

In this equation ϕ stands for a molecular orbital, c_α is the basis function (from 1 to M) and χ_α is the expansion coefficient [78]. Similar to the exchange correlation functionals, hundreds of basis sets are available and applied aiming to balance chemical accuracy and computational power [79]. Although using Slater type orbitals (STOs) [80] to construct atomic orbitals would have been indicated from the solution of the Schrödinger equation to the hydrogen atom, due to its limitations and the high computational effort required, Gaussian type orbitals (GAOs) are most commonly used [78]. While choosing an appropriate basis set, two significant errors may occur: the basis set incompleteness (BSIE) [81] and superposition (BSSE) [82] errors [60]. The first error refers to the size of the basis set, basis sets can be characterized based on the independent functions to construct a valence atomic orbital [60]. Therefore we distinguish valence double, triple or quadruple zeta basis sets [60]. The second error term stems from the non-covalent intermolecular interactions and can be eliminated using correction terms [60]. A polarization function can be added to basis sets for further performance improvement as the primary basis set may not be adequate to describe electron density distribution accurately [83]. For this study the Gaussian type double zeta def2-SVPP basis set was utilized due to its general applicability and accurate performance in TM complexes [84].

Dispersion corrections

The standard Kohn–Sham DFT theory does not take the London dispersion forces into account [85]. However, when dealing with large molecules, such as TM complexes, these forces have a significant impact on the total energy of the system and therefore a correction term should be subtracted from the KS DFT energy [85]. The dispersion interaction energy depends on the molecular distance by the following rate [86]:

$$E_{dispersion} \propto -\frac{C_6}{R^6} \tag{2.6}$$

Where C_6 stands for the dispersion coefficient and R the atomic distance [86]. In this study the DFT-D3 dispersion correction [87] was used as it is providing an accurate representation of the dispersion energy for almost all elements of the periodic table including transition metals [86].

Potential energy surface

DFT is a widely applied tool to perform molecular geometry optimization [88, 89]. After providing and initial geometry, by minimizing the potential energy gradient, the molecular geometries belong to the local minima on the potential energy surface (PES) can be obtained. To validate whether the found molecular structure is indeed a local minimum, the Hessian matrix can be analyzed containing the second partial derivative of the potential energy. In case there is an imaginary frequency found, the structure is not the actual local minima and it has to be recalculated with a slightly modified initial geometry provided. As performing these calculations is rather expensive [90], easily reachable minima should be ensured by providing the sufficient initial molecular geometries.

2.2.3 (Semi-) empirical computational methods

While DFT calculations are often considered the most accurate computational method affordable for the design of homogeneous catalysts, they demand significant computational resources, making them very expensive on larger scales [21]. Thus, relying solely on DFT calculations is not feasible for high-throughput screening of large molecules [21]. To address this, several methods have been developed to strike a balance between chemical accuracy and computational efficiency.

Force field methods

Force field methods (FF) are computationally significantly less demanding than DFT as they rely on simple energy calculations to predict the structural and thermodynamic parameters of molecules [91]. The main advantages of FF methods lies in the fact that as atoms (nuclei and electrons) and bonds are treated as fixed balls and strings, instead of dealing with the Schrödinger equation, the Newtonian mechanics can be utilized [81]. Force field methods are reported to require the lowest CPU-time out of the commonly used methods for TM complexes [92, 93]. A major challenge of the development of FF methods are generating parameters based on the elements and bonding information [19]. Many FF methods do not provide parameters for a broad range of elements and cannot be used for transition metal complexes [40]. For the current study, the universal force field (UFF) was applied as it is parameterised for all elements of the periodic table [94] and widely used for TM complexes [19].

GFNn-xTB methods

Another alternative tool for molecular modeling involves using quantum chemistry based semiempirical tight-binding models, which offer relatively accurate results at reduced computational costs [95]. The Geometry, Frequency, Noncovalent, eXtended Tight-Binding (GFN-xTB) method was developed aiming to provide accurate results and universal applicability (atomic number = 1-86), including metals [96]. Despite many successful applications, the versions of GFN-xTB preceding the GFN2-xTB method struggled to describe systems that are highly polar and involve strong hydrogen bonds [97]. To tackle this challenge, the GFN2-xTB was developed and designed to show accurate results for organic, organometallic and biomolecules [97]. This low-cost quantum chemical method has shown robust and accurate results from chemical space exploration and conformer searching [98].

2.2.4 Cheminformatics

To efficiently use data driven predictive models, descriptors should be calculated in a high-throughput, automated manner [40]. Cheminformatics emerged as a significant field of informatics and made impact not only on catalyst discovery, but also drug design, material science or computational chemistry [99]. Cheminformatics intends to provide useful tools for scientists to store, analyse, manipulate and manage the vast amount of chemical data in an automated as systematic manner [40, 100]. Several computational packages are being developed and optimized for various purposes: for instance molecular structure and reaction representation, storage and use of chemical databases, descriptor calculations, similar molecular characteristics search and identification, data visualisation or generation of new structures [100, 101]. In the domain of homogeneous catalysis, main cheminformatics tools include functions for molecular featurization, databases, conformer searching algorithms and they often include ML techniques as well [102]. Most cheminformatic workflows are integrated in Python and use the open-source RDKit [103] and OpenBabel [99] packages as a backend [40]. These toolkits are primarily designed for handling organic molecules, accepting various input formats, and including simple FF optimization tools [19].

Conformer searching

The quantification of catalyst properties is a major challenge in the field, as accurate representation is linked to the success of the models [24, 104]. A main focus point of cheminformatics is to provide a tool for accurate representation of molecules [100]. The highest level of representation (4D), instead of focusing on a static molecular geometry, includes the whole conformer ensemble [40]. Several conformational parameters are essential to describe the structural characteristics of the molecules such as conformational energies or steric effects [23, 105]. In data-driven models for homogeneous catalyst design, there are generally two approaches for including conformers. The first approach involves taking only the most reasonable geometries into account. This selection is based on chemical intuition introducing human bias, often leading to inaccurate representations [106] or ignoring

all conformational effects [26]. The second approach is a broad exploration of conformational outcomes significantly raising computational costs [26]. Many conformer searching tools are developed and utilized with diverse optimization goals, ranging from searching for local minima to finding bioactive conformers [107]. The three main underlying geometry optimization methodologies of these conformer searching tools are empirical (FF), semiempirical, and ML potential based [107]. ML based conformer searching tools, such as auto3D use neural network potentials that are not yet parameterized for TM complexes [108]. Both OpenBabel and RDKit can be utilized for conformer searching using different FF methods such as UFF or MMFF [103, 109]. Generally force field methods are characterized as fast speed and low computational power tools with questionable accuracy and reliability [108]. Semiempirical methods on the other hand are reported to be more accurate but requires more computational time [107, 108]. A widely applied semiempirical tool, CREST (Conformer–rotamer sampling tool) is using GNFn-xTB tight-binding methods and therefore aims to find balance between high accuracy and low computational costs [110]. Finding the appropriate conformer searching tool for TM complexes is still an active area of research, as there is no definitive conclusion on the best performing method [93, 107].

2.2.5 Machine Learning in homogeneous catalysis

Artificial intelligence (AI) is one of the most rapidly developing domain of computer science impacting numerous fields of chemistry including homogeneous catalyst design [19]. Several ML models are trained to provide accurate predictions of catalyst performance or feature selection [6, 21] from white-box to black-box models [40]. An overview of general ML techniques and their classification is represented in Figure 2.4 [40]. White box models

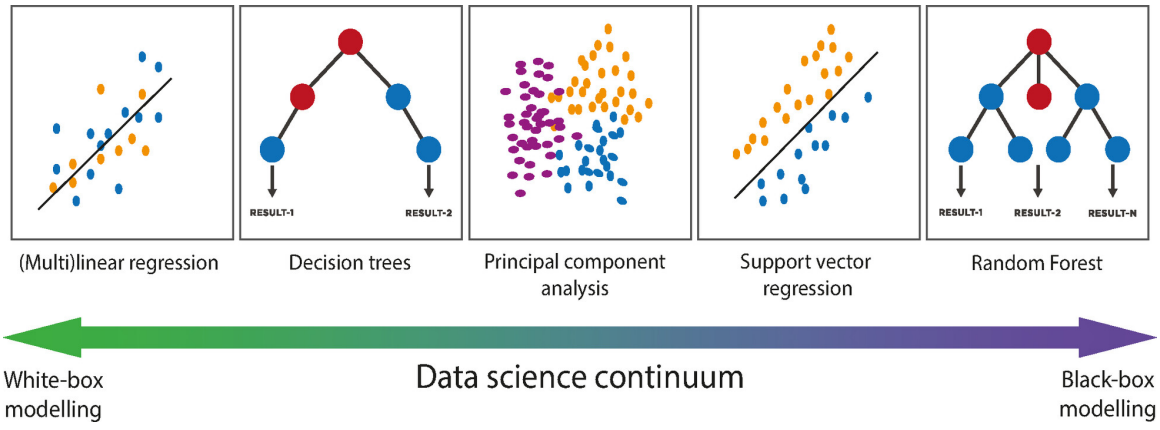


Figure 2.4: Commonly used ML techniques of the data science continuum ranging from white box to black box models. Image taken from [40].

are based on traditional statistical approaches identifying transparent correlations between features and output labels with a clear model architecture [40, 111]. Black box models on the other hand consist of multiple internal hidden non-linear functionalities, making

them too complex for straightforward interpretation or analysis [112]. Therefore, while white box models can be used to capture catalyst descriptor-performance relationships, black box models provide higher accuracy in predictions [40]. Example of typical white box models includes simple regressions models while the most commonly used black box models are random forests or artificial neural networks [112]. Both white box and black box algorithms can be classified into the three main groups of ML models: supervised, unsupervised, and reinforcement learning [113]. The main characteristics and application potentials of ML methods that are relevant in homogeneous catalysis are presented below.

Supervised learning

Supervised ML models are trained to provide accurate predictions of continuous (regression) or discrete (classification) output labels from certain predictors [113]. To build a supervised model, the dataset has to contain both input features (descriptors, reaction parameters) and output values (activity, selectivity) [102]. Many of these techniques have been successfully implemented to homogeneous catalyst design. Derek et al. investigated the application of decision tree and linear regression models for reaction yield prediction [114]. The ML model was trained using palladium catalyst descriptors to predict the performance in Buchwald-Hartwig (C-N cross-coupling) amination of 4-methylaniline with aryl halides [102, 114]. The random forest algorithm was reported to provide more accurate performance with the root mean square error (RMSE) of 0.78 and the R^2 of 0.92 [114]. Enantioselectivity ($\Delta\Delta G^{\ddagger}$) prediction by support vector regression was investigated by Zahrt et al [115]. The ML model was applied on the formation reaction of selective N,S-acetals using phosphoric acid catalysts and similarly led to highly accurate results [115]. Logistic regression models classifying catalysts into active and inactive categories based on conversion rate also showed a high prediction accuracy of 0.95 [17]. In this case, the model system included bisphosphine ligand-catalyzed hydroformylation reactions [17].

However, most models present many drawbacks and limitations such as their transferability, making them yet unsuitable for general application purposes [102, 116].

Unsupervised learning

Unsupervised ML algorithms aim to find general patterns and segments of a dataset without explicitly providing output labels. Although most research in TM complexes is related to supervised learning, different clustering and dimentionality reduction approaches are often included in catalyst design workflows [102]. Feature dimentionality reduction tools are developed due to the complexity of high-dimensional datasets aiming to find a hidden more simple structure and translate it to a low-dimensional space [117]. The most common dimentionality reduction tools applied on a set of chemical descriptors are principal component analysis (PCA) [17] and Uniform Manifold Approximation and Projection (UMAP) [118, 119]. Clustering is an efficient way to classify the data into subgroups based on certain similarities [120]. The most used clustering algorithm on descriptors for catalyst

design is K-means clustering [118, 121]. Clustering based on the Root Mean Square Deviation (RMSD) and the energy values is also commonly used in the field of bioinformatics and protein discovery [122, 123].

Deep learning

Deep neural networks can be either supervised or unsupervised, aiming to mimic the behavior of biological neurons through several connected layers [124, 125]. Although the application of deep learning techniques in chemistry is still limited, these methods are gaining increasing attention in the field of life sciences [126, 127]. Transformer models [128] provide a useful tool for various applications in chem- and bioinformatics, such as molecular property prediction, generation, and optimization [127, 129]. These natural language processing models can embed a sequence of input data, such as chemical SMILES [45] or SELFIES [130] into latent space resulting into a sequential output [127, 129]. One of the major success of transformer models is linked to protein structure prediction with an AlphaFold model [131, 132]. In the domain of cheminformatics, the principal applications are connected to organic chemistry. A chemical yield reaction predictive model was developed by Schwaller et al. using a BERT encoder transformer model [133] followed by a regression step [134]. For atom-mapping -a mapping of the rearrangement of reactant atoms to products during a chemical reaction [135]- an ALBERT model [136] was implemented [137]. Both application examples used chemical SMILES as input data representation.

3

Computational methods

This chapter provides a comprehensive overview outlining the selected setting and configuration parameters for each computational method used. Section 3.1 describes the overall workflow followed, while the subsequent sections dive into the specifics of each step in the workflow.

3.1 Workflow

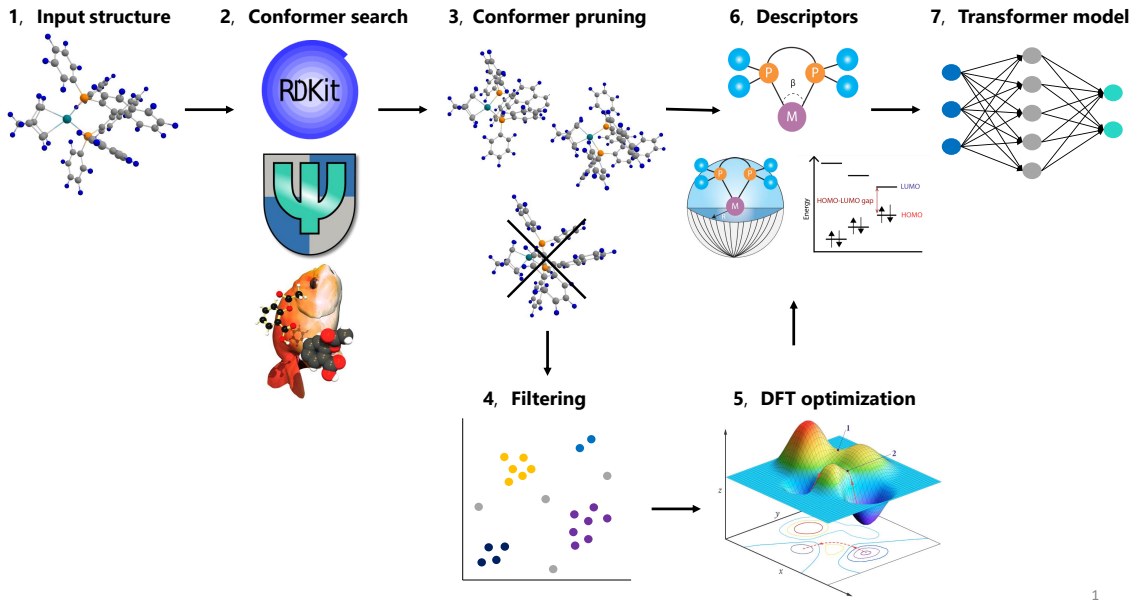


Figure 3.1: An overview of the general workflow of the research: The workflow begins with a library of input catalyst structures. The next steps involve conformer searching and pruning. A filtering approach then selects a subset of conformers to undergo DFT optimization. Descriptors are calculated from both sets of conformers. Finally, a transformer model is trained. Figure created using figures from [22, 99, 138, 139]

Figure 3.1 represents an overview of the workflow, illustrating the development of a

Python script for searching and processing conformers. The input catalyst structures (step 1) were used to perform conformer searching via various conformer search engines (step 2): CREST [98], RDKit [139] and OpenBabel [99]. In step 3, the obtained conformers were pruned to eliminate molecules from the ensembles, which chirality changed during conformer search. After selecting the best performing conformer searching method, a subset of conformers was chosen (step 4) to undergo further geometry optimization through DFT calculations (step 5). Note that during the development of the filtering algorithm, steps 4 and 5 were swapped: a subset of ensembles was selected for DFT optimization to serve as input data for the filtering approach. In step 6, the conformer ensembles obtained were subjected to descriptor calculation via the *OBeLiX* [40] computational workflow. Statistical tools such as ANOVA [140] were used to analyze and evaluate the data. In the final step, it was intended to use the obtained descriptor data to train a transformer model for accurate catalyst performance predictions and conformer selection. However, this could not be accomplished within the limited time-frame of the research.

Certain parts of the calculations required the power of supercomputers. CREST calculations were performed using the DelftBlue supercomputer [141], while the Snellius supercomputer [142] was utilized for DFT geometry optimization and single-point calculations.

3.2 Input structures

192 catalyst structures were used as the starting point of the project, each containing a rhodium ion at its centre. The oxidation state of the Rh ion was +1 in all cases. A norbornadiene (NBD) model substrate was bonded to the metal-centre to ensure a catalyst geometry capable of accommodating the binding of the actual substrate. Catalyst structures with various commercially available and tested ligand families were investigated. A comprehensive description of the catalyst structures is available in Appendix A.

The catalyst structures were initially represented using Cartesian coordinates in *.xyz file format. To determine the most accurate representation for the conformer searching engines under investigation, different input formats were also analyzed. A MDL MOL v.2000 (*.mol) file was generated for each input structure using the ChemCraft 1.8 software package [143]. RDKit mol objects and OpenBabel mol objects were created using the RDKit and OpenBabel python packages.

3.3 Conformer searching and pruning

The conformer generation and evaluation process can be seen in Figure 3.2. Various software packages were utilized to facilitate conformer generation, and the ensembles were further pruned using the Morfeus package [144].

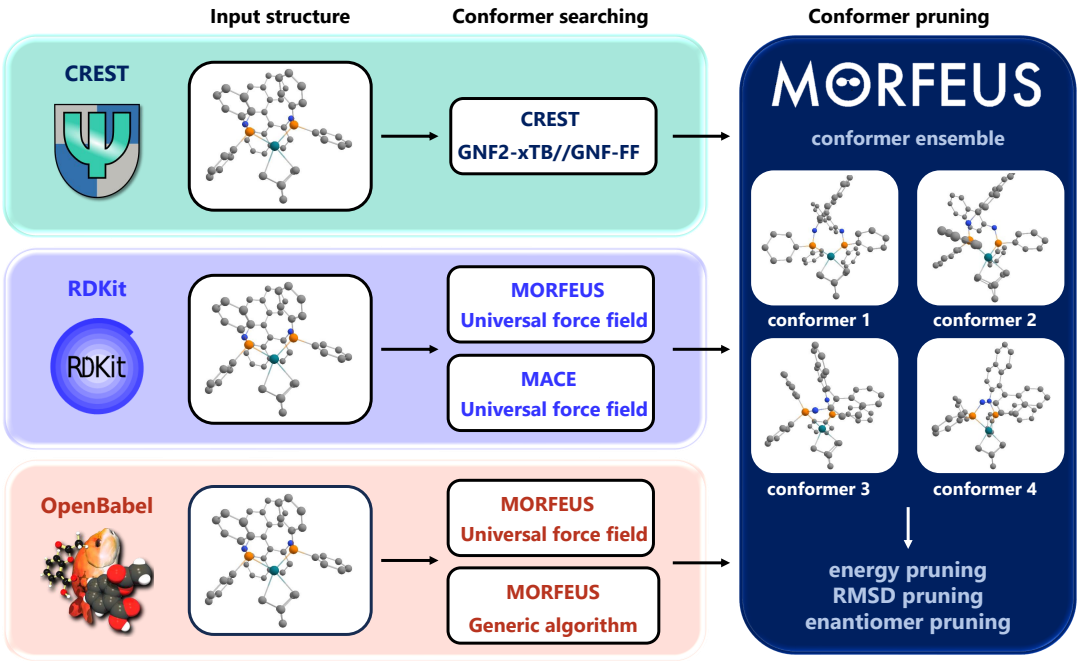


Figure 3.2: Conformer searching workflow: All conformer searching approaches start with an input structure (left). In the middle, the various packages and methods are presented. Finally, the conformers are all collected in Morfeus for further pruning (right).

3.3.1 Conformer searching

CREST

The conformer-rotamer ensemble sampling tool (CREST) software [110, 145] version 2.12, and xTB version 6.4.0 were used for conformer generation and exploration. CREST calculation were performed on all 192 Rh-based structures using Cartesian coordinates (*.xyz file) as input geometries for conformer ensemble creation. The GFN2-xTB//GFN-FF hybrid potential was chosen for its accurate performance at reasonable computational costs and universal applicability [146]. This approach utilizes the GFN-FF method for generating and optimizing conformational geometries, followed by a GNF2-xTB single point calculation on the obtained conformers [138]. Due to the cationic nature of the Rh ion and the lowest energy-state of the molecules, the charge and multiplicity were both set to 1.

The resulting CREST folder, containing the Cartesian coordinates of the conformers and their corresponding energy values, was loaded into the Morfeus Python package for further pruning and preprocessing.

RDKit and OpenBabel

The RDKit and OpenBabel conformer searching engines are by default integrated in the Morfeus python package. Conformer ensembles can be generated from the correct input format, such as chemical SMILES, RDKit, or OpenBabel mol objects. The UFF force field

was chosen for its general applicability and broad coverage of chemical elements. For OpenBabel, conformer searching using its genetic algorithm was also conducted.

The MACE python package [147] was also utilized for conformational exploration via RDKit. MACE is an open-source python library developed to generate 3D structures in a fully automated manner. Similar to Morfeus, MACE conformer generation is also performed via RDKit-UFF, but additional parameters of bonds and angles applicable to TM complexes are implemented [147]. For example, torsional angles that account for rotations around dative bonds are included [147].

3.3.2 Conformer pruning

One of the advantageous built-in features of Morfeus is its capability of conformer pruning based on various criteria: energy pruning, enantiomer pruning and Root Mean Square Deviation (RMSD) pruning. The default settings were applied for all pruning algorithms. These algorithms are not suitable for pruning conformers directly from *.xyz file formats. Therefore, for the ensembles obtained using CREST, an explicitly added connectivity matrix was required, which was extracted from the MDL MOL file.

Energy pruning involves eliminating conformers above a certain energy threshold based on their relative energy within the ensemble. Enantiomer pruning is applied when the catalyst structure contains chiral atoms. By performing enantiomer pruning, conformers with changed chirality from the original input structure can be eliminated. Morfeus can also be utilized to select and eliminate duplicates due to its RMSD pruning function. RMSD is commonly used to quantify the structural differences between two molecules. The RMSD value of a conformer relative to another one can be calculated by the following expression:

$$RMSD_{A,B} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left((A_{ix}-B_{ix})^2+(A_{iy}-B_{iy})^2+(A_{iz}-B_{iz})^2\right)} \tag{3.1}$$

Where A and B represent two conformer molecules, N is the number of atoms in each molecule, and x, y, z are the Cartesian coordinates of the i-th atom in the molecules. The overlap of two molecules is visually represented in Figure 3.3.

With the Morfeus RMSD pruning function, a matrix is calculated that contains the RMSD values of every conformer relative to every other conformer in the ensemble. By extracting the first column of this matrix, the RMSD values relative to the first (lowest energy) conformer can be obtained.

Due to the limitations of the Morfeus RMSD pruning function, the RMSD Python package [148] was also used to compute the RMSD deviation between two molecules from *.xyz file formats. The main difference is that the RMSD Python package allows the utilization of a reordering algorithm to align atoms in molecules that have undergone rotation during conformer search, and it provides an option to exclude hydrogen atoms from the calculations. Both algorithms were used with their default settings.

