Data, Representation, Models and Analysis: the four horsemen of machine learning for homogeneous catalysis

by

Tai Hong Chow

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended on 25th of February, 2025 at 15:00.

Performed at:
Inorganic Systems Engineering
Faculty of Applied Sciences

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

Student number:	5309395	
Project duration:	2nd of September, 2024 – 25th of February, 2025	
Thesis committee:	Prof. Dr. E. A. Pidko,	TU Delft, supervisor
	Dr A. M. Schweidtmann,	TU Delft
	Dr J. M. Weber,	TU Delft

Abstract

Bidentate ligand-coordinated transition metal complexes are often used as homogeneous catalyst, as they have the ability to produce enantioselective compounds. These compounds are of high interest in the pharmaceutical and food industries. However, identifying high performing catalysts relies on trial-and-error approaches, which is time consuming and costly. The use of data-driven predictive models could improve this process significantly, by shifting most of the process from experimental work to computational work.

Previous work from the group has attempted to develop such predictive model using Machine Learning (ML), a representation of a manually generated static structure and a database generated through High-Throughput Experimentation (HTE). However, these models faced challenges in terms of model performance and consistency between different substrates. This research aims to enhance these models, by improving the representations used in ML to achieve more accurate predictions.

To bring the representations closer to reality, both dynamic and new static approaches are tested, using conformer ensembles (CEs) generated by CREST. These structures were then used in DFT calculations to obtain accurate properties of these complexes. Additionally, new HTE data, which is closer to the complexes used in the simulation, was incorporated to improve training data for the ML models.

The investigated reaction is the hydrogenation of norbornadiene (NBD) using Rh-NBD complexes. The performance of both classification and regression was compared across different representations, a cheap topological connectivity fingerprint (ECFP), semi-empirical DFT representations and expensive fully DFT optimized representations. The results conclude that none of the DFT-based representation outperforms the cheap topological fingerprint for this specific reaction. The study also highlights the importance of high quality data in training the models. Ultimately, while the representation was improved, the much simpler topological method was the most effective for prediction of catalyst performance.

Contents

List of Figures iv

1 Introduction 1

2 Supporting theory 3
- 2.1 Transition metal (TM) complexes 3
  - 2.1.1 (Bidentate) ligands 3
- 2.2 Catalyst design and Machine Learning 4
- 2.3 Density Functional Theory 5
  - 2.3.1 Theoretical framework 5
  - 2.3.2 Basis sets 7
  - 2.3.3 Potential energy surface 7
  - 2.3.4 DFT Optimization 7
  - 2.3.5 DFT Single-Point 7
- 2.4 Exchange-correlation functional 8
- 2.5 Conformer searching with semi-empirical methods 8
  - 2.5.1 GFNn-xTB 9
  - 2.5.2 CREST 9
- 2.6 Properties vs barcode 9
  - 2.6.1 Descriptors for DFT-based representations 9
  - 2.6.2 Descriptors for ECFP representation 11
- 2.7 Representations 12
  - 2.7.1 DFT-based representation 12
  - 2.7.2 ECFP 12
- 2.8 Machine Learning 13
  - 2.8.1 Supervised learning 13
  - 2.8.2 Unsupervised learning 14

3 Methods 15
- 3.1 Experimental data 15
- 3.2 Workflow 16
  - 3.2.1 Input structures 16
  - 3.2.2 Conformer searching and pruning 16
  - 3.2.3 Filtering for DFT 17
  - 3.2.4 DFT calculation 17
  - 3.2.5 Descriptor calculation 17
  - 3.2.6 Model selection 17
  - 3.2.7 Representations 17

4 Results & Discussion 18
- 4.1 Data generation 18
- 4.2 Experimental data 20
- 4.3 Original substrate models 21
- 4.4 Sensitivity analysis 22
  - 4.4.1 Model 22
  - 4.4.2 Data 23
- 4.5 Static representation (DFT-NBD) 23
  - 4.5.1 Removal stereochemistry 25
  - 4.5.2 Removal free ligand descriptors 26

- 4.6 CREST-based representations (CREST-NBD & CREST-DFT-NBD) . . . 27
    - 4.6.1 Cheap vs Expensive E min selection . . . 28
    - 4.6.2 Analysis of increased energy structures after CREST . . . 30
    - 4.6.3 Cheap vs Expensive Boltzmann-averaging . . . 32
    - 4.6.4 Analysis descriptors . . . 34
    - 4.6.5 Boltzmann-averaging vs E min selection . . . 34
    - 4.6.6 Analysis descriptors . . . 35
- 4.7 Comparison all models . . . 36
- 4.8 ECFP analysis . . . 38
    - 4.8.1 Feature Importance . . . 38
    - 4.8.2 PCA . . . 38

5 Conclusion & Outlook 40
- 5.1 Conclusion . . . 40
- 5.2 Outlook . . . 41

Bibliography 43

Appendices

A Appendix A 1
- A.1 Methods . . . 2
- A.2 Results and Discussion . . . 18
    - A.2.1 Sensitivity analysis . . . 18
    - A.2.2 Removal stereochemistry . . . 19
    - A.2.3 ECFP . . . 19
- A.3 ChatGPT . . . 20
    - A.3.1 Examples . . . 21

List of Figures

2.1 Reaction scheme of hydrogenation of norbornadiene (NBD) to the products norbornene (NBEN) and norbornane (NBAN). . . . 3
2.2 Example where different ligands can change complex properties a lot, these two complexes have a very different groups resulting in different bite angles between the donor atoms. The differences in properties such as these can lead to differences in catalytic activity. . . . 4
2.3 The four steps of creating a Machine Learning Model, including Data generation, Representation for algorithms, Machine Learning Models, and Analysis and evaluation of results. In our case, Data and Representation steps are mostly varied, as these steps can be improved the most. 5
2.4 Example of a Potential Energy Surface (PES), illustrating how the energy varies given the atomic positions. The red arrow shows the pathway between two stable conformers. Figure adapted from LibreText Physical Chemistry. [31] . . . 7
2.5 Jacob's ladder showing the different levels of exchange-correlation functionals. Left of the ladder the dependencies of each level is given, starting with electron density ρ, gradient of the electron density $\nabla\rho$, kinetic energy density τ/ Laplacian of the electron density $\nabla^2\rho$, occupied orbitals ϕ_i and unoccupied (virtual) orbitals ϕ_a. . . . 8
2.6 Visualization of the cone angle (θ) and bite angle (β). Figure adapted from Chemical Communications. [43] . . . 10
2.7 Example of how circular iterations are taken in a molecule. Figure adapted from Medium. [46] . 12
2.8 Example of new features that are added after each iteration. Figure adapted from Medium. [46] 13
2.9 Example of hashing and bit collision in the final ECFP representation. The values are binary, and there is 1 bit collision where two different identifiers refer to the same point in the representation. Figure adapted from Research Gate. [47] . . . 13

3.1 Interpolated conversion curves of the experimental data on Rh-NBD complexes. Reaction conditions: 303.15 K, 3 bar H_2. The data point are at 15 and 30 minutes. . . . 15
3.2 Complete workflow. Figure adapted from Sára's thesis. [49] . . . 16

4.1 Distribution plot of the average 15 minute conversion results for the first 96 Rh-NBD complexes, showing most data points in either the low or high conversion regime. The median conversion 0.317. . . . 20
4.2 Distribution plot of the 30 minute conversion results for the first 96 Rh-NBD complexes, showing most data points in either low or high conversion regime. The median conversion is 0.321. . 21
4.3 Box plot of regression performances on conversion using the Original Substrate model (96 data points). Five substrates are shown with representations: OG-DFT-NBD, ECFP and random-36. . 22
4.4 Sensitivity analysis of initialization of the Machine Learning Model. Box plots are shown of the performance of the ML models using classification (left) and regression (right). Both methods show small changes in performance in two different representations. . . . 22
4.5 Sensitivity analysis of data splitting of the train/test data. Box plots are shown of the performance of the ML models using classification (left) and regression (right). While classification shows a small change in performance, regression shows a much larger change. . . . 23
4.6 Box plot of model performances for classification (left) and regression (right) for average conversion at 15 minutes. Static DFT-NBD representation is used, in combination with the NBD dataset. . . . 23
4.7 Box plot of model performances for classification and regression for conversion at 30 minutes. Static DFT-NBD representation is used, in combination with the NBD dataset. . . . 24
4.8 Comparison of performances between DFT-NBD representations using the new NBD dataset and the five substrate datasets. SM**X** refers to the different substrates used. . . . 24

4.9 Feature importance plot of all descriptors used in DFT-NBD. Mostly electronic and descriptors and the bite angle (geometric descriptor) show dominating contribution with the lowest influence from steric descriptors. 25
4.10 Effect of removing the stereochemistry descriptors on the NBD average 15 minute dataset. On the left the classification performances are shown for DFT-NBD with stereochemistry and DFT-NBD without stereochemistry. Performances are similar in both representations. 25
4.11 Effect of substituting free ligand descriptors for complex descriptors on the average 15 minute conversion NBD dataset. On the left the classification performances are shown for DFT-NBD without stereochemistry and DFT-NBD without stereochemistry and free ligand descriptors. Performances are similar. 26
4.12 Effect of substituting free ligand descriptors for complex descriptors on the NBD 30 minute conversion NBD dataset. On the left the classification performances are shown for DFT-NBD without stereochemistry and DFT-NBD without stereochemistry and free ligand descriptors. Performances are similar. 27
4.13 Example of a destroyed ferrocene group by CREST. This ferrocene group in the front of Figure 4.13a makes bonds with other atoms outside of the cyclopentadienyl groups, such as phosphor and carbon. 28
4.14 Comparison of performances of minimum energy representations from conformer ensembles on the average 15 minute dataset. The conformers in the left and middle representation are based on the energy obtained from xTB geometries. The representation on the right is based on the energy obtained from the DFT OPT geometry (cheapest representation on the left tot the most expensive on the right). 28
4.15 Comparison of performances of minimum energy representations from conformer ensembles on the 30 minute dataset. The conformers in the left and middle representation are based on the energy obtained from xTB geometries. The representation on the right is based on the energy obtained from the DFT OPT geometry (cheapest representation on the left to most expensive to the right). 29
4.16 Energy analysis of CREST-NBD vs static DFT-NBD. A lower dE means a more stable structure has been found after CREST, minimum energy selection based on xTB energies and reoptimization using DFT OPT. 29
4.17 Energy analysis of CREST-DFT-NBD vs static DFT-NBD. A lower dE means a more stable structure has been found after CREST, DFT OPT and minimum energy selection based on DFT OPT energies. 30
4.18 Example of a structure where the ferrocene group has an extremely noticeable different orientation, compared to the handmade structure. In the CREST structure this ferrocene group is directed towards the metal center, resulting in a more compact complex, whereas in the handmade structure this group is oriented outwards. The inward orientation in the CREST structure causes steric hindrance, resulting in the higher observed energy. 31
4.19 Example of a structure where an increase in energy is caused by CREST's inability to find all conformers. A clear difference is observed between the handmade DFT optimized structure and CREST generated DFT optimized structure. The atoms are displayed in colors: Rh (greenish blue), P (orange), N (dark orange), O (blue), C (gray), H (white). 31
4.20 Example of a structure where a small change in orientation causes an increase in energy. When viewed straight through Rh-NBD, it is clear that the handmade structure is much more symmetric compared to the CREST generated structure. CREST is unable find the correct rotation for one of the benzene rings, preventing the complex to be completely symmetrical. The atoms are color-coded as follows: Rh(greenish blue), P (orange), N (dark orange), O (blue), C (gray), H (white). 32
4.21 Comparison of performances of Boltzmann averaged representations from conformer ensembles on the average 15 minute dataset. The cheap representation (left) show very consistent performances, while the expensive representation (right) show a broader IQR. However, the absolute performance does not change. 33
4.22 Comparison of performances of Boltzmann averaged representations from conformer ensembles on the 30 minute dataset. The cheap (left) and expensive (right) representation show similar performances and IQR's. 33

4.23 Example of a few descriptor scatter plots, where the values of the same descriptor over a static and dynamic representation is compared (DFT-NBD vs CREST-DFT-NBD). The red line shows the identity line, where the values. 34
4.24 Box plot comparison of performances of all conformer ensemble representations for average 15 minute dataset. Classification (left) and regression (right). 35
4.25 Box plot comparison of performances of all conformer ensemble representations for 30 minute dataset. Classification (left) and regression (right). 35
4.26 Two example of scatter plots, where the values of the same descriptor over minimum energy structure vs the entire CE is compared (CREST-DFT-NBD [boltz. avg.] vs CREST-DFT-NBD [E min.]). The red line shows the identity line, where the values. 36
4.27 Distribution of highest Boltzmann weights in a conformer ensemble. 36
4.28 Overview all no stereochemistry representations average 15 minute dataset. 37
4.29 Overview all no stereochemistry representations 30 minute dataset. 37
4.30 Example of a Feature Importance figure of ECFP, where the highest descriptors are shown. This figure is zoomed in and contains only the highest 10 features. 38
4.31 PCA of DFT-NBD (top) vs ECFP (bottom) for average 15 minute conversion dataset (random state 66). Both models have points where each outperforms the other. 39
4.32 PCA of DFT-NBD (top) vs ECFP (bottom) for 30 minute conversion dataset (random state 66). Both models have points where each outperforms the other. 39

A.1 Sensitivity analysis of initialization of the Machine Learning model. Box plots are shown of the performance of the models using classification (left) and regression (right). Classification show no change after different initializations and regression show small changes. 18
A.2 Sensitivity analysis of data splitting of the train/test data. Boxplots are shown of the performance of the ML models using classification (left) and regression (right). While classification show a small change in performance, regression show a much larger change. 18
A.3 Effect of removing the stereochemistry descriptors on the NBD 30 min dataset. On the left the classification performances are shown for DFT-NBD with stereochemistry and DFT-NBD without stereochemistry. Performances are similar 19
A.4 ECFP Figure Importance plot of random state 20. 19
A.5 ECFP Figure Importance plot of random state 30. 20
A.6 ECFP Figure Importance plot of random state 62. 20
A.7 ECFP Figure Importance plot of random state 66. 20
A.8 Example where I inputted a paragraph to see if I missed punctuation and changed structure if it is better. 21
A.9 Example where I checked spelling of words. 21
A.10 Example where I used ChatGPT for finding synonyms. 22
A.11 Example where I asked how to create a sns.barplot, when I have a pd.DataFrame. 22

Acronyms

AI	Artificial Intelligence
BA	Balanced accuracy
BO	Born-Oppenheimer
Boltz. avg.	Boltzmann averaged
CE(s)	Conformer ensemble(s)
COMP	Complex (chemistry definition)
CREST	Conformer-Rotamer Ensemble Sampling Tool
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DFT	Density functional theory
DFT OPT	DFT Optimization
DFT SP	DFT Single-Point
(i)FF	Integrated force field
ECFP	Extended-Connectivity Fingerprint
E min.	Lowest energy (conformer)
FI	Feature Importance
FL	Free ligand
Free lig.	Free ligand
FN	False Negative
FP	False Positive
GC-FID	Gas Chromatography Flame Ionization Detector
GFN-xTB	Geometry, Frequency, Non-covalent, extended Tight-Binding
GGA	Generalized Gradient Approximation
GTO	Gaussian Type Orbital
HF	Hartree-Fock
HK	Hohenberg-Kohn
HOMO	Highest Occupied Molecular Orbital
HTE	High-Throughput Experimentation
IQR	Interquartile range
KS	Kohn-Sham
L(S)DA	Local (Spin) Density Approximation
LUMO	Lowest Unoccupied Molecular Orbital
ML	Machine learning
MORFEUS	**Mo**lecula**r** **fe**at**u**re**s** for machine learning
MSE	Mean square error
NBAN	Norbornane
NBD	Norbornadiene
NBEN	Norbornene
NBO	Natural Bond Orbital
OBeLiX	Open Bidentate Ligand eXplorer
PBE	Perdew-Burke-Enzerhof
PCA	Principal Component Analyis
PES	Potential energy surface
R^2	R-squared
RMSD	Root mean square deviation
RSS	Residual sum square
SMILES	Simplified Molecular-Input Line-Entry System
StereoC.	Stereochemistry
STO	Slater Type Orbital
TM	Transition metal
TN	True Negative
TP	True Positive
XC	Exchange-correlation

1

Introduction

Artificial Intelligence (AI) is becoming increasingly indispensable in our lives. From self-driving cars to advanced language models like ChatGPT, AI is the latest technological advancement, driven by our desire to use automation to maximize safety, efficiency, profit and many more things. According to current trends, AI is expected to grow exponentially in the upcoming years, with Machine Learning, the fastest growing field within AI, to quintuple in market value by 2030. [1] This growth enables potential applications of Machine Learning across numerous sectors, as it might be more beneficial to use in the future, in terms of knowledge, costs and computational power.

Chemists are increasingly interested in utilizing AI for chemical exploration of molecules, materials and catalysts due to challenges associated to their discovery and optimization. The major issue in this field is that our knowledge is largely based on experimental findings, often without understanding the full underlying principles. [2]

Catalyst design, for instance, involves complex mechanisms and correlations, making it difficult to directly connect catalyst properties to their performance. Despite decades of research, this field still relies heavily on trial-and-error methods to discover new or improved compounds. [3] While High-Throughput Experimentation (HTE) has improved efficiency by enabling parallel testing, it remains an exhaustive trial-and-error process, as each potential catalyst is experimentally evaluated. [2]

Machine Learning (ML) can offer a solution to both the knowledge gap and inefficiency of the process, as this data-driven approach is excellent in identifying similarities and patterns, which ultimately can be used in predictive models and relate properties to performances. [4] This approach also changes the majority of the process from manual experimentation to computational analysis, allowing for faster wider and less expensive screening. [5]

In general, this process of using ML for chemical exploration can be outlined in four fundamental steps: **Data** (generation and preprocessing), **Representation** (input for the ML models), **Models** (pattern recognition and predictions) and **Analysis** (evaluation of results). While literature has shown successful applications, many studies used relatively simple methods for the Data and Representation step. [4][6][7] We aim to improve these two steps, as the potential impact to improve catalyst design is the highest here.

Developing a universal predictive model however is unrealistic, given catalysis is a diverse field with a variety of categories and characteristics. Catalysts can be divided into three main categories: heterogeneous, homogeneous and bio-catalyst. [8][9] Each type of catalyst has its own distinct properties and applications, requiring specialized approaches for design and optimization. Our focus is on a specific type of homogeneous catalysts: bidentate ligand coordinated transition metal (TM)-based catalysts. [10][11][12] These catalysts are utilized to yield stereoselective compounds, which are of high interest in the pharmaceutical and food industries. [13][14][15]

In previous research from our group, such predictive ML models for bidentate coordinated TM-based catalysts were investigated, with as main objective to investigate their performance both In- and Out-Domain. [16] A case of asymmetric hydrogenation of imines was performed for five different substrates using 192 ligands, while a generalized precatalyst complex was simulated to obtain numeric features for each ligand (descriptors). Although, the predictive models were accurate in very specific cases, the performance was not consistent for all substrates. We suspect this might have been caused by underdevelopment in the 'Data' and 'Representation' steps of these models. Since the reaction involved asymmetric hydrogenation, stereochemistry might have been important, making the use of simulating a general complex too simple. Therefore, two main modifications were made to address this problem and continue on this research.

First, in the previous study, experiments were performed on multiple different substrate-Rh complexes and ligand descriptors were derived from a precatalyst-Rh complex, while substrate specific descriptors were supplemented separately. Now a new reaction is performed, where the norbornadiene (NBD) moiety in the precatalyst structure is hydrogenated. Thus creating a more closely related relation between simulated data and experimental data.

Second, an alternative structural representation method is explored. A dynamic representation, incorporating conformer searching. This approach might provide more accurate descriptors, since it considers all possible conformations, leading to representations that are closer to reality.

This leads to the main objective of this thesis, which is about investigating the effect on the performance of predictive ML models, when different types of representations are used. To answer the main objective a variety of things will be done:

- Different representations will be compared, including variations in: DFT methods, static and dynamic structures, and descriptor selection.
- Analysis will be done on prediction results and descriptors.

The key changes made compared to previous research, are mostly focused on improving the least developed steps Data and Representation. These changes include:

- The usage of a new experimental dataset of hydrogenation of the NBD moiety. Simulated and experimental data will have a closer relation, as both are using NBD as substrate now.
- The usage of dynamic representation, which involves conformer searching. The idea is to bring the descriptors closer to reality by consideration of all possible conformers.
- Consideration of In-Domain modeling and conversion only, instead of In- and Out-Domain modeling, and conversion and selectivity. This will simplify the prediction task.