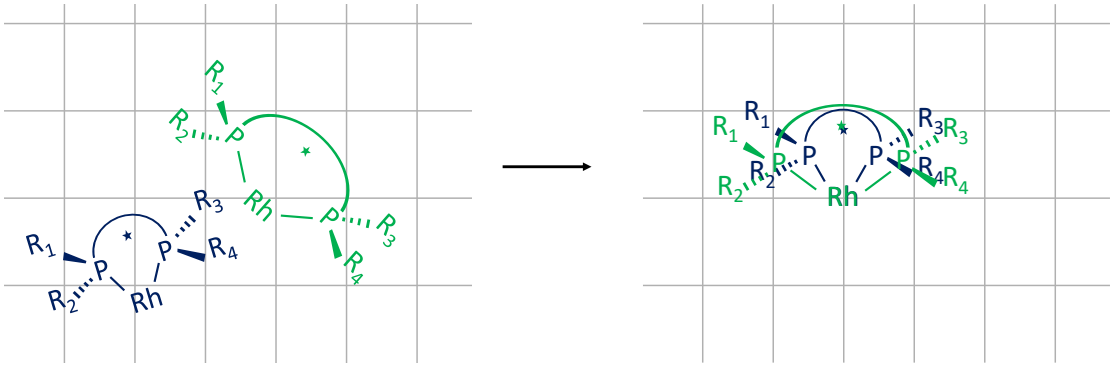


Figure 3.3: Overlapping two molecules for RMSD calculation.

3.4 DFT calculations

24 conformer ensembles were selected for further DFT geometry optimization. The ensembles were previously preprocessed to remove wrong enantiomers, and all conformers in the ensembles served as initial input geometries for DFT calculations. DFT calculations were performed using Gaussian 16 C.0239 on the Snellius supercomputer. The PBE0-D3(BJ)/def2-SVPP [149–151] level of theory was applied as it is reported to show accurate results at relatively low computational effort for TM complexes [146, 152, 153]. After frequency analysis via the Hessian matrix, the pyQRC python script [154, 155] (version 1.0.3) was used for conformers with imaginary frequencies to provide new input geometries. The new geometries were additionally optimized with the same computational parameters.

DFT single point calculations were also performed on all of the conformers of the 192 obtained ensembles to compute electronic properties of the obtained conformers.

3.5 Filtering for DFT

One of the main challenges of this study was to find a subset of conformers to accurately represent the ensemble. The workflow for investigating and evaluating different approaches is represented in Figure 3.4.

In step 1, a training set of 9 different ensembles was chosen to undergo further DFT optimization. A categorization approach (step 2) was utilized to mark the key conformers serving as output labels for the model training and evaluation. By the categorical nature of the targets, a classification problem was generated with the conformer parameters before DFT refinement as predictors. In step 3, algorithms based on molecular descriptor selection and clustering were created to make accurate predictions on the conformers. Three clustering methods were tested; K-means for its general applicability [156], K-medoids as it takes data points as cluster centres [157] and Density Based Spatial Clustering of Applications with Noise (DBSCAN) as it is primarily designed for data having higher amount

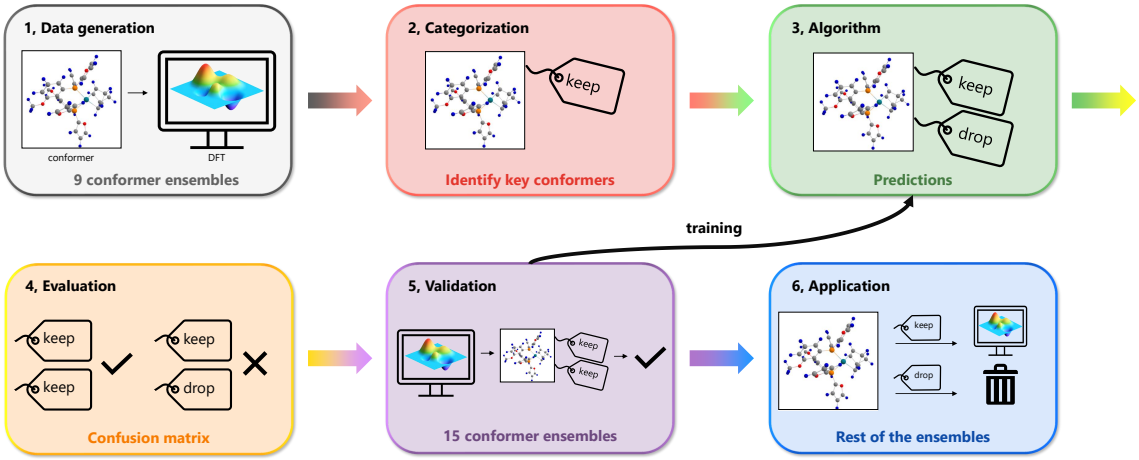


Figure 3.4: Overview of the approach of conformer selection for further DFT geometry optimization: In step 1, nine ensembles are chosen for DFT refinement. The true labels are assigned in step 2. An algorithm is developed for label predictions in step 3. After that, the algorithm is evaluated in step 4. An additional set of ensembles is used for validation, and finally, the prediction algorithm is applied to the rest of the ensembles.

of noise [158]. For K-means and K-medoids models, the minimum number of clusters (k) was set to 1. For DBSCAN, the minimum number of samples in a cluster parameter was set to 2 and the distance to centroid parameter (ϵ) was further optimized based on the model performance.

In step 4, the investigated methods were primarily evaluated by a confusion matrix. A confusion matrix is a commonly used evaluation technique for ML classification models, where binary categories are present [159]. The obtained prediction results are divided into four subsets:

- **True negative (TN)**: prediction model correctly predicted category 1
- **False negative (FN)**: prediction model predicted category 1, although its true value is category 2
- **False positive (FP)**: prediction model predicted category 2, although its true value is category 1
- **True positive (TP)**: prediction model correctly predicted category 2

TN and TP are the subsets that represent correct predictions by the model, while FN and FP are the subsets where the predicted category of the datapoints does not match their true category. An imagine of the confusion matrix can be seen in Figure 3.5. Certain parameters, such as precision ($\frac{TP}{TP+FP}$), recall ($\frac{TP}{TP+FN}$) and accuracy ($\frac{TP+TN}{TP+FP+TN+FN}$) are commonly used for model assessment. However, in this study using the $\frac{TN}{FN}$ ratio as a model assessment parameter was more suitable. A model is considered to perform better than another when this parameter is higher.

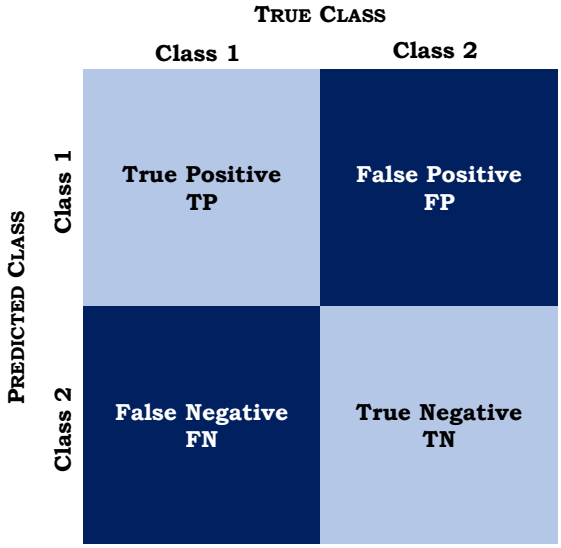


Figure 3.5: Visual representation of a confusion matrix. The dataset is divided into 4 subgroups based on their true and predicted values.

In step 5, an additional set of 15 conformers, containing structures with ligands from distinct ligand families was selected to undergo DFT refinement. This additional data served as a tool to validate the universal application of the chosen algorithm across the dataset. This algorithm was further optimized using the extended dataset including both the training and validation sets (24 conformer ensembles) in step 6. The final step involved applying the model to the remaining 168 conformer ensembles.

3.6 Descriptor calculation through OBeLiX

Calculating molecular descriptors was a key part of the study. To ensure the accurate representation of relevant descriptors, a universal automated computational approach, OBeLiX was used. With OBeLiX several steric, geometric and electronic descriptors can be obtained from various input formats such as *.xyz file format, CREST output folder or DFT output file (*.log file format). For conformer ensembles, the script calculates Boltzmann-averaged descriptor values. The script was further modified to accommodate a Morfeus conformer ensemble object as input and to provide the descriptor values for each individual conformer as well as the Boltzmann averaged values. A total of 37 descriptors were obtained. A detailed list of descriptors and their calculation method can be seen in Appendix B.

3.6.1 ANOVA

The obtained descriptors from calculations at different levels of theory were analyzed using a one way ANOVA test. The ANOVA test was performed handling the computational chemical method as the independent variable and the descriptor values as the dependent variable. Therefore an F-statistics and a p-value was calculated for each descriptor. The

evaluation was conducted using a significance level of 0.05 meaning that there is 5% chance that the null hypothesis (no substantial difference between the methods) is rejected while true.

4

Results & Discussion

This chapter starts with the results and discussion of different conformer searching tools that were investigated. Secondly, the correlations between the conformers from these tools and DFT refinements were highlighted with the possible selection algorithms for further DFT optimization. Finally, a descriptor comparison of both techniques is also included.

4.1 Conformer searching

When investigating various conformer tools, the following criteria were chosen to focus on:

- **Automation**: Is the investigated method able to generate conformers in a high-throughput and automated manner, without the need for manual modification?
- **Accuracy**: Is the level of theory applied by the conformer searching engine sufficient to provide accurate low-energy conformer geometries from the conformational space?
- **Reliability:** Can conformers reliably generated using all different ligand structures? Can the conformer searching tool be applied reliably across a large set of data containing diverse ligand properties?
- **Practical applicability:** Are fast calculations feasible at reasonable computational costs, or is extensive computational power required?

4.1.1 Input structure representation

One of the main purposes of this research was to find a dynamic representation of catalyst structures for data-driven prediction models, making the use of an accurate input representation critical. As mentioned in Section 2.1, due to the unique properties of TM complexes, it is challenging to find a robust and automated way to capture their complex stereochemistry, bond strengths, and orders. Therefore, before feeding the catalyst structures to the conformer searching engines, it is essential to ensure that the input data format

accurately encodes the structural information. A study was conducted on various input data formats relevant to TM complexes, highlighting the main advantages and disadvantages of each molecular representation method. A summary of the key findings regarding the benefits and drawbacks of these input formats is presented in Table 4.1.

Table 4.1: A summary of the advantages and drawbacks of the input formats investigated.

	Input format		
	***.xyz file**	***.mol file**	**RDKit and OpenBabel mol objects**
Advantages	Simple Compact Universally applied	Contains bonding information Can easily be generated from an *.xyz file	Contains bonding information Can be directly used for conformer searching
Disadvantages	No bonding information Cannot be directly used for conformer searching	Cannot be directly used for conformer searching Manual generation is often the most reliable	Generation from *.xyz file is often unreliable

***.xyz file**

The input data was initially stored in an *.xyz file format, which includes only the atomic numbers, symbols, and their Cartesian (x,y,z) coordinates. This file format is simple, compact and commonly used in chemistry. However, a primary limitation of this format is its lack of information regarding bonding orders and atomic charges. Furthermore, the Python packages utilized, which employ RDKit and OpenBabel as their backend, can only initiate conformer search with either chemical SMILES or RDKit and OpenBabel molecular objects as inputs. Therefore, this format cannot be directly used to initiate conformer searching with these engines. Although conformer searching can be performed via CREST using this file format, an explicitly added connectivity matrix with bonding information cannot be avoided for further pruning and preprocessing. Thus, an *.xyz file by itself is not suitable as input for any of the conformer searching tools investigated in this study, and further conversion is necessary.

RDKit and OpenBabel mol objects

An evident approach would be to utilize the appropriate input format of RDKit and OpenBabel: either SMILES or RDKit/OpenBabel mol objects directly. While SMILES is a universally applied method to represent molecular structures, it has limitations when deal-

ing with TM complexes and indicating their stereochemistry (Section 2.2.4). The SMILES representation was hence disregarded, and RDKit/OpenBabel mol objects were further investigated. In addition to their direct applicability for conformer searching algorithms, another key advantage to mention is their capacity to store bonding and charge information.

However, their main limitation lies in their generation from other representation formats, such as an *.xyz file in the case of this study. Since there is no bonding information in an *.xyz representation, these cheminformatic packages are expected to accurately identify the bonding of the atoms. The database of TM complexes were used to test their conversion accuracy from *.xyz files. The OpenBabel mol objects generated by the Openbabel package were lacking crucial bonding information of the molecules. An example of the failed conversion is presented in Figure 4.1, where both the actual bonding (a) and the bonding identified by OpenBabel (b) between the atoms of structure 1 (ligand: SL-J001-1) is depicted. It can be observed that OpenBabel failed to recognize several bonds, highlighting its unreliable performance in structure generation. Although a more accurate conversion can be

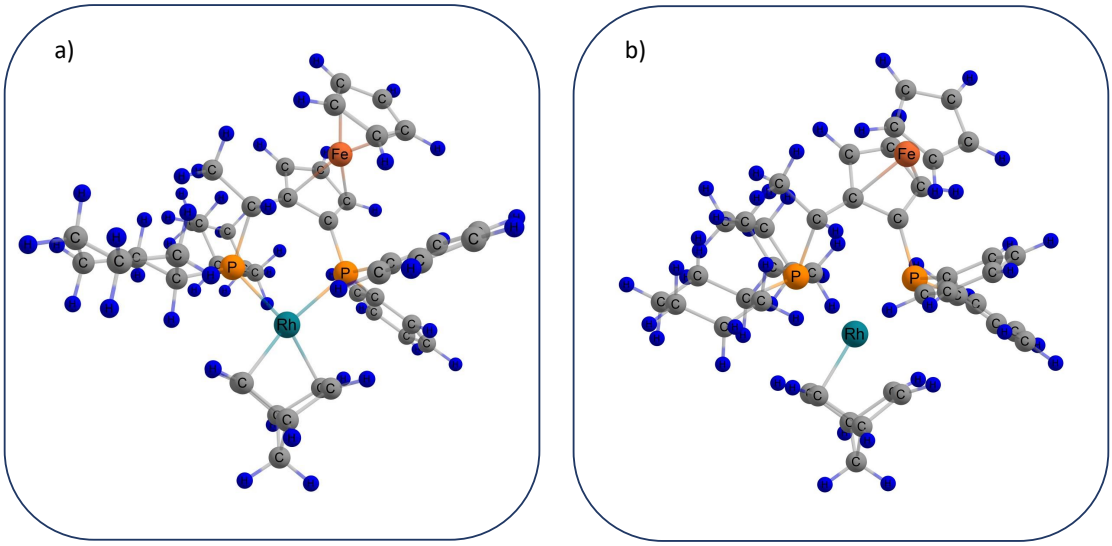


Figure 4.1: The original bonding of structure 1 (ligand: SL-J001-1) (a) and its OpenBabel representation (b).

achieved using the RDKit Python package, the process still remained unreliable. As illustrated in Figure 4.2, RDKit could not accurately encode the correct bonding information in certain cases. For instance, in the presented structure (structure 174, ligand: SL-J681-1) a hydrogen atom is erroneously connected to both a phosphorus and a carbon atom (b), instead of correctly identifying the phosphorus-carbon and hydrogen-carbon bonds (a). Hence, relying solely on RKDit's conversion of the current TM database could lead to many errors and false structure representations.

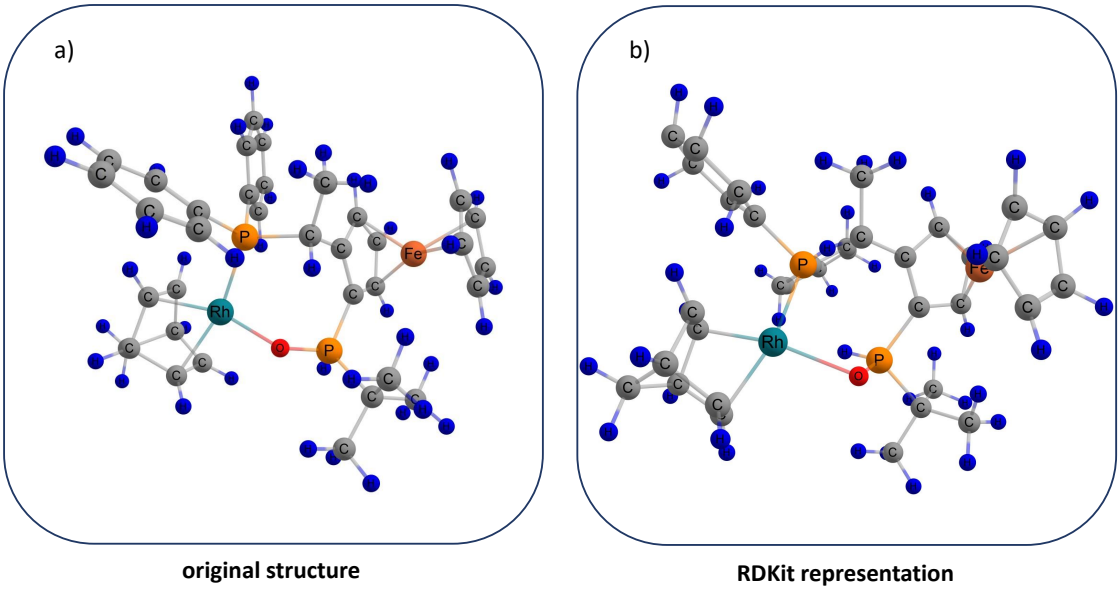


Figure 4.2: The original bonding of structure 174 (ligand: SL-J681-1) (a) and its RDKit representation (b).

MDL MOL file

None of the analyzed packages were able to automatically recognize the correct bonding information of the examined catalyst structures from their *.xyz files. Therefore a manual bond modification step is inevitable. Generating a *.mol file is a convenient tool for this. A *.mol file contains bonding information, and can be further converted to an OpenBabel or RDKit mol object. Its convenience lies in its generation capability via the Chemcraft program. Visual representation of the structures can be attained, facilitating an intuitive way to identify incorrect bonds between the atoms. These bonds can manually removed or added to the structure and the updated bonding information can be stored in a *.mol file. In addition to the drawback of manual generation, another key limitation is the inability to directly fed into the conformer searching engines. However, once correct bonding information is stored, automatic conversion to both RDKit and OpenBabel mol objects becomes feasible.

Therefore the following workflow was implemented: all 192 *.xyz files were converted manually to *.mol files via the Chemcraft program. They were loaded to a Python script for further conversion to RDKit and an Openbabel mol objects. This served as input for RDKit and OpenBabel conformer searching scripts and was suitable for extracting bonding information in the form of a connectivity matrix. CREST conformer searching was initiated using *.xyz file format and pruned utilizing the connectivity matrix generated by RDKit.

4.1.2 Conformer searching via RDKit

As mentioned in Section 3.3, MACE and Morfeus python packages were used to initiate conformer search using the RDKit conformer searching engine. The generated RDKit mol object was used as the input of the script, and the charge of the rhodium was modified to 1. In the MACE script, the bonds of the metal centre were changed to dative. The Morfeus script resulted into several warnings and errors during conformer search such as rhodium in the current coordination (coordination number 4) cannot be recognized. The script could not generate conformer ensembles from the input structures. The MACE script successfully created complexes but most of the cases failed to set up the universal force field. This occurred because during screening the molecule, and angle bend was identified as impossible and terminated the conformer searching program. However, in some exceptional cases this error was avoided and conformer generation was successful. A possible explanation for this behaviour could be that the input Cartesian coordinates were manually determined, potentially introducing slight deviations from the actual configuration. Hence, a dataset containing the structure coordinates after further DFT refinement was also explored. Nonetheless, the same error persisted. Thus RDKit was unsuitable for generating conformers with the current settings and catalyst structures.

4.1.3 Conformer searching via OpenBabel

Another method investigated was OpenBabel conformer searching tool, implemented in Morfeus. A script was created to use both the OpenBabel genetic algorithm and the universal force field. Consequently, both methods were evaluated. The input data format for the conformer searching tools was an OpenBabel mol object. The genetic algorithm failed to set up the stereochemistry of the stucture and did not yield any result. Meanwhile the UFF method only produced one conformer each for each input catalyst structure. Therefore once again, OpenBabel is found to be not suitable to be utilized for conformer searching with the current configuration and settings.

4.1.4 Conformer searching via CREST

The Morfeus script was not suitable for initiating a conformer search via CREST, but the obtained CREST folder could be explicitly added for further pruning. CREST conformer searching successfully produced conformer ensembles for all 192 metal-ligand complexes.

However, in two of the cases, the obtained conformer structures did not maintain the original bonding mode. Figure 4.3 illustrates the original structure (a) and the generated CREST conformer (b) of structure 19 (ligand: SL-T002-2). This ferrocenylphosphine ligand was observed to lose its original biphosphane characteristic and become hemilabile after conformer searching with CREST. While one phosphorus donor maintained its bond with the Rh ion, the other phosphorus detached and distanced itself from the metal centre. Additionally, a carbon atom from the ferrocene moiety formed a new bond with the Rh, preserving its original coordination. This led to a less bulky complex near the active site

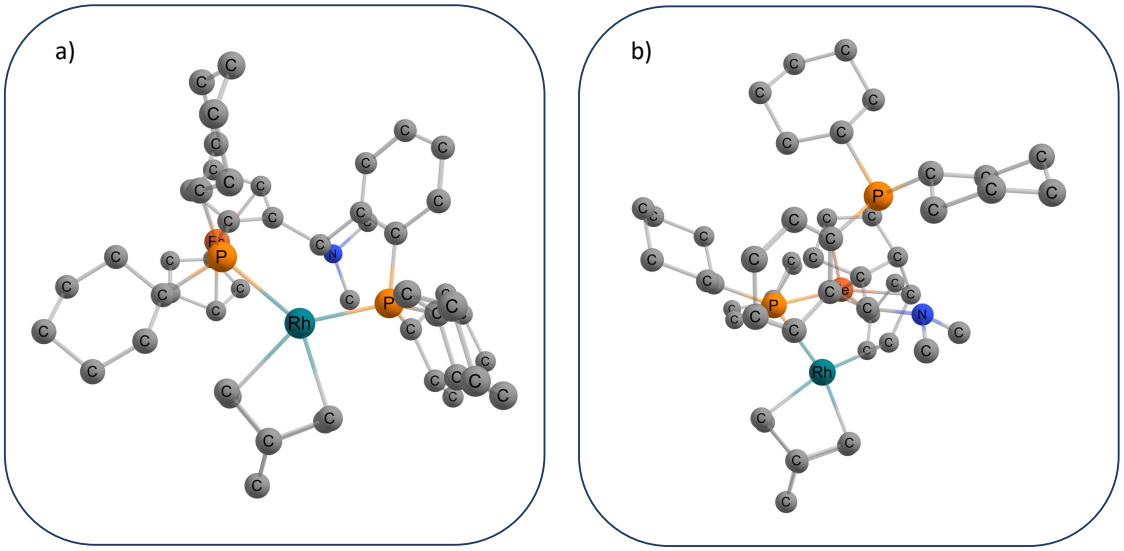


Figure 4.3: Structures of structure 19 before (a) and after (b) CREST calculations.

potentially influencing its catalytic performance. In contrast, the rest of the complexes containing the same Taniaphos backbone retained their original bidentate bonding after conformer searching via CREST.

Similar behaviour can be observed at the generated conformers of structure 186 (ligand: [(2R,3R)-4(9-Anthracenyl)-3-(1,1-dimethylethyl) -2,3-dihydro-1,3- benzoxaphosphol-2-yl]pyridine); Figure 4.4 represents the original structure (a) and the CREST conformer (b). This originally biphospane ligand consists of two symmetrical ligand parts with pyrim-

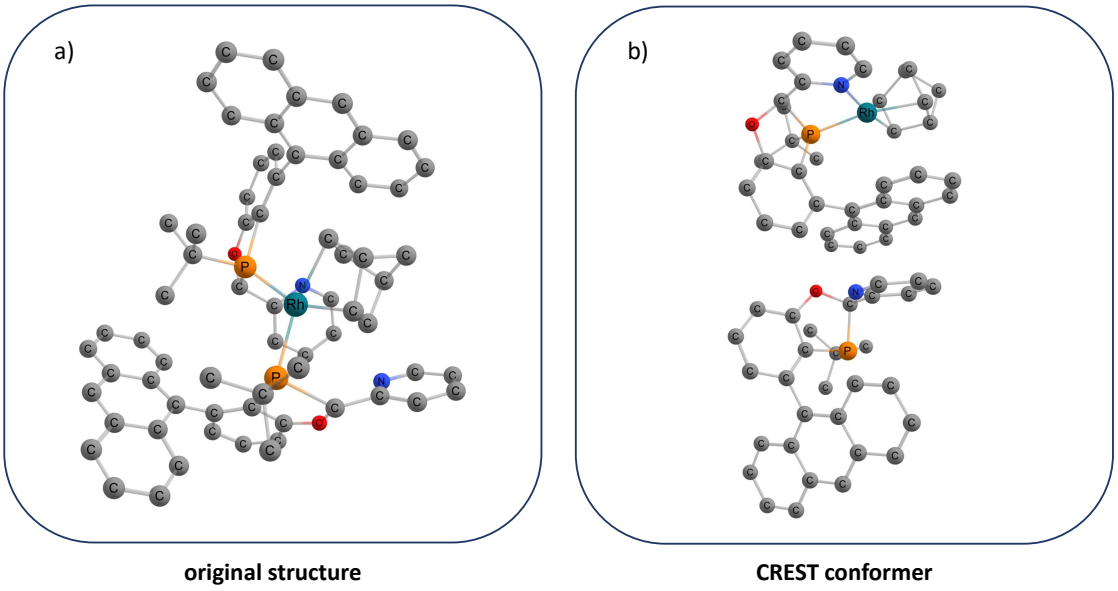


Figure 4.4: Structures of structure 186 before (a) and after (b) CREST calculations.

idine cycles in proximity to the metal-substrate contact. However, in the obtained conformer, one of the P-Rh bond is absent, leading to the separation of the complex into two distinct molecular fragments. To maintain its oxidation state, the Rh atom formed a bond with the N atom from the pyrimidine heterocycle. The resulting complex positions one of the anthracene rings near active site, potentially modifying the catalytic behaviour. Once again, this behaviour appears unique to this specific case as it is not observed for any other complexes with pyrimidine heterocycle.