This thesis begins with a theory section to provide the necessary context, followed by a methods section that describes the tools and techniques used. The main section presents the results and discussion, concluding with a final section on the conclusion and outlook.

2

Supporting theory

2.1. Transition metal (TM) complexes

Transition metal (TM) complexes are one of the most important types within homogeneous catalysis. [17] These complexes consist of a central metal atom surrounded by ligands, which stabilize the complex and can alter the characteristics of the complex, such as geometry or electronic properties. [18]

These metal centers are usually transition metals, which are located in the d- block of the periodic table. The partially filled *d* sub-shell causes the unique properties in these complexes. These properties include: (1) a range of oxidation states, (2) the presence of unpaired electrons, and (3) the ability to coordinate with four to six electron-donating ligands. These properties contribute to the tunability of the complexes, making a versatile and optimizable catalyst. [19]

However, not all transition metals are well-suited for homogeneous catalysis. Metals in effective catalysts depend on multiple factors, but in general they must be reactive, tuneable, stable, and reasonably affordable. [19] That is why depending on application, usually transition metals such as rhodium, cobalt, nickel or iron are used. [20] In this thesis rhodium-based catalysts are investigated, as multiple cases show its capability to reach high activity and excellent enantioselectivity in combination with phosphorus-based ligands. [21] A few examples of well-known use cases of rhodium are (asymmetric) hydrogenation, cross-coupling and hydroformylation reactions. [11][19] The specific reaction investigated, is the rhodium-catalyzed hydrogenation reaction of norbornadiene (NBD), which is a precatalyst. [16][22] The reaction scheme is given in Figure 2.1, where norbornene (NBEN) and norbornane (NBAN) are the formed products of the substrate norbornadiene (NBD) in presence of hydrogen.

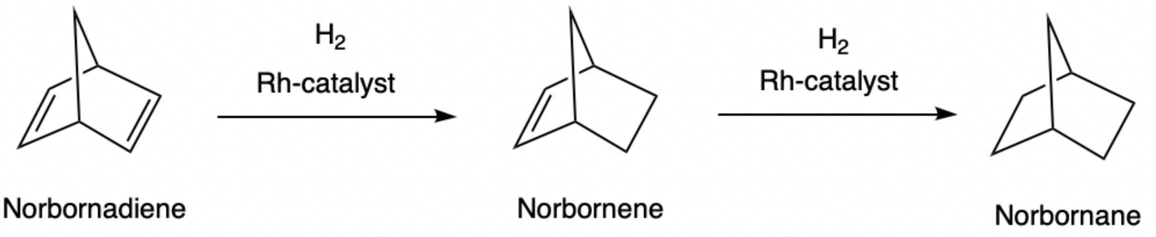


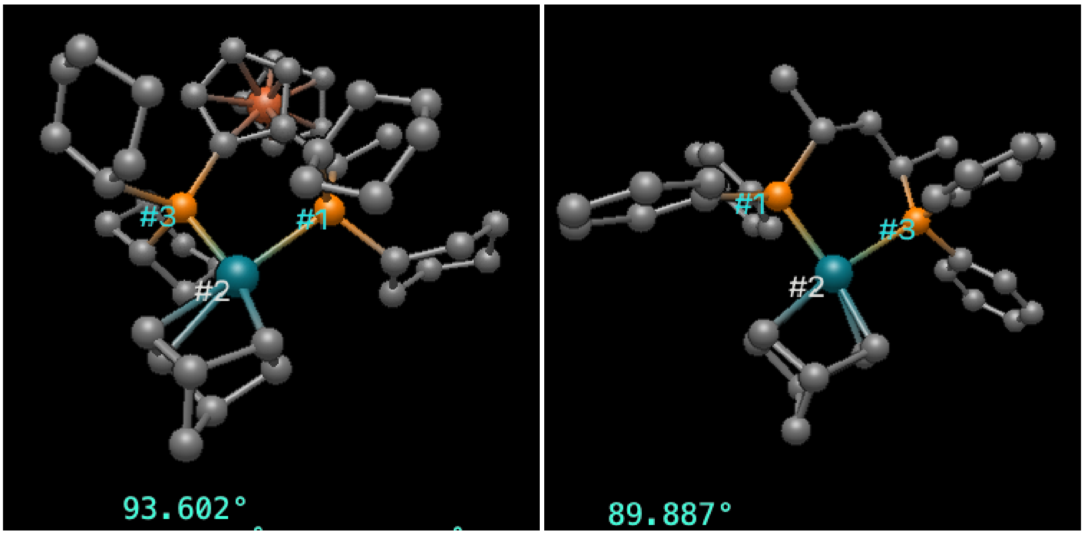
Figure 2.1: Reaction scheme of hydrogenation of norbornadiene (NBD) to the products norbornene (NBEN) and norbornane (NBAN).

2.1.1. (Bidentate) ligands

The ligands in these complexes are lewis bases, they are typically molecules or ions with one or more free electron pairs that can be donated to a metal center to form a strong bond. These ligands are categorized based on the number of coordination bonds they form with the metal atom: monodentate (one bond), bidentate (two bonds) or polydentate (multiple bonds) ligands. Thermodynamically, more bonds with the metal atom will usually lead to a stronger bond between ligand and metal. [19] A wide variety of molecules or ions can be used as ligands, each influencing the electronic and steric parameters of the complex in a different way. In homogeneous catalysis, these ligands are often phosphorus-based due to their high tunability, unique properties and strong ligation (soft on lewis acid scale). [11][23]

As previous work of the group involved a case of asymmetric hydrogenation, mostly bidentate ligands were tested. [16] The reason being that bidentate ligands not only offer greater bond strength compared to monodentate ligands, but also enable the ability to produce stereoselective or enantiopure compounds when asymmetric bidentate ligands are used. The importance and demand in stereoselective or enantiopure compounds can be linked to the pharmaceutical and food industries, where strict regulations on purity and quality are active. The thalidomide tragedy emphasizes the need for pure compounds, as the wrong enantiomer can lead to very different effects. [24][25]

In this work, the same ligand selection was used as in previous work. This selection contains mostly phosphorus-based ligands, and a few hetero donor ligands like PN or PO or phosphorus-based monodentate ligands. These additional ligands are included as literature have demonstrated some specific applications in asymmetric hydrogenation using these type of ligands. [26]



(a) XYZ structure of L3, where all hydrogen atoms are hided. NBD is located at the bottom. The angle between L-M-L is 93.6 degrees.

(b) XYZ structure of L77, where all hydrogen atoms are hided. NBD is located at the bottom. The angle between L-M-L is 89.9 degrees.

Figure 2.2: Example where different ligands can change complex properties a lot, these two complexes have a very different groups resulting in different bite angles between the donor atoms. The differences in properties such as these can lead to differences in catalytic activity.

2.2. Catalyst design and Machine Learning

As mentioned in the introduction, the design of catalysts can be significantly enhanced through the integration of computational techniques. This approach would requires a high-performing predictive model, which would be employed to assess the performance of numerous ligands and identify the most promising ones. These selected ligands would then proceed to the next phase of the process and undergo lab testing. This method would reduce a lot of resources and labor required, as only a selection would make it to the lab testing phase instead of all complexes.

To develop such high-performing model, four well-established steps are important: **Data** (generation), **Representation** (for the algorithms), (Machine Learning) **Models** and **Analysis** (and evaluation of results).

Data generation consist of creating digital representations of the complexes and obtaining experimental data to be used in the models. This data is stored in a database and preprocessed before it can be used to train the model. Experimental data is obtained through High-Throughput Experimentation (HTE), a technique that screens multiple samples simultaneously. This approach allows us to quickly build a large database. Digitally generated data mostly includes calculations using Density Functional Theory (DFT) (Section 2.3) and CREST (Conformer-Rotamer Ensemble Sampling Tool) (Section 2.5.2).

Representation, the generated data is represented in a way that the algorithms can use the data to find patterns, correlations and more. The complexes are represented through descriptors that are calculated in the DFT calculations. (Section 2.7)

Machine Learning **Models**, use an algorithm to find patterns and correlations in different selections of the descriptors. Together with the experimental data, predictions on (multi-class) classification (Section 2.8.1) and regression (Section 2.8.1) are made.

Analysis and evaluation of the results are done to determine the performance of the models. The analysis methods include Balanced Accuracy (BA) (Section 2.8.1) for classification and r-squared (r^2) (Section 2.8.1) for regression.

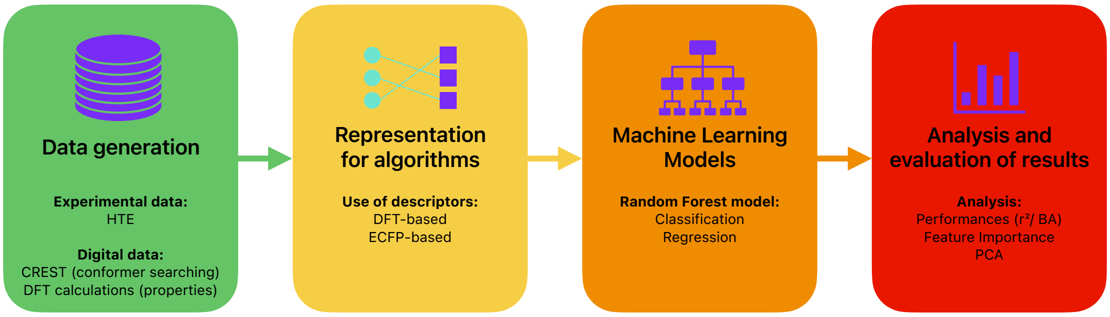


Figure 2.3: The four steps of creating a Machine Learning Model, including Data generation, Representation for algorithms, Machine Learning Models, and Analysis and evaluation of results. In our case, Data and Representation steps are mostly varied, as these steps can be improved the most.

2.3. Density Functional Theory

Density Functional Theory (DFT) is a quantum mechanical modeling method, used in specializations, such as computational physics or computational science. It is used to obtain molecular properties of compounds, by solely using equations to calculate parameters. It is mostly used for calculating electronic parameters, but can also be used to optimize geometries or calculate frequencies of a molecule or complex. In this work, DFT is used to generate accurate descriptions of the different complexes. Properties are calculated, which are used by the ML algorithms to find patterns and similarities.

The information provided in this section is based on Chapter 2 of Dr. V. Sinha PhD thesis and the books: 'A Chemist's Guide to Density Functional Theory' and 'Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory'. [27][28][29]

2.3.1. Theoretical framework

DFT calculations are done on the electron density at molecular level and stem from the Kohn-Sham equations, which are based on the Hohenberg-Kohn theorems. To understand these, we can start from the foundation of quantum mechanics, which is the non-relativistic, time-independent Schrödinger equation, given in Equation 2.1.

$$\hat{H}\Psi_i(\vec{x}_1,\vec{x}_2,...,\vec{x}_n,\vec{R}_1,\vec{R}_2,...,\vec{R}_n) = E_i\Psi_i(\vec{x}_1,\vec{x}_2,...,\vec{x}_n,\vec{R}_1,\vec{R}_2,...,\vec{R}_n) \tag{2.1}$$

Where $\hat{H}$ is the Hamiltonian operator for a system with N electrons and M nuclei, Ψ the wave function with x_i denoting the X,Y,Z-coordinates, and E the energy Eigenvalue.

The Hamiltonian operator can be written as the kinetic energy + the potential energy for both electron and nucleus, which is shown in Equation 2.2. The terms in this equation describe the kinetic energy of the electrons, kinetic energy of the nuclei, Coulomb attraction between nucleus and electron, Coulomb repulsion between electrons and Coulomb repulsion between the nuclei.

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

Where ∇_i^2 is the Laplacian operator for index i, Z_A the charge of index A, r_{iA} the distance between electron i and nucleus A, r_{ij} the inter-electronic distance, and R_{AB} the inter-nuclear distance.

However, this equation requires approximations, since an exact solution can only be obtained for single-electron systems. The Born-Oppenheimer approximation simplifies Equation 2.2, based on the significantly faster motion of much lighter electrons compared to nuclei. This approximation assumes that the kinetic energy of the nuclei is negligible (approximately zero) and that the nucleus-nucleus coulomb interaction remains constant constant. These assumptions result in a simplified Hamiltonian shown in Equation 2.3.

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

Where ∇_i^2 is the Laplacian operator for index i, Z_A the charge of index A, r_{iA} the distance between electron i and nucleus A, and r_{ij} the inter-electronic distance.

The simplification of Equation 2.3 modifies the original time-independent Schrödinger equation (Equation 2.1), resulting in a wave function that becomes parametrically dependent on the nuclear coordinates. The Schrödinger equation for electrons is given in Equation 2.4.

$$\begin{aligned} \hat{H}_{el}\Psi_{el} &= E_{el}\Psi_{el} \\ \Psi_{el} &= \Psi_i(\vec{x}_1, \vec{x}_2, ..., \vec{x}_n | \{\vec{R}_1, \vec{R}_2, ..., \vec{R}_n\}) \end{aligned} \tag{2.4}$$

Where $\hat{H}_{el}$ is the Hamiltonian operator for electrons and Ψ_{el} is parametrically dependent on the nuclear coordinates.

The Born-Oppenheimer (BO) approximation, however, encounters challenges in accurately describing the kinetic energy term, caused by the correlated motion of electrons. This issue can be addressed by using the electron density $\rho(r)$, which creates a mean-field approximation.

The use of electron density in the BO approximation reduces the degrees of freedom from a number of spatial 3N variables for a system of N electrons to 3 spatial variables. This idea was based on two Hohenberg and Kohn (HK) theorems. 'The ground state of any interacting many particle system is a unique functional of the electron density' and 'The ground state of any interacting many particle system minimizes the ground state energy'. Kohn-Sham combined this idea to the Hartree-Fock (HF) energy, given in Equation 2.5, to find an expression for the hard to describe terms.

$$E_{HF}[\rho(\vec{r})] = V_{MM} + V_{NM}[\rho(\vec{r})] + J[\rho(\vec{r})] + T[\rho(\vec{r})] + E_{XC}^{1}[\rho(\vec{r})] \tag{2.5}$$

Where V_{MM} is the nuclei-nuclei repulsion term, V_{NM} is the electrons-nuclei attraction term, J is the mean-field electron-electron interaction, T is the kinetic energy term, E_{XC}^{1} is the exchange-correlation functional term with non-classical effects of the mean-field approximation and $\rho(\vec{r})$ is the electron density.

The Kohn-Sham (KS) energy (Equation 2.6) solves the unknown terms, kinetic energy J and exchange-correlation functional E_{XC}^{1}, in the HF energy. The exchange correlation correction is explicitly given in Equation 2.7.

$$E_{KS}[\rho(\vec{r})] = V_{MM} + V_{NM}[\rho(\vec{r})] + J[\rho(\vec{r})] + T_{non-interacting}^{KS}[\rho(\vec{r})] + E_{XC}[\rho(\vec{r})] \tag{2.6}$$

$$E_{XC}[\rho(\vec{r})] = E_{XC}^{1}[\rho(\vec{r})] + T[\rho(\vec{r})] - T_{non-interacting}^{KS}[\rho(\vec{r})] \tag{2.7}$$

Where $T_{non-interacting}^{KS}$ is the kinetic energy functional of a non-interacting system, E_{XC} is the exchange and correlation correction to the potential energy and $\rho(\vec{r})$ is the electron density.

2.3.2. Basis sets

Basis sets are required to solve equations such as the Schrödinger equation for a many-bodied system (similar to Equation 2.4), such as molecules. While this equation theoretically is solvable using numerical integration methods, no practical techniques using such approaches are available for molecules. The solution is to use linear algebra and matrix methods to transform these wave functions (differential equations) to a basis set of algebraic equations. The wave function can then be expressed as in Equation 2.8.

$$\Psi_i = \sum_{j=1}^{K} c_{ji}\phi_j(r) \tag{2.8}$$

Where K is the amount of basis functions, c_{ji} coefficients and $\phi_j(r)$ the known function.

These known functions are based on atomic orbitals, typically Slater-Type Orbitals (STOs) and Gaussian-Type Orbitals (GTOs). In theory, an infinite K should yield the exact solution to the differential equation. However, this is not the case, and will always result in an energy higher than the true ground state energy E_0. Therefore, a basis set is chosen that balances accuracy with computational cost. In this thesis, the def2-Split-Valence Polarization-enhanced (def2-SVPP) basis set is used, which provides additional basis functions for polarization in heavier atoms. This results in more accurate calculations compared to the normal def2-SVP.

2.3.3. Potential energy surface

Since DFT is used for geometry optimization, it must be able to determine the lowest energy geometries. The potential energy surface (PES) is een representation of the energy landscape of a given molecule or complex, where the energy is plotted as a function of the atomic positions. [30] Local minima found on the PES correspond to stable structures, while local maxima correspond to the transition states.

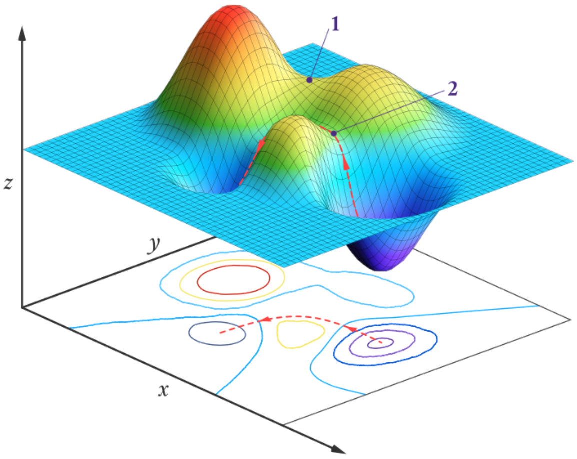


Figure 2.4: Example of a Potential Energy Surface (PES), illustrating how the energy varies given the atomic positions. The red arrow shows the pathway between two stable conformers. Figure adapted from LibreText Physical Chemistry. [31]

2.3.4. DFT Optimization

DFT optimization (DFT OPT) is a method within DFT that optimizes the input structure to find a local minimum on the PES, which corresponds to the local most stable configuration. [32] This method is computationally intensive, as the structure is iteratively adjusted until convergence is reached.

2.3.5. DFT Single-Point

DFT Single-Point (DFT SP) is a method that calculates the energy from an input structure without adjusting the input structure. [33] It represents the energy calculation of a specific point on the PES. Since the structure is not modified, it involves less parameters compared to DFT OPT making it computationally much cheaper.

2.4. Exchange-correlation functional

Exchange-correlation (XC) functionals approximate the quantum mechanical effects of electron exchange and correlation. These functionals use either non-empirical or semi-empirical methods, spanning a range from low to high accuracies. While more refined functionals offer greater accuracy, they also come with a higher computational cost. These functionals are categorized in tiers based on their accuracy and computational cost. An overview of all XC functional tiers are shown in the Jacob's ladder given in Figure 2.5.

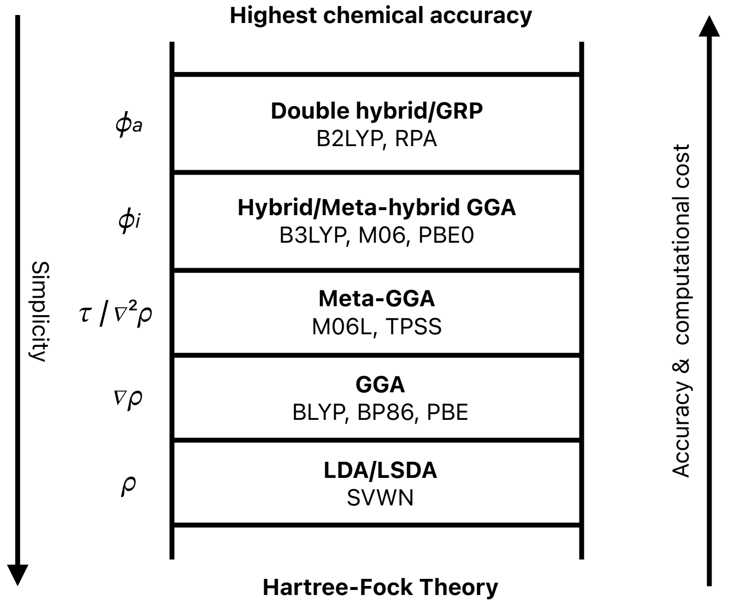


Figure 2.5: Jacob's ladder showing the different levels of exchange-correlation functionals. Left of the ladder the dependencies of each level is given, starting with electron density ρ, gradient of the electron density $\nabla\rho$, kinetic energy density τ / Laplacian of the electron density $\nabla^2\rho$, occupied orbitals ϕ_i and unoccupied (virtual) orbitals ϕ_a.

The Jacob's ladder shows the HF theory at the bottom, where the XC-functional is equal to zero. The tiers are arranged in increasing order of accuracy and computational cost. The Local (Spin) Density Approximation (L(S)DA) occupies the lowest tier, this approximation assumes homogeneous gas behavior for the electrons. The second tier is the Generalized Gradient Approximation (GGA), which is an improvement of the LDA. It accounts for inhomogeneity by including spatial derivatives of the electron density. The third tier is meta-GGA, which further refines the approximation by incorporating the kinetic energy density. Hybrid functionals are in the fourth tier, combining empirical parameters to linearly fit a fraction of the exact non-local exchange from the HF theory. Meta-hybrid GGA which are even more refined are also in the same tier. It incorporate a fraction of the HF exchange to the meta-GGA functionals. The current highest tier are the double hybrid and generalized random phase approximation, which introduces more higher order correlations to the system, such as unoccupied orbitals. In this thesis the hybrid functional 'PBE0' (also known as 'PBE1PBE') is used for all DFT calculations. PBE0 is a hybrid function, which combines the exact HF energy with the Perdew-Burke-Enzerhof functional. Which should give improved accuracy for a higher computational cost. This functional is specifically chosen, because studies on similar complexes have shown good results. [34] [35]

2.5. Conformer searching with semi-empirical methods

While DFT is excellent at calculating properties and locating minima or maxima, it is also very resource intensive. Locating all conformers in a conformer ensemble requires rotation of all bonds and optimization of all conformers, which makes using DFT inefficient and unreasonable. Semi-empirical methods offer a more practical solution. These methods are like DFT based on quantum mechanics, however it uses a lot of approximations, either derived from experimental data or simplifying phenomena to make the calculations quicker but also less accurate. This allows for quick exploration of the entire PES to identify stable conformers (local minima). The resulting approximate structures can be refined and optimized using DFT. This approach accelerates the process by a lot and reduces computational cost by using semi-empirical methods for initial screening and DFT for final optimization.

2.5.1. GFNn-xTB

Semi-empirical tight-binding models are considered an efficient alternative to DFT calculations. One of these models is called the Geometry, Frequency, Non-covalent, extended Tight-Binding (GFN-xTB) method, which provide reasonably accurate results for a fraction of the cost of DFT OPT calculations. [36] GFNn-xTB methods have multiple methods tailored for different needs in system, such as just optimization of geometries or adding improved accuracy for electronic parameters. The current derived GFN-xTB methods are: GFN1-xTB, GFN2-xTB, GFN-iFF and GFN-FF. GFN-xTB is solely based on quantum mechanical tight binding models, while GFN-(i)FF are based on quantum mechanical integrated or generalized force field methods. GFN2-xTB, the successor of GFN1-xTB, has improved accuracy, particularly for larger complexes, on polarization, dispersion and interactions. This increased accuracy however, comes with a higher computational cost. The GFN-iFF and GFN-FF are often used for larger systems, specifically for molecular dynamics, but also conformer searching. GFN-iFF is more accurate in transfer polarization and other electronic parameters within the systems. [36][37][38][39]

2.5.2. CREST

CREST (Conformer-Rotamer Ensemble Sampling Tool) is an automatized algorithm designed for efficient conformer exploration using these GFNn-xTB methods. This tool offers a variety of functions allowing for generation of a diverse conformer ensemble, which includes finding different conformations, aggregates and isomers. CREST identifies different conformers by rotating covalent bonds. It calculates the root mean square deviations (RMSD) to distinguish duplicates and isomers. xTB energies are used to identify stable conformers by locating minima on the PES. [40][41] In this thesis, CREST was used for conformer searching in combination with GFN2-xTB calculations. GFN2-xTB is specifically chosen for its improved accuracy when dealing with polar and complex systems. [36]

2.6. Properties vs barcode

ML algorithms require a machine-readable format of the complexes, to identify similarities and patterns within the data. This done through descriptors, as they allow us to better understand the relationship between found patterns and real chemical properties. Descriptors have a different meaning for DFT-based and ECFP representations.

2.6.1. Descriptors for DFT-based representations

DFT-based descriptors represent the properties of a complex by using values from the prior DFT calculations. The descriptors were calculated in two ways. In representations where only one conformer is used, such as the manually generated structure (handmade) or the minimum energy conformer from a CE, values are directly taken from the DFT calculations. In representations where the entire CE is considered, the descriptors are Boltzmann averaged values from all conformers in the ensemble. The Boltzmann weights are determined by the energies calculated for each conformer, with a lower energy corresponding to a higher Boltzmann weight. As a result, more stable conformers on the PES have greater influence on the descriptors. These descriptors were calculated and retrieved using OBeLiX, an automated Python workflow developed by our group. [42] A more detailed explanation of the process can be found in Section 3.2.5.

The descriptors can be categorized in geometric, steric and electronic parameters, with the complete list given in shown in Table 2.1, which also highlights the differences between the three descriptor selections used.

The **36 descriptors selection** includes all descriptors used, with some descriptors using values obtained from calculations using the free ligand (FL) geometries. These FL geometries were retrieved from DFT Optimization calculations, and used DFT SP calculations to obtain FL descriptors. DFT SP calculations ensured that the FL geometries remained consistent with their geometry in the complex. This approach allows for the calculation of properties solely influenced by the ligand itself, without the interference from the rest of the complex. **The 24 descriptor selection** differs by excluding steric descriptors related to stereochemistry.

The 21 descriptor selection, additionally, differs to this with using only complex descriptors, where the other ones use free ligand descriptors. This 21 descriptors selection, is used by geometries which only use DFT SP, as it was not possible to extract the FL geometries from these calculations.

Steric descriptors

Steric descriptors represent the steric values for each quadrant (quad) and octant. The buried volume is used to quantify in percent how much of the metal center is occupied by the ligand. [43] Donor min and max refer to the two electron-donating atoms that forms bonds with the metal center.

Geometric descriptors

Geometric descriptors were related to the bite angle, cone angle, dihedral angles of the complex. The bite angle represents the angle between the two electron-donating atoms and the metal center. The cone angle describes the angle including the most outer atoms of the ligand, with the metal center positioned at one corner. The dihedral angle refers to the angle of two plane through two sets of three atoms.

Electronic descriptors

Electronic descriptors are related to the electronic properties of the complex, where energies and charges are calculated. The abbreviations in these descriptors are NBO (Natural Bonding Orbital), HOMO (Highest Occupied Molecular Orbital), LUMO (Lowest Unoccupied Molecular Orbital).

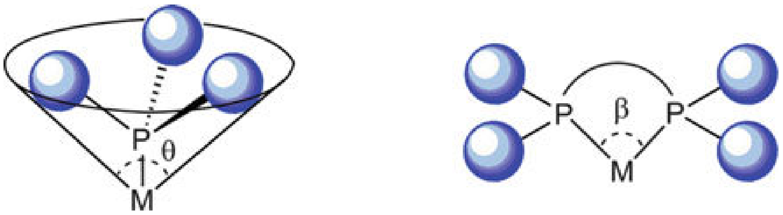


Figure 2.6: Visualization of the cone angle (θ) and bite angle (β). Figure adapted from Chemical Communications. [43]

Table 2.1: Overview of the different descriptor selections used and on which structures they are based on. FL refers to free ligand geometry is used, COMP refers to complex geometry is used.

Name	**36 descriptors (stereoC., free ligand)**	**24 descriptors (no stereoC., free ligand)**	**21 descriptors (no stereoC., complex)**
Steric			
'NE_quad'	Yes	-	-
'NW_quad'	Yes	-	-
'SW_quad'	Yes	-	-
'SE_quad'	Yes	-	-
'+,+,+_octant'	Yes	-	-
'-,+,+_octant'	Yes	-	-
'-,-,+_octant'	Yes	-	-
'+,-,+_octant'	Yes	-	-
'+,-,-_octant'	Yes	-	-
'-,-,-_octant'	Yes	-	-
'-,+,-_octant'	Yes	-	-
'+,+,-_octant'	Yes	-	-
'buried_volume_Rh_3.5A'	Yes	Yes	Yes
'buried_volume_donor_max'	Yes	Yes	Yes
'buried_volume_donor_min'	Yes	Yes	Yes
Geometric			
'bite_angle_sin'	Yes	Yes	Yes
'bite_angle_cos'	Yes	Yes	Yes
'cone_angle_sin'	Yes	Yes	Yes
'cone_angle_cos'	Yes	Yes	Yes
'dihedral_angle_1'	Yes	Yes	Yes
'dihedral_angle_2'	Yes	Yes	Yes
'distance_Rh_max_donor_gaussian'	Yes	Yes	Yes
'distance_Rh_min_donor_gaussian'	Yes	Yes	Yes
Electronic			
'distance_pi_bond_1'	Yes	Yes	Yes
'distance_pi_bond_2'	Yes	Yes	Yes
'sum_electronic_and_free_energy_dft'	Yes	Yes	-
'dipole_moment_dft'	FL	FL	COMP
'lone_pair_occupancy_min_donor_dft'	FL	FL	COMP
'lone_pair_occupancy_max_donor_dft'	FL	FL	COMP
'dispersion_energy_dft'	FL	FL	COMP
'nbo_charge_Rh_dft'	FL	FL	COMP
'nbo_charge_min_donor_dft'	FL	FL	COMP
'nbo_charge_max_donor_dft'	FL	FL	COMP
'homo_lumo_gap_dft'	FL	FL	COMP
'lone_pair_occupancy_max_donor_dft_abs_diff'	Yes	Yes	-
'lone_pair_occupancy_min_donor_dft_abs_diff'	Yes	Yes	-

2.6.2. Descriptors for ECFP representation

ECFP descriptors tell essentially nothing about electronic or geometric properties of the complex, but act essentially as a barcode to describe the topology of the complex. The values in these descriptors are boolean, and how they are created in described in more detail in Section 2.7.2. Patterns and similarities between different complexes can be found as these descriptors are created in such systematic way that values within a descriptors refer to the same group of atoms.

2.7. Representations

Several representations are used for the ML algorithms to identify patterns and similarities. These representations vary in simulation methods or selection of descriptors. These representations are essentially a list of selected descriptors, where each descriptor has a value for each complex.

In the ML algorithms, these representations are linked to the experimental data (outcome). The descriptors within these representations are crucial for the model's performance, as the ML algorithm finds patterns and similarities among them. The goal of using these different representations is to assess whether increasing the accuracy of descriptors leads to improved model performance.

2.7.1. DFT-based representation

The DFT simulation based descriptors, vary in geometry calculation method, conformers selection and descriptors used.

Geometries were calculated using three methods: a handmade structure which is DFT optimized, a handmade structure which is xTB optimized in CREST, and a handmade structure which is first xTB optimized in CREST and then DFT optimized.

Two conformer selection methods were used (in case of a conformer ensemble): the minimum energy structure, which considers only the most stable conformer, and the Boltzmann averaged structure, which accounts for the entire conformer ensemble.

Three descriptor selection sets were used: a 36-descriptor set that includes stereochemistry and DFT optimized free ligand parameters, a 24-descriptor set containing DFT optimized free ligand parameters, and a 24-descriptor set containing DFT SP complex descriptors.

2.7.2. ECFP

Extended Connectivity Finger Prints (ECFP) are circular topological fingerprints used to characterize molecular complexes. These fingerprints contain a lot of fingerprint parameters, which are boolean values and describe a part of the molecular structure. These fingerprints are systematically made in a way, which allows for very precise and structural mapping of parameters. These allow for excellent similarity searching and clustering if they have the same fixed-length bit string. [44][45]

Creation

The creation of ECFP is based on a simple and structured concept. It starts with user-provided inputs such as a readable chemical structure and fixed-bit string. These readable structure is typically given in SMILES (.smi), structure-data file (.sdf), MOL/MOL2 (.mol) or XML files (.xml). The fixed-bit string defines the number of 'parameters' these fingerprint consist of. This is typically chosen based on the complexity of the compounds, as too much bits can lead to overfitting and too few bits result in loss of data. These bits are generated using hashed data, where each hash contains information of the atomic number, heavy atom neighbor count, valance (minus the number of hydrogen atoms), ring membership and atomic charge.

In a molecule or complex, the data of each atom processed iteratively. Each atom is considered as starting point and its n-th neighbors are considered n iterations from the starting point, creating a circular iteration. This process goes iteratively through the entire molecule, where each unvisited neighbor updates the current identifiers, similarly like in graph traversal. Only non-duplicate combinations are added as new information to the identifiers, and where identifiers are found in different iterations, only the one with the lowest iteration count is kept. once all non-duplicate identifiers are found, they are then hashed, each referring to specific (sub)groups of atoms. An example is shown in Figure 2.7, 2.8 and 2.9.

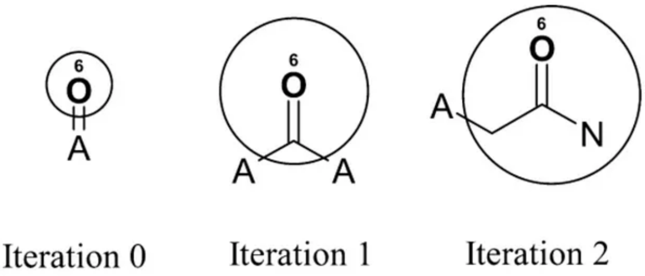


Figure 2.7: Example of how circular iterations are taken in a molecule. Figure adapted from Medium. [46]

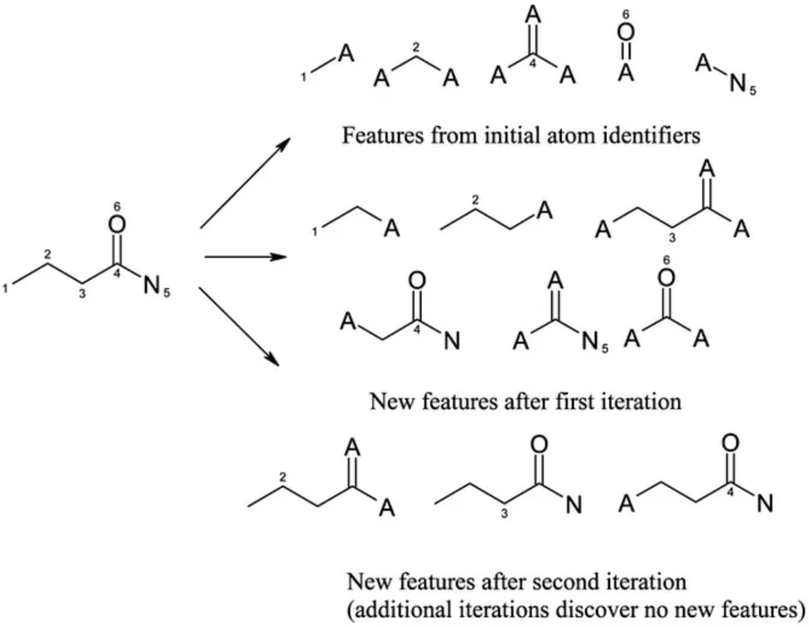

Features from initial atom identifiers
New features after first iteration
New features after second iteration
(additional iterations discover no new features)

Figure 2.8: Example of new features that are added after each iteration. Figure adapted from Medium. [46]

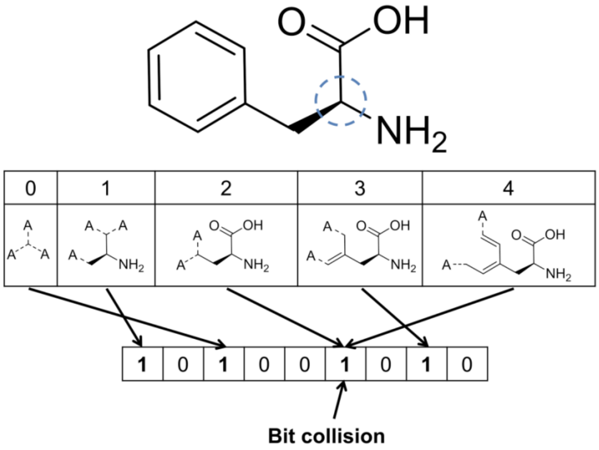


Figure 2.9: Example of hashing and bit collision in the final ECFP representation. The values are binary, and there is 1 bit collision where two different identifiers refer to the same point in the representation. Figure adapted from Research Gate. [47]

2.8. Machine Learning

Machine learning is big part of AI and can be scribed as "The field of study that gives computers the ability to learn without being explicitly programmed." – Arthur Samuel. Within Machine Learning there are also categories to distinguish different variant of ML. There is supervised, unsupervised and reinforced learning. With supervised and unsupervised being a relevant for the models we want to use.

2.8.1. Supervised learning

Supervised learning, uses labeled data to train algorithms to classify or predict data. [48] Labeled data is data with the target value known. These target values can be either a categorical or continuous value. This method is widely used in the chemical sector, as it relies on known target values. These values are often values we derive from observations in experiments. Supervised learning provides predictions that can be validated against real-world data. In context of catalyst design, it can find correlations between the conversion (target) and properties of the catalysts, making the development of more effective catalyst better.

Classification

Classification is a component of supervised learning and uses categorical data, where the categories are defined based on thresholds set by the user. These classes can be binary or multi-class. Classification typically has a higher tolerance for errors, require less data to achieve good predictions and have simpler evaluation metrics compared to regression. For these reasons classification often outscores regression and is considered easier to predict. We can classify catalysts in 'bad' or 'good' catalysts, or in case of multi-class classification, we add 'deactivation' or other categories to it. In this thesis, the Balanced Accuracy (BA) will be used to evaluate the classification performance of the model as the data split can lead to unbalanced datasets. This method labels the predictions results as True Positive (TP), False Negative (FN), True Negative (TN), False Positive (FP).

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{2.9}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.10}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.11}$$

Regression

Regression only uses continuous data, where a value can only be correct if the real and predicted value are exactly the same. Therefore, different evaluation metrics are used compared to regression. Typically in regression, metrics are used that are related to the distance between the real and predicted value. Some examples are mean squared error (MSE), residual sum squared (RSS) and R-squared (R^2). In this report R^2 will be used as evaluation metric for regression results. The formula for R^2 is given in Equation 2.12.

$$\mathrm{R}^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{2.12}$$

Where y_i, $\hat{y}_i$ and $\bar{y}$ are the observed value, predicted value and mean observed value respectively.

2.8.2. Unsupervised learning

Unsupervised learning uses data without explicit target values (unlabeled data) to train algorithms by finding patterns or relationships between the data.

Clustering

Clustering is a form of unsupervised learning, where data is clustered based on similarities between data, common methods include k-means clustering or DBSCAN. K-means clustering divides the data points in k amount of clusters and is based on distances from the centroid points, while DBSCAN groups these points based on density.

In this thesis DBSCAN is used for Filtering for DFT, where conformers are clustered based on their likelihood to converge to the same minimum. Previous work has tested these methods in combination with an algorithm, and DBSCAN works better for this purpose. [49] Filtering for DFT is explained in more detail in Section 3.2.3.

Dimensionality reduction

Dimensionality reduction is another part of unsupervised learning, where the algorithm analyzes the data and reduces the dimensionality to retain valuable information and be more interpretable.

In this thesis Principal Component Analysis (PCA) is used, which is a linear based model. The data is usually shown in 2D or 3D, where the Principal Components shown on each axis are the components with the highest variance. This allows for the largest separation of points.

3

Methods

3.1. Experimental data

Experimental data was obtained by Johnson & Johnson. [50] High-Throughput experimentation on Rh-NBD complexes was conducted and the first half of 192 different complexes was provided. The full list can be found in Appendix A.1. The experiments were performed twice, with conversion data recorded at 15 minutes for both runs and at 30 minutes for 'run I'. This resulted in two experimental datasets: one containing the average conversion from both runs at 15 minutes ('run I' and 'run II') and another with the conversion at 30 minutes ('run I'). The reaction conditions were T = 303.15 K and P = 3 bar H_2. The samples contained 0.2 mol% Rh-catalyst complex and a total volume of 150 μl (10.2 μl NBD and 139.8 μl THF). The analysis was done using GC-FID in MeCN.

The company also provided an interpolated conversion plot between 0 and 30 minutes, which is shown in Figure 3.1. With most curves already having reached their final conversion at 15 minutes. Some curves show decrease of conversion after 30 minutes.