Conformer pruning

Among the three built-in pruning options offered by Morfeus - energy pruning, RMSD pruning, and enantiomer pruning as described in Section 3.3 - only enantiomer pruning was utilized. As chiral ligands are often present in the complexes, enantiomer pruning was implemented to eliminate conformers where the chirality of the ligand has changed.

The relative energies of the conformers compared to the lowest energy conformer of the ensemble are provided by the CREST calculations and can be used to prune conformers whose relative energy exceeds a certain threshold. However, these energies are calculated using the semiempirical GFN2-xTB level of theory, which is lower than the DFT level. Therefore, performing a pruning step based on these energy values carries the risk of eliminating conformers that are deemed low-energy at a higher level of theory.

RMSD pruning is a useful tool for removing duplicates and similar structures from the ensemble. However, more accurate calculations, such as DFT geometry optimization could lead to additional structural differences. Therefore, this pruning step also carries the risk of eliminating significant conformer geometries from the ensembles.

Conformer ensembles

The collection of the pruned conformer ensembles consisted of a total of 7024 conformers, averaging 37 conformers per ensemble. A relatively high deviation was observed: the smallest conformer ensembles contained only 1 conformer, while the largest ensemble contained 807 conformers. Structure 8 (ligand: SL-J008-1), 144 (ligand:(S)-MorfPhos) and 154 (ligand: SL-N009-2) each resulted in a single conformer, meanwhile the 807 conformers were generated from structure 96 (ligand: (S,S)-DIPSKEWPhos).

This observation is in line with the expectations for structure 144 and 96, illustrated in Figure 4.5 (a) and (b) respectively. Structure 144 contains multiple rigid ring structures that do not allow for rotations that could lead to different conformers. In contrast, structure 96 consists of a larger ligand with reduced steric hindrance between atoms. This allows atoms to rotate freely without significant energy differences, making the ligand more flexible.

In the case of ligand 8 and 154 (Figure 4.6 (a) and (b) respectively) many trifluoromethyl groups are free to rotate over the C-C bonds (indicated by yellow arrows), resulting in slightly different orientations of the F atoms. Therefore, it is surprising that these rotational effects did not generate conformers. Additionally, structure 154 contains a rotatable C-C bond that could change the orientation of carbon atoms (indicated by a blue

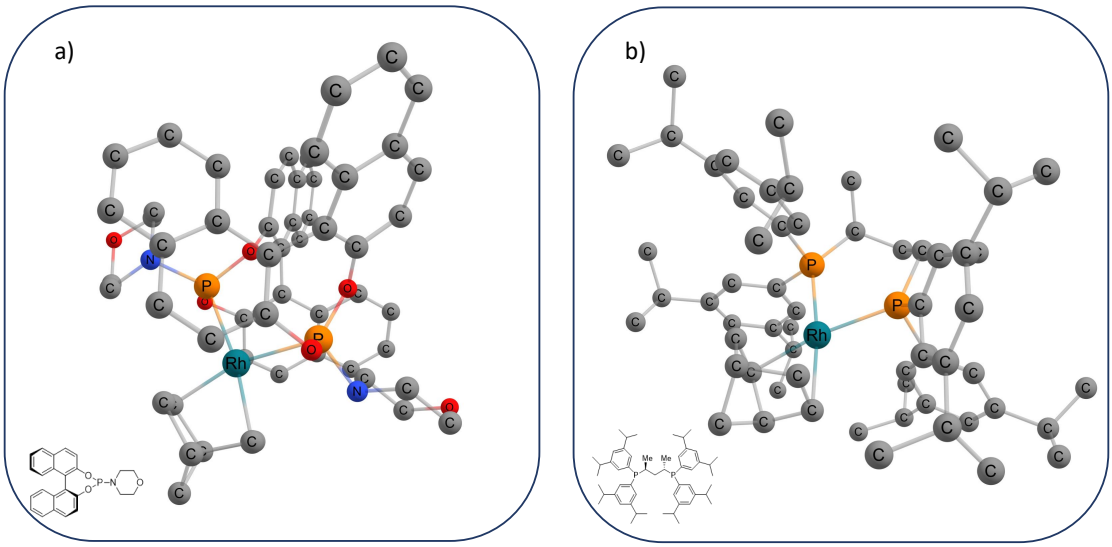


Figure 4.5: Examples of conformers from ensembles 144 (a) and 96 (b). In the left corner, the 2D drawing of their ligand is presented.

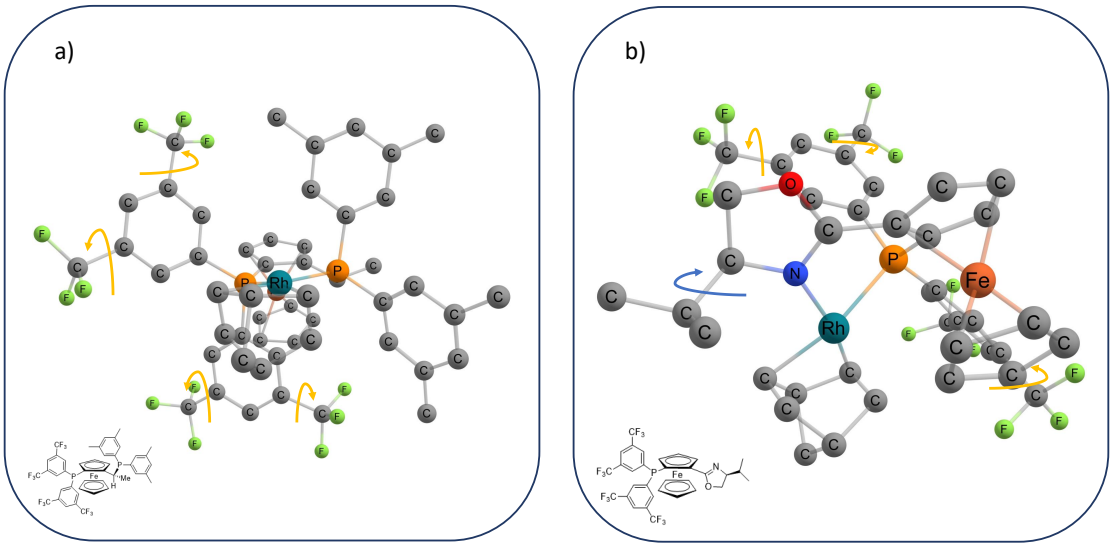


Figure 4.6: Examples of conformers from ensembles 8 (a) and 154 (b). In the left corner, the 2D drawing of their ligand is presented.

arrow). However, given that this group is located close to the model substrate, even slight rotations could lead to significant energy changes. Since CREST considers conformers within a 6 kcal/mol range, the energy increase may prevent this conformer from appearing in the ensemble.

Conformer and rotamers by CREST

A limitation of CREST for the current dataset was observed regarding to the classification of conformers and rotamers by CREST. During the conformer search process, CREST stores the generated conformer-rotamer ensemble (CRE). Since rotamers are considered "degenerate forms of their respective conformers" [138], the final conformer ensemble generated by CREST may not include many rotamers with identical energies. However, it is often found that structures with substantial structural and energy deviations are still considered rotamers of the same conformer and therefore do not appear in the conformer ensemble. For example, the conformer ensemble of structure 154 contained only one conformer. In the rotamer collection, another structure appears with an energy difference of around 8 kJ/mol and an RMSD of 2.25 $\AA$. The structural differences are shown in Figure 4.7 (a), where the two rotamers are overlapped. Conversely, in the conformer ensemble of structure 96, conformers 11 and 12 shows an energy difference of less than 1 kJ/mol and an RMSD of 0.70 $\AA$. The structures of conformers 11 and 12 are overlapped in Figure 4.7 (b).

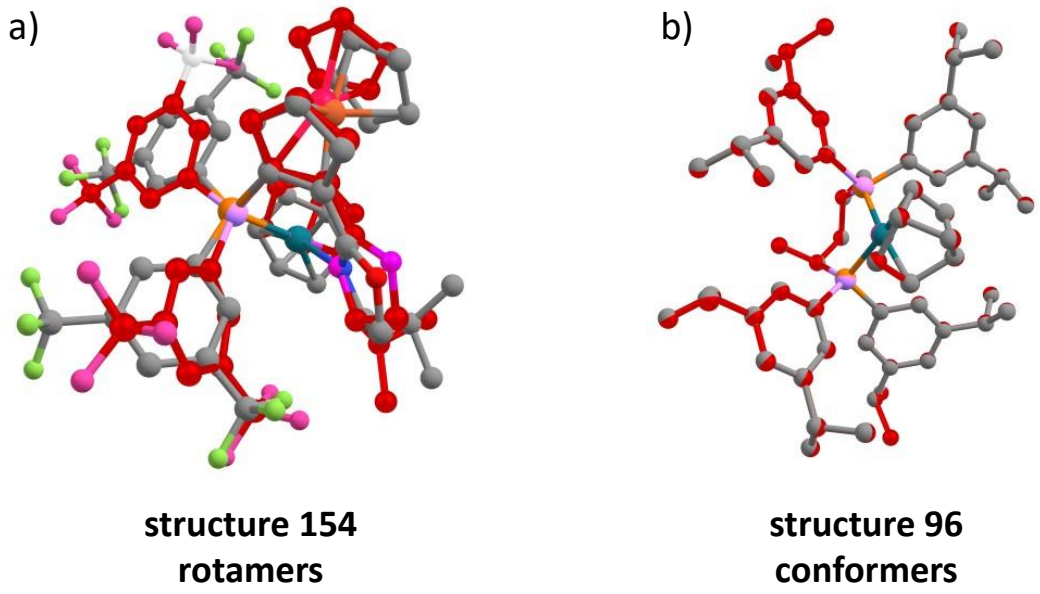


Figure 4.7: Two rotamers of structure 154 (a) and conformer 11 and 12 of structure 96 (b).

It is important to note that since the conformer ensembles are generated at the GNF2-xTB level of theory, it is possible that multiple conformers lay in the same energy minimum at a higher level of theory. This can change the nature of the conformer ensembles. To investigate this, further DFT refinement can be performed and analyzed.

4.2 Filtering for DFT

After obtaining 7024 conformers by CREST calculations, it is clear that doing further DFT geometry optimization for all conformers would consume an outstandingly high amount of computational resources. To quantify this, standard billing units (SBUs) can be calculated. SBU is a standardized parameter used to track computational usage and normalize computational costs [142]. An SBU is expressed by multiplying the number of used cores by the number of computational hours [142]. As DFT geometry optimization for a single complex uses 32 cores for approximately 48 hours, the SBUs charged for running all 7024 conformer calculations is around 11 million. As 1 SBU can be estimated to cost 1 euro cent, the total cost would be 110,000 euros.

Therefore finding a way to reduce the resources consumed without losing significant information is key for the study. This can be achieved by developing a method to select a subset of conformers based on GFN2-xTB calculations that can accurately represent the DFT-based conformer ensemble. The subset should include all conformers that are refined to distinct DFT minima on the potential energy surface and exclude those that converge to the same energy minimum. In order to choose and train a suitable selection algorithm, it is necessary to establish a mapping between the conformers obtained through DFT and CREST calculations.

During the method development (detailed in Section 3.5), a total of 24 ensembles were selected for DFT optimization. These ensembles were chosen to include various ligand families and characteristics, ensuring universal observations and no bias towards any particular ligand family. Structures 19 and 186 were excluded from this selection due to their structural changes during CREST calculations, as described in Section 4.1.4. After DFT calculations, imaginary frequencies were found in 19 conformers. As described in Section 3.5, these conformers underwent additional optimizations. While most of the imaginary frequencies were successfully eliminated, 9 conformers still showed them. These conformers were subsequently excluded from further analysis. The complexes from which these conformers were derived showed large conformer ensembles (around 80 conformers per ensemble), thereby reducing the risk of overlooking key conformers with this action. To examine the behavior of these ensembles, the relative energies of the conformers obtained at both levels of theory and the RMSD values to their corresponding geometries can be analyzed.

4.2.1 Energy analysis

A mapping between DFT and GFN2-xTB can be achieved by analyzing the relative conformational energies within the ensembles obtained by both quantum chemical calculations. The relative energies in respect to conformer 1 (lowest GFN2-xTB energy conformer) of both the DFT and CREST conformers were compared. Figure 4.8 illustrates these relative energy plots for 4 conformer ensembles: ensemble 17 (a), ensemble 80 (b), ensemble 192 (c), and ensemble 139 (d). The cyan-coloured dots indicate each conformer of the ensemble,

marked in ascending order based on their CREST relative energy.

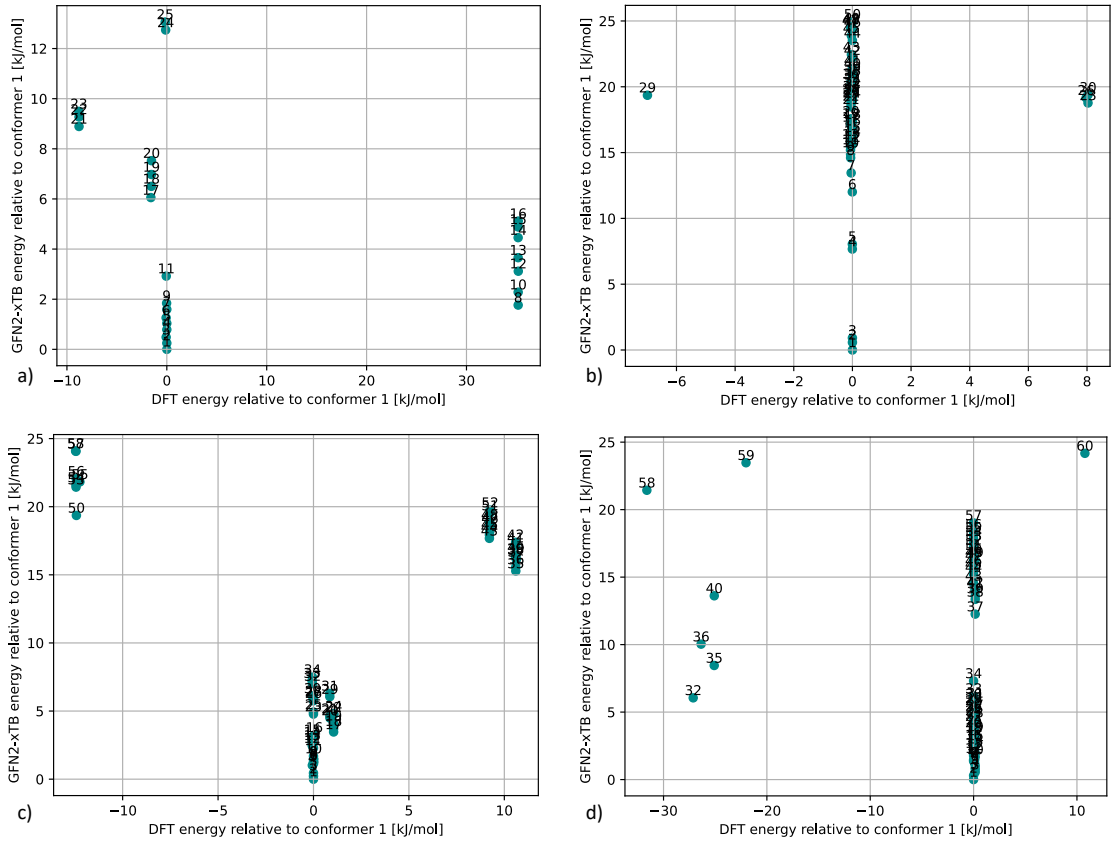


Figure 4.8: DFT and GFN2-xTB energies relative to conformer 1 of ensemble 17 (a), ensemble 80 (b), ensemble 192 (c) and ensemble 139 (d).

Local minima

As it can be seen in Figure 4.8, many conformers that were recognized as distinct energy conformers by CREST (GFN2-xTB) converged to the same energy minimum after DFT optimization. As a result, significantly fewer conformers were distinguished by their energy at the DFT level compared to the GFN2-xTB level. For example, ensemble 17 (Figure 4.8 (a)) included 25 conformers with distinct energy values according to CREST calculations, but only 4 distinct energy minima were present at the DFT level. This trend was even more pronounced for ensemble 80 (Figure 4.8 (b)), where the number of energy levels dropped from 58 to just 3. Based on the 24 examined ensembles, the average of 23 conformers per ensemble at the GFN2-xTB level was reduced to an average of 4 conformers per ensemble at the DFT level.

Lowest energy conformer

A difference is also observable in which conformer is ranked as the lowest energy conformer in the ensemble by the two calculation methods. Since identifying the local minima is often the goal of conformer searching tools, this serves as a relevant indicator of their performance. By observing the energy plots in Figure 4.8, it is visible that conformer 1 (lowest energy by CREST) is rarely the lowest DFT energy conformer. To present a few examples, in ensemble 139, conformer 58 has the lowest DFT energy within the ensemble, holding an approximately 32 kJ/mol energy difference to conformer 1. However, conformer 58 was identified to have approximately 21 kJ/mol higher energy than conformer 1 via CREST. Another example is ligand 80, where conformer 29 is the lowest DFT energy conformer. It has approximately 7 kJ/mol lower DFT energy and 19 kJ/mol higher CREST energy than conformer 1.

Energy ranges

After analyzing the relative energy plots, it is noticeable that the trend of discrete DFT energy classes created from the continuous CREST energy values does not appear for four ensembles (ensembles 57, 110, 172, and 177). The CREST-DFT relative energy plots for ensembles 172 and 110 are presented in Figure 4.9 (a) and (b) respectively. In these cases, the energy values from both methods remain continuous, which could indicate the presence of more DFT local minima than in the rest of the ensembles. However, these energy differences are very low, approximately 0.05 and 0.10 kJ/mol, respectively, which makes the differences negligible. Thus, the conformers of these ensembles all converged into one DFT minimum. Although in the case of ensemble 172, CREST calculations also resulted in a relatively narrow energy range (around 6 kJ/mol), for ensemble 110, the CREST ensemble still presents a large energy range (around 25 kJ/mol).

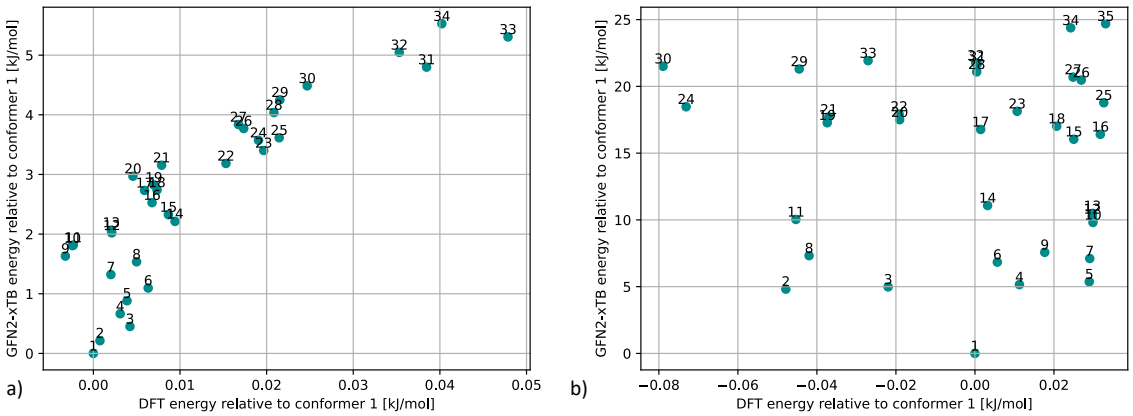


Figure 4.9: DFT and GFN2-xTB energies relative to conformer 1 of ensemble 172 (a), and ensemble 110 (b).

Nevertheless, in the case of ensemble 7, despite the narrow energy range of approximately 1 kJ/mol at the DFT level, discrete energy categories are still distinguishable. Fig-

ure 4.10 (a) shows the CREST-DFT relative energy plot, demonstrating that the conformers cluster into three distinct DFT energy values. To investigate whether actual structural differences cause slight deviations in energy or if it is simply an artifact, conformers 6 and 12 were compared. The comparison in Figure 4.10 (b) reveals a slight positional difference in two methyl groups (indicated by black arrows). However, this structural variance occurs away from the active site, potentially resulting in negligible impact on catalyst performance.

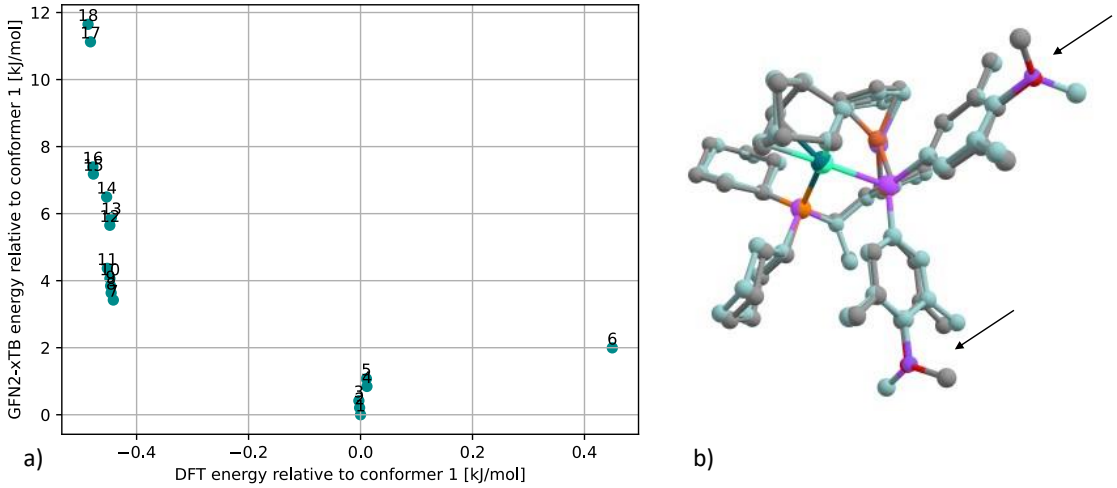


Figure 4.10: DFT and GFN2-xTB energy relative to conformer 1 of ensemble 7 (a), structural overlap of conformer 6 and 12 from ensemble 7 (b).

No correlation between CREST and DFT energy ranges can be identified from these results. On average, CREST ensembles show a higher energy difference than DFT ensembles (approximately 19 kJ/mol and 13 kJ/mol, respectively). DFT calculations also show a greater deviation (minimum: 0.01 kJ/mol, maximum: 44 kJ/mol) compared to CREST (minimum: 6 kJ/mol, maximum: 25 kJ/mol).

4.2.2 RMSD analysis

CREST conformer structures

From the CREST-DFT energy plots discussed in Section 4.2.1, it is visible that conformers converging to the same DFT minimum often have similar GFN2-xTB energy values. These energy differences may be assigned to minor structural deviations. This phenomenon could be utilized to identify and eliminate conformers generated by CREST that will land in the same minimum after DFT refinement. The RMSD values in respect to conformer 1 were used to capture and analyze structural differences. Figure 4.11 illustrates the plots of GNF2-xTB relative energy against the RMSD value to conformer 1 for the CREST structures.