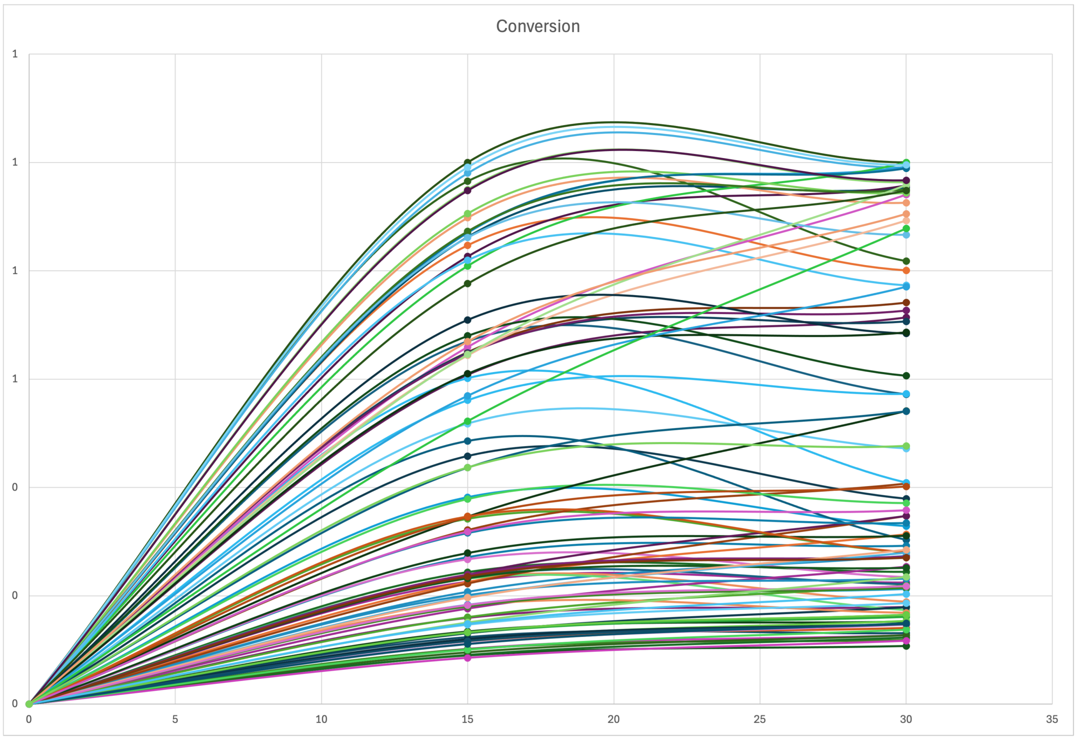


Figure 3.1: Interpolated conversion curves of the experimental data on Rh-NBD complexes. Reaction conditions: 303.15 K, 3 bar H_2. The data point are at 15 and 30 minutes.

3.2. Workflow

The complete workflow for thesis is adapted from previous work [49]. It begins with the creation of input structures, which are then processed through conformer searching and pruning packages in Python. The resulting conformer ensembles are reduced further by filtering before undergoing DFT calculations.

In this thesis, this workflow is finished by doing DFT calculations, calculating descriptors, running these ML algorithms and analyzing the results.

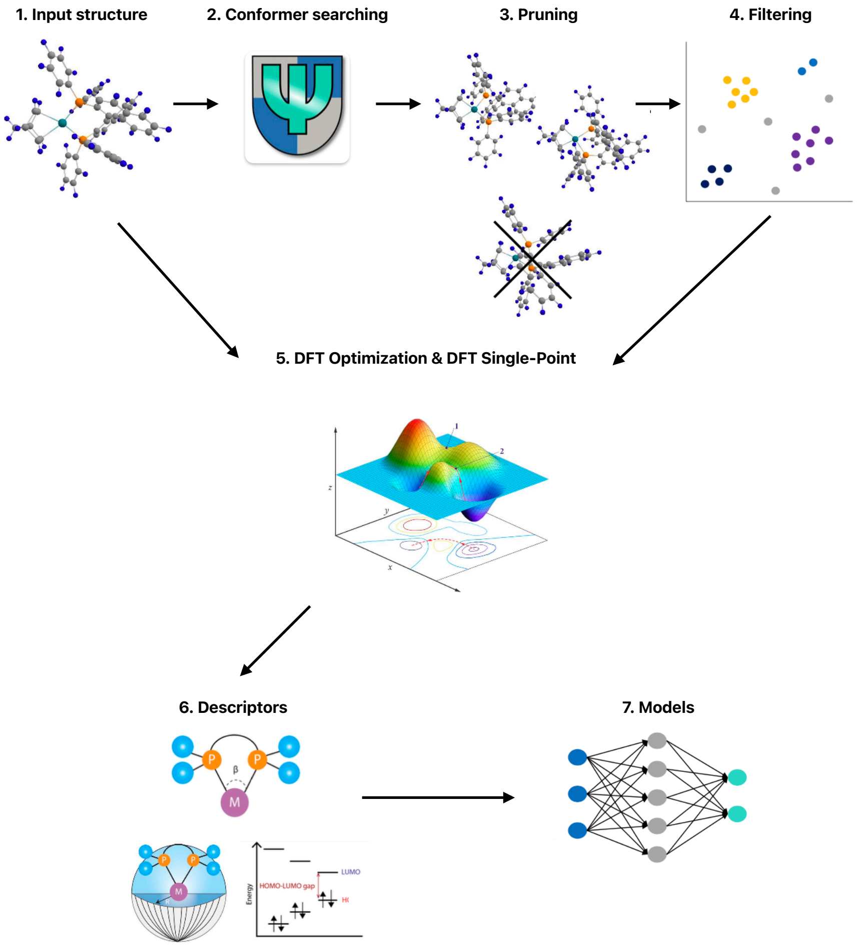


Figure 3.2: Complete workflow. Figure adapted from Sára's thesis. [49]

3.2.1. Input structures

Input structures were created from 192 rhodium-NBD complexes. These complexes contain rhodium with +1 oxidation state and a molecular moiety that resembles the precatalyst structure. Different ligand families were investigated, which are specified in Appendix A.1. The catalyst structures are represented using Cartesian coordinates in (.xyz) files. These initial structures were handmade. These structures are the starting point for DFT calculations and CREST to find conformers and optimize these structures.

3.2.2. Conformer searching and pruning

The input structures were used in CREST to find conformers and pruned using MORFEUS.

CREST

CREST version 2.12 and xTB version 6.4.0 was used to generate and find conformers. The hybrid potential GFN2-xTB//GFN-FF was used, for the improved accuracy and relatively low computational cost. The GFN-

FF method optimizes conformational geometries using Force Field calculations, then GFN2-xTB single point calculations are done which are more accurate in our case since we have heavier atoms. The multiplicity and charge were set to 1, since the total complex has a charge of +1, Rh (1), NBD (0), ligands (0).

MORFEUS

The MORFEUS Python package reduced and preprocessed these conformer ensembles by conformer pruning. The MORFEUS package has multiple pruning tools: energy pruning, enantiomer pruning and root mean square deviation (RSMD) pruning. All were used to reduce the conformer ensemble. Energy pruning, removed conformers above a certain energy threshold, based on relative energies within the ensemble. Enantiomer pruning, removed conformers with a different chirality compared to the original input structure. RSMD pruning, removed duplicate conformers by comparing the difference between molecules.

3.2.3. Filtering for DFT

Filtering for DFT is required, as CREST generates a large number of 'stable' conformers, many of these are very close on the PES. After DFT Optimization, these closely packed conformers would likely converge to the same minimum. Filtering these conformers would reduce the computational cost significantly, as for example DFT Optimization would be done over 500 structures instead of 3000.

This filtering was performed using a classification algorithm trained on a subset of the data, which was analyzed and improved using Density-Based clustering (DBSCAN). A more detailed description of how this algorithm works can be found in Sára's thesis. [49] The ϵ used is 0.19.

3.2.4. DFT calculation

DFT Optimization (DFT OPT) and DFT Single-Point (DFT SP) calculations were done for all xTB geometries generated in CREST. The calculations were done on Gaussian 16 (C.0239). PBE0-D3(BJ)/def2-SVPP level of theory was used to minimize computational cost, while being very accurate. PBE0 refers to the hybrid functional, D3(BJ) accounts for dispersion corrections with a high accuracy for non-covalent bonds and def2-SVPP is a basis set optimized for polarization accuracy for heavier atoms. For DFT OPT, an additional frequency analysis was done to identify imaginary frequencies. In case of an imaginary frequency, PyQRC was used to create a slightly modified geometry, which is DFT optimized again to end up in a stable minimum.

3.2.5. Descriptor calculation

The descriptors were calculated using the OBeLiX workflow. For DFT SP calculations of an xTB-optimized geometry, the descriptor_calculator.py script was employed to obtain the descriptors. Python notebooks contain code to generate descriptors sets from the DFT OPT calculations. The calculation of descriptor sets for the free ligand consisted of more extensive steps such as the extraction of the free ligand geometry, however code to generate descriptors sets from DFT SP calculations do not have these extraction of free ligand geometry.

3.2.6. Model selection

Previous work used Random Forest models with hyperparameter optimization using grid search to obtain more robust and reliable results. Three different train/test splits were evaluated and each grid search contained a predefined parameter space and used k-fold cross-validation, to prevent overfitting of the data. Ultimately an 80/20 train/test split was chosen as optimal configuration for the ML models. [16]

Since we have a small dataset of only 192 or even 96 samples, cross-validation allowed us to reuse the samples and to get 'maximum' data from this dataset. We partitioned the data in 5, which increased our dataset by five. This approach improved the efficiency and allowed us to confidently say we obtain the best performing model for each random state.

3.2.7. Representations

Different representations are used to test different levels of computational power and importance of descriptors. DFT-based descriptors are chosen by ourselves and the main target. As alternative method we have added ECFP, with a fixed-bit string of 512 parameters. 1024 and 2048 bits were also tested, but previous work chose 512 to cause the least amount of useless bits (bits that are uniform for all samples). For baseline testing, two models were made which contains 24 or 36 descriptors with random floats between -100 and 100. The amount of descriptors and 'ligands' were matched to the DFT-based descriptors to keep a fair comparison. The representation are described in more detail in Section 4.1.

4

Results & Discussion

This chapter begins with an overview of generated data and available representations along with their corresponding abbreviations. It then analyzes the new obtained experimental data on the Rh-catalyst with NBD as substrate. Next, it revisits previous work, describing the results from the five different substrates with the original representation (Original Substrate model). Further on, this original representation will be combined with the new experimental data and on the representations will be altered. These variations, are either variations in method used (DFT, ECFP, random), method of representation of structure (static, dynamic) and/or descriptors chosen.

4.1. Data generation

For the results, we have generated missing data from previously investigated representations, created five new representations for the ML models, and tested three different descriptor selections. We evaluated the performance of the ML models using the different representations and descriptor selections, and also analyzed descriptors, energies and ECFP. Everything is done for the 192 different Rh-complexes, except the fully DFT Optimized conformer ensemble (CE). Here we have done the DFT OPT for all complexes but only processed the first 96 complexes as time was limited and we only have experimental data for the first 96 complexes.

There are multiple representations used in this thesis, each with varying levels of computational cost and accuracy. Tables 4.1 and 4.2 give a clear overview of all representations with their description and key differences. There are 12 different representations used, 9 of them are related to DFT-based descriptors, 1 using ECFP and 2 baseline models containing 36 or 24 descriptors with randomly generated values. In the DFT-based representations, variations were introduced in the input structure, DFT methods and descriptor selection.

The input structures refer to the geometries used in the DFT calculations. These structures were varied to improve the descriptors and bring them closer to reality. In previous work, a handmade structure was used as input structure, but now additional representations were tested, where CREST generated these structures. CREST generated the entire CE, from which we created two types of representation from. One representation which only uses the most stable structure of the CE (lowest energy conformer) as input, and another where DFT calculations are done for all conformers in the CE and combined using Boltzmann weighting. The idea of using CREST is to create a more realistic representation of the complexes, as the CE can introduce the dynamic aspect of a complex or find the most structure with the lowest energy.

Two DFT methods were used: DFT OPT and DFT SP calculations. Given the high computational cost of DFT OPT, DFT SP calculations, which are based on xTB calculated geometries from CREST were evaluated to determine whether the additional cost of DFT OPT results in a significant advantage.

Various descriptor selections were used to test their influence. Specifically, descriptors related to stereochemistry and those where free ligand geometries were used. These were tested, as stereochemistry should have minimal impact as our substrate is symmetrical. Additionally, free ligand descriptors were included in previous work, as a generalized substrate was used. This may no longer be necessary, as both experimental and simulated data now corresponds to the exact same compounds.

These changes can also be recognized in their names:

- Representations without **'CREST'** refer to the static representations where an initial handmade structure is used as input for DFT calculations.
- Representations with **'DFT'** refer to DFT Optimized structures and calculations, while those without **'DFT'** corresponds to DFT SP calculations.
- Representations with **'CREST'** & **'[E min.]'** refer to the static representations where only the lowest energy conformer from the CREST generated CE is used.
- Representations with **'CREST'** & **'[boltz. avg.]'** correspond to the dynamic representations, where the entire CREST generated CE is used, considering the Boltzmann-weighted average.

The descriptor selection is specified by the terms in parentheses:

- No brackets indicates all 36 descriptors are used, with free ligand geometries applied for certain descriptors.
- **'(no stereoC.)'** refers to the 24 descriptor selection, where stereochemistry related descriptors are excluded.
- **'(no stereoC., no free lig.)'** excludes stereochemistry descriptors and replaces free ligand descriptors with complex descriptors, as the free ligand geometries could not be extracted from the DFT SP calculations.

Table 4.1: Overview of all representation names, with their more detailed description. All DFT representations start from a handmade structure, which are DFT Optimized or passed through CREST to generate a conformer ensemble (CE). When a CE is applicable, structures are included based on the Boltzmann averaged energy or the lowest energy conformer is selected and then used in DFT calculations.

Name	**Description**
OG-DFT-NBD	Handmade structure + DFT Optimization (Original Substrate model)
DFT-NBD	Handmade structure + DFT Optimization (Original NBD model)
DFT-NBD (no stereoC.)	Handmade structure + DFT Optimization (stereochemistry descriptors removed)
DFT-NBD (no stereoC., no free lig.)	Handmade structure + DFT Optimization (stereochemistry removed, free ligand replaced for complex descriptors)
CREST-NBD [boltz. avg.] (no stereoC., no free lig.)	Handmade structure + CREST + DFT Single Point + Boltzmann averaged (stereochemistry removed, free ligand replaced for complex descriptors)
CREST-NBD [E min.] (no stereoC. no free lig.)	Handmade structure + CREST + DFT Single Point + Mimimum energy selection (stereochemistry removed, free ligand replaced for complex descriptors)
CREST-NBD [E min.] (no stereoC.)	Handmade structure + CREST + DFT Single Point + Minimum energy selection + DFT Optimization (stereochemistry removed)
CREST-DFT-NBD [boltz. avg.] (no stereoC.)	Handmade structure + CREST + DFT Optimization + Boltzmann averaged (stereochemistry descriptors removed)
CREST-DFT-NBD [E min.] (no stereoC.)	Handmade structure + CREST + DFT Optimization + Minimum energy selection (stereochemistry descriptors removed)
ECFP	Extended Connectivity Fingerprint using fixed-length bit string 512
Random-24	Randomized values (between -100 – 100) for 24 descriptors (Baseline)
Random-36	Randomized values (between -100 – 100) for 36 descriptors (Baseline)

Table 4.2: Overview of all representations with their characteristics. Representation names containing 'CREST' used CREST generated structures, if the final geometry is xTB-based only DFT SP was used.

Name	Experimental dataset	Final geometry	Energy based on	Descriptors
OG-DFT-NBD	5 different substrates	DFT-based	DFT OPT geometry	36
DFT-NBD	NBD	DFT-based	DFT OPT geometry	36
DFT-NBD (no stereoC.)	NBD	DFT-based	DFT OPT geometry	24
DFT-NBD (no stereoC., no free lig.)	NBD	DFT-based	DFT OPT geometry	21
CREST-NBD [boltz. avg.] (no stereoC., no free lig.)	NBD	xTB-based	xTB geometry	24
CREST-NBD [E min.] (no stereoC., no free lig.)	NBD	xTB-based	xTB geometry	21
CREST-NBD [E min.] (no stereoC.)	NBD	DFT-based	xTB geometry	21
CREST-DFT-NBD [boltz. avg.] (no stereoC.)	NBD	DFT-based	DFT OPT geometry	24
CREST-DFT-NBD [E min.] (no stereoC.)	NBD	DFT-based	DFT OPT geometry	24
ECFP	NBD	-	-	447
Random-24	NBD	-	-	24
Random-36	NBD	-	-	36

4.2. Experimental data

Beginning with the new obtained experimental data on Rh-catalysts with NBD as substrate. As mentioned in Chapter 3, this data was generated in collaboration with Johnson & Johnson. [50] They conducted High-Throughput experimentation on Rh-NBD complexes with 96 different ligands. The experiments were performed twice, with conversion data recorded at 15 minutes for both runs and at 30 minutes for 'run I'. This resulted in two experimental datasets: one containing the average conversion from both runs at 15 minutes ('run I' and 'run II') and another with the conversion at 30 minutes ('run I'). The reaction conditions were T = 303.15 K and P = 3 bar H_2.

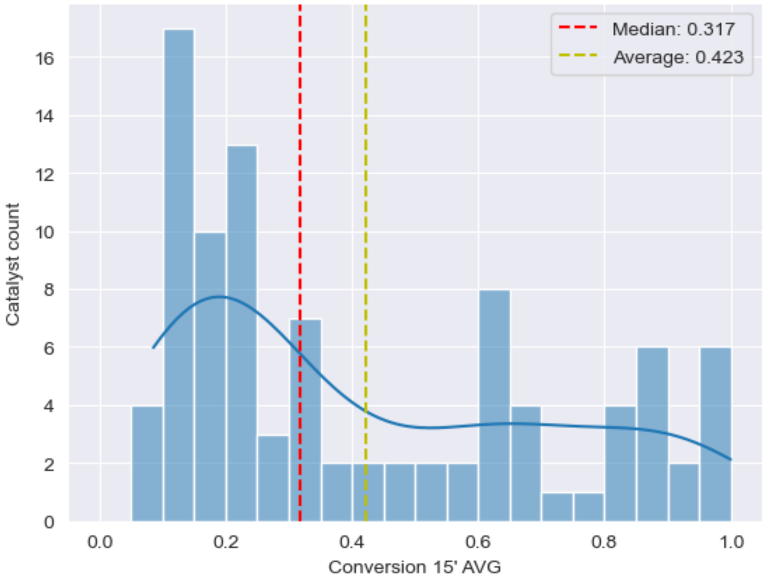


Figure 4.1: Distribution plot of the average 15 minute conversion results for the first 96 Rh-NBD complexes, showing most data points in either the low or high conversion regime. The median conversion 0.317.

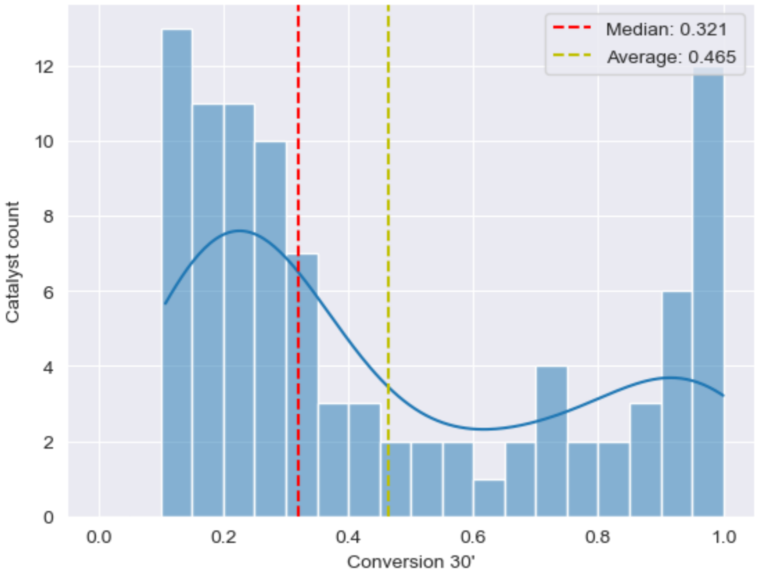


Figure 4.2: Distribution plot of the 30 minute conversion results for the first 96 Rh-NBD complexes, showing most data points in either low or high conversion regime. The median conversion is 0.321.