In the plots of ensembles 7 and 17 (Figure 4.11 (a) and (b), respectively), clusters can be

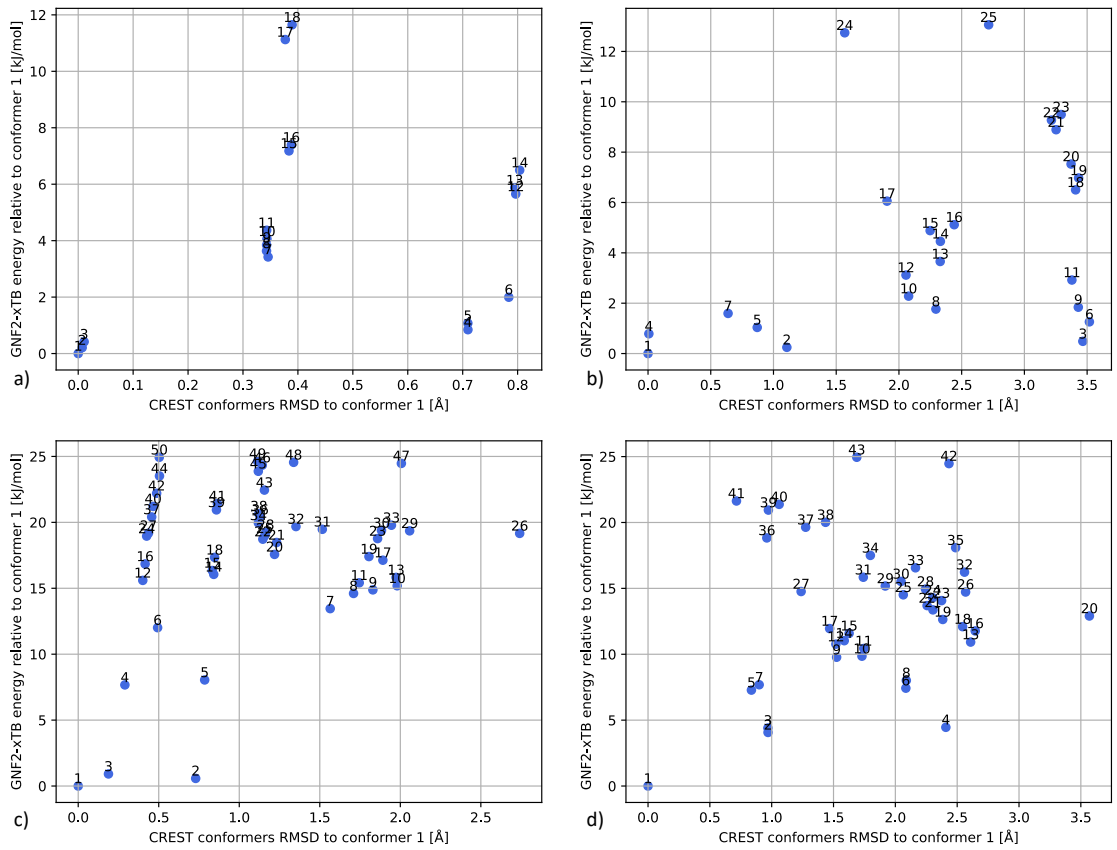


Figure 4.11: CREST relative energy - RMSD to conformer 1 plots of ensemble 7 (a), ensemble 17 (b), ensemble 80 (c) and ensemble 108 (d).

identified. It is noticeable that there are certain conformers present in these ensembles with very close energy and RMSD values. For instance, ensemble 7 shows six distinguishable energy-RMSD clusters, with only one conformer not belonging to any cluster (conformer 6). While more individual points are noted for ensemble 17, clusters are still observable. However, some ensembles show plots with higher noise, such as ensemble 80 and 108 (Figure 4.11 (c) and (d), respectively) where clusters cannot be as clearly separated and identified.

DFT conformer structures

After DFT geometry optimization, many newly obtained DFT geometries are found to be degenerate. Rotamers with equal energy levels are expected to show only very slight structural variations. Once again, the RMSD values in respect to conformer 1 can be used to quantify these structural differences. In Figure 4.12, the DFT relative energies are plotted against the RMSD value to conformer 1 for the DFT structures of ensembles 80 (a) and 7 (b).

For ensemble 7, the observed energy-RMSD correlation aligns with expectations. While not all degenerate geometries are identical, the rotamers exhibit minor RMSD differences

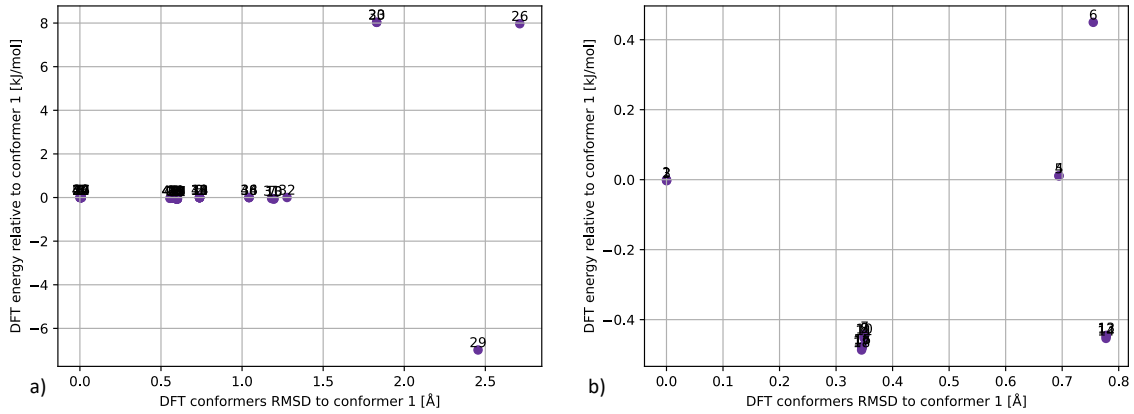


Figure 4.12: DFT relative energy - RMSD to conformer 1 plots of ensemble 80 (a) and ensemble 7 (b).

(0.7 Å). Ensemble 80, on the other hand, presents a more significant RMSD difference between the geometries within the lowest energy DFT minimum. The two most distinct degenerate geometries —conformer 1 and conformer 32— show an RMSD difference of approximately 1.3 Å. To investigate the source of these structural differences, the structures were overlaid as illustrated in Figure 4.13 and 4.14. Conformers 13, 26 and 32 were selected to assess their structural deviations from conformer 1.

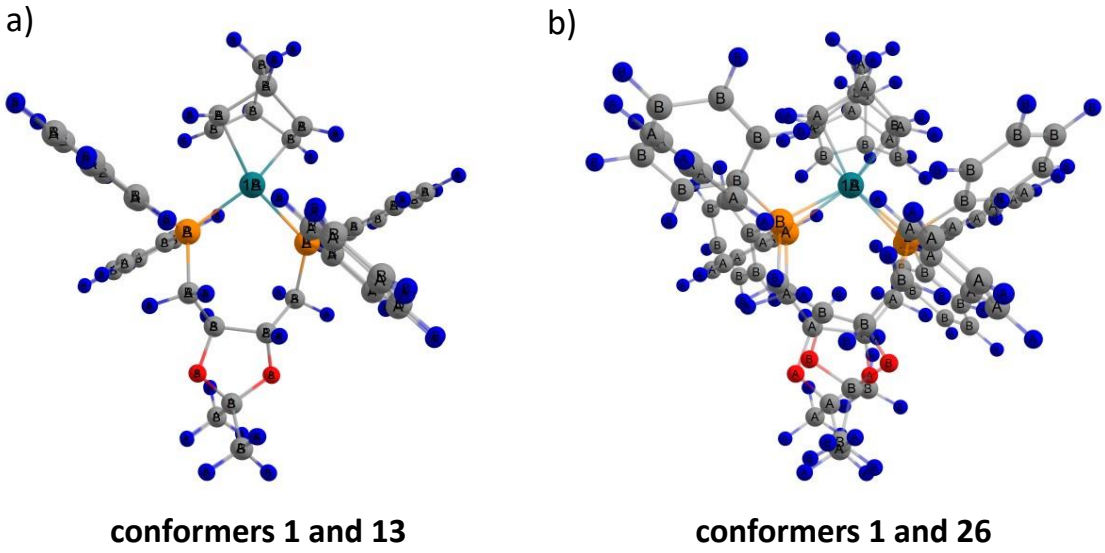


Figure 4.13: Overlap of DFT geometries: (a) conformer 1 (molecule A) and conformer 13 (molecule B), and (b) conformer 1 (molecule A) and conformer 26 (molecule B).

According to Figure 4.12 (a), structure 13 exhibits no energy or RMSD difference relative to conformer 1. Consistent with this observation, Figure 4.13 (a) shows no structural differences between these conformers. Conformer 26 was selected to represent structures

with substantial energy and RMSD differences compared to the first conformer. Figure 4.13 (b) demonstrates a significant structural deviation in this case, aligning with expectations. The overlapped structures of conformers 1 and 32, which resulted in unexpectedly

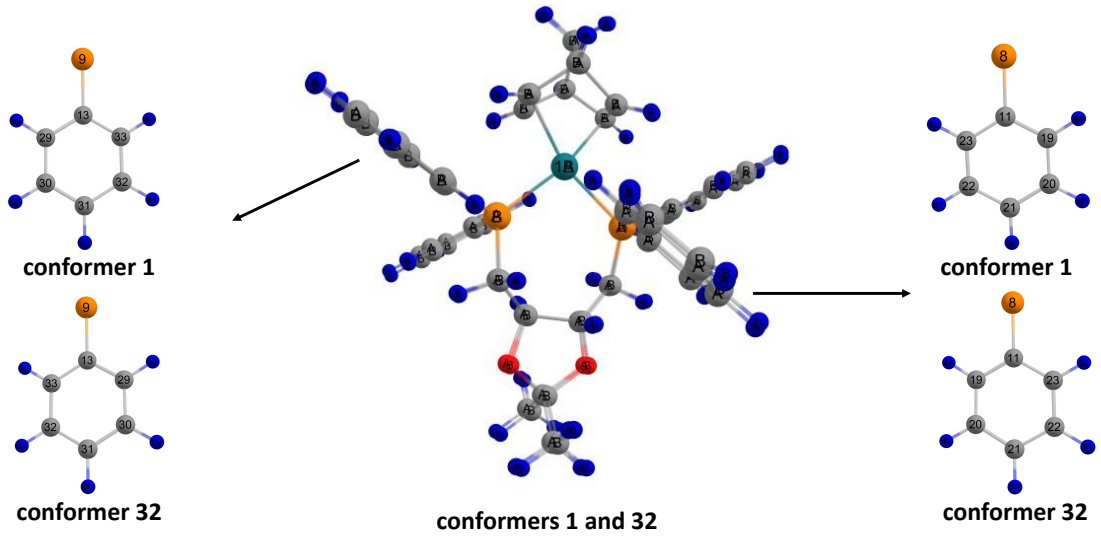


Figure 4.14: Overlap of DFT geometries: conformer 1 (molecule A) and conformer 13 (molecule B).

high RMSD deviations despite being degenerate, are presented in Figure 4.14. Visually, very minor structural differences are observable, suggesting that the RMSD difference may arise from the limitations of the calculation method. This is further confirmed by the fact that RMSD is calculated according to Equation 3.1, making the atom numbers significant. Since different initial (CREST) geometries rotated to the same DFT geometry, these rotational effects cause atoms in identical positions to be treated separately. For example, in conformer 32, the positions of two carbon atoms—atom 29 and atom 33—are swapped compared to conformer 1, causing a false RMSD difference. After manually redistributing the atom numbers and excluding hydrogen atoms from the calculations, an RMSD value of 0.6 Åis obtained. This value is much more consistent with the structural differences observed.

The RMSD Python package [148] was also investigated, as it allows reordering of atom numbers in an automated way, resulting in an RMSD of 0.8 Å. However, it should be noted that this package uses *.xyz files for input, which do not include bonding information between atoms.

4.2.3 Algorithm architecture

DFT geometry refinement changes the nature of the conformer ensemble as many CREST conformers fall into the same local minimum on the DFT PES. Therefore, it is reasonable to develop a method that can select conformers from the CREST ensemble to cover the whole DFT ensemble. This subset of conformers should include at least one conformer from each

DFT minimum, avoiding redundant conformers that converge into the same minimum.

This challenge was transformed into a classification problem as illustrated in Figure 4.15. Certain parameters of the CREST conformers were used as input features. Output

Figure 4.15: Classification problem using the CREST conformer parameters as input features and the DFT local minimum values as output labels. Binary predictions are made to either keep or eliminate conformers.

labels was assigned based on the local minima obtained by DFT calculations, indicated in Figure 4.16. In these plots, the relative CREST and DFT energies of the conformers are presented, and conformers are marked with different colours based on their DFT energy. For ensemble 17 (Figure 4.16 (a)) 4, whereas for ensemble 80 (Figure 4.16 (b)) 3 DFT energy

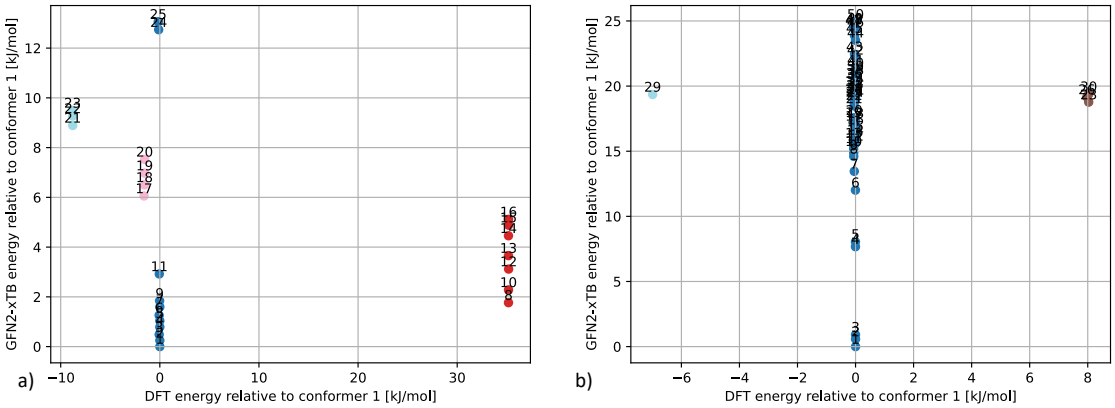


Figure 4.16: The identification of DFT local energy minima for ensemble 17 (a) and 80 (b) is illustrated with their CREST-DFT relative energy plots. Conformers classified to the same DFT minimum are indicated with the same colour.

minima were identified. To assign output labels, a dynamic threshold was utilized: the DFT energies were scaled with a mean of 0, and a standard deviation of 1. Conformers with the scaled DFT energies within a 0.06 range were categorized as being in the same energy minimum. The algorithms were therefore developed to identify patterns of input parameters leading to a subset of conformers. As the nature of the algorithms was unsupervised, output labels were used to evaluate of the accuracy of this subset.

After selecting a suitable algorithm, predictions can be made to each CREST ensemble. Each conformer in the CREST ensemble will be classified as either kept in the selected subset (1), or eliminated (0). Based on the conformer parameters used as input features, two distinct algorithms were investigated.

4.2.4 Feature selection

Chemical intuition

One approach was to capture the nature of the DFT ensemble using certain independent CREST conformer properties, such as relative energy, RMSD difference, cone angle and buried volume, illustrated in Figure 4.18. The relative energies and RSMD values were

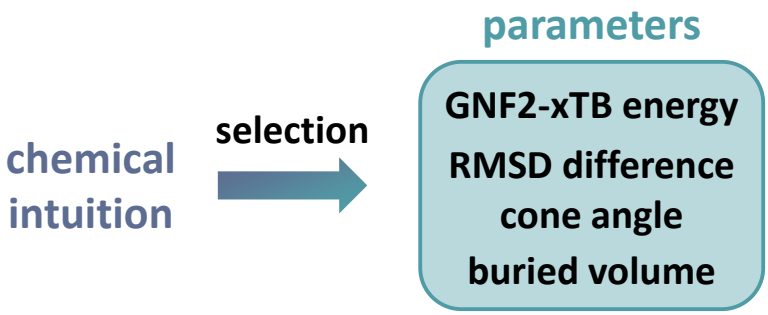


Figure 4.17: Selection criteria approach: CREST conformer parameters are chosen based on chemical intuition.

mainly used to eliminate conformers from the ensemble, while the descriptor properties aimed to capture conformers that will land in distinct DFT energy minimum.

Morfeus' built-in tools were utilized to perform energy and RMSD pruning on the CREST ensembles. Pruning based on the GFN2-xTB energy of the conformers may reduce redundant conformers that would fall into the same minimum as other, lower energy conformers. RMSD pruning was performed to drop conformers with similar geometries, as it is more probable that they are optimized by DFT to the same structure and therefore energy minimum. However, after performing these pruning options, it was observed that not all DFT minima are captured by the new subset of conformers in both the RMSD and the energy pruning cases. Figure 4.18 (a) shows the GFN2-xTB - DFT energy plot of ensemble 71, with the conformers eliminated by energy pruning marked in cyan. Figure 4.18 (b) shows the GFN2-xTB - DFT energy plot of ensemble 7, where the conformers eliminated by RMSD pruning are marked in cyan. In both cases, one DFT minimum was not captured. However, in 8 out of the 9 tested ensembles, both method captured all DFT minima. An alternative approach to eliminating conformers based on energy is to drop the conformers with the highest n % of energy in the ensemble instead of using a universal threshold. This approach was also tested.

To cover the whole DFT ensemble, these pruning functions should not eliminate key conformers from the ensemble. Certain parameters could be used to capture conformer structures that will likely land in different DFT minima and add them to the selected subset. A method of determining distinct conformer geometries revolves around the geometric and steric descriptors of conformers within the ensemble. Prior research by Gensch et al. [160] on monodentate ligands has represented the whole ensemble using the conform-

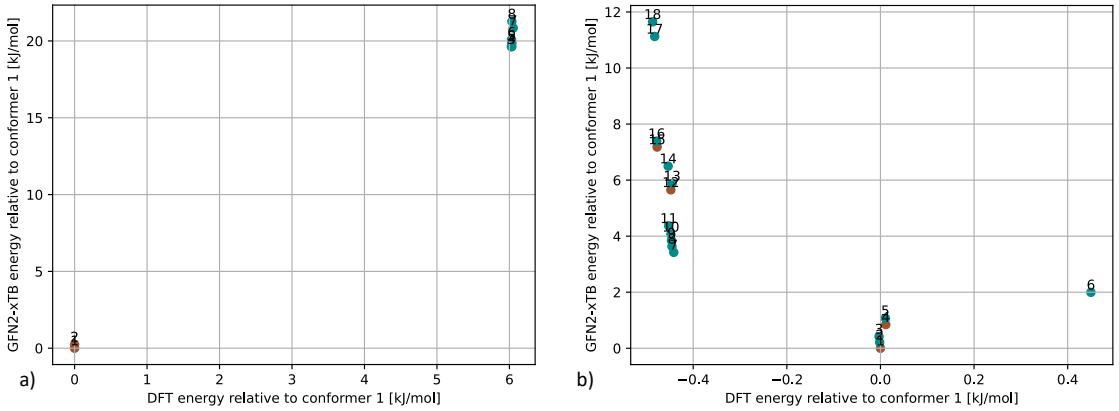


Figure 4.18: DFT and GFN2-xTB energy relative to conformer 1 of ensemble 71 with energy pruning (a), and ensemble 7 with RMSD pruning (b). Eliminated conformers are indicated by cyan dots.

ers with the highest and lowest buried volume. Conformers with the minimum and maximum buried volume (with radius 4Å) and cone angle were selected to remain in the new subset. However, as illustrated in Figure 4.19, for some tested ensembles, these conformers were still optimized into the same DFT minimum. In Figure 4.19 (a), the highest and

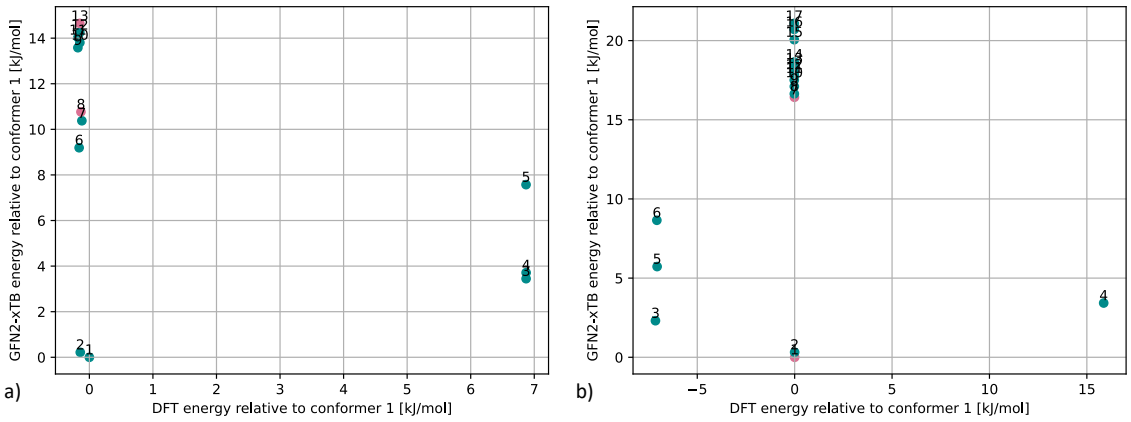


Figure 4.19: DFT and GFN2-xTB energy relative to conformer 1 of ensemble 44 with highest and lowest cone angle conformers (a), and ensemble 33 with highest and lowest buried volume (at 4 Å) (b). Conformers with the aforementioned descriptor properties are highlighted in pink.

lowest cone angle conformers of ensemble 44 are marked in pink. After DFT refinement, both of these conformers landed in the lowest DFT energy minimum. Similar observations can be made about the conformers with the minimum and maximum buried volume of ensemble 33: as illustrated in Figure 4.19 (b), they are degenerate in the DFT ensemble. In total, conformers with the highest and lowest cone angle and buried volume fall into different DFT categories in 7 and 8 cases from the training set, respectively.

The final method included a pruning function additionally including conformers with the aforementioned descriptor properties in the new subset.

Clustering

As discussed in Section 4.2.2, many CREST conformers show very similar GNF2-xTB and RMSD values. Figure 4.11 showed distinct energy-RMSD clusters within the conformer ensembles. The hypothesis is that conformers with similar geometries and close energy values will likely to be optimized to the same DFT local minimum. Therefore, instead of considering the energy and RMSD values independently, their correlation was also utilized to describe the DFT ensemble. This was achieved using a clustering algorithm as presented in Figure 4.20. After identifying the conformer clusters, the conformer classification was

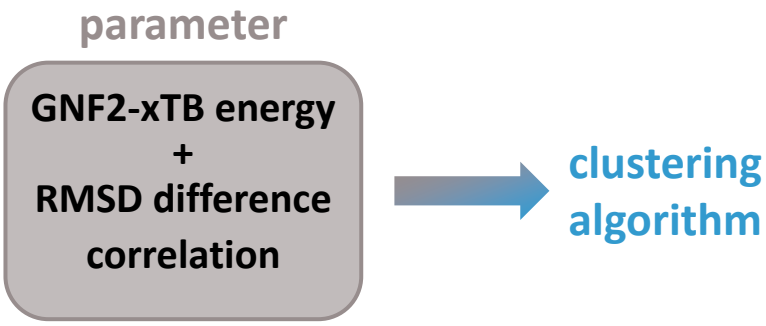


Figure 4.20: Clustering approach: the correlation of GN2-xTB - RMSD is used by clustering algorithms.

as follows: conformers closest to the cluster centres, as well as individual points outside of clusters, were included in the subset of selected conformers, while the rest were eliminated.