The data distribution is presented in Figures 4.1 and 4.2. These figures show a broad, asymmetric distribution of the data. Figure 4.1 illustrates a higher concentration of data points at at lower conversions, while Figure 4.2 shows a small shift towards higher conversions (with the average increasing from 0.423 to 0.465). However, the median remains almost unchanged (0.317 vs 0.321), suggesting that most, if not all, of these low performing catalysts remained below this threshold. Visually, it is also noticeable as the bars left of the median are now more evenly distributed, while the right side side increases to conversions closer to 1.

The threshold used for classification in previous work ('Original Substrate model')(threshold = 0.8) will result to unbalanced training datasets for the ML algorithms, as the majority of the data is below this value. The threshold is set to the median conversion for each dataset, to create a more balanced classification. However, since these thresholds are different in different datasets, it is not entirely 'fair' to compare results from different datasets. This will be avoided, and classification will only be compared if it involves the same dataset.

4.3. Original substrate models

To understand the decisions made in this thesis, it is important to first reevaluate the results obtained in the model developed in previous work of the group. Their research focused on a non-specific predictive model for a case of asymmetric hydrogenation of imines. The ligand descriptors were thus separately supplemented with descriptors of the substrate. This model was used to predict conversion through classification and stereoselectivity through regression. However, as previously mentioned, the results were inconsistent across the five investigated substrates, with only 'SM4' and 'SM5' showing cases with high performance. These models and substrates are described in more detail in [16].

In this thesis, an experimental dataset for NBD is used, shifting the model from a non-specific to a substrate-specific approach. Additionally, only conversion is analyzed, as NBD is symmetrical and eliminates any stereoselectivity in the complexes. Since conversion is now the sole prediction task, both classification and regression will be performed on conversion. This requires generation of regression results for conversion using the 'Original Substrate model', as previous work only applied regression to stereoselectivity. Additionally, to maintain consistency in training data across all representations, only the first 96 ligands from the substrate datasets were used, as they correspond with the first 96 ligands from the NBD dataset. However, due to structural issues caused by CREST, which will be explained later, ligands L19 and L24 were excluded, leaving 94 usable data points in the datasets. This set will be still referred as the first 96 ligands. The generated data is presented in Figure 4.3 and will serve as a benchmark for model performance.

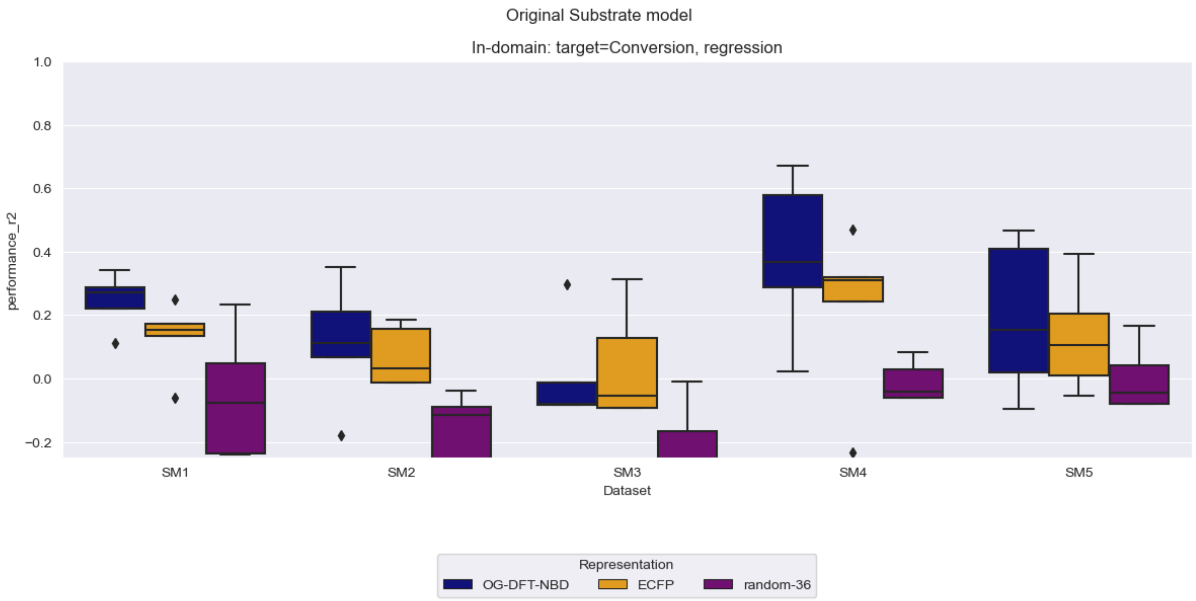


Figure 4.3: Box plot of regression performances on conversion using the Original Substrate model (96 data points). Five substrates are shown with representations: OG-DFT-NBD, ECFP and random-36.

Figure 4.3 shows that SM4 and SM5 have cases where the performance is very high, but also cases of very low performance. This results in the box plots having a large range. SM1, SM2 and SM3 show a small spread of the performances, but in all cases low performance ($R^2 < 0.4$).

4.4. Sensitivity analysis

In the complete ML model workflow two things can be randomized: the initialization of the ML model and the train/test data split used in the ML algorithms. A sensitivity analysis is done to determine the importance and influence of these randomizations.

4.4.1. Model

Three different model initializations were tested over the two datasets and showed no significant change in model performance for both classification and regression (Figures 4.4 and A.1).

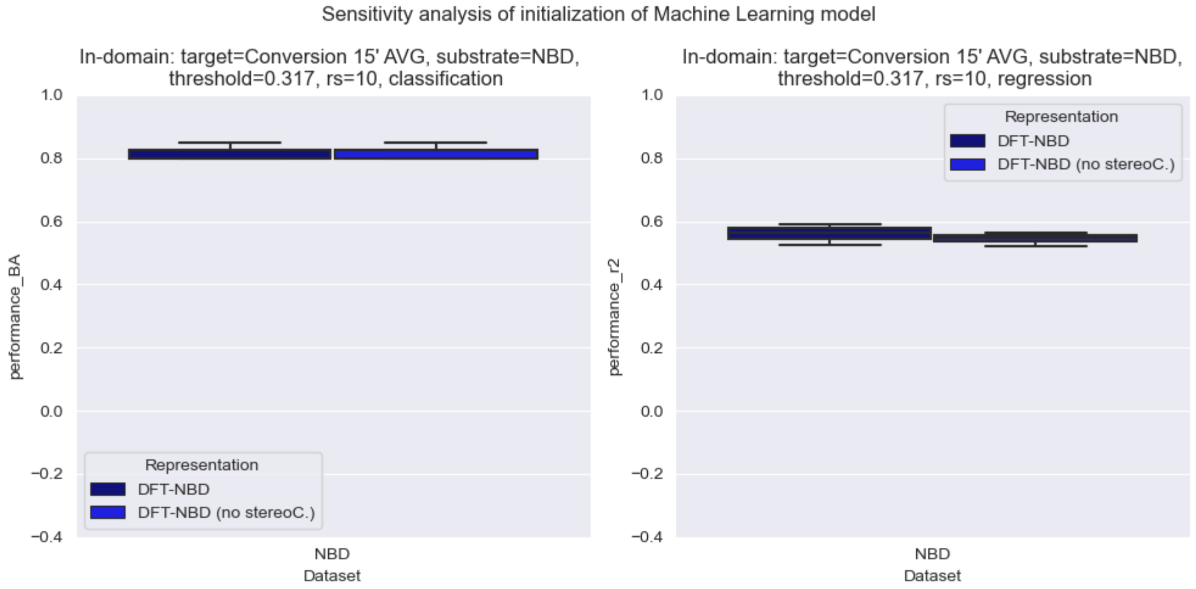


Figure 4.4: Sensitivity analysis of initialization of the Machine Learning Model. Box plots are shown of the performance of the ML models using classification (left) and regression (right). Both methods show small changes in performance in two different representations.

4.4.2. Data

Three different test/train data splits were tested over the two datasets and showed a noticeable change in performance (Figures 4.5 and A.2).

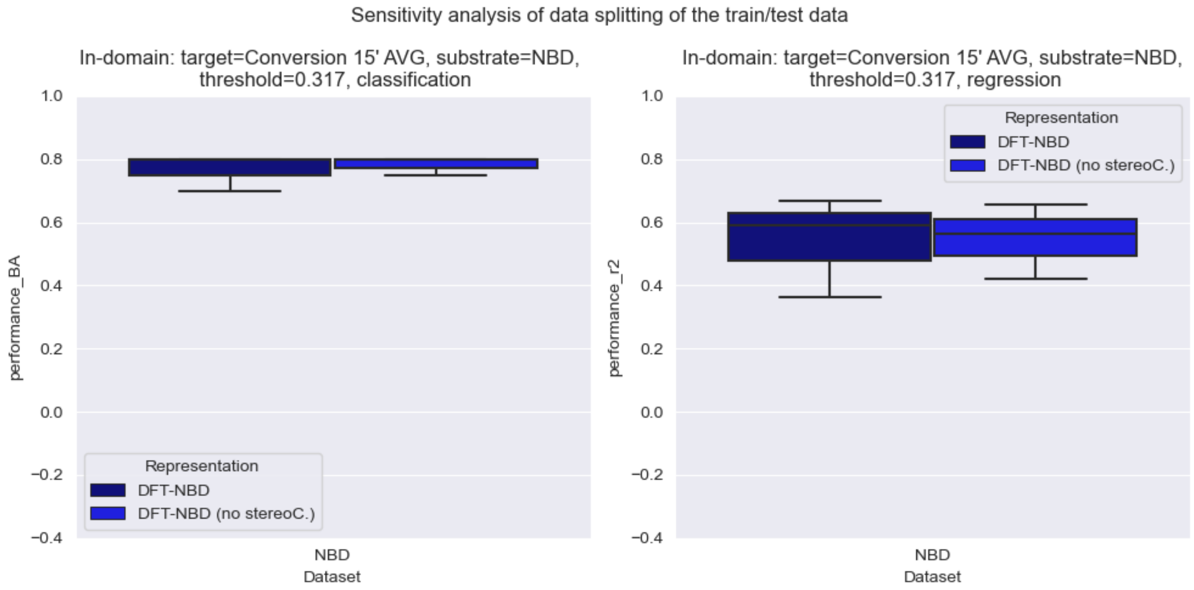


Figure 4.5: Sensitivity analysis of data splitting of the train/test data. Box plots are shown of the performance of the ML models using classification (left) and regression (right). While classification shows a small change in performance, regression shows a much larger change.

The data split is dominant over the initialization, so the most change in performance would be obtained by parametrization of the data split. The ML model initialization random state is kept constant since the influence is negligible, and the data split is randomized using five random states [8 20 30 62 66]. This sensitivity analysis also shows that the train/test split influences the performance a lot.

4.5. Static representation (DFT-NBD)

Moving on to the first model representation of this thesis. In this model representation, the same descriptors are used as the Original Substrate model, however the new experimental dataset is introduced. The results of classification and regression are shown in Figure 4.6 and 4.7.

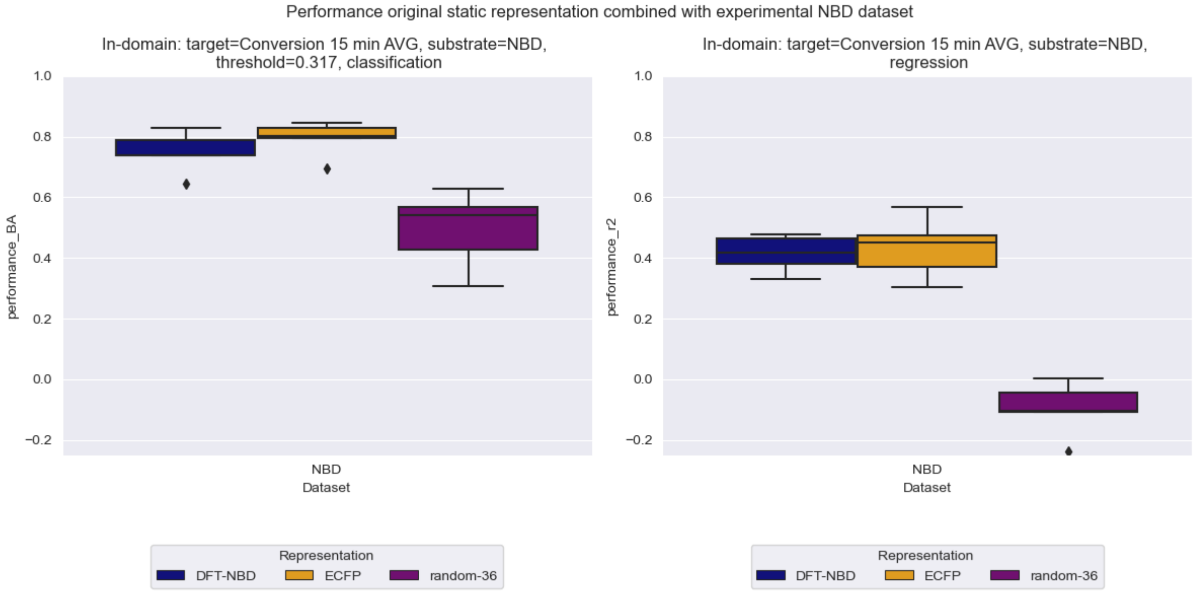


Figure 4.6: Box plot of model performances for classification (left) and regression (right) for average conversion at 15 minutes. Static DFT-NBD representation is used, in combination with the NBD dataset.

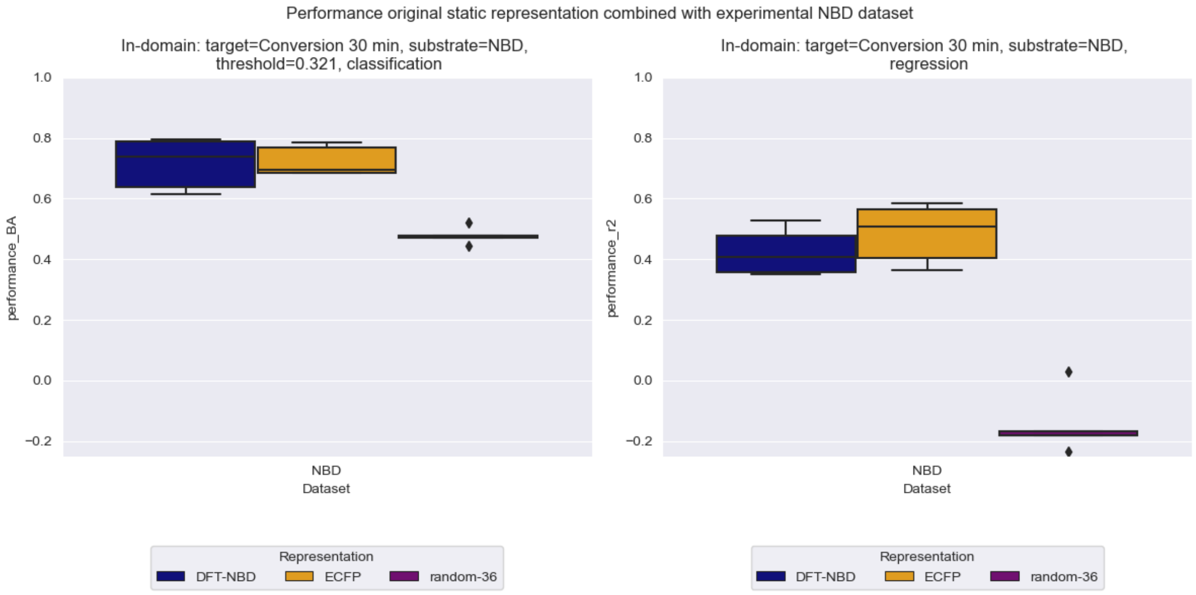


Figure 4.7: Box plot of model performances for classification and regression for conversion at 30 minutes. Static DFT-NBD representation is used, in combination with the NBD dataset.

Both the average 15 minute and 30 minute conversion figures show similar results, with high performance for classification and regression score around 0.4. In all cases the DFT-NBD performances are comparable to ECFP, and significantly higher than the random baseline. This suggests that the models are able to find some patterns using the DFT-NBD and ECFP representations.

Original NBD model vs Original substrate model

The static DFT-NBD representation, using the new NBD dataset, is compared to the five substrate datasets in Figure 4.8. The figure highlights the differences between all these models. The average 15 minute dataset is chosen to represent the NBD dataset, as the performances for 15 minute averaged and 30 minute datasets were very similar. The addition of the new dataset resulted in a much smaller spread in the DFT-NBD representation compared to the OG-DFT-NBD representation in the substrate representations. While the new representation did not necessarily improve on absolute performance, it did improve on consistency. Surprisingly, the ECFP representation also showed improvement compared to the representations in the substrate models.

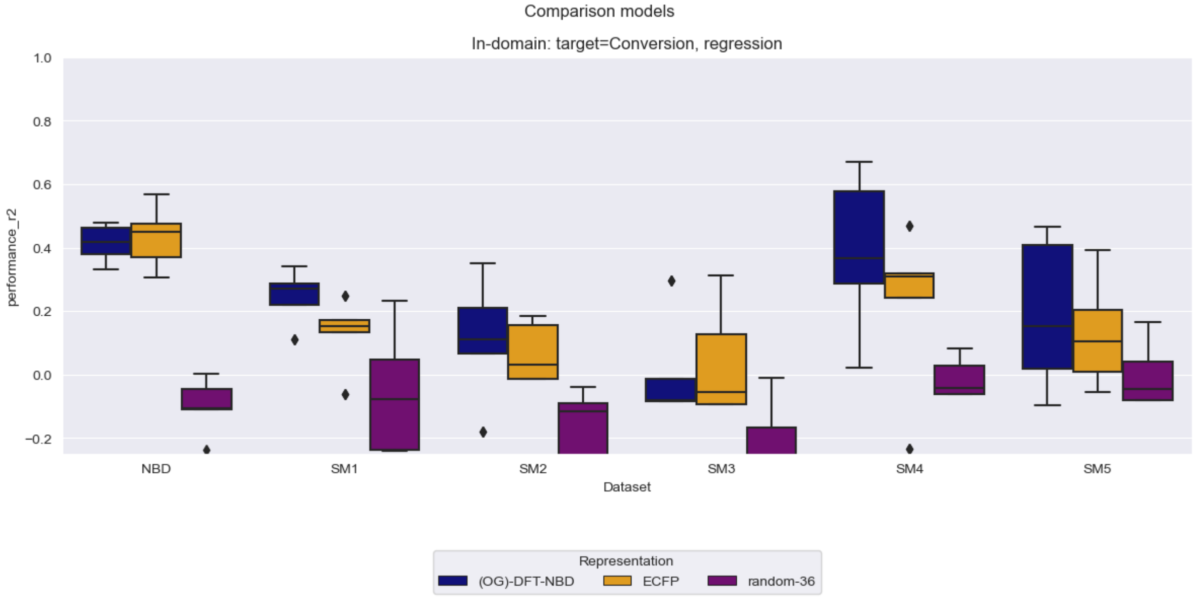


Figure 4.8: Comparison of performances between DFT-NBD representations using the new NBD dataset and the five substrate datasets. SM**X** refers to the different substrates used.

Feature Importance

Feature Importance (FI) plots reveal which descriptors contributed the most on the decision making of ML algorithms. Figure 4.9 shows the averaged FI over the five random states used for DFT-NBD. It can be seen that the descriptors which lead to the highest performance were mostly electronic descriptors except SW quadrant (steric) and bite angle (geometric). The bite angle specifically had the highest importance. These descriptors are reasonable as other studies have that ligands with a certain bite angle can influence the catalytic activity. [51][52]

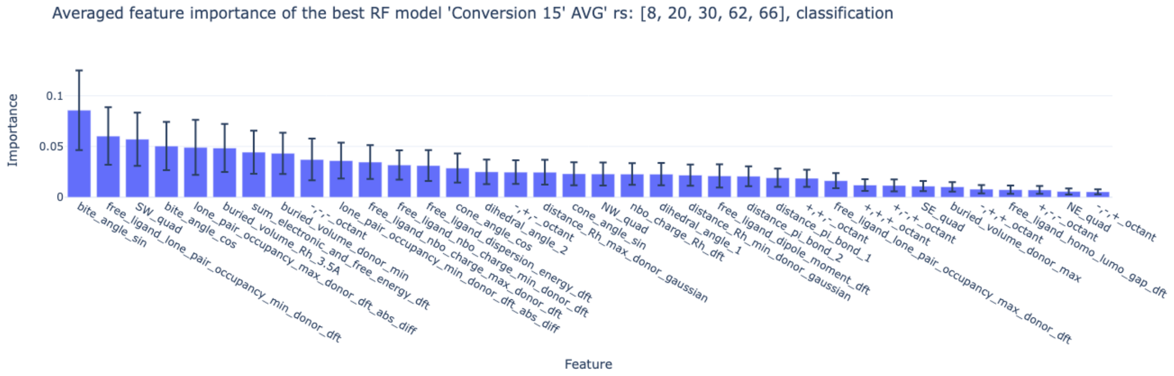


Figure 4.9: Feature importance plot of all descriptors used in DFT-NBD. Mostly electronic and descriptors and the bite angle (geometric descriptor) show dominating contribution with the lowest influence from steric descriptors.