To choose a suitable clustering approach for this data, three clustering algorithms were initially considered: K-means, K-medoids, and DBSCAN clustering. After an initial study, it became clear that due to the nature of our data and the objective of the clustering method, DBSCAN is the most beneficial to move forward with. K-medoids and K-means often resulted in broader clusters, increasing the chance of potentially eliminating key conformers from the ensemble. On the other hand, DBSCAN is designed to deal with data having higher amount of noise, thus it only groups conformers into the same clusters if they are actually close in RMSD and energy. An example of this behavior is shown in Figure 4.21, where K-medoids (a) and DBSCAN (b) algorithms were tested on the conformers of ensemble 87. It is visible that the K-medoids method groups conformers 32-35 into one cluster, while DBSCAN identified a cluster that only contains conformers 32, 33, and 35, leaving 34 as an outlier. Although these conformers present a really close energy value, the maximum RMSD difference within the K-medoids conformers is much higher than the DBSCAN cluster (0.72 $\mathring{A}$ and 0.30 $\mathring{A}$ respectively). Since not losing any conformers that converge into distinct DFT minima is crucial, it was decided to move forward with the

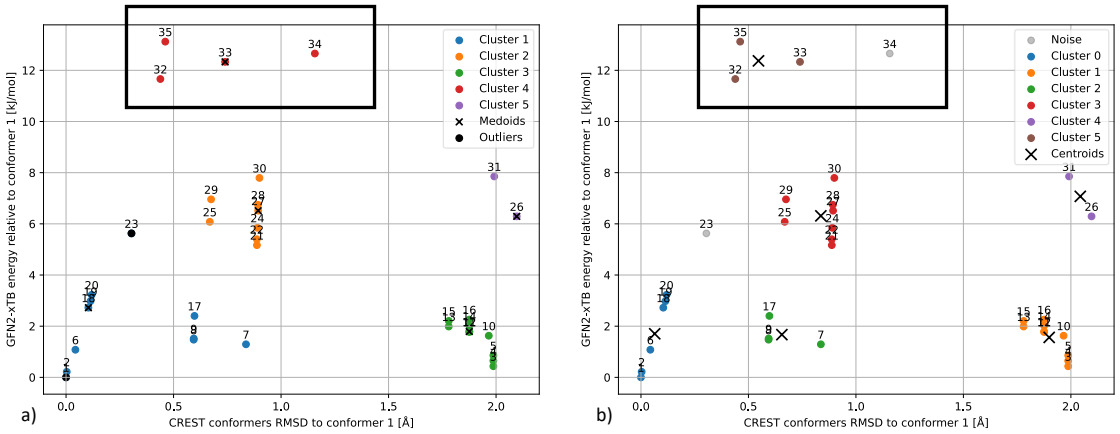


Figure 4.21: RMSD-GFN2-xTB energy clustering of ensemble 87 via K-medoids (a) and DBSCAN (b).

DBSCAN method.

4.2.5 Algorithm evaluation

Confusion matrix

Unlike typical classification problems, there is no 1:1 correspondence between the output labels and the model predictions. The algorithms assign binary categories to conformers, deciding whether to include or exclude each conformer in a subset. The output labels on the other hand, indicate which DFT local minimum the conformer was refined into. The DFT ensemble is accurately captured if at least one conformer from each energy minimum is included in the predicted subset of conformers. Given that multiple conformers can converge to the same energy minimum, this can be achieved by various combination of conformers. Hence, the implementation of the confusion matrix to evaluate the performance of the algorithm is not straightforward. The following approach was used to determine the parameters of the confusion matrix:

- **True negative (TN)**: The number of conformers that are correctly eliminated by the algorithm: their DFT minima are already represented by other conformers in the predicted subset, making them redundant to cover the DFT ensemble.
- **False negative (FN)**: The number of conformers that are incorrectly eliminated by the algorithm: their DFT minima are not represented by other conformers in the predicted subset, making them necessary to cover the DFT ensemble.
- **False positive (FP)**: The number of conformers that are incorrectly included in the predicted subset by the algorithm: their DFT minima are already represented by other conformers, making them redundant to cover the DFT ensemble.

- **True positive (TP)**: The number of conformers that are correctly included in the predicted subset by the algorithm: their DFT minima are not represented by other conformers, making them necessary to cover the DFT ensemble.

Due to the specific nature of the problem, instead of using the commonly used parameters for assessment (such as precision, recall, and accuracy), it was decided to evaluate the models based on their $\frac{TN}{FN}$ ratio. In a well-performing model, the TN value is maximized (all redundant conformers are eliminated), while the FN value is minimized (no DFT minimum is missed), resulting in a high $\frac{TN}{FN}$ ratio.

A practical example of analyzing algorithm performance through constructing a confusion matrix is presented in Figure 4.22.

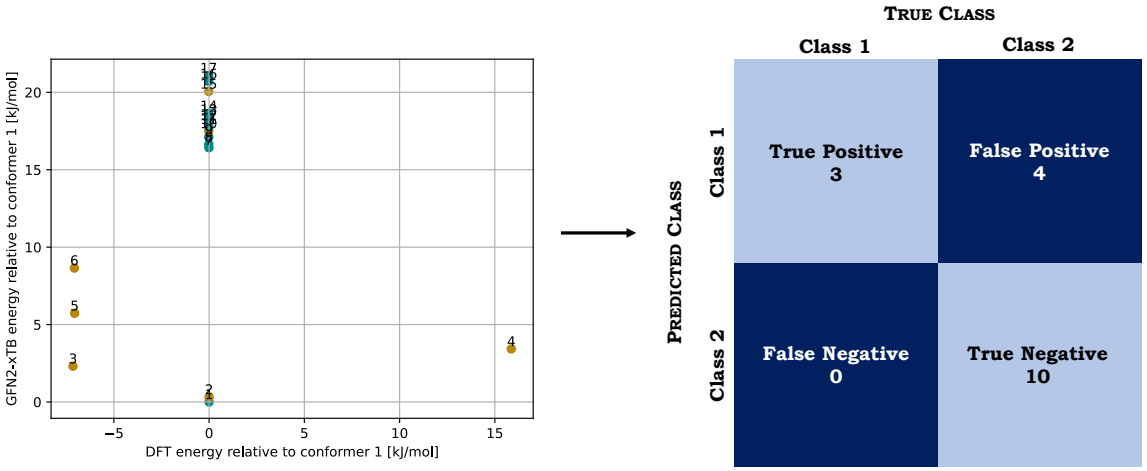


Figure 4.22: An example of confusion matrix implementation: on the left, the plot of DFT and GFN2-xTB energies relative to conformer 1 of ensemble 117 is displayed. On the right, the constructed confusion matrix based on retained and eliminated conformers is presented.

The retained (gold dots) and eliminated (cyan dots) conformers on the GFN2-xTB - DFT energy plot of ensemble 117 is shown utilizing DBSCAN clustering (ϵ = 0.2). As at least one conformer from all DFT energy minima is captured, the TP equals the maximum number of DFT minima (3) and FN is 0 (no minimum is missed). The model successfully eliminated 10 conformers (TN). Since the model included a total of 7 conformers in the predicted subset, and only 3 DFT local minima are present, 4 conformers from this subset are redundant (FN).

Best performing algorithm

Since multiple algorithms were tested, the best-performing model was selected by comparing the confusion matrices. After testing the initial 9 ensembles, most algorithms showed very similar results. Therefore, an additional set of 15 ensembles was added for evaluation. This validation set consisted of ensembles with different ligand families, allowing the test of the algorithms' universal applicability for our Rh-based TM complexes. In Figure

4.23, the total number of missed key conformers (FN) against the number of eliminated redundant conformers (TN) is shown for all investigated methods based on the results of all 24 ensembles.

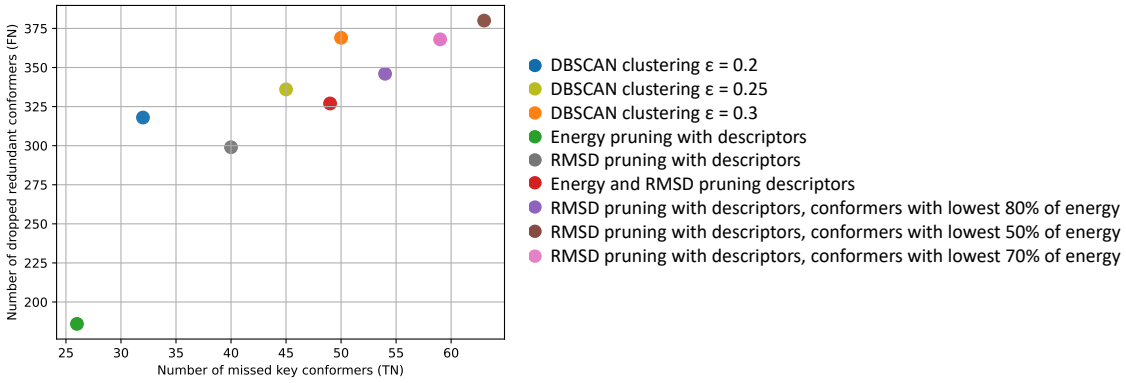


Figure 4.23: The total number of missed key conformers (FN) against the number of eliminated redundant conformers (TN) for all approaches based on 24 ensembles.

Table 4.2: Values of used assessment parameters (FN, TN, FN/TN ratio) for all investigated algorithms.

Algorithm	**FN**	**TN**	$\frac{TN}{FN}$ **ratio**
DBSCAN clustering $\epsilon = 0.2$	32	318	9.94
DBSCAN clustering $\epsilon = 0.25$	45	336	7.47
DBSCAN clustering $\epsilon = 0.3$	50	369	7.38
Energy pruning with descriptors	26	186	7.15
RMSD pruning with descriptors	40	299	7.48
Energy and RMSD pruning with descriptors	59	368	6.24
RMSD pruning with descriptors, conformers with lowest 80% of energy	49	327	6.67
RMSD pruning with descriptors, conformers with lowest 70% of energy	54	346	6.41
RMSD pruning with descriptors, conformers with lowest 50% of energy	63	380	6.03

Table 4.2 shows that DBSCAN clustering with a distance to centroid parameter of 0.2 achieved the highest $\frac{TN}{FN}$ ratio of 9.94. This method successfully eliminated 318 redundant conformers but missed 32 key conformers. In contrast, all other tested algorithms have $\frac{TN}{FN}$ ratios ranging from 6 to 7.5.

Another point to note is that the evaluation did not bias any assessment parameter, it treated the loss of a DFT minimum and the removal of a redundant conformer equally. However, from a chemical perspective, capturing all DFT minima (minimize FN) might be

more crucial than dropping all redundant conformers (maximize TN). Furthermore, since the CREST ensemble contains significantly more conformers than the DFT ensemble, the total number of redundant conformers is higher than the number of DFT minima. The clustering method (ϵ = 0.2) also performed well in terms of identifying DFT minima. Among the methods investigated, only one method -energy pruning with descriptors- eliminated fewer key conformers. However, this method was found to be less selective as it removed 132 fewer redundant conformers compared to clustering one.

4.2.6 Algorithm optimization

From the results in Table 4.2, it is evident that the choice of ϵ significantly influences the clustering model's performance. Increasing ϵ values from 0.2 to 0.25 and 0.3 resulted in a decrease in the $\frac{TN}{FN}$ ratio from 9.94 to 7.47 and 7.38, respectively. Consequently, further optimization of this parameter was done to determine the optimal value and achieve the best clustering performance. In Figure 4.24, the number of eliminated redundant conformers is plotted against the number of missed key conformers for each clustering algorithms with different ϵ, based on the 24 ensembles. A table containing the assessment parameters for all

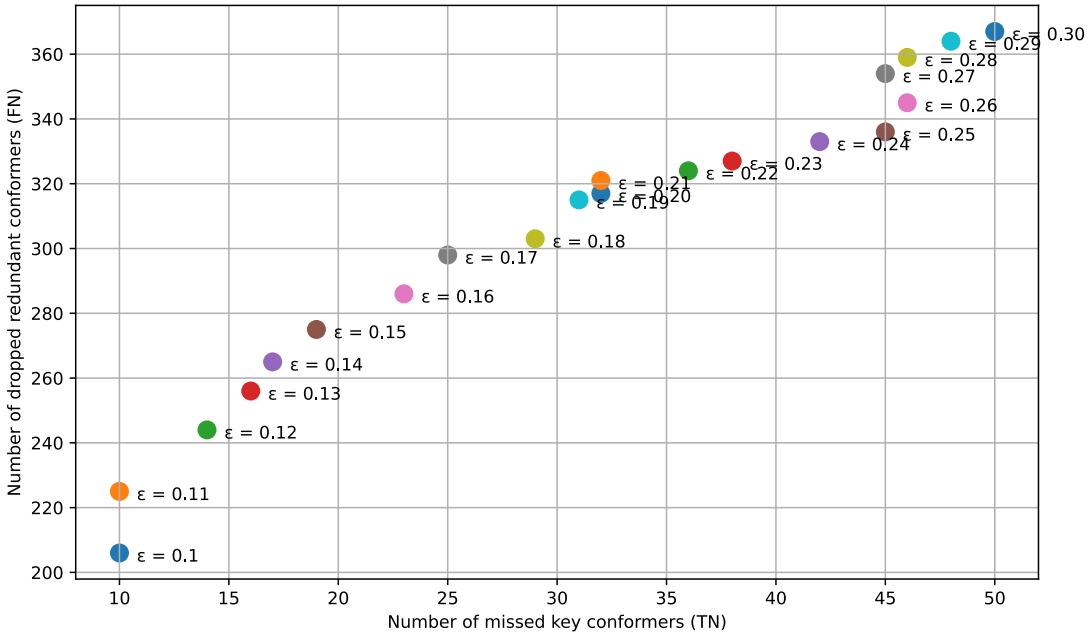


Figure 4.24: The total number of missed key conformers (FN) against the number of eliminated redundant conformers (TN) all tested ϵ values.

investigated ϵ values can be found in Appendix D. Based on these values, the model with $\epsilon = 0.11$ showed the highest FN/TN ratio of 22.5. This algorithm successfully eliminated 225 conformers while losing only 10 DFT minima. These DFT minima originated from three DFT ensembles: ensemble 13, 149, and 172. After examining the DFT energy ranges

of these ensembles, it became apparent that they fall within a very narrow energy range: within less than 1 kJ/mol for ensembles 13 and 172 and within 4 kJ/mol for ensemble 149. Due to the dynamic threshold applied to distinguish DFT minima, they appear as separate DFT energy conformers despite being degenerate. Therefore, the loss of these conformers by the algorithm does not reflect an actual loss of DFT minima but rather arises from the limitation of the model.

4.2.7 Transferability test

To ensure a universally applicable selection algorithm, the developed model (DBSCAN clustering on GFN2-xTB-RMSD of CREST geometries with $\epsilon = 0.11$) was tested on a new dataset [146]. This new dataset contained the same backbone structure (ligand and Rh metal centre), but instead of using the precatalyst form with an NBD model substrate, methyl 2-acetamidoacrylate substrate was attached to Rh. Based on the ligand-substrate configurations, four different coordination modes are possible: two of them are more sterically restricted, and two are less sterically restricted [146]. These coordination modes are presented in Appendix C. Consequently, the algorithm was tested on 44 CREST ensembles from 11 different ligands.

Compared to the original precatalyst dataset, three observation can be made. First, the RMSD clusters based on the CREST geometries are less distinguishable. Second, many CREST conformers converge to the same DFT minimum after DFT geometry optimization. And finally, the applied clustering algorithm eliminated significantly less redundant conformers. An example of this behaviour is shown in Figure 4.25: (a) DBSCAN clustering is

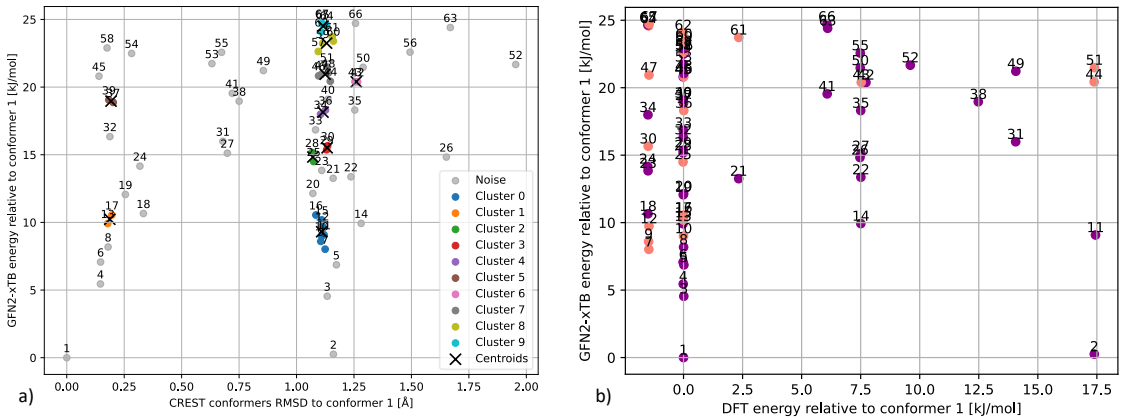


Figure 4.25: Ensemble 49 (substrate in minor 2 coordination): DBSCAN clustering on the GFN2-xTB - RMSD plot (a) and DFT and GFN2-xTB energy relative to conformer 1 (b) highlighting the eliminated conformers by magenta.

visualized on the GFN2-xTB - RMSD plot of ensemble 49 (minor 2 substrate coordination), and (b) the GFN2-xTB - DFT energy plot, with the retained conformers marked in salmon and the eliminated conformers marked in magenta. In this case, out of the 59 redundant conformers, the algorithm only eliminated 23. The final comparison of the evaluation ma-

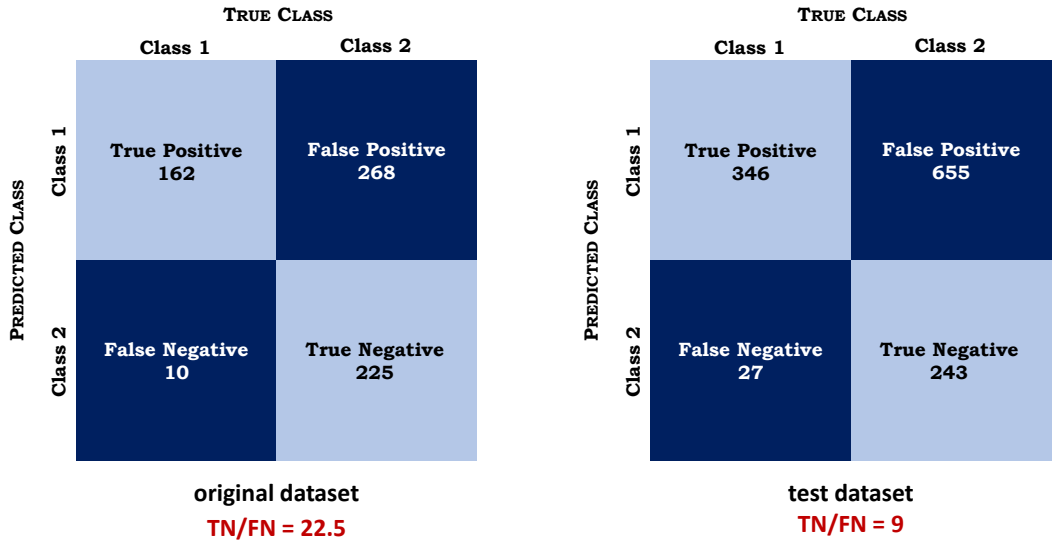


Figure 4.26: Comparison of confusion matrices from DBSCAN clustering ($\epsilon = 0.11$ on both the original dataset (left) and the test set (right).

trix obtained for both the original and the test dataset is shown in Figure 4.26 (a) and (b) respectively. The original dataset shows a significantly higher $\frac{TN}{FN}$ ratio of 22.5 compared to the $\frac{TN}{FN}$ ratio of 9 that is obtained for this test set. Another main performance indicator to mention is the dropping ratio. Since the total number of redundant conformers differ from the two dataset, the $\frac{TN}{TN+FP}$ ratio would also serve as a useful assessment tool. The ultimate goal is to drop all redundant conformers, therefore FP = 0, resulting in the $\frac{TN}{TN+FP} = 1$. In the original dataset, this ratio is 0.46, indicating that 46% of the redundant conformers are recognized by the model. In the case of the test set, however, this ratio dropped to 0.27, meaning that the model recognized 19% fewer redundant conformers. Therefore, based on the current assessment parameters, better results are achieved on the original dataset.

4.2.8 Application

The obtained algorithm (DBSCAN clustering with $\epsilon = 0.11$) was used to predict a subset of conformer for DFT geometry optimization on all the 192 CREST conformer ensembles. Out of the 7024 conformers, the method identified only 3796 for further DFT refinement. By reducing the amount of calculations needed by around 46%, approximately 5 million less SBUs are required for the calculations saving 50,000 euros on computational costs.

4.3 Descriptor calculation

The OBeLiX workflow was modified to calculate individual conformer descriptors for a conformer ensemble. With both the CREST and DFT conformers obtained for 24 conformer

ensembles, this resulted in CREST and DFT descriptors for a total of 665 conformers. The steric and geometric descriptors were subjected to ANOVA to determine which descriptors significantly differ based on the applied level of theory. Out of the 23 descriptors analyzed, significant differences were observed in 7 cases.

Out of the four analyzed geometric descriptors, two were found to be significantly different based on the applied quantum chemical calculations. These were the two dihedral angles between the hydrogen and carbon atoms of the NBD-metal donor atoms. The other two geometric descriptors, the cone angle and the bite angle, did not show significant differences based on the p-values obtained from the ANOVA test.

The set of steric descriptors contained six buried volume values: four were calculated with the metal as the centre of the sphere, while two were calculated around the two donor atoms of the ligand. Out of the four metal-centred buried volumes, three showed low p-values and were therefore marked as significant: buried volumes with radii of 3.5 Å, 4 Å, and 5 Å. It is noteworthy that the buried volume with the highest examined radii (6 Åand 7Å) did not show significant differences between the CREST and DFT ensembles. The buried volumes around one of ligand donor atoms resulted in a low p-value and therefore analyzed as significantly different between the GFN2-xTB and DFT levels of theory, while the other one did not show a significant difference. The last descriptor that is marked as significantly different based on the level of theory according the ANOVA is one of the octants of buried volume at 3.5Å. A full ANOVA table, including the F-statistics and p-values for all analyzed descriptors, is available in Appendix E.

As it is especially observable from the buried volume results, most of the significantly differing descriptors describe the catalyst around the metal centre. Since this is where the substrate-metal contact takes place, the ANOVA analysis indicates that the differences between CREST and DFT calculations for these features are not negligible. However, it is key to mention that the impact of the level of theory cannot be fully determined without knowing the feature importance in the ML model. It is possible that the descriptors showing significant differences based on the level of theory may not be important in the ML model, and therefore, the level of theory might not have a significant impact.

5

Conclusion & Outlook

5.1 Conclusion

The aim of this study was to explore the possibility of capturing the dynamic behavior of catalyst structures in data-driven predictive models using conformer ensembles. Various approaches and methods were evaluated on a dataset consisting mainly of Rh-based TM complexes with bidentate ligands. The main conclusions can be categorized into three sections as presented below: conclusions regarding the tested conformer searching tools, the effects of DFT geometry optimization, and the developed algorithm to select conformers for further DFT refinements.

Conformer searching

The evaluation of various conformer searching engines —CREST, RDKit, and OpenBabel— on our TM dataset showed several key findings. None of these cheminformatics programs were able to correctly recognize bonding information from the Cartesian coordinates of the atoms, necessitating a manual bond modification step. The conformer searching algorithms in RDKit and OpenBabel were found unsuitable for our purposes with the current configurations. In contrast, the CREST conformer searching engine successfully generated conformer ensembles for all examined structures. However, a limitation of CREST is found as conformers were often identified and stored as rotamers, hence do not appear in the final conformer ensemble. Additionally, during the calculations, two structures deviated from their original configuration, losing their biphosphane nature and becoming hemilabile.

DFT geometry optimization

In total, 24 CREST conformer ensembles were subjected to further DFT geometry optimization. After DFT refinement, many of the CREST conformers converged into the same DFT local minimum, leading to significantly fewer conformers in the DFT ensemble than in the CREST ensemble. Furthermore, a limitation of the RMSD calculation algorithm of

Morfeus was found, resulting in higher RMSD values between the conformer geometries than the actual value. By analyzing the obtained descriptors with a one-way ANOVA test, most geometric and steric descriptors did not show a significant difference based on the applied level of theory.

Filtering approach

The final collection of conformer ensembles contained 7024 conformers, raising the costs of DFT geometry optimization to around 110,000 euros. Multiple algorithms were tested to identify a subset of conformers based on GFN2-xTB parameters that can accurately represent the entire DFT ensemble. These unsupervised algorithms used the parameters of the CREST conformers to assign binary classes: include or eliminate conformer from the new subset. The approaches mainly differed in the conformer parameters that were taken into account as input features. After assessing the eliminated redundant conformers as well as the missed DFT minima, the best performing model was found to be DBSCAN clustering ($\epsilon = 0.11$) on GFN2-xTB energy and RMSD of the conformer geometries. The model successfully eliminated 46% of the redundant conformers. A dataset of 44 structures containing methyl 2-acetamidoacrylate substrate with the same metal-ligand backbone was used to test the transferability of the model. However, less accurate performance was observed with only eliminating 27% of redundant conformers. The model missed 7% of the total DFT minima of the ensembles. By applying the obtained classification approach to the full dataset of 192 conformer ensembles, the costs for DFT calculations can be reduced by 50,000 euros.