4.5.1. Removal stereochemistry

The FI plot in Figure 4.9 showed almost no influence of steric descriptors (quadrants and octants). These steric descriptors were important in case of stereochemistry, as the steric effects might prevent conversion towards a specific enantiomer. In the new experimental dataset about Rh-NBD complexes this cannot be the case, as NBD is symmetrical and eliminates any stereoselectivity. Including these steric descriptor might lead to over-specification, which can introduce correlations that may exist but do not have any influence on the conversion. Removing these descriptors related to stereochemistry results in 12 less descriptors, described in Table 2.1 in Chapter 2.

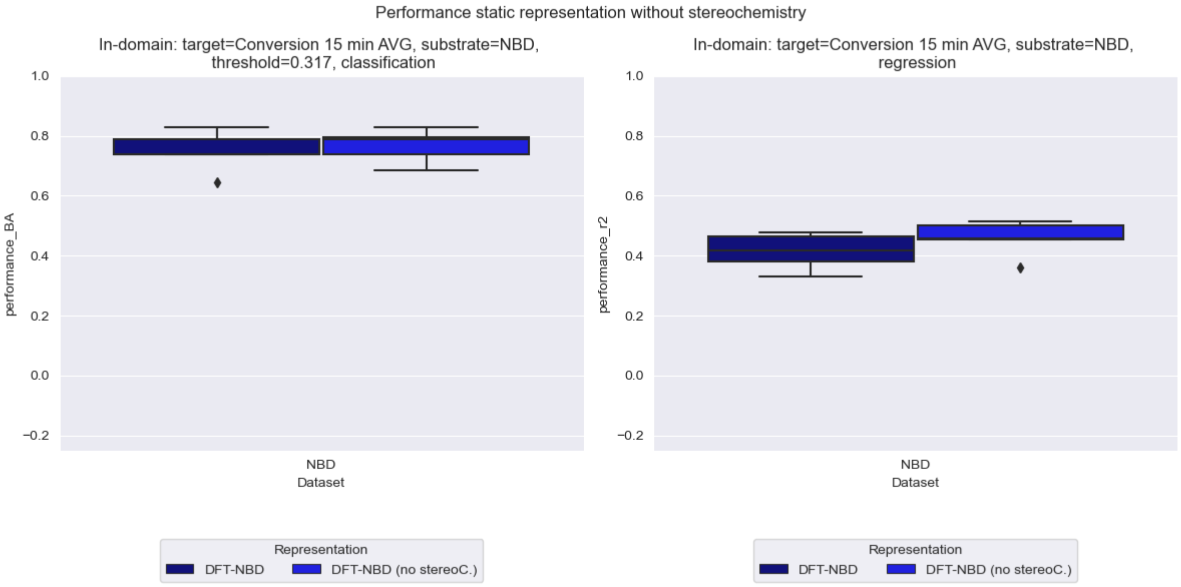


Figure 4.10: Effect of removing the stereochemistry descriptors on the NBD average 15 minute dataset. On the left the classification performances are shown for DFT-NBD with stereochemistry and DFT-NBD without stereochemistry. Performances are similar in both representations.

Figure 4.10 shows the removal of stereochemistry in the average 15 minute dataset. The effect in both classification and regression is minimal, making the addition of stereochemistry negligible as expected. The results for 30 minutes are shown in Figure A.3, and share the same observation, where the performance is in the same range and not significantly different.

Since the removal of stereochemistry is logical and also show minimal change in the results, it has been decided not to include stereochemistry in the following sections (unless specified otherwise).

4.5.2. Removal free ligand descriptors

In previous work, free ligand descriptors were used to represent the ligands, as complex descriptors could be influenced by the other atoms in the complex, such as the simulated NBD moiety which was not present in the other substrates. In this thesis, we focus solely on Rh-NBD complexes, both in simulations and experimentally. Therefore, the effect of substituting these free ligand descriptors for complex descriptors was evaluated. Eight electronic descriptors based on free ligand geometries were substituted for electronic descriptors calculated using the complex geometry. Additionally, this substitution led to the removal of three electronic descriptors that used both geometries.

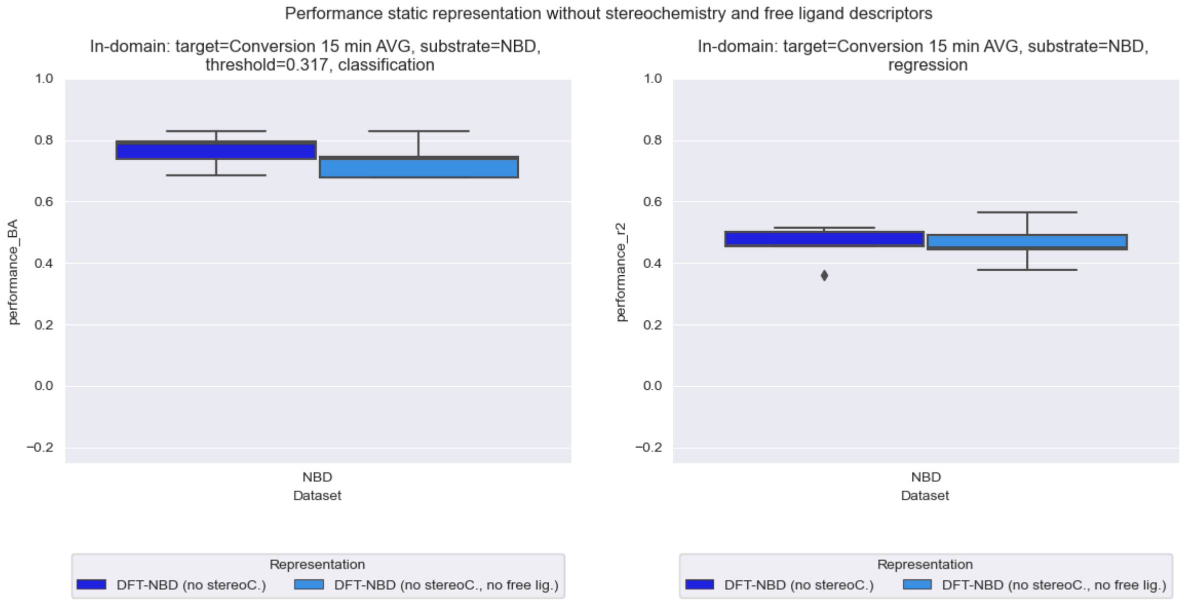


Figure 4.11: Effect of substituting free ligand descriptors for complex descriptors on the average 15 minute conversion NBD dataset. On the left the classification performances are shown for DFT-NBD without stereochemistry and DFT-NBD without stereochemistry and free ligand descriptors. Performances are similar.

Figure 4.11 shows the performances of the DFT-NBD representation with stereochemistry removed and the free ligand descriptors substituted for the complex descriptors for the average 15 minute dataset. The substitution resulted in small decrease in classification performance and a performance in the same range as the DFT-NBD representation with free ligand descriptors.

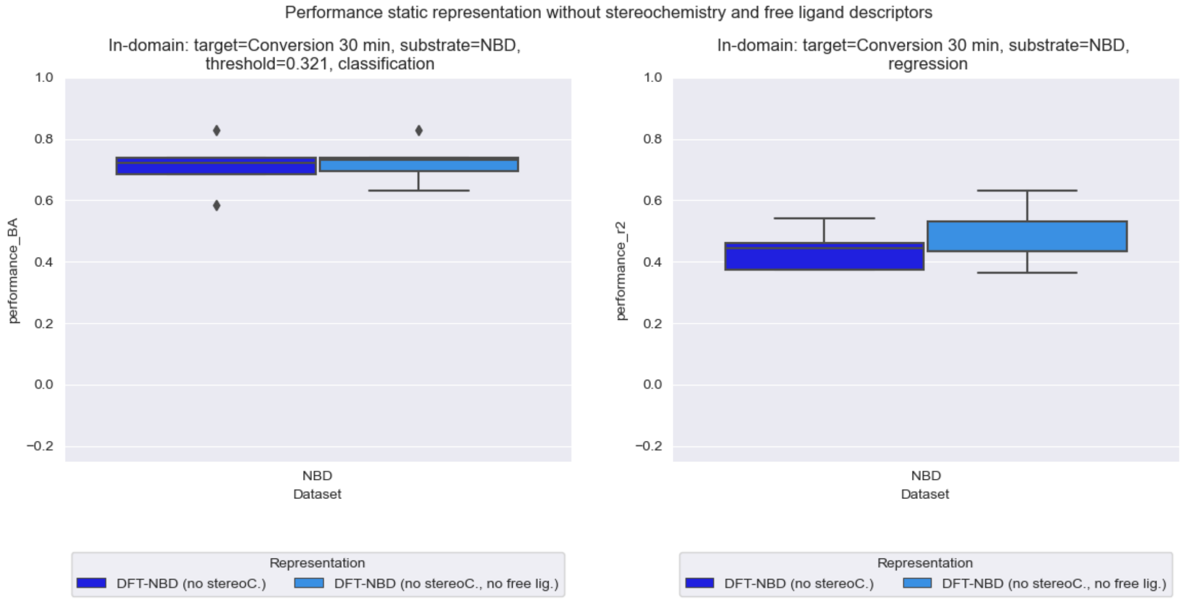


Figure 4.12: Effect of substituting free ligand descriptors for complex descriptors on the NBD 30 minute conversion NBD dataset. On the left the classification performances are shown for DFT-NBD without stereochemistry and DFT-NBD without stereochemistry and free ligand descriptors. Performances are similar.

Figure 4.12 shows the performance of the same representations as in Figure 4.11, but with the 30 minute NBD dataset. The classification performance is almost identical, with the slightest improvement in regression performance. This improvement is not significant and also was not present in the average 15 minute dataset results, leading to the conclusion that the performance remains within the same range.

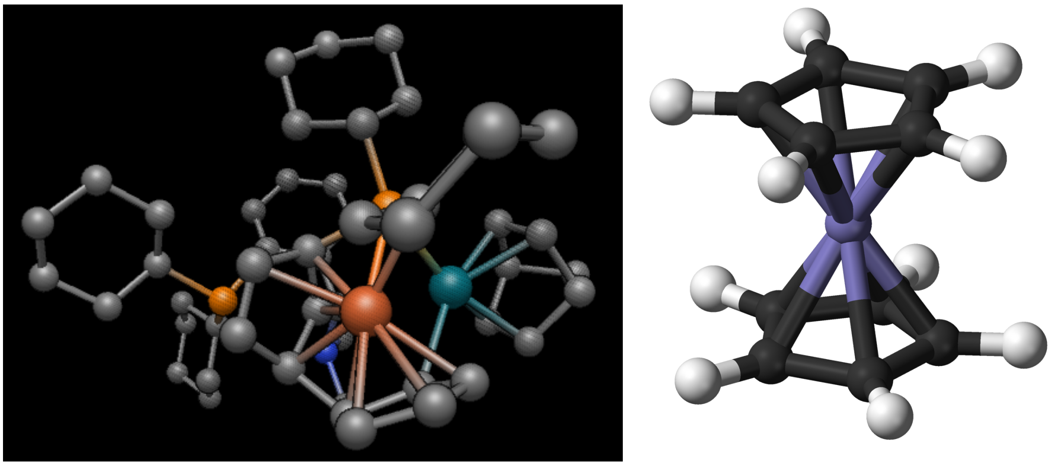
4.6. CREST-based representations (CREST-NBD & CREST-DFT-NBD)

Since the static representation shows similar performance between the DFT models and ECFP, a different approach is chosen to represent the complex structure. This approach includes searching the entire conformer ensemble. The representations are progressively increased in accuracy, ranging from cheaper to more expensive models.

Starting with CREST-NBD, the xTB geometries obtained from CREST were used to do DFT SP calculations. We obtained descriptors for each conformer in a CE. Three different representations were created for these ensembles: one where all descriptors were combined using the Boltzmann average to describe all conformers based on their energy, one where the lowest energy (most stable) conformer was selected, and one where the lowest energy conformer was chosen, followed by additional DFT (re)optimization to obtain free ligand descriptors.

Representations with the highest level of theory are also investigated. These are the CREST-DFT-NBD models, where xTB geometries from CREST are used to do DFT OPT calculations. From these DFT optimized structures two representations were created, one representing all conformers by using the Boltzmann average, and one using the lowest energy conformer.

Initially, CREST generated over 6000 structures, which were reduced to 2756 structures after applying Sára's filtering algorithm. [49] However, not all of these geometries were usable. 7 of the 192 conformer ensembles (CEs) contained structures with destroyed ferrocene groups, which formed bonds with other atoms. These affected CEs corresponding to ligand: 19, 24, 115, 140, 141, 176 and 188. These CEs were excluded, which reduced the structures further. Ultimately, this ferrocene issue impacted only two CE, as we only had experimental data for the first 96 ligands.



(a) DFT Optimized structure of L19 generated by CREST. Fe shown in dark orange, Rh (greenish blue), N (orange), C (gray) and O (blue). Hydrogen atoms are not displayed to keep overview.

(b) Ferrocene molecular structure, Fe center (blue) sandwiched by two cyclopentadienyl groups (C (black) and H (white)). [53]

Figure 4.13: Example of a destroyed ferrocene group by CREST. This ferrocene group in the front of Figure 4.13a makes bonds with other atoms outside of the cyclopentadienyl groups, such as phosphor and carbon.

4.6.1. Cheap vs Expensive E min selection

There are three representations based on the lowest energy conformer in the conformer ensembles (CEs). Two of these are based on the energy obtained from the xTB geometries, while the third is based on the energy obtained from the DFT optimized geometries. The first representation CREST-NBD, is the cheapest and only utilizes DFT SP calculations and complex descriptors. A more expensive representation is re-optimized CREST-NBD. This representation uses an additional DFT OPT calculation to be comparable with DFT OPT level representations (CREST-NBD [E min.] (no stereoC.)). The highest level of accuracy minimum energy representation is CREST-DFT-NBD. IN this representation every conformer in a CE is DFT Optimized and the most stable structure is found afterwards. These representation models are run and compared in Figures 4.14 and 4.15.

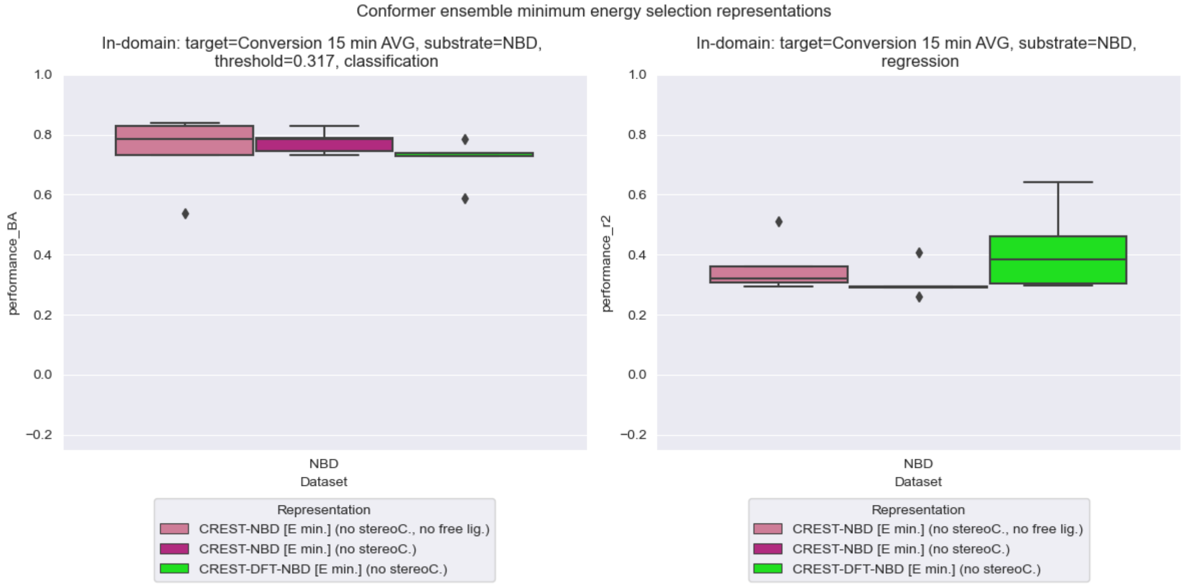


Figure 4.14: Comparison of performances of minimum energy representations from conformer ensembles on the average 15 minute dataset. The conformers in the left and middle representation are based on the energy obtained from xTB geometries. The representation on the right is based on the energy obtained from the DFT OPT geometry (cheapest representation on the left tot the most expensive on the right).

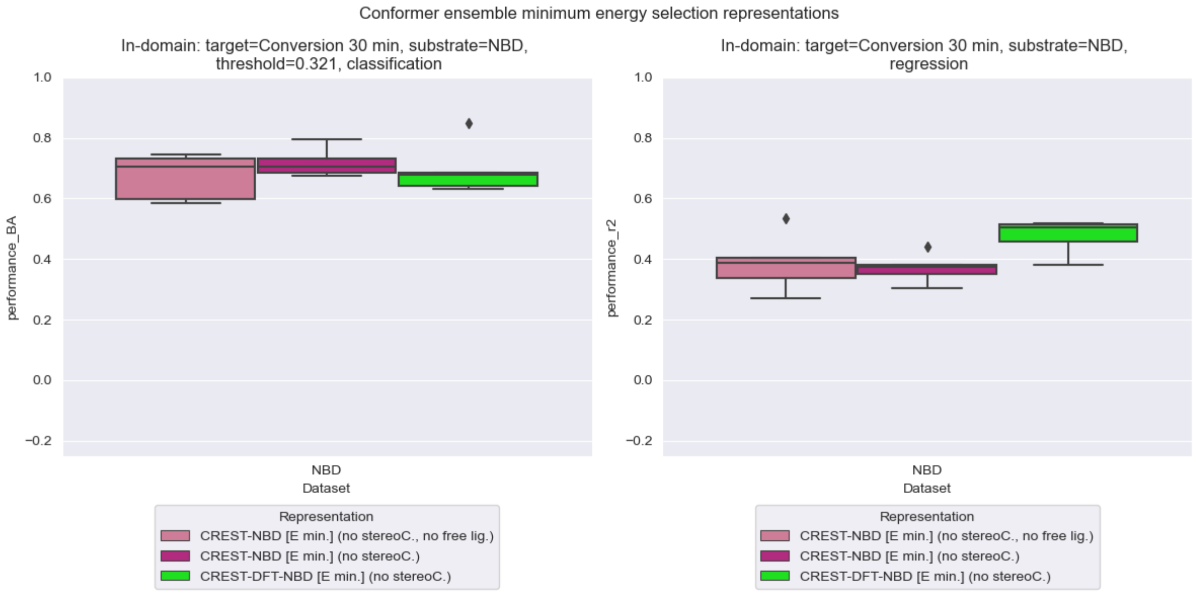


Figure 4.15: Comparison of performances of minimum energy representations from conformer ensembles on the 30 minute dataset. The conformers in the left and middle representation are based on the energy obtained from xTB geometries. The representation on the right is based on the energy obtained from the DFT OPT geometry (cheapest representation on the left to most expensive to the right).

These figures show the comparison between the different lowest energy representations for the average 15 minute and 30 minute datasets. It shows clear differences for classification and regression. The classification results are in the same range, however it can be seen that the cheapest representation (CREST-NBD [E min.] (no stereoC., no free lig.)) has the largest interquartile range (IQR), meaning it is the least consistent and has smaller boxes for the DFT Optimized structures. For regression, it can be seen that the most expensive CREST-DFT-NBD representation performs the best, however the middle representation is more consistent for both datasets, but with the lowest performance.

Energy analysis

Energies of these lowest energy conformers were analyzed, as CREST-NBD [E min.] (no stereoC.) and CREST-DFT-NBD [E min.] (no stereoC.) give different performances, suggesting that different minimum energy structures were used. These energies were compared to the energies of static representation (DFT-NBD) and shown in Figures 4.16 and 4.17.

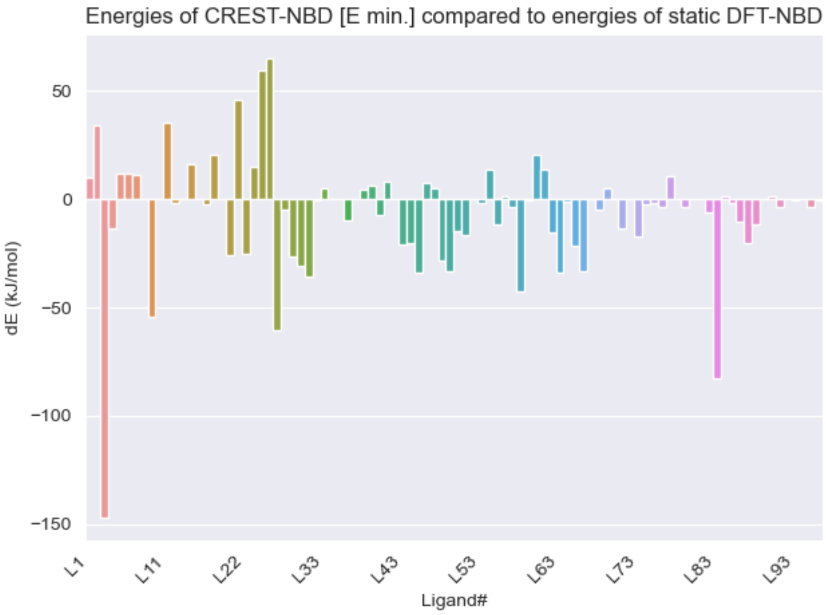


Figure 4.16: Energy analysis of CREST-NBD vs static DFT-NBD. A lower dE means a more stable structure has been found after CREST, minimum energy selection based on xTB energies and reoptimization using DFT OPT.

Figure 4.16 presents the energies of the lowest energy conformers in the CREST-NBD model compared to the static DFT-NBD model. It can be noted that some ligands have an increased energy, suggesting either that the lowest energy conformer based on the xTB energy does not corresponds to the true minimum energy conformer or that CREST is unable to find all possible conformers.

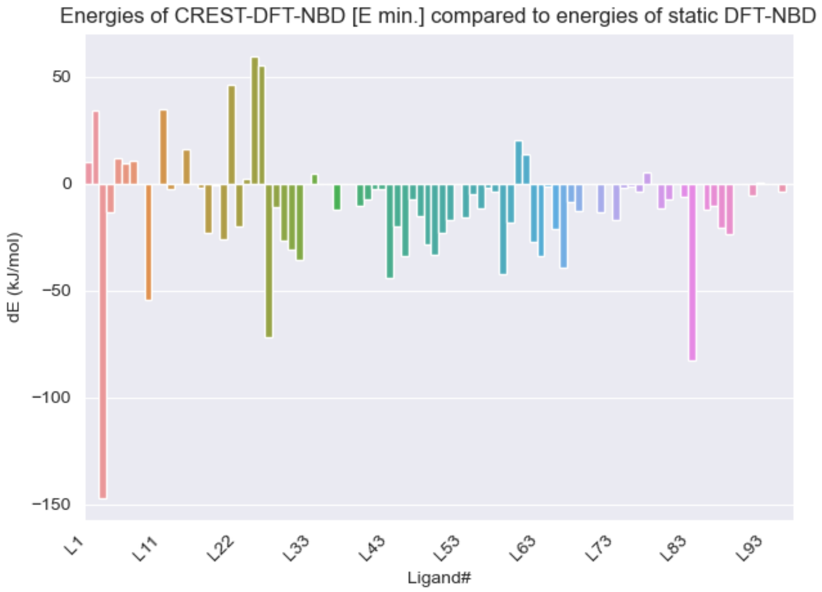


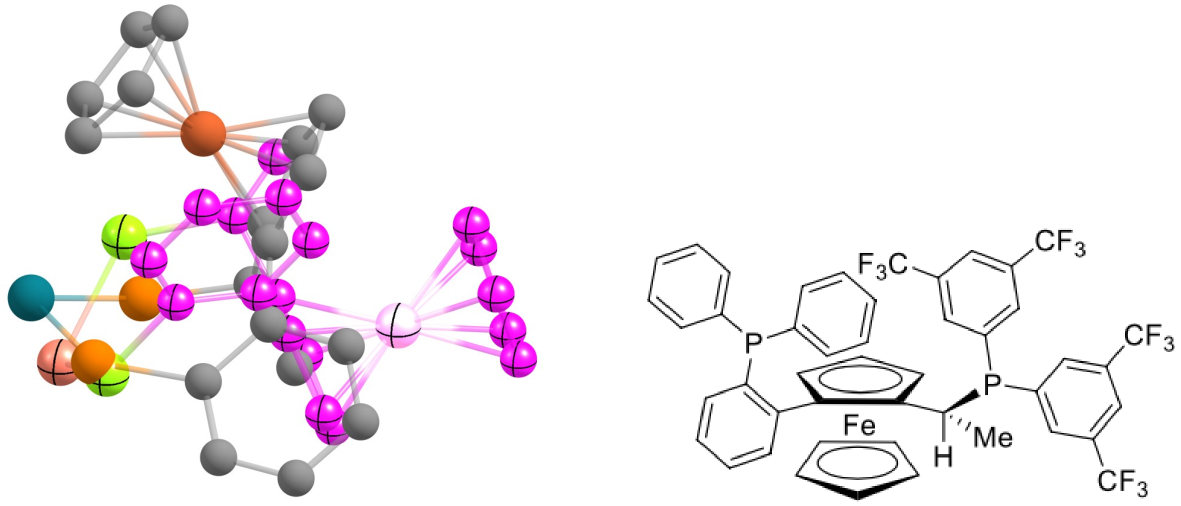
Figure 4.17: Energy analysis of CREST-DFT-NBD vs static DFT-NBD. A lower dE means a more stable structure has been found after CREST, DFT OPT and minimum energy selection based on DFT OPT energies.

Figure 4.17 presents the energies of the lowest energy conformers in the CREST-DFT-NBD representation compared to the static DFT-NBD representation. Several observations can be made. First, this figure still shows increased energies compared to the the DFT Optimized handmade structure (DFT-NBD), suggesting that CREST not only fails to find all conformers in a CE, but also is unable to identify the important, most favorable structures in some cases. Second, Figure 4.17 displays equal or lower energies than those in Figure 4.16, with for example L23, L26 or L77 showing a lower energy. This indicates that different conformers are used, which means that the lowest xTB energy conformers do not always corresponds to the lowest energy after DFT Optimization.

The CEs with higher energy compared to the initial DFT-NBD representation are ligands [**1, 2, 5, 6, 7, 10, 11, 13, 14, 17, 21, 23, 25, 26, 33, 34, 35, 37, 38, 39, 41, 46, 47, 52, 54, 56, 60, 61, 69, 77, 80, 84**], which accounts for 34% of the CEs. The ligands that cross 10 kJ/mol are considered extreme cases and are further analyzed. These extremes are **[2, 5, 6, 7, 11, 14, 17, 21, 25, 26, 54, 60, 61, 77].**

4.6.2. Analysis of increased energy structures after CREST

The .xyz structures with higher energy after CREST-DFT-NBD were analyzed using ChemCraft and Avogadro2. Output structures files from DFT Optimization revealed several causes for the observed energy differences. While all cases retained the correct bonds and atoms, CREST failed in identifying the optimal or logical orientations. The largest energy increases were primarily observed in ligands containing ferrocene groups, while in the less extreme cases CREST did not find the most optimal conformer, such as perfect symmetry, or avoiding steric hindrance. Small energy increases were caused by small variations in angles but nothing structurally significant.

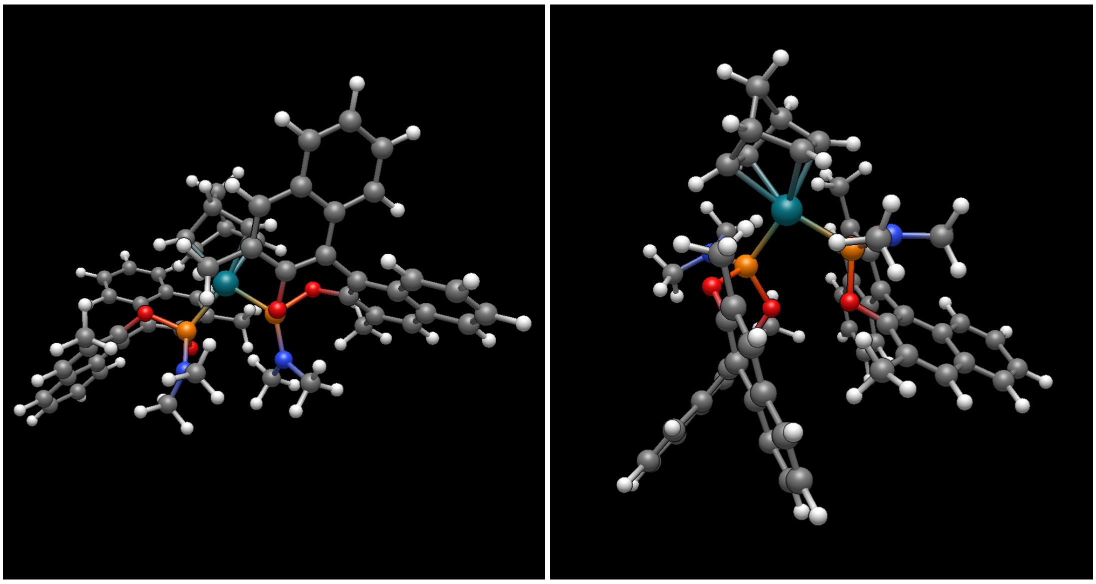


(a) Overlay of L25 DFT-NBD (crossed atoms) and CREST-DFT-NBD (unmarked atoms), illustrating Fe (dark orange, light blue), Rh (greenish blue, shiny orange), N (orange, yellow) and C (gray, pink), respectively. To remain clarity, only the backbone is shown. Hydrogen, NBD and the end groups attached to phosphor are ommited.

(b) Complete 2D structure of the Rh-NBD complex L25. [**?**]

Figure 4.18: Example of a structure where the ferrocene group has an extremely noticeable different orientation, compared to the handmade structure. In the CREST structure this ferrocene group is directed towards the metal center, resulting in a more compact complex, whereas in the handmade structure this group is oriented outwards. The inward orientation in the CREST structure causes steric hindrance, resulting in the higher observed energy.

Figure 4.18a shows an example where the ferrocene group causes an energy increase in the CREST generated structure. The bulk of the ferrocene is oriented towards the metal center, whereas in the handmade DFT Optimized structure this ferrocene is oriented outwards from the metal center. This steric presence of this bulky group leads to unfavorable interactions between electrons, resulting in higher energies. CREST seems to struggle a lot with these ferrocene groups, as multiple cases are found where CREST either destroys or fails to identify correct conformers. Other examples where this is the case are L11, L19, L24 or L26. However, CREST also led to cases where the structure was improved, such as L3, where a decrease of 150 kJ/mol is observed. In this particular case, no weird distortions or illogical positioning was observed, CREST simply found a much better conformer.

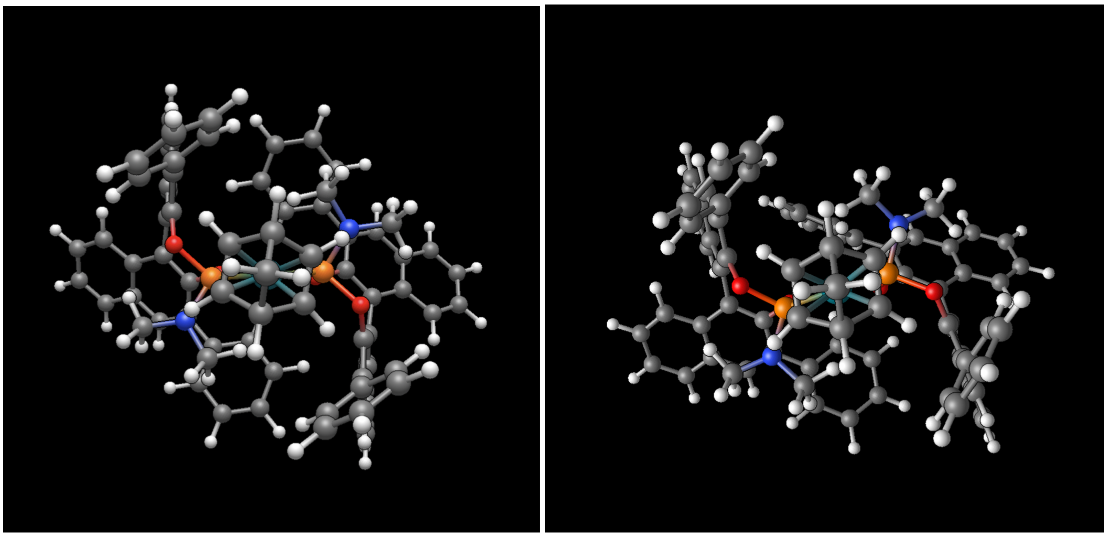


(a) DFT Optimized handmade structure (DFT-NBD) of L60, with the bulky subgroups attached to nitrogen oriented to the sides and smaller subgroups attached to oxygen oriented down.

(b) DFT Optimized structure generated by CREST (CREST-DFT-NBD) of L60, with the bulky subgroups attached to nitrogen oriented down and smaller subgroups attached to oxygen oriented to the sides.

Figure 4.19: Example of a structure where an increase in energy is caused by CREST's inability to find all conformers. A clear difference is observed between the handmade DFT optimized structure and CREST generated DFT optimized structure. The atoms are displayed in colors: Rh (greenish blue), P (orange), N (dark orange), O (blue), C (gray), H (white).

Figure 4.19 shows the structure for L60 in the DFT Optimized handmade structure and CREST generated structure. There is a clear difference in geometries, where the handmade structure orients the bulkier subgroups to the sides, while in the CREST structure these groups oriented downward. CREST was unable to find this orientation, which would have led to a difference of 20.4 kJ/mol.



(a) Top view of the DFT Optimized handmade structure (DFT-NBD) of L61. It can be noted that this structure is complete symmetrical.

(b) Top view of the DFT Optimized structure generated by CREST (CREST-DFT-NBD) of L61. It can be noted that this structure is not symmetrical, as the cyclohexane group at the bottom middle does not have a mirrored orientation in the cyclohexane group at the top middle.

Figure 4.20: Example of a structure where a small change in orientation causes an increase in energy. When viewed straight through Rh-NBD, it is clear that the handmade structure is much more symmetric compared to the CREST generated structure. CREST is unable find the correct rotation for one of the benzene rings, preventing the complex to be completely symmetrical. The atoms are color-coded as follows: Rh(greenish blue), P (orange), N (dark orange), O (blue), C (gray), H (white).

Figure 4.20 shows L61, where the two structures are very similar, but a difference in the orientation of one benzene ring has caused the large energy increase. In the handmade DFT optimized structure these groups are arranged in perfect symmetry, while this not the case in the CREST structure.

4.6.3. Cheap vs Expensive Boltzmann-averaging

The Boltzmann-averaged representations are also compared, these differ in DFT method used. The cheap variant, CREST-NBD, uses DFT SP calculations and complex descriptors, as the code used was unable to extract free ligand .xyz geometries from DFT SP calculations. The expensive CREST-DFT-NBD variant, uses DFT OPT calculations for each conformer and uses the free ligand descriptors.

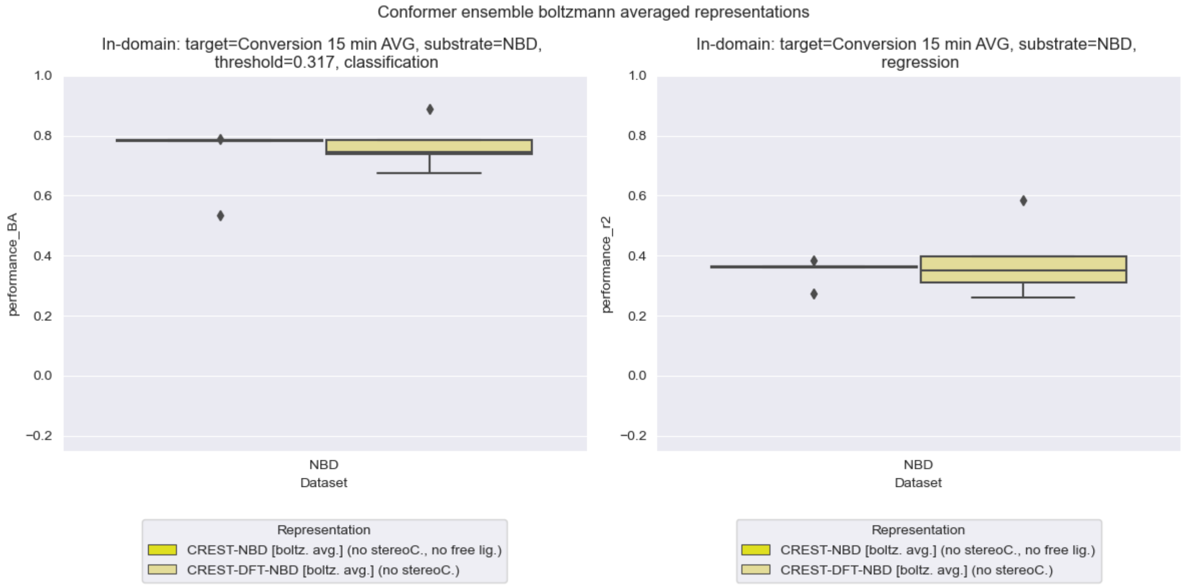


Figure 4.21: Comparison of performances of Boltzmann averaged representations from conformer ensembles on the average 15 minute dataset. The cheap representation (left) show very consistent performances, while the expensive representation (right) show a broader IQR. However, the absolute performance does not change.

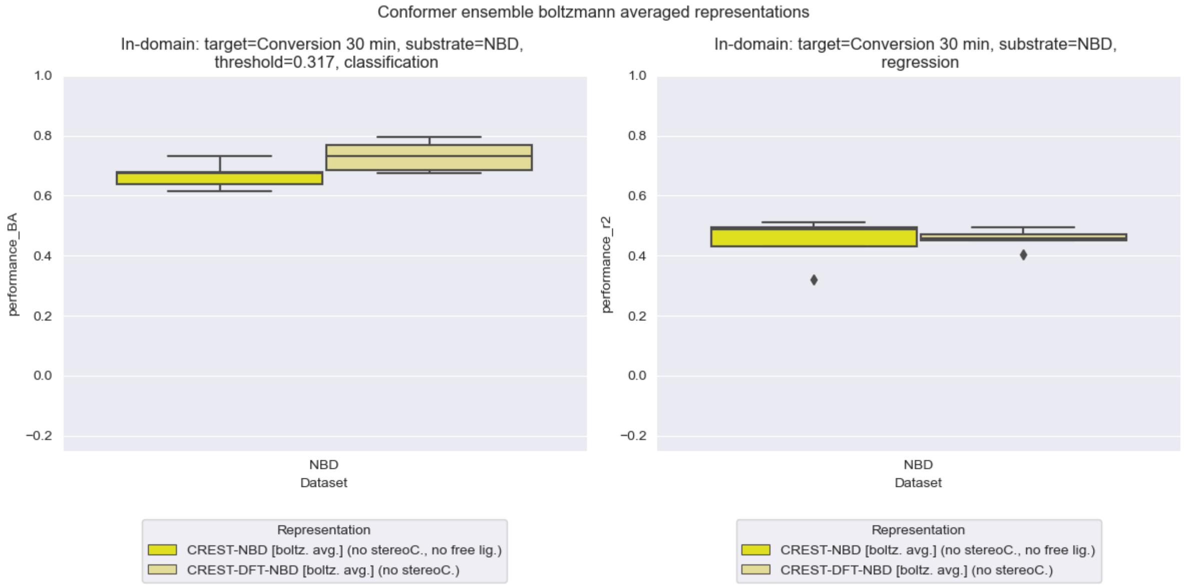


Figure 4.22: Comparison of performances of Boltzmann averaged representations from conformer ensembles on the 30 minute dataset. The cheap (left) and expensive (right) representation show similar performances and IQR's.

Figure 4.21 shows the Boltzmann averaged representations without stereochemistry for the average 15 minute dataset. The prediction results are very similar, with small but almost not noticeable improvements for the expensive representation. The expensive representation loses on consistency compared to the cheap variant for this specific case. However, this effect is not seen in Figure 4.22, where the same representations are compared but with a the 30 minute dataset. As the performance is similar, the only conclusion that can be made is CREST-DFT-NBD is not performing better, despite the expensive DFT OPT calculations.