5.2 Outlook

The general approach of an automated catalyst design workflow can be completed in two steps related to conformer searching, as illustrated in Figure 5.1.

After the initial digital representation of the input structure, a conformer searching step can be implemented using CREST. The developed clustering algorithm can then be utilized to select a subset of conformers for further DFT refinement. In this manner, the featurization of the catalyst will not be based on a static molecule, but will consider the geometric and energetic differences between the conformers. This approach aims to produce more accurate predictive models that better describe the correlation between the structure and the catalytic performance. For smooth application of these steps, the following modifications and improvements would be advantageous:

Automation

The current conformer searching tools are inadequate for these TM complexes without requiring manual corrections. Ideally, a fully automated, high-throughput conformer searching algorithm is desired. To achieve this, several improvements can be implemented, enhancing efficiency and reliability of the existing method.

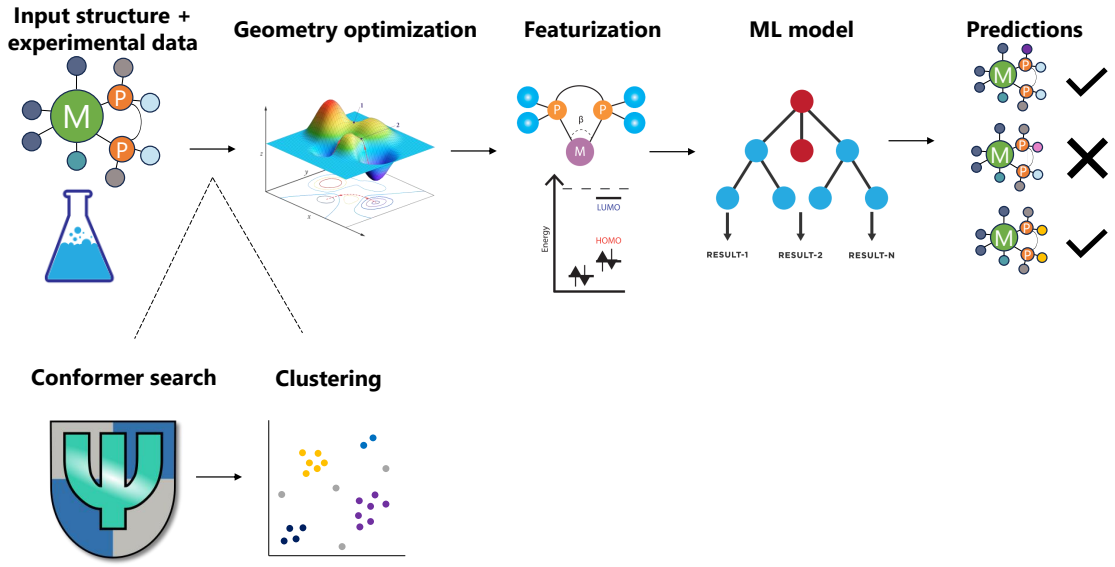


Figure 5.1: Additional steps for a predictive model: conformer searching via CREST and clustering for conformer filtering for DFT.

Firstly, a Python script should be developed to correct erroneous bonding information from cheminformatics packages. Since RDKit showed better performance in bond recognition than OpenBabel, it should be used to generate a mol object from an *.xyz file. The script would then screen atoms and detect incorrect bonds based on the atoms' coordination numbers. By analyzing neighboring atoms' coordinations, the script could accurately identify and remove erroneous bonds, facilitating the conversion of *.xyz files into representations with correct bonding details without additional chemical expertise.

Secondly, a script could be designed to screen the CREST conformer and rotamer lists and access their relative energies. By reselecting conformers based on these energy differences, it could address issues caused by CREST's tendency to store conformers in the romater list.

Finally, another script could be developed to recognize detached ligands. By setting a distance threshold that two original molecules should not exceed after conformer search, this script would flag cases where the structure of conformers significantly deviate from the original structure.

Algorithm optimization

Several modifications can enhance the performance of the developed method. First of all, the descriptors of conformers are currently stored in a large dataframe. However, storing this data in a tensor offers several advantages. This multidimensional matrix would enable more effective data manipulation and structured representation in a compact format that is compatible with ML models.

Secondly, the conformer classification method can be further optimized by replacing

the currently used dynamic energy threshold with a fixed one to distinguish conformers based on their DFT minima. This change would address the current limitation that conformers within a very narrow energy range of the DFT ensemble are mistakenly identified as distinct conformers in the DFT ensemble.

Furthermore, improving the RMSD algorithm could enhance the clustering accuracy, as the current models either do not reorder atom numbers or do not consider bonding information. A new script could be developed to reorder atoms by pairing atoms that are closest in distance and share the same bonding information. This approach would help ensure that atoms are correctly aligned before computing RMSD values.

Lastly, the current feature selection method relies on chemical intuition. Dimensionality reduction tools such as PCA or UMAP could be applied to the complete set of parameters from CREST conformers, including RMSD and descriptors. This approach would allow clustering in the PCA space that is based on the entire parameter set rather than just xTB energy-RMSD relationships. This method allows could result in more robust, data-driven insights based on a comprehensive dataset that would reduce the bias that stems from relying solely on chemical intuition.

Transformer model

The final part of the study originally aimed to design and build a transformer model on the data. The approach for applying this model to our problem is illustrated in Figure 5.2. 5.1. The model is designed to utilize the descriptors of the CREST conformers as input to

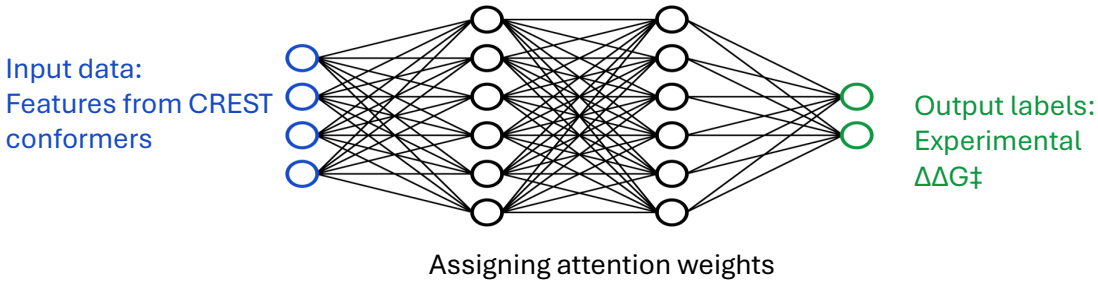


Figure 5.2: An example of a transformer model architecture that could be used on the combination of computational and experimental data. On the left (in blue): the conformer descriptors obtained by CREST is used as input features. The middle part shows the hidden layers where attention weights are assigned. On the right (in green) the experimentally obtained enantioselectivity ($\Delta\Delta G^{\ddagger}$) is utilized as output. The figure of the transformer model is taken from [161].

predict experimental labels, select significant conformers, and identify important features.

The input data of CREST conformer descriptors was transformed into a sequential format where all features of conformers within an ensemble were concatenated together (Figure 5.3. All conformers underwent DFT single-point calculations to use more accu-

Ligand 1	Conformer 1 features	Conformer 2 features	Conformer 3 features	Label
Ligand 2	Conformer 1 features	Conformer 2 features	Conformer 3 features	Label
Ligand 3	Conformer 1 features	Conformer 2 features	Conformer 3 features	Label
Ligand 4	Conformer 1 features	Conformer 2 features	000000000	Label

Figure 5.3: Structure of input data: all conformer features of a structure are concatenated into one line.

rate electronic descriptors. Due to the uneven number of conformers across the ensembles, padding was applied to ensure uniformity. Each structure was experimentally tested with five different substrates, this information was incorporated into the feature list using one-hot encoding. The enantioselectivity was chosen as catalyst performance indicator parameter and therefore used as output label. To standardize the dataset, a standard scaling procedure was applied to the descriptors, ensuring uniform data ranges with a mean of 0 and a standard deviation of 1. A three way data splitting was utilized: 70% training data - 20% validation data - 10% test data. Unfortunately, due to the time limit of the project, further steps were not completed. The planned architecture involved a BERT model, as it was successfully applied in other fields of chemistry (Section 2.2.5). Hyperparameter optimization would have been conducted using the validation set, and the test set was designed for final performance evaluation. A main advantage of this approach is the ability to select conformers within the ensembles that notably influence the catalyst performance indicator (output label) by analyzing the attention weights assigned by the model.

Additionally, descriptors from the 24 ensembles subjected to DFT geometry optimization could have been used for transfer learning. This approach would include the fine-tuning of the pretrained BERT model on the DFT descriptors. By evaluating the prediction accuracy, the the impact of DFT descriptors can be observed.

Acknowledgements

This thesis has been quite a journey. First of all, I would like to thank my professor, Evgeny Pidko, for his support throughout the process. Thank you for the endless assistance, valuable advice, and always being there for me.

Special thanks go to my daily supervisor, Adarsh Kalikadien, who was always ready to discuss to my newest ideas and results. Thank you for accepting me as your Master's student and taught me so so much along the way. Without you, this thesis wouldn't have been possible.

I would also like to thank everybody in the ISE research group for their feedback, discussions, and for creating a fun and supportive environment.

I would like to thank to all my friends and family to share my passion for coding and listen to my struggles.

Lastly, I would like to thank the facilities at TU Delft for enabling me to finally learn to program with an English keyboard. This has been a huge improvement for me.

Bibliography

[1] Piet WNM Van Leeuwen. *Homogeneous catalysis: understanding the art*. Springer Science & Business Media, 2006.

[2] James T Richardson. *Principles of catalyst development*. Springer, 2013.

[3] José Luís Figueiredo Joaquim Faria, Mariette M Pereira, and Joaquim Faria. *Catalysis from Theory to Application: An Integrated Course: An Integrated Course*. Imprensa da Universidade de Coimbra/Coimbra University Press, 2008.

[4] RV Chaudhari. "Fundamentals of homogeneous catalysis". In: *Industrial Catalytic Processes for Fine and Specialty Chemicals*. Elsevier, 2016, pp. 17–39.

[5] John N Armor. "A history of industrial catalysis". In: *Catalysis Today* 163.1 (2011), pp. 3–9.

[6] Ademola Soyemi and Tibor Szilvási. "Trends in computational molecular catalyst design". In: *Dalton Transactions* 50.30 (2021), pp. 10325–10339.

[7] Erica Farnetti, Roberta Di Monte, and Jan Kašpar. "Homogeneous and heterogeneous catalysis". In: *Inorganic and bio-inorganic chemistry* 2.6 (2009), pp. 50–86.

[8] Ana I Benítez-Mateos, Martina L Contente, David Roura Padrosa, and Francesca Paradisi. "Flow biocatalysis 101: design, development and applications". In: *Reaction chemistry & engineering* 6.4 (2021), pp. 599–611.

[9] Sumit Bhaduri and Doble Mukesh. "Chemical industry and homogeneous catalysis". In: *Homogeneous Catalysis* (2014), pp. 1–21.

[10] Emad L Izake. "Chiral discrimination and enantioselective analysis of drugs: an overview". In: *Journal of Pharmaceutical Sciences* 96.7 (2007), pp. 1659–1676.

[11] Barbara Kasprzyk-Hordern. "Pharmacologically active compounds in the environment and their chirality". In: *Chemical Society Reviews* 39.11 (2010), pp. 4466–4503.

[12] Carl A Busacca, Daniel R Fandrick, Jinhua J Song, and Chris H Senanayake. "Transition metal catalysis in the pharmaceutical industry". In: *Applications of Transition Metal Catalysis in Drug Discovery and Development: An Industrial Perspective* (2012), pp. 1–24.

[13] Nicolas Fleury-Brégeot, Verónica de la Fuente, Sergio Castillón, and Carmen Claver. "Highlights of Transition Metal-Catalyzed Asymmetric Hydrogenation of Imines". In: *ChemCatChem* 2.11 (2010), pp. 1346–1371.

[14] Jianmin Mao and David C Baker. "A chiral rhodium complex for rapid asymmetric transfer hydrogenation of imines with high enantioselectivity". In: *Organic Letters* 1.6 (1999), pp. 841–843.

[15] Diego Ghislieri and Nicholas J Turner. "Biocatalytic approaches to the synthesis of enantiomerically pure chiral amines". In: *Topics in Catalysis* 57 (2014), pp. 284–300.

[16] Albert Cabré, Xavier Verdaguer, and Antoni Riera. "Recent advances in the enantioselective synthesis of chiral amines via transition metal-catalyzed asymmetric hydrogenation". In: *Chemical Reviews* 122.1 (2021), pp. 269–339.

[17] Jordan J Dotson, Lucy van Dijk, Jacob C Timmerman, Samantha Grosslight, Richard C Walroth, Francis Gosselin, Kurt Püntener, Kyle A Mack, and Matthew S Sigman. "Data-driven multi-objective optimization tactics for catalytic asymmetric reactions using bisphosphine ligands". In: *Journal of the American Chemical Society* 145.1 (2022), pp. 110–121.

[18] Derek J Durand and Natalie Fey. "Computational ligand descriptors for catalyst design". In: *Chemical reviews* 119.11 (2019), pp. 6561–6594.

[19] Aditya Nandy, Chenru Duan, Michael G Taylor, Fang Liu, Adam H Steeves, and Heather J Kulik. "Computational discovery of transition-metal complexes: from high-throughput screening to machine learning". In: *Chemical Reviews* 121.16 (2021), pp. 9927–10000.

[20] John M Newsam and F Schüth. "Combinatorial approaches as a component of high-throughput experimentation (HTE) in catalysis research". In: *Biotechnology and bioengineering* 61.4 (1999), pp. 203–216.

[21] Marco Foscato and Vidar R Jensen. "Automated in silico design of homogeneous catalysts". In: *ACS catalysis* 10.3 (2020), pp. 2354–2377.

[22] Stephen Lower and Tom Neils. *The Potential Energy Surface Can Be Calculated Using Quantum Mechanics*. 2024. URL: https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Physical_Chemistry_(LibreTexts)/30%3A_Gas-Phase_Reaction_Dynamics/30.10%3A_The_Potential-Energy_Surface_Can_Be_Calculated_Using_Quantum_Mechanics.

[23] Alexandre V Brethome, Stephen P Fletcher, and Robert S Paton. "Conformational effects on physical-organic descriptors: the case of sterimol steric parameters". In: *ACS Catalysis* 9.3 (2019), pp. 2313–2323.

[24] Liliana C Gallegos, Guilian Luchini, Peter C St. John, Seonah Kim, and Robert S Paton. "Importance of engineered and learned molecular representations in predicting organic reactivity, selectivity, and chemical properties". In: *Accounts of Chemical Research* 54.4 (2021), pp. 827–836.

[25] Katrin Köhnke, Niklas Wessel, Jesús Esteban, Jing Jin, Andreas J Vorholt, and Walter Leitner. "Operando monitoring of mechanisms and deactivation of molecular catalysts". In: *Green Chemistry* 24.5 (2022), pp. 1951–1972.

[26] Rubén Laplaza, Jan-Grimo Sobez, Matthew D Wodrich, Markus Reiher, and Clémence Corminboeuf. "The (not so) simple prediction of enantioselectivity–a pipeline for high-fidelity computations". In: *Chemical Science* 13.23 (2022), pp. 6858–6864.

[27] Christopher Masters. *Homogeneous transition-metal catalysis: a gentle art*. Springer Science & Business Media, 2012.

[28] James Keeler and Peter Wothers. *Chemical structure and reactivity: an integrated approach*. Oxford University Press, USA, 2013.

[29] Paul CJ Kamer and Piet WNM van Leeuwen. *Phosphorus (III) ligands in homogeneous catalysis: design and synthesis*. John Wiley & Sons, 2012.

[30] Jack Halpern. "Homogeneous catalysis by coordination compounds". In: ACS Publications, 1968.

[31] Yves Jean. *Molecular orbitals of transition metal complexes*. OUP Oxford, 2005.

[32] Susan Lühr, Jens Holz, and Armin Börner. "The synthesis of chiral phosphorus ligands for use in homogeneous metal catalysis". In: *ChemCatChem* 3.11 (2011), pp. 1708–1730.

[33] J Carles Bayón, Carmen Claver, and Anna M Masdeu-Bultó. "Homogeneous catalysis with transition metal complexes containing sulfur ligands". In: *Coordination Chemistry Reviews* 193 (1999), pp. 73–145.

[34] Fabienne Fache, Emmanuelle Schulz, M Lorraine Tommasino, and Marc Lemaire. "Nitrogen-containing ligands for asymmetric homogeneous and heterogeneous catalysis". In: *Chemical Reviews* 100.6 (2000), pp. 2159–2232.

[35] David J Berrisford, Carsten Bolm, and K Barry Sharpless. "Ligand-accelerated catalysis". In: *Angewandte Chemie International Edition in English* 34.10 (1995), pp. 1059–1070.

[36] Adarsh V Kalikadien, Cecile Valsecchi, Robbert van Putten, Tor Maes, Mikko Muuronen, Natalia Dyubankova, Laurent Lefort, and Evgeny A Pidko. "Probing Machine Learning Models Based on High Throughput Experimentation Data for the Discovery of Asymmetric Hydrogenation Catalysts". In: *submitted* ().

[37] Michael P Maloney, Brock A Stenfors, Paul Helquist, Per-Ola Norrby, and Olaf Wiest. "Interplay of Computation and Experiment in Enantioselective Catalysis: Rationalization, Prediction, and Correction?" In: *ACS Catalysis* 13.21 (2023), pp. 14285–14299.

[38] Simone Gallarati, Raimon Fabregat, Rubén Laplaza, Sinjini Bhattacharjee, Matthew D Wodrich, and Clemence Corminboeuf. "Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts". In: *Chemical Science* 12.20 (2021), pp. 6879–6889.

[39] Takashi Toyao, Zen Maeno, Satoru Takakusagi, Takashi Kamachi, Ichigaku Takigawa, and Ken-ichi Shimizu. "Machine learning for catalysis informatics: recent applications and prospects". In: *Acs Catalysis* 10.3 (2019), pp. 2260–2297.

[40] Adarsh V Kalikadien, Adrian Mirza, Aydin Najl Hossaini, Avadakkam Sreenithya, and Evgeny A Pidko. "Paving the road towards automated homogeneous catalyst design". In: *ChemPlusChem* (2024), e202300702.

[41] Kendall N Houk and Paul Ha-Yeon Cheong. "Computational prediction of small-molecule catalysts". In: *Nature* 455.7211 (2008), pp. 309–313.

[42] Jesús Jover and Natalie Fey. "The computational road to better catalysts". In: *Chemistry–An Asian Journal* 9.7 (2014), pp. 1714–1723.

[43] Wenhong Yang, Timothy Tizhe Fidelis, and Wen-Hua Sun. "Machine learning in catalysis, from proposal to practicing". In: *ACS omega* 5.1 (2019), pp. 83–88.

[44] Daniel S Wigh, Jonathan M Goodman, and Alexei A Lapkin. "A review of molecular representation in the age of machine learning". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 12.5 (2022), e1603.

[45] David Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *Journal of chemical information and computer sciences* 28.1 (1988), pp. 31–36.

[46] Heather J Kulik. "Making machine learning a useful tool in the accelerated discovery of transition metal complexes". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 10.1 (2020), e1439.

[47] Ajnabiul Hoque and Raghavan B Sunoj. "Deep learning for enantioselectivity predictions in catalytic asymmetric β-C–H bond activation reactions". In: *Digital Discovery* 1.6 (2022), pp. 926–940.

[48] Giuseppe Antinucci, Busra Dereli, Antonio Vittoria, Peter HM Budzelaar, Roberta Cipullo, Georgy P Goryunov, Pavel S Kulyabin, Dmitry V Uborsky, Luigi Cavallo, Christian Ehm, et al. "Selection of low-dimensional 3-D geometric descriptors for accurate enantioselectivity prediction". In: *ACS Catalysis* 12.12 (2022), pp. 6934–6945.

[49] Yovani Marrero-Ponce, Oscar Martínez Santiago, Yoan Martínez López, Stephen J Barigye, and Francisco Torrens. "Derivatives in discrete mathematics: a novel graph-theoretical invariant for generating new 2/3D molecular descriptors. I. Theory and QSPR application". In: *Journal of computer-aided molecular design* 26.11 (2012), pp. 1229–1246.

[50] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors*. John Wiley & Sons, 2008.

[51] Alan R Katritzky and Ekaterina V Gordeeva. "Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research". In: *Journal of chemical information and computer sciences* 33.6 (1993), pp. 835–857.

[52] Boegel Horst. "Molecular Descriptors and the Electronic Structure". In: *Statistical Modelling of Molecular Descriptors in QSAR/QSPR* 2 (2012), pp. 245–292.

[53] Florbela Pereira, Kaixia Xiao, Diogo ARS Latino, Chengcheng Wu, Qingyou Zhang, and Joao Aires-de-Sousa. "Machine learning methods to predict density functional theory B3LYP energies of HOMO and LUMO orbitals". In: *Journal of chemical information and modeling* 57.1 (2017), pp. 11–21.

[54] Marzieh Miar, Abolfazl Shiroudi, Khalil Pourshamsian, Ahmad Reza Oliaey, and Farhad Hatamjafari. "Theoretical investigations on the HOMO–LUMO gap and global reactivity descriptor studies, natural bond orbital, and nucleus-independent chemical shifts analyses of 3-phenylbenzo [d] thiazole-2 (3 H)-imine and its para-substituted derivatives: Solvent and substituent effects". In: *Journal of Chemical Research* 45.1-2 (2021), pp. 147–158.

[55] Adrián Gómez-Suárez, David J Nelson, and Steven P Nolan. "Quantifying and understanding the steric properties of N-heterocyclic carbenes". In: *Chemical communications* 53.18 (2017), pp. 2650–2660.

[56] Peter Dierkes and Piet WNM van Leeuwen. "The bite angle makes the difference: a practical ligand parameter for diphosphine ligands". In: *Journal of the Chemical Society, Dalton Transactions* 10 (1999), pp. 1519–1530.

[57] Jenna A Bilbrey, Arianna H Kazez, Jason Locklin, and Wesley D Allen. "Exact ligand cone angles". In: *Journal of computational chemistry* 34.14 (2013), pp. 1189–1197.

[58] Chadwick A Tolman. "Steric effects of phosphorus ligands in organometallic chemistry and homogeneous catalysis". In: *Chemical reviews* 77.3 (1977), pp. 313–348.

[59] Ernest R Davidson. "Computational transition metal chemistry". In: *Chemical reviews* 100.2 (2000), pp. 351–352.

[60] Markus Bursch, Jan-Michael Mewes, Andreas Hansen, and Stefan Grimme. "Best-practice DFT protocols for basic molecular computational chemistry". In: *Angewandte Chemie International Edition* 61.42 (2022), e202205735.

[61] Jeremy N Harvey, Fahmi Himo, Feliu Maseras, and Lionel Perrin. "Scope and challenge of computational methods for studying mechanism and reactivity in homogeneous catalysis". In: *Acs Catalysis* 9.8 (2019), pp. 6803–6813.

[62] Maria Besora and Feliu Maseras. "Microkinetic modeling in homogeneous catalysis". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 8.6 (2018), e1372.

[63] Yuta Hori and Tsukasa Abe. "Theoretical Approach to Homogeneous Catalyst of Methane Hydroxylation: Collaboration with Computation and Experiment". In: *Direct Hydroxylation of Methane: Interplay Between Theory and Experiment* (2020), pp. 151–165.

[64] Jialing Lan, Xin Li, Yuhong Yang, Xiaoyong Zhang, and Lung Wa Chung. "New insights and predictions into complex homogeneous reactions enabled by computational chemistry in synergy with experiments: isotopes and mechanisms". In: *Accounts of Chemical Research* 55.8 (2022), pp. 1109–1123.

[65] Valeria Butera. "Density Functional Theory Methods applied to Homogeneous and Heterogeneous Catalysis: a Short Review and a Practical User Guide". In: *Physical Chemistry Chemical Physics* (2024).

[66] Libero J Bartolotti and Ken Flurchick. "An introduction to density functional theory". In: *Reviews in computational chemistry* (1996), pp. 187–216.

[67] Wolfram Koch and Max C Holthausen. *A chemist's guide to density functional theory*. John Wiley & Sons, 2015.

[68] Attila Szabo and Neil S Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.

[69] Mohammad Reza Akbari, Sara Akbari, and Esmaeil Kalantari. "The schrödinger nonlinear partial differential equation solution in quantum physic by new approach aym". In: *Quantum Journal of Engineering, Science and Technology* 2.2 (2021), pp. 40–46.

[70] Yusuke Nomura and Ryosuke Akashi. "Density functional theory". In: *arXiv preprint arXiv:2210.07647* (2022).

[71] Frank Neese. "Prediction of molecular properties and molecular spectroscopy with density functional theory: From fundamental theory to exchange-coupling". In: *Coordination Chemistry Reviews* 253.5-6 (2009), pp. 526–563.

[72] Narbe Mardirossian and Martin Head-Gordon. "Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals". In: *Molecular physics* 115.19 (2017), pp. 2315–2372.

[73] John P Perdew. "Climbing the ladder of density functional approximations". In: *MRS bulletin* 38.9 (2013), pp. 743–750.

[74] Paul Ziesche, Stefan Kurth, and John P Perdew. "Density functionals from LDA to GGA". In: *Computational materials science* 11.2 (1998), pp. 122–127.

[75] Jeremy N Harvey. "On the accuracy of density functional theory in transition metal chemistry". In: *Annual Reports Section" C"(Physical Chemistry)* 102 (2006), pp. 203–226.

[76] Michael Buhl, Christoph Reimann, Dimitrios A Pantazis, Thomas Bredow, and Frank Neese. "Geometries of third-row transition-metal complexes from density-functional theory". In: *Journal of chemical theory and computation* 4.9 (2008), pp. 1449–1459.

[77] John P Perdew, Kieron Burke, and Matthias Ernzerhof. "Generalized gradient approximation made simple". In: *Physical review letters* 77.18 (1996), p. 3865.

[78] Balazs Nagy and Frank Jensen. "Basis sets in quantum chemistry". In: *Reviews in Computational Chemistry* 30 (2017), pp. 93–149.

[79] Ronit Sarangi, Marta L Vidal, Sonia Coriani, and Anna I Krylov. "On the basis set selection for calculations of core-level states: Different strategies to balance cost and accuracy". In: *Molecular Physics* 118.19-20 (2020), e1769872.

[80] John C Slater. "Atomic shielding constants". In: *Physical review* 36.1 (1930), p. 57.

[81] Frank Jensen. *Introduction to computational chemistry*. John wiley & sons, 2017.

[82] Roman M Balabin. "Enthalpy difference between conformations of normal alkanes: Intramolecular basis set superposition error (BSSE) in the case of n-butane and n-hexane". In: *The Journal of chemical physics* 129.16 (2008).

[83] Milena Palhares Maringolo, Ana Cristina Mora Tello, Amanda Ribeiro Guimaraes, Júlia Maria Aragon Alves, Francisco das Chagas Alves Lima, Elson Longo, and Albérico Borges Ferreira da Silva. "On polarization functions for Gaussian basis sets". In: *Journal of Molecular Modeling* 26 (2020), pp. 1–7.

[84] Neirigelson Ferreira de Barros Leite, Rosemarie Brandim Marques, Antonio Macedo-Filho, Gerd Bruno Rocha, and Evandro PS Martins. "Evaluation of DFT methods for predicting geometries and NMR spectra of Bi (III) dithiocarbamate complexes with antitumor properties". In: *Journal of Molecular Modeling* 30.6 (2024), pp. 1–11.

[85] Stefan Grimme. "Density functional theory with London dispersion corrections". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.2 (2011), pp. 211–228.

[86] Lars Goerigk. "A comprehensive overview of the DFT-D3 London-dispersion correction". In: *Non-covalent interactions in quantum chemistry and physics* (2017), pp. 195–219.

[87] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu". In: *The Journal of chemical physics* 132.15 (2010).

[88] J Ulises Reveles and Andreas M Köster. "Geometry optimization in density functional methods". In: *Journal of computational chemistry* 25.9 (2004), pp. 1109–1116.

[89] Sandro E Schönborn, Stefan Goedecker, Shantanu Roy, and Artem R Oganov. "The performance of minima hopping and evolutionary algorithms for cluster structure prediction". In: *The Journal of chemical physics* 130.14 (2009).

[90] SK Lai and Wafa Maftuhin. "An efficient optimization algorithm that hybridizes DFTB and DFT theories both operated within the modified basin hopping method". In: *Computer Physics Communications* 236 (2019), pp. 164–175.

[91] Francesco Fracchia, Gianluca Del Frate, Giordano Mancini, Walter Rocchia, and Vincenzo Barone. "Force field parametrization of metal ions from statistical learning techniques". In: *Journal of chemical theory and computation* 14.1 (2018), pp. 255–273.

[92] Anthony K Rappé, Carla J Casewit, KS Colwell, William A Goddard III, and W Mason Skiff. "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations". In: *Journal of the American chemical society* 114.25 (1992), pp. 10024–10035.

[93] Yury Minenkov, Dmitry I Sharapa, and Luigi Cavallo. "Application of semiempirical methods to transition metal complexes: Fast results but hard-to-predict accuracy". In: *Journal of Chemical Theory and Computation* 14.7 (2018), pp. 3428–3439.

[94] AK Rappe, KS Colwell, and CJ Casewit. "Application of a universal force field to metal complexes". In: *Inorganic Chemistry* 32.16 (1993), pp. 3438–3450.

[95] Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. "Extended tight-binding quantum chemistry methods". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 11.2 (2021), e1493.

[96] Stefan Grimme, Christoph Bannwarth, and Philip Shushkov. "A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (Z= 1–86)". In: *Journal of chemical theory and computation* 13.5 (2017), pp. 1989–2009.

[97] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. "GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions". In: *Journal of chemical theory and computation* 15.3 (2019), pp. 1652–1671.

[98] Stefan Grimme. "Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations". In: *Journal of chemical theory and computation* 15.5 (2019), pp. 2847–2862.

[99] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. "Open Babel: An open chemical toolbox". In: *Journal of cheminformatics* 3 (2011), pp. 1–14.

[100] Thomas Engel. "Basic overview of chemoinformatics". In: *Journal of chemical information and modeling* 46.6 (2006), pp. 2267–2277.

[101] Andrew R Leach and Valerie J Gillet. *An introduction to chemoinformatics*. Springer, 2007.

[102] Nil Sanosa, David Dalmau, Diego Sampedro, Juan V Alegre-Requena, and Ignacio Funes-Ardoiz. "Recent advances of machine learning applications in the development of experimental homogeneous catalysis". In: *Artificial Intelligence Chemistry* (2024), p. 100068.

[103] Greg Landrum. "Rdkit documentation". In: *Release* 1.1-79 (2013), p. 4.

[104] Lauren C Burrows, Luke T Jesikiewicz, Gang Lu, Steven J Geib, Peng Liu, and Kay M Brummond. "Computationally guided catalyst design in the type I dynamic kinetic asymmetric Pauson–Khand reaction of allenyl acetates". In: *Journal of the American Chemical Society* 139.42 (2017), pp. 15022–15032.

[105] Markus Bursch, Andreas Hansen, Philipp Pracht, Julia T Kohn, and Stefan Grimme. "Theoretical study on conformational energies of transition metal complexes". In: *Physical Chemistry Chemical Physics* 23.1 (2021), pp. 287–299.

[106] Ruben Laplaza, Matthew D Wodrich, and Clemence Corminboeuf. "Overcoming the Pitfalls of Computing Reaction Selectivity from Ensembles of Transition States". In: (2024).

[107] Susanta Das and Kenneth M Merz Jr. "Molecular Gas-Phase Conformational Ensembles". In: *Journal of Chemical Information and Modeling* (2023).

[108] Zhen Liu, Tetiana Zubatiuk, Adrian Roitberg, and Olexandr Isayev. "Auto3d: Automatic generation of the low-energy 3d structures with ANI neural network potentials". In: *Journal of Chemical Information and Modeling* 62.22 (2022), pp. 5373–5382.

[109] Paolo Tosco, Nikolaus Stiefl, and Gregory Landrum. "Bringing the MMFF force field to the RDKit: implementation and validation". In: *Journal of cheminformatics* 6 (2014), pp. 1–4.

[110] Philipp Pracht, Stefan Grimme, Christoph Bannwarth, Fabian Bohle, Sebastian Ehlert, Gereon Feldmann, Johannes Gorges, Marcel Müller, Tim Neudecker, Christoph Plett, et al. "CREST—A program for the exploration of low-energy molecular chemical space". In: *The Journal of Chemical Physics* 160.11 (2024).

[111] Pak L Fung, Martha A Zaidan, Hilkka Timonen, Jarkko V Niemi, Anu Kousa, Joel Kuula, Krista Luoma, Sasu Tarkoma, Tuukka Petäjä, Markku Kulmala, et al. "Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration". In: *Journal of aerosol science* 152 (2021), p. 105694.

[112] Michael Affenzeller, Bogdan Burlacu, Viktoria Dorfer, Sebastian Dorl, Gerhard Halmerbauer, Tilman Königswieser, Michael Kommenda, Julia Vetter, and Stephan Winkler. "White box vs. black box modeling: On the performance of deep learning, random forests, and symbolic regression in solving regression problems". In: *Computer Aided Systems Theory–EUROCAST 2019: 17th International Conference, Las Palmas de Gran Canaria, Spain, February 17–22, 2019, Revised Selected Papers, Part I 17*. Springer. 2020, pp. 288–295.

[113] Eduardo F Morales and Hugo Jair Escalante. "A brief introduction to supervised, unsupervised, and reinforcement learning". In: *Biosignal processing and classification using computational learning and intelligence*. Elsevier, 2022, pp. 111–129.

[114] Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. "Predicting reaction performance in C–N cross-coupling using machine learning". In: *Science* 360.6385 (2018), pp. 186–190.

[115] Andrew F Zahrt, Jeremy J Henle, Brennan T Rose, Yang Wang, William T Darrow, and Scott E Denmark. "Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning". In: *Science* 363.6424 (2019), eaau5631.

[116] Matthew Welborn, Lixue Cheng, and Thomas F Miller III. "Transferability in machine learning for electronic structure via the molecular orbital basis". In: *Journal of chemical theory and computation* 14.9 (2018), pp. 4772–4779.

[117] Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou. "Feature dimensionality reduction: a review". In: *Complex & Intelligent Systems* 8.3 (2022), pp. 2663–2693.

[118] Stavros K Kariofillis, Shutian Jiang, Andrzej M Żurański, Shivaani S Gandhi, Jesus I Martinez Alvarado, and Abigail G Doyle. "Using data science to guide aryl bromide substrate scope analysis in a Ni/photoredox-catalyzed cross-coupling with acetals as alcohol-derived radical sources". In: *Journal of the American Chemical Society* 144.2 (2022), pp. 1045–1055.

[119] Isaiah O Betinol, Junshan Lai, Saumya Thakur, and Jolene P Reid. "A data-driven workflow for assigning and predicting generality in asymmetric catalysis". In: *Journal of the American Chemical Society* 145.23 (2023), pp. 12870–12883.

[120] Lior Rokach and Oded Maimon. "Clustering methods". In: *Data mining and knowledge discovery handbook* (2005), pp. 321–352.

[121] Julian A Hueffel, Theresa Sperger, Ignacio Funes-Ardoiz, Jas S Ward, Kari Rissanen, and Franziska Schoenebeck. "Accelerated dinuclear palladium catalyst identification through unsupervised machine learning". In: *Science* 374.6571 (2021), pp. 1134–1140.

[122] Lu Zhang and Haiyan Liu. "Exploring binding positions and backbone conformations of peptide ligands of proteins with a backbone-centred statistical energy function". In: *Journal of Computer-Aided Molecular Design* 37.10 (2023), pp. 463–478.

[123] Janka Mátrai, Willem Lammens, Abel Jonckheer, Katrien Le Roy, Anja Rabijns, Wim Van den Ende, and Marc De Maeyer. "An alternate sucrose binding mode in the E203Q Arabidopsis invertase mutant: An X-ray crystallography and docking study". In: *Proteins: Structure, Function, and Bioinformatics* 71.2 (2008), pp. 552–564.

[124] Nithin Buduma, Nikhil Buduma, and Joe Papa. *Fundamentals of deep learning*. " O'Reilly Media, Inc.", 2022.

[125] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[126] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. "Accurate prediction of protein structures and interactions using a three-track neural network". In: *Science* 373.6557 (2021), pp. 871–876.

[127] Afnan Sultan, Jochen Sieg, Miriam Mathea, and Andrea Volkamer. "Transformers for molecular property prediction: Lessons learned from the past five years". In: *arXiv preprint arXiv:2404.03969* (2024).

[128] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. "Transformer models for text-based emotion detection: a review of BERT-based approaches". In: *Artificial Intelligence Review* 54.8 (2021), pp. 5789–5829.

[129] Jiazhen He, Eva Nittinger, Christian Tyrchan, Werngard Czechtizky, Atanas Patronov, Esben Jannik Bjerrum, and Ola Engkvist. "Transformer-based molecular optimization beyond matched molecular pairs". In: *Journal of cheminformatics* 14.1 (2022), p. 18.

[130] Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, et al. "SELFIES and the future of molecular string representations". In: *Patterns* 3.10 (2022).

[131] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589.

[132] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. "Highly accurate protein structure prediction for the human proteome". In: *Nature* 596.7873 (2021), pp. 590–596.

[133] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[134] Philippe Schwaller, Alain C Vaucher, Teodoro Laino, and Jean-Louis Reymond. "Prediction of chemical reaction yields using deep learning". In: *Machine learning: science and technology* 2.1 (2021), p. 015016.

[135] Mario Latendresse, Jeremiah P Malerich, Mike Travers, and Peter D Karp. "Accurate atom-mapping computation for biochemical reactions". In: *Journal of chemical information and modeling* 52.11 (2012), pp. 2970–2982.

[136] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. "Albert: A lite bert for self-supervised learning of language representations". In: *arXiv preprint arXiv:1909.11942* (2019).

[137] Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. "Extraction of organic chemistry grammar from unsupervised learning of chemical reactions". In: *Science Advances* 7.15 (2021), eabe4166.

[138] *CREST Documentation*. https://crest-lab.github.io/crest-docs/. [Accessed: 2024-06-13]. 2024.

[139] *RDKit*. URL: https://www.rdkit.org/.

[140] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. "An anova test for functional data". In: *Computational statistics & data analysis* 47.1 (2004), pp. 111–122.

[141] *Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase 1)*. https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1. 2022.

[142] *Snellius: de Nationale Supercomputer*. https://www.surf.nl/en/dutch-national-supercomputer-snellius. 2023.

[143] *ChemCraft - graphical software for visualization of quantum chemistry computations*. Version 1.8, build 682. 2024. URL: https://www.chemcraftprog.com.

[144] Kjell Jorner, Tobias Gensch, Pascal Friedrich, and Gabriel dos Passos Gomes. *Morfeus: Molecular Features for Machine Learning*. Version 0.7.2. 2022.

[145] Philipp Pracht, Fabian Bohle, and Stefan Grimme. "Automated exploration of the low-energy chemical space with fast quantum chemical methods". In: *Physical Chemistry Chemical Physics* 22.14 (2020), pp. 7169–7192.

[146] Margareth S Baidun, Adarsh V Kalikadien, Laurent Lefort, and Evgeny A Pidko. "Impact of Model Selection and Conformational Effects on the Descriptors for In Silico Screening Campaigns: A Case Study of Rh-Catalyzed Acrylate Hydrogenation". In: *The Journal of Physical Chemistry C* (2024).

[147] Ivan Yu Chernyshov and Evgeny A Pidko. "MACE: Automated Assessment of Stereochemistry of Transition Metal Complexes and Its Applications in Computational Catalysis". In: *Journal of Chemical Theory and Computation* (2024).

[148] *RMSD Python Package*. https://github.com/charnley/rmsd. Accessed: 2024.

[149] Carlo Adamo and Vincenzo Barone. "Toward reliable density functional methods without adjustable parameters: The PBE0 model". In: *The Journal of chemical physics* 110.13 (1999), pp. 6158–6170.

[150] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. "Effect of the damping function in dispersion corrected density functional theory". In: *Journal of computational chemistry* 32.7 (2011), pp. 1456–1465.

[151] Florian Weigend and Reinhart Ahlrichs. "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy". In: *Physical Chemistry Chemical Physics* 7.18 (2005), pp. 3297–3305.

[152] Vivek Sinha, Jochem J Laan, and Evgeny A Pidko. "Accurate and rapid prediction of p K a of transition metal complexes: semiempirical quantum chemistry with a data-augmented approach". In: *Physical Chemistry Chemical Physics* 23.4 (2021), pp. 2557–2567.

[153] Adarsh V Kalikadien, Evgeny A Pidko, and Vivek Sinha. "ChemSpaX: exploration of chemical space by automated functionalization of molecular scaffold". In: *Digital Discovery* 1.1 (2022), pp. 8–25.

[154] Jonathan M Goodman and Marıa A Silva. "QRC: a rapid method for connecting transition structures to reactants in the computational analysis of organic reactivity". In: *Tetrahedron letters* 44.45 (2003), pp. 8233–8236.

[155] María A Silva and Jonathan M Goodman. "Aziridinium ring opening: a simple ionic reaction pathway with sequential transition states". In: *Tetrahedron letters* 46.12 (2005), pp. 2067–2069.

[156] Gopal Behera and Ashok Kumar Bhoi. "General Applicability of K-means Algorithm with Enhanced Centroids". In: *International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS)* 7 (), pp. 201–205.

[157] Hae-Sang Park and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering". In: *Expert systems with applications* 36.2 (2009), pp. 3336–3341.

[158] Kamran Khan, Saif Ur Rehman, Kamran Aziz, Simon Fong, and Sababady Sarasvady. "DBSCAN: Past, present and future". In: *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*. IEEE. 2014, pp. 232–238.

[159] Jingsai Liang. "Confusion matrix: Machine learning". In: *POGIL Activity Clearinghouse* 3.4 (2022).

[160] Tobias Gensch, Gabriel dos Passos Gomes, Pascal Friederich, Ellyn Peters, Théophile Gaudin, Robert Pollice, Kjell Jorner, AkshatKumar Nigam, Michael Lindner-D'Addario, Matthew S Sigman, et al. "A comprehensive discovery platform for organophosphorus ligands for catalysis". In: *Journal of the American Chemical Society* 144.3 (2022), pp. 1205–1217.

[161] Victor Zhou. *Neural Networks From Scratch*. https://victorzhou.com/series/neural-networks-from-scratch/. Accessed: 2024-07-04. 2019.

A

Ligand database

A comprehensive list of the 192 investigated ligands can be found in Table A.1.

Table A.1: List of ligands

Ligand number	Ligand name	CAS	Formula
L1	SL-J001-1	155806-35-2	$C_{36}H_{44}FeP_2$
L2	SL-J002-1	155830-69-6	$C_{32}H_{40}FeP_2$
L3	SL-J003-1	167416-28-6	$C_{36}H_{56}FeP_2$
L4	SL-J004-1	158923-09-2	$C_{36}H_{44}FeP_2$
L5	SL-J005-1	184095-69-0	$C_{40}H_{40}FeP_2$
L6	SL-J006-1	292638-88-1	$C_{40}H_{40}F_{12}FeP_2$
L7	SL-J007-1	360048-63-1	$C_{42}H_{56}FeO_2P_2$
L8	SL-J008-1	166172-63-0	$C_{44}H_{36}F_{12}FeP_2$
L9	SL-J009-1	158923-11-6	$C_{32}H_{52}FeP_2$
L10	SL-J011-1	246231-79-8	$C_{34}H_{38}F_6FeP_2$
L11	SL-J013-1	187733-50-2	$C_{38}H_{52}FeO_2P_2$
L12	SL-J212-1	849924-41-0	$C_{28}H_{36}FeO_2P_2$
L13	SL-J404-1	851308-40-2	$C_{48}H_{44}FeP_2$
L14	SL-J418-1	849924-45-4	$C_{46}H_{52}FeO_2P_2$
L15	SL-J452-1	849924-73-8	$C_{34}H_{32}FeO_2P_2$
L16	SL-J502-1	223120-71-6	$C_{32}H_{40}FeP_2$
L17	(R)-BINAM-P	74974-14-4	$C_{44}H_{34}N_2P_2$
L18	SL-J505-1	849924-76-1	$C_{34}H_{44}FeP_2$
L19	SL-T002-2	914089-00-2	$C_{43}H_{63}FeNP_2$
L20	SL-M001-1	174467-31-3	$C_{52}H_{50}FeN_2P_2$
L21	SL-M003-1	494227-36-0	$C_{60}H_{42}F_{24}FeN_2P_2$
L22	SL-M004-1	494227-37-1	$C_{64}H_{74}FeN_2O_4P_2$
L23	SL-M009-1	793718-16-8	$C_{60}H_{66}FeN_2P_2$
L24	SL-T001-2	850444-36-9	$C_{43}H_{39}FeNP_2$

L25	SL-W001-1	387868-06-6	$C_{46}H_{32}F_{12}FeP_2$
L26	SL-W002-1	388079-58-1	$C_{42}H_{36}FeP_2$
L27	SL-W003-2	849925-19-5	$C_{42}H_{48}FeP_2$
L28	SL-W005-2	849925-20-8	$C_{52}H_{44}F_{12}FeO_2P_2$
L29	SL-W008-2	849925-22-0	$C_{46}H_{44}F_{12}FeP_2$
L30	SL-W009-1	894771-28-9	$C_{50}H_{52}FeP_2$
L31	SL-F356-1	952586-19-5	$C_{42}H_{53}Fe_2NP_2$
L32	(R)-BINAP	76189-55-4	$C_{44}H_{32}P_2$
L33	(R)-BTFM-GarPhos	1365531-84-5	$C_{48}H_{28}F_{24}O_4P_2$
L34	(R)-Tol-BINAP	99646-28-3	$C_{48}H_40P_2$
L35	(R)-Xyl-BINAP	137219-86-4	$C_{52}H_{48}P_2$
L36	(R)-H8-BINAP	139139-86-9	$C_{44}H_40P_2$
L37	(S)-SegPhos	210169-54-3	$C_{38}H_{28}O_4P_2$
L38	(S)-Xyl-SegPhos	210169-57-6	$C_{46}H_{44}O_4P_2$
L39	(S)-DTBM-SegPhos	210169-40-7	$C_{74}H_{100}O_8P_2$
L40	(R)-Cl-MeO-BIPHEP	185913-97-7	$C_{38}H_{30}Cl_2O_2P_2$
L41	SL-A109-1	352655-61-9	$C_{74}H_{104}O_6P_2$
L42	SL-A120-1	394248-45-4	$C_{46}H_{48}O_2P_2$
L43	SL-A107-1	352655-40-4	$C_{70}H_{100}N_4O_2P_2$
L44	SL-A108-2	145214-59-1	$C_{30}H_{24}O_6P_2$
L45	SL-A102-2	133545-25-2	$C_{42}H_{40}O_2P_2$
L46	SL-A121-1	192138-05-9	$C_{70}H_{96}O_2P_2$
L47	SL-A104-1	256390-47-3	$C_{50}H_{56}O_{14}P_2$
L48	(R)-GarPhos	1365531-75-4	$C_{40}H_{36}O_4P_2$
L49	(R)-Xyl-GarPhos	1365531-89-0	C48H52O4P2
L50	(R)-DTBM-GarPhos	1365531-98-1	C76H108O8P2
L51	(S)-iPr-BIPHEP	150971-43-0	C26H40O2P2
L52	(R)-C3-TunePhos	301847-89-2	C39H32O2P2
L53	(S,S)-iPr-BPE	528854-34-4	C22H44P2
L54	(R,R,R)-SPIRAP	NA	C43H38O2P2
L55	(R,R,S,S)-DuanPhos	528814-26-8	C24H32P2
L56	(R,R)-DiPamp	55739-58-7	C28H28O2P2
L57	(R)-iPr-PHOX	164858-78-0	C24H24NOP
L58	SL-F131-1	899811-43-9	C50H54Fe3N2P2
L59	(R)-Xyl-SDP	917377-75-4	C49H50P2
L60	(S)-DM-MonoPhos	185449-86-9	C24H22NO2P
L61	(R)-Ph-Monophos	936010-61-6	C34H26NO2P
L62	(S)-NEt2-MonoPhos	252288-04-3	C24H22NO2P
L63	(R,R,R)-Xyl-SKP	1429939-35-4	C52H54O2P2
L64	(R,R)-Ph-BPE	528565-79-9	C34H36P2
L65	(S,S)-ChiraPhos	64896-28-2	$C_{28}H_{28}P_2$