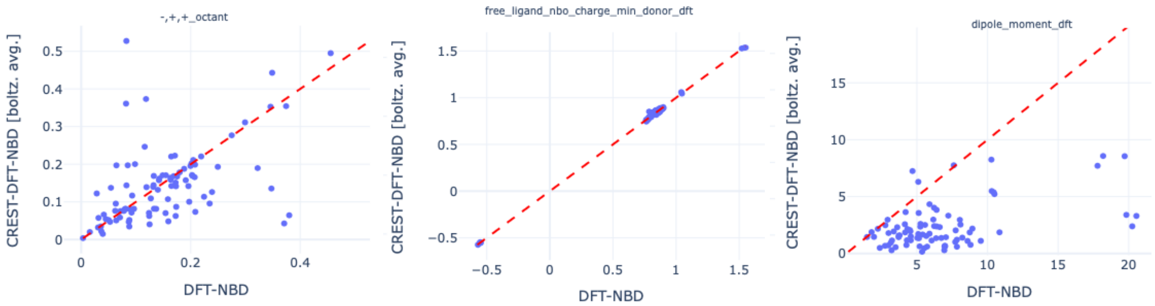
4.6.4. Analysis descriptors

Boltzmann averaged descriptors are compared to the static DFT-NBD descriptors, to investigate the effect of CREST. Scatter plots are made for each descriptor with on the x-axis the values of DFT-NBD and on the y-axis the values of CREST-DFT-NBD [boltz. avg.]. The diagonal x=y line is the identity line, which implies that the values equals the same. The interactive descriptor figures can be found in the supplementary files.

Analyzing these figures has shown that steric and geometric descriptors change a lot after conformer searching. These descriptors changed as they are very structure dependent. Considering that the structures in both representations are not identical, it is expected that these descriptors will differ. In the electronic descriptors only lone pair and dipole moment descriptors showed similar variation. The variation is evenly spread between above and below the identity line, showing no bias in all cases except ’dipole_moment_dft’. This figure shows an overestimation of DFT-NBD or underestimation of CREST-DFT-NBD, where all points in CREST-DFT-NBD show (much) lower values.

As the lowest energy after conformer searching is not always lower than the static DFT-NBD, it is hard to say if the CREST descriptors are as good as the DFT-NBD descriptors for the structures. Further analysis of the outliers in these figures, confirmed that they were not necessarily from CEs where CREST was unable to find the lowest energy conformer. This result led us believe that these variations in descriptors could be caused by the conformational flexibility that CREST has included.

Descriptors that pretty much stayed using these static and dynamic method were: ’sum_electronic_and_free_energy’, ’nbo_charge_min_donor’, ’nbo_charge_max_donor’, ’free_ligand_nbo_charge_min_ donor’, ’free_ligand_nbo_charge_max_ donor’, ’free_ligand_dispersion_energy’. They showed that changing the geometry did not alter the results.



(a) Scatter plot of descriptor ’-,+,+_octant’, a broad but evenly distributed variation is observed over the two representations.

(b) Scatter plot of descriptor ’free_ligand_nbo_charge_min_donor_dft’, almost identical values are found using a different approach.

(c) Scatter plot of descriptor ’dipole_moment_dft’, figure shows a overestimation in the CREST-DFT-NBD representation.

Figure 4.23: Example of a few descriptor scatter plots, where the values of the same descriptor over a static and dynamic representation is compared (DFT-NBD vs CREST-DFT-NBD). The red line shows the identity line, where the values.

4.6.5. Boltzmann-averaging vs E min selection

The last comparison that can be made is the difference in performance for the Boltzmann averaged and minimum energy structures. While the lowest energy conformer is often considered a sufficiently accurate representation, recent studies have suggested taking into account CE can be beneficial. [54] To compare these two, might show us if the expensive Boltzmann averaged representation is much better than the minimum energy structure representation.

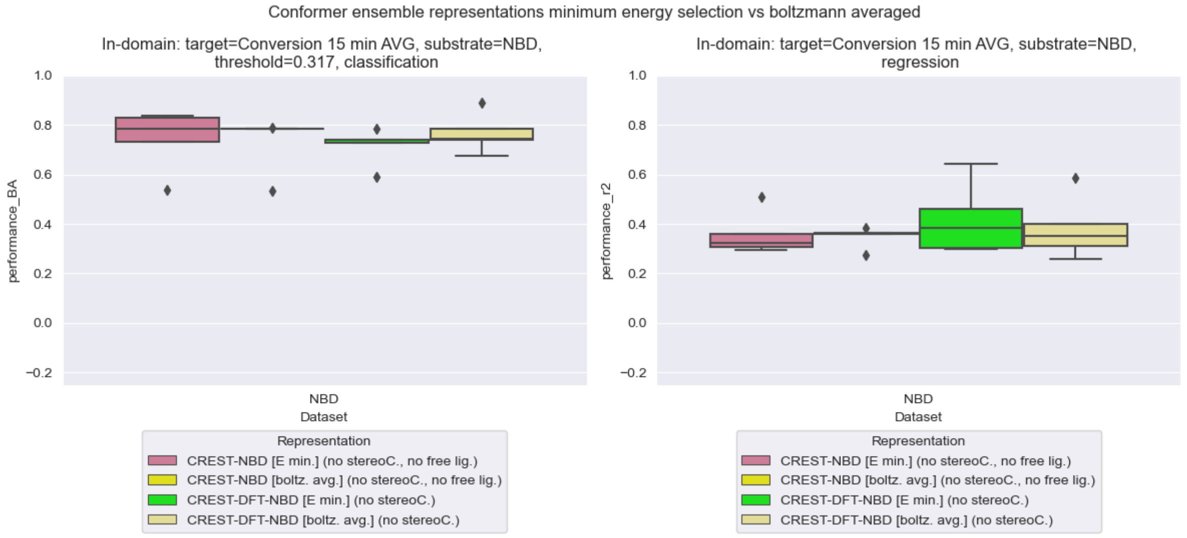


Figure 4.24: Box plot comparison of performances of all conformer ensemble representations for average 15 minute dataset. Classification (left) and regression (right).

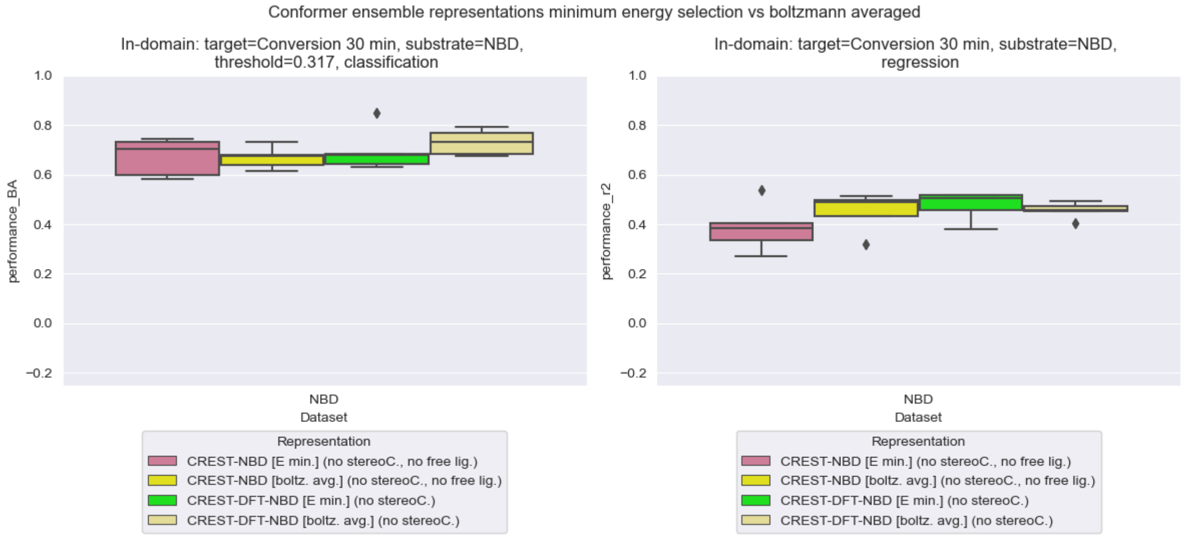
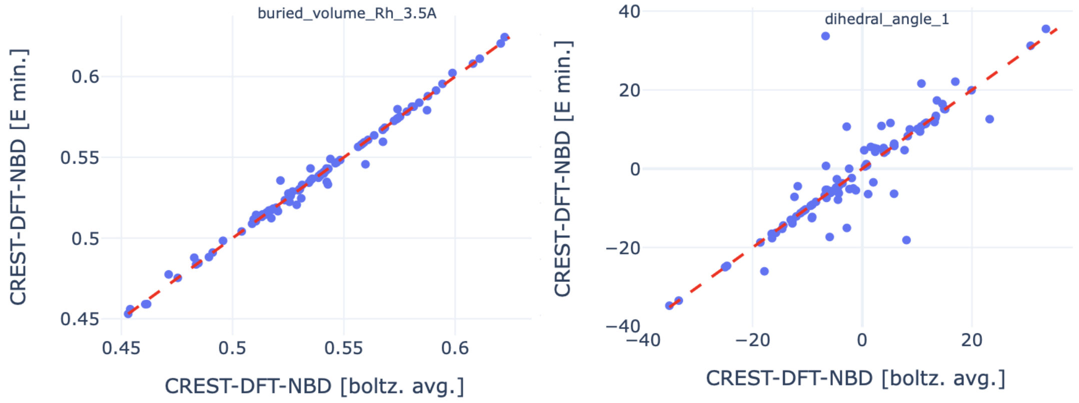


Figure 4.25: Box plot comparison of performances of all conformer ensemble representations for 30 minute dataset. Classification (left) and regression (right).

Figure 4.24 shows the cheap lowest energy representation compared to the cheap Boltzmann averaged representation and Figure 4.25 shows the expensive lowest energy representation and the expensive Boltzmann averaged representation. Both of these figures show visible differences, but none of them shows clearly which is better. Results in classification show high performances, with varying range. Regression show more consistency for Boltzmann averaged representations, but performances are still in the same range, suggesting that the minimum energy structure is a good representation for the entire CE in this case.

4.6.6. Analysis descriptors

The descriptors of these representations are compared to further analyze the differences between the minimum energy conformer and the entire CE, and almost no difference is found between these descriptors. Most descriptors show plots of the values being only on or close to the identity line. Some figures related to steric descriptors and or angles show some noticeable outliers, but these are the descriptors which are very sensitive to changes within the geometry of the structure.



(a) Scatter plot of descriptor 'buried_volume_Rh_3.5A'. Most of the plots are similar almost all points are on the identity line.

(b) Scatter plot of descriptor 'dihedral_angle_1', this figure is the most scattered plot.

Figure 4.26: Two example of scatter plots, where the values of the same descriptor over minimum energy structure vs the entire CE is compared (CREST-DFT-NBD [boltz. avg.] vs CREST-DFT-NBD [E min.]). The red line shows the identity line, where the values.

Additionally, the highest Boltzmann weights for each CE is displayed in Figure 4.27. It can be seen that in more than 70% of the CEs, the highest weight is below 50%, suggesting that even in cases where the CE is flexible, the consideration of multiple structures does not lead to large differences in the descriptor. Based on this observation, for this specific case, we found that the minimum energy conformer representation is almost the same as the Boltzmann averaged representation.

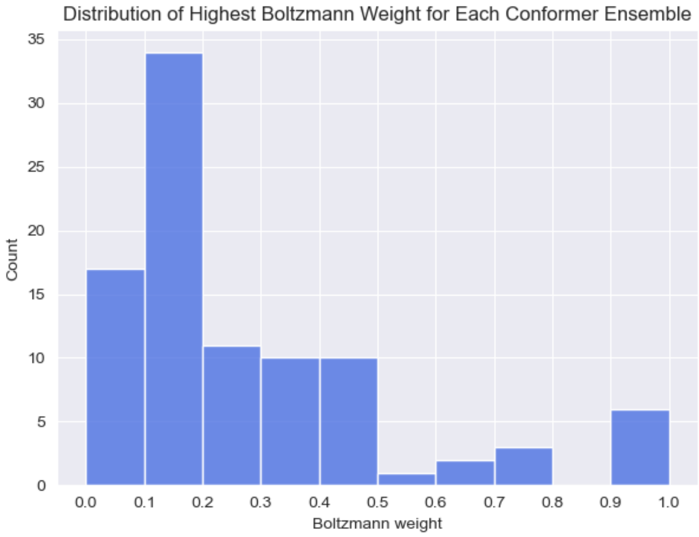


Figure 4.27: Distribution of highest Boltzmann weights in a conformer ensemble.

4.7. Comparison all models

Finally, all no stereochemistry models are compared and the results are presented in Figures 4.28 and 4.29.

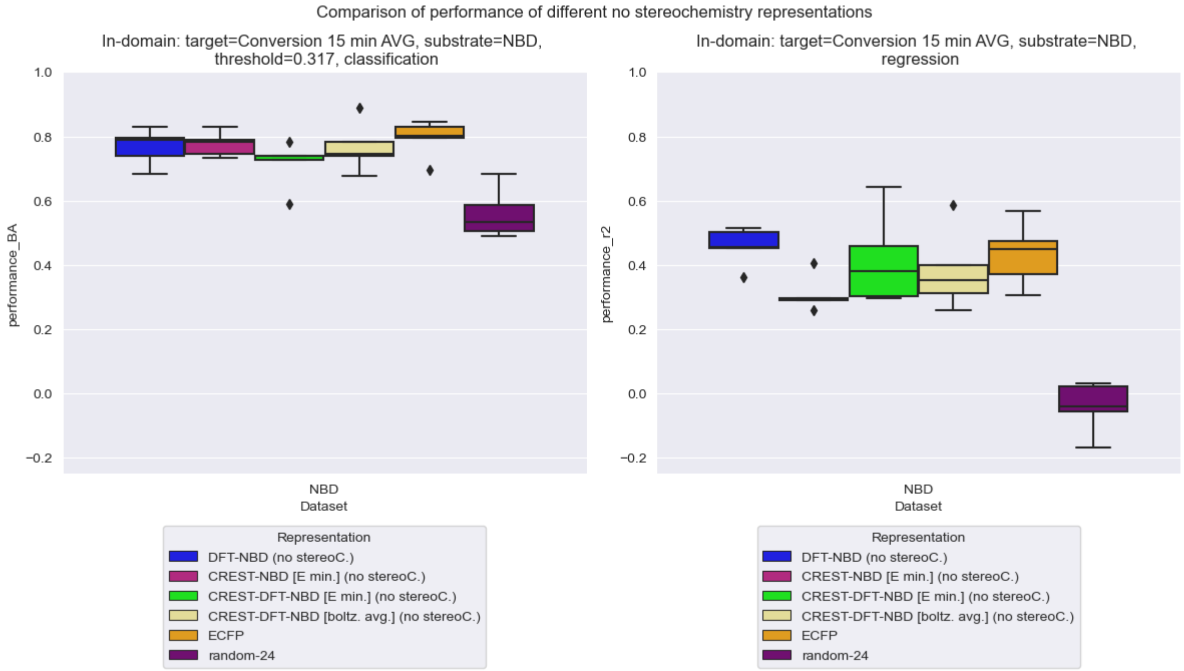


Figure 4.28: Overview all no stereochemistry representations average 15 minute dataset.

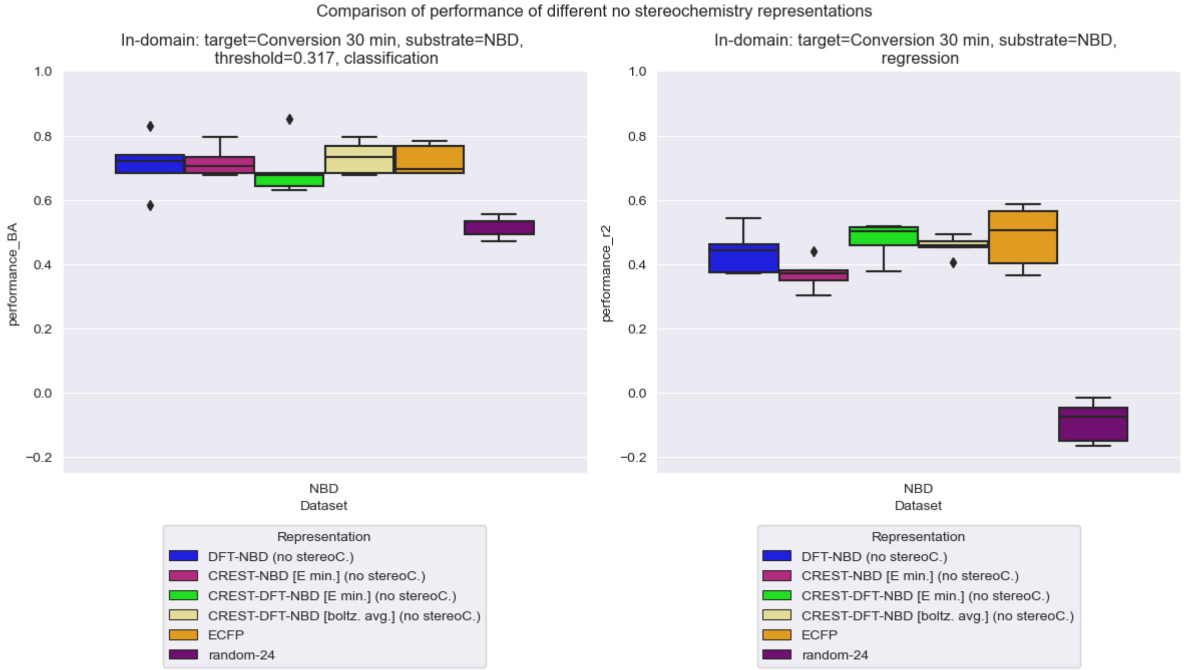


Figure 4.29: Overview all no stereochemistry representations 30 minute dataset.

It can be seen that all models are very comparable for classification and regression, with them being in the same range. However, the size of the range differs a lot between models with no clear trend. Furthermore, ECFP scores similar or even better compared to the DFT models. This suggests that the ECFP method, which is based on topology, is as good, if not better, compared to the expensive DFT calculated descriptors. As a result ECFP will be further analyzed.

4.8. ECFP analysis

4.8.1. Feature Importance

The FI plots were analyzed, revealing high importance of fingerprints parameters fp57, fp333, fp467 and fp91. However, these parameters cannot be directly traced to specific groups or atoms in the complex, as the process is done iteratively starting from the center. As a result, these parameters can point to different subgroups of atoms in different complexes, but also multiple different subgroups can point the the same bit (bit collision). This makes further analysis challenging. If these bits can be connected to the original (sub)groups in the future it could provide new insights into what groups and combinations makes ECFP work this well.

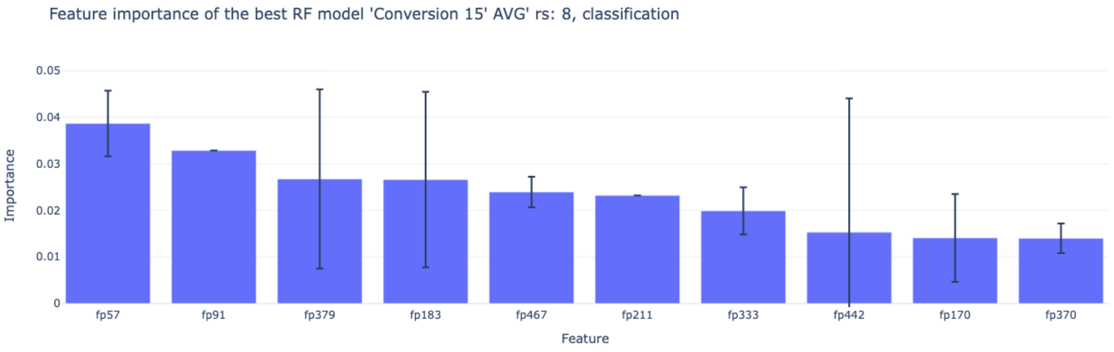


Figure 4.30: Example of a Feature Importance figure of ECFP, where the highest descriptors are shown. This figure is zoomed in and contains only the highest 10 features.

4.8.2. PCA

There is no method to compare the regression results of the DFT models and ECFP, but for classification, accuracy results can be compared. PCA is used to reduce the dimensionality of the descriptors for each ligand and to represent them in a 2D chemical space. DFT-NBD is compared to ECFP, as this has the same level of theory in the original substrate model and as the results of the other DFT models are not significantly different. FN, FP, TN, TP values were compared over the five random states analyzed. We looked