L66	(R,R)-Et-BPE	136705-62-9	C18H36P2
L67	(R)-QuinoxP	866081-62-1	C18H28N2P2
L68	(R,R)-Et-DuPhos	136705-64-1	C22H36P2
L69	(R,R)-Me-DuPhos	147253-67-6	C18H28P2
L70	(S)-PhanePhos	192463-40-4	C40H34P2
L71	(S)-Me-iPr-PHOX	1152313-76-2	C26H28NOP
L72	SL-N003-2	163169-29-7	C28H28FeNOP
L73	(S)-NeoPHOX	1199225-38-1	C22H28NOP
L74	(R,R)-Me-BoPhoz	406680-94-2	C37H35FeNP2
L75	(R)-Xyl-PhanePhos	325168-89-6	C48H50P2
L76	(S,S)-f-Binaphane	544461-38-3	C54H40FeP2
L77	(R,R)-BDPP	96183-46-9	C29H30P2
L78	(R,R)-NorPhos	71042-55-2	C31H28P2
L79	(R,S)-BPPFA	74311-56-1	C38H37FeNP2
L80	(R,R)-DIOP	32305-98-9	$C_{31}H32O_2P_2$
L81	(S)-Tol-tBu-PHOX	218460-00-5	C27H30NOP
L82	(S,S)-DPE-Phos	2119686-55-2	C38H32O3P2
L83	(S)-NMDPP	43077-29-8	C22H29P
L84	(S,S)-BABIBOP	2207601-04-3	C22H28O2P2
L85	(S,S,S,S)-Me-BABIBOP	2207601-10-1	C24H32O2P2
L86	(S,S,S,S)-iPr-BABIBOP	2207601-12-3	C28H40O2P2
L87	(R,R,R,R)-Me-BIBOP	1884680-48-1	C38H44O6P2
L88	(R,R)-PPM	77450-05-6	C29H29NP2
L89	SL-A101-2	133545-16-1	C38H32O2P2
L90	(S)-MeO-F12-BIPHEP	116008-37-6	C38H20F12O2P2
L91	(R)-MeO-F16-BIPHEP	NA	C42H24F16O2P2
L92	(R)-MeO-py-F12-BIPHEP	NA	C38H24F12N4O2P2
L93	(R)-MeO-F20-BIPHEP	NA	C42H20F20O2P2
L94	(R)-MeO-BFPy-BIPHEP	NA	C42H20F24N4O2P2
L95	(S,S)-XylSKEWPhos	551950-92-6	C37H46P2
L96	(S,S)-DIPSKEWPhos	NA	C53H78P2
L97	SL-W022-1	849925-29-7	C44H48FeP2
L98	catASium D(R)	99135-95-2	C35H33NP2
L99	(2R)-1-[(1S)-1-Aminoethyl]-2-(diphenylphosphino)ferrocene	607389-84-4	C24H24FeNP
L100	SL-W012-1	565184-30-7	C38H44FeP2
L101	SL-W030-1	1854067-62-1	C34H52FeP2
L102	(S,S)-Et-FerroTANE	290347-66-9	C24H36FeP2
L103	SL-W029-1	1854067-50-7	C38H56FeP2
L104	(S)-NMe2-MonoPhos	157488-65-8	C22H18NO2P
L105	SL-F103-1	55700-44-2	C26H28FeNP

L106	(R)-Xyl-P-Phos	442905-33-1	C46H50N2O4P2
L107	(S)-2-(Diphenylphosphinomethyl)pyrrolidine	60261-46-3	C17H20NP
L108	(R)-ProPhos	67884-32-6	C27H26P2
L109	(3R)-3-(1,1-Dimethylethyl)-2,3-dihydro-4-(2-methoxyphenyl)-1,3-benzoxaphosphole	1338454-28-6	C18H21O2P
L110	(2S,3R)-2-[Bis(1,1-dimethylethyl)phosphino]-3-(1,1-dimethylethyl)-2,3-dihydro-4-methoxy-1,3-benzoxaphosphole	1215081-28-9	C20H34O2P2
L111	(R,R)-BenzP*	919778-41-9	C16H28P2
L112	SL-J216-1	849924-43-2	C40H44FeP2
L113	(S,S)-1-Naphthyl-DiPamp	256469-70-2	C34H28P2
L114	(S,R)-PPFA	55650-58-3	C26H28FeNP
L115	SL-F173-1	166172-70-9	C30H24F12FeNP
L116	(R)-Xyl-SDP Oxide	1462321-89-6	C49H50OP2
L117	(R)-SITCP	856407-37-9	C25H23P
L118	(R,R,R)-Tol-SKP	1429939-32-1	C48H46O2P2
L119	(R,R)-BCPM	114751-47-2	C34H49NO2P2
L120	(R)-DiFluorPhos	503538-69-0	C38H24F4O4P2
L121	(R,R)-Me-BPE	129648-07-3	C14H28P2
L122	(R)-SynPhos	445467-61-8	C40H32O4P2
L123	(R)-SIPhos	443965-14-8	C19H20NO2P
L124	(3R,8R)-Tetrahydro-N,N,2,2-tetramethyl-4,4,8,8-tetraphenyl-1,3-dioxolo[4,5-e][1,3,2]dioxaphosphepin-6-amine	213843-90-4	C33H34NO4P
L125	(R)-SDP	917377-74-3	C41H34P2
L126	(R,R,R,R)-Ph-BIBOP	2301856-53-9	C34H36O2P2
L127	(R,S)-Ph-Bn-SIPHOX	2074610-05-0	C39H34NOP
L128	(R,R)-iPr-BPF	849950-54-5	C30H48FeP2
L129	(R)-Tol-SDP	528521-87-1	C45H42P2
L130	(R)-DMM-GarPhos	1365531-93-6	C52H60O8P2
L131	8-[(3R)-3-(1,1-Dimethylethyl)-2,3-dihydro-1,3-benzoxaphosphol-4-yl]benzo[1,2-b:5,4-b']difuran	1835717-07-1	C21H23O3P
L132	(S)-PipPhos	284472-79-3	C25H22NO2P
L133	(R)-An-PhanePhos	364732-86-5	C44H42O4P2
L134	(S)-BINAPINE	528854-26-4	C52H48P2
L135	(S)-H8-MonoPhos	389130-06-7	C22H26NO2P
L136	(R,R)-Me-Ferrocelane	540475-45-4	C22H32FeP2
L137	(R,R)-Et-Ferrocelane	147762-89-8	C26H40FeP2

L138	(S,S,S,S)-MeO-BIBOP	1202033-19-9	C24H32O4P2
L139	(R)-CTH-BINAM	208248-67-3	C44H42N2P2
L140	(2R)-1-[(R)-Aminophenylmethyl]-2-(diphenylphosphino)ferrocene	498580-48-6	C29H26FeNP
L141	(1R,2S)-TaniaPhos-OH	851308-43-5	C41H34FeOP2
L142	2-[2-[(2R,5R)-2,5-Dimethyl-1-phospholanyl]phenyl]-1,3-dioxolane	1044256-04-3	C15H21O2P
L143	(R,R)-BPPM	72598-03-9	C34H37NO2P2
L144	(S)-MorfPhos	185449-81-4	C24H20NO3P
L145	(R,R,R)-Ph-SKP	1360823-43-3	C44H38O2P2
L146	(S,R)-N-PINAP	1173836-08-2	C38H30N3P
L147	(R)-CTH-P-Phos	221012-82-4	C38H34N2O4P2
L148	(R)-SIPHOS-PE	500997-69-3	C33H32NO2P
L149	(R)-Tol-GarPhos	1365531-81-2	C44H44O4P2
L150	(R)-DTB-SpiroSAP-Ph	1809609-38-8	C53H66NPS
L151	SL-N004-1	1226898-27-6	C29H30FeNOP
L152	SL-N011-2	950201-43-1	C36H32FeNOP
L153	(S,S,S,S)-BIBOP	1202033-17-7	C22H28O2P2
L154	SL-N009-2	706814-27-9	C32H24F12FeNOP
L155	SL-J408-1	950982-69-1	C44H48FeP2
L156	(2R,2R)-2,2-bis(diphenylphosphino)-1,1-biferrocene	136274-57-2	C44H36Fe2P2
L157	(R)-Cy-GarPhos	2829282-18-8	C40H60O4P2
L158	(R)-DTB-SpiroPAP-6-Me	1298133-26-2	C52H65N2P
L159	Exo-4-Methoxyphenyl Kwon [2.2.1] Bicyclic Phosphine	1975180-37-0	C19H22NO3PS
L160	Endo-4-Methoxyphenyl Kwon [2.2.1] Bicyclic Phosphine	1883493-01-3	C19H22NO3PS
L161	(R,R)-(Diphenylphosphino)-phenylbenzeneethanamine	1091606-68-6	C26H24NP
L162	(1R,2R)-2-(Diphenylphosphino)-2,3-dihydro-1H-inden-1-amine	1091606-70-0	C21H20NP
L163	(S,S)-tBuPh-SKEWPhos	911415-22-0	C45H62P2
L164	(R,R)-(S,S)-PhTRAP	137096-37-8	C48H44Fe2P2
L165	(R)-BINAPhane	253311-88-5	C50H36P2
L166	(1R)-8-(Diphenylphosphino)-1,2,3,4-tetrahydro-1-naphthalenamine	960128-64-7	C22H22NP
L167	(R,R)-iPr-DuPhos	136705-65-2	C26H44P2
L168	(3R)-4-[2,6-Bis(1-methylethoxy)phenyl]-3-(1,1-dimethylethyl)-2,3-dihydro-1,3-benzoxaphosphole	1338454-38-8	C23H31O3P

L169	SL-M002-1	494227-35-9	C52H74FeN2P2
L170	(S)-DTBM-BINAP	541502-07-2	C80H104O4P2
L171	(S,S,S,S)-Et-BABIBOP	2415751-83-4	C26H36O2P2
L172	(R,R,R,R)-WingPhos	1884680-45-8	C50H44O2P2
L173	2-[(2S,3S)-3-(1,1-Dimethylethyl)-2,3-dihydro-4-methoxy-1,3-benzoxaphosphol-2-yl]pyridine	2565792-52-9	C17H20NO2P
L174	SL-J681-1	1221745-90-9	C28H32FeOP2
L175	(S,Sp)-p-Tol-TaniaPhos	NA	C47H47FeNP2
L176	(R,Rp)-2-Furyl-TaniaPhos	NA	C35H31FeNO4P2
L177	(R)-DM-MorfPhos	864529-90-8	C27H26NO2P
L178	(R)-C2-TunePhos	301847-88-1	C38H30O2P2
L179	(R)-QUINAP	149341-34-4	C31H22NP
L180	SL-J015-1	649559-65-9	C36H36FeO2P2
L181	SL-J403-1	166172-60-7	C40H28F12FeP2
L182	SL-J425-1	849924-49-8	C44H48FeO2P2
L183	(R,R)-CyPP	70774-28-6	C32H34P2
L184	(R,R)-MeO-BoQPhos	1542796-16-6	C18H22NO3P
L185	2-[(2R,3R)-4-(2,6-Dimethoxyphenyl)-3-(1,1-dimethylethyl)-2,3-dihydro-1,3-benzoxaphosphol-2-yl]-6-methoxypyridine	2565792-77-8	C25H28NO4P
L186	2-[(2R,3R)-4-(9-Anthracenyl)-3-(1,1-dimethylethyl)-2,3-dihydro-1,3-benzoxaphosphol-2-yl]pyridine	1542796-14-4	C30H26NOP
L187	(S)-SunPhos	765312-54-7	C42H36O4P2
L188	(1R)-1-[Bis[3,5-bis(1,1-dimethylethyl)-4-methoxyphenyl]phosphino]-2-[(1R)-1-(dicyclohexylphosphino)ethyl]ferrocene	1453803-83-2	C54H80FeO2P2
L189	(1R,4R)-1,4-dimethyl-1,4-butanediylbis(diphenylphosphine)	142494-67-5	C30H32P2
L190	(2R,3R)-4-(9-Anthracenyl)-3-(1,1-dimethylethyl)-2,3-dihydro-2-(1-methylethyl)-1,3-benzoxaphosphole	1891002-60-0	C28H29OP
L191	(S,S)-XantPhos	2119686-35-8	C41H36O3P2
L192	(3R)-3-(1,1-Dimethylethyl)-4-(2,6-diphenoxyphenyl)-2,3-dihydro-1,3-benzoxaphosphole	1441830-74-5	C29H27O3P

B

Descriptors

A comprehensive list of the utilized steric, geometric and electronic descriptors can be found in Table B.1.

Table B.1: List of descriptors

Descriptor	Category	Description
cone angle	geometric	cone angle of metal-ligand, ignores NBD
buried volume Rh 6A	steric	buried volume at metal centre with radius 6 $\AA$, ignores NBD
bite angle	geometric	bite angle between donor max - metal centre - donor min
NE quad	steric	quadrant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
SE quad	steric	quadrant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
buried volume Rh 4A	steric	buried volume at metal centre with radius 4 $\AA$, ignores NBD
NW quad	steric	quadrant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
buried volume donor min	steric	buried volume at donor min with radius 3.5 $\AA$, ignores NBD
dihedral angle 2	geometric	dihedral angle between H-C central of NBD-metal-donor
buried volume Rh 7A	steric	buried volume at metal centre with radius 7 $\AA$, ignores NBD

dihedral angle 1	geometric	dihedral angle between H-C central of NBD-metal-donor
buried volume donor max	steric	buried volume at donor max with radius 3.5 $\AA$, ignores NBD
buried volume Rh 5A	steric	buried volume at metal centre with radius 5 $\AA$, ignores NBD
buried volume Rh 3.5A	steric	buried volume at metal centre with radius 3.5 $\AA$, ignores NBD
SW quad	steric	quadrant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
-,+,- octant	steric	octant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
+,+,- octant	steric	Octant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
+,-,+ octant	steric	octant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
+,-,- octant	steric	octant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
-,-,- octant	steric	octant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
+,+,+ octant	steric	octant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
-,-,+ octant	steric	octant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
-,+,+ octant	steric	octant of 3.5 $\AA$ buried volume orientated such that average of position of donors define Z-axis and donor max defines the XZ plane, ignores NBD
HOMO LUMO gap gfn2 xtb	electronic	HOMO LUMO gap calculated via single point GFN2-xTB
nucleofugality gfn2 xtb	electronic	nucleofugality
nucleophilicity gfn2 xtb	electronic	nucleophilicity

ea gfn2 xtb	electronic	electron affinity
dispersion p int Rh gfn2 xtb		
electrofugality gfn2 xtb	electronic	electrofugality
distance pi bond 2	electronic	distance of C=C NBD pi bond that coordinates to metal centre
distance pi bond 1	electronic	distance of C=C NBD pi bond that coordinates to metal centre
sasa gfn2 xtb	electronic	solvent accessible surface area
dipole gfn2 xtb	electronic	Debye dipole moment
dispersion p int donor min gfn2 xtb	electronic	p int dispersion descriptor on donor min
dispersion p int donor max gfn2 xtb	electronic	p int dispersion descriptor on donor max
ip gfn2 xtb	electronic	ionization potential
electrophilicity gfn2 xtb	electronic	electrophilicity

C

Minor and major substrate coordination

Figure C.1 illustrates the four possible coordinations of methyl 2-acetamidoacrylate substrate to the metal-ligand complex [146].

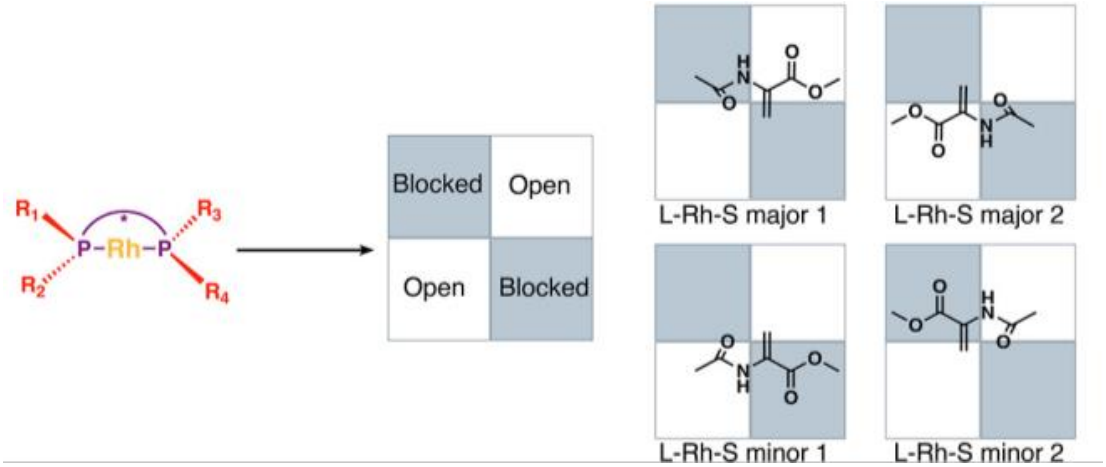


Figure C.1: Major 1, major 2, minor 1, and minor 2 potential coordinations of the methyl 2-acetamidoacrylate substrate to the metal-ligand complex. Image taken from [146].

D

Assessment parameters for ϵ

The obtained assessment parameters for utilizing DBSCAN with different ϵ values can be found in Table D.1.

Table D.1: Assessment parameters for ϵ

ϵ	**FN**	**TN**	**FN/TN**
0.1	10	206	20.6
0.11	10	225	22.5
0.12	14	244	17.4
0.13	16	256	16.0
0.14	17	265	15.6
0.15	19	275	14.5
0.16	23	286	12.4
0.17	25	298	11.9
0.18	29	303	10.4
0.19	31	315	10.1
0.20	32	317	9.9
0.21	32	321	10.0
0.22	36	324	9.0
0.23	38	327	8.6
0.24	42	333	7.9
0.25	45	336	7.5
0.26	46	345	7.5
0.27	45	354	7.9
0.28	46	359	7.8
0.29	48	364	7.6
0.30	50	367	7.3

E

ANOVA test on steric and geometric descriptors

The results of the ANOVA test on steric and geometric descriptors are presented in Table E.1.

Table E.1: Results of ANOVA test on steric and geometric descriptors

Descriptor	F-statistic	p-value
cone angle	1.828753397	0.177038886
buried volume Rh 6A	1.592676775	0.207679469
bite angle	0.627925759	0.428586891
NE quad	0.051780275	0.820110184
-,-,+ octant	0.724126302	0.395302846
SE quad	0.007123389	0.932780493
buried volume Rh 4A	11.69286453	0.000691909
-,+,+ octant	1.136624739	0.287009741
NW quad	0.153997934	0.694952716
buried volume donor min	5.208492775	0.023001177
dihedral angle 2	10.55158115	0.001258905
buried volume Rh 7A	0.118368761	0.730991267
dihedral angle 1	5.734888507	0.017090451
buried volume donor max	1.32235219	0.250857388
buried volume Rh 5A	5.205407744	0.023041487
buried volume Rh 3.5A	17.82528672	0.00003
SW quad	0.06209379	0.077932811
-,+,- octant	3.123509991	0.077932811
+,+,- octant	0.006684279	0.934880494
+,-,+ octant	0.165663627	0.6842125
+,-,- octant	2.46201837	0.117419516
-,-,- octant	1.669628539	0.197054296
+,+,+ octant	6.120438503	0.013776909

F

Use of generative AI tools

During this thesis I have used 2 generative AI tools: ChatGPT and Canva AI image generator. ChatGPT was mainly used for writing assistance and easily programming purpouses, while the image of the cover page was taken from Canva. Since Canva was only used for that one image, I will focus on describing the use of ChatGPT in this chapter.

F.1 Writing assistance

I never asked ChatGPT to write or rewrite sections for my thesis. However, I did used ChatGPT to help me improve my writing. I usually provided my sentences and asked ChatGPT to comment on the grammar and readability of it. ChatGPT gave me recommendation of different wording and corrected my grammar. I often found these recommendations and feedback very valuable and implemented them in my text, a few examples are presented below.

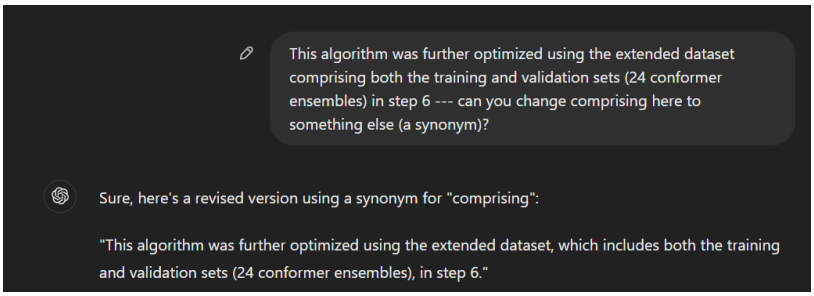


Figure F.1: I often used ChatGPT to provide me synonyms. I provided the full sentence, so ChatGPT can understand the context. I was making sure that the new word or expression does not change the overall language style I am using.

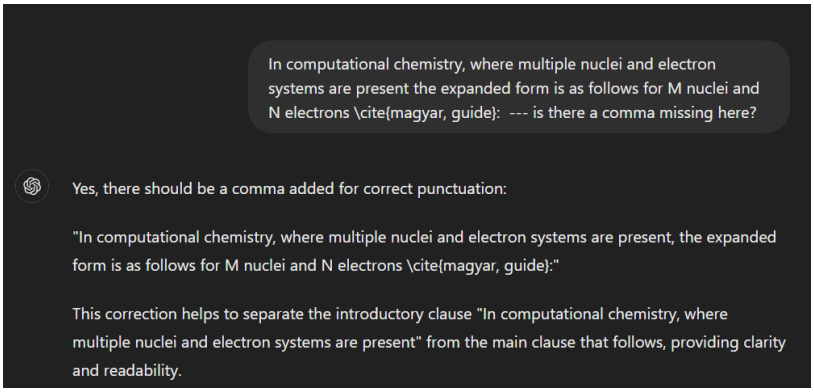


Figure F.2: I often asked ChatGPT to correct the grammar of my sentences.

F.2 Coding assistance

Overall, I barely used any generative AI tools for my code, as I enjoy writing my own code. Only ChatGPT was used for very easy coding tasks, that mainly required the usage of a single function. An example is provided below.

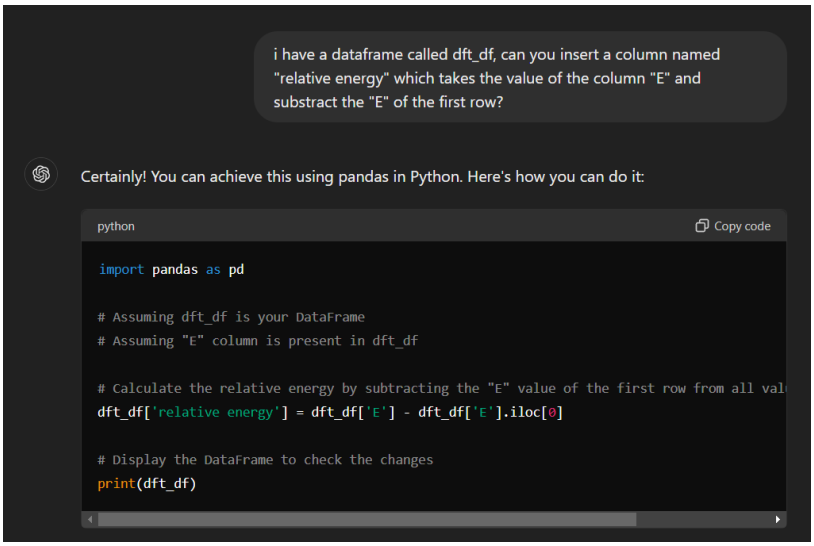


Figure F.3: I often used ChatGPT to handle easy coding tasks.

I did not find ChatGPT suitable to handle more complicated programming tasks and I noticed an overall decrease in ChatGPT's coding ability during my thesis. Therefore, for the second part of my thesis I did not use ChatGPT's coding assistance anymore.

Cover image created by https://www.canva.com/ai-image-generator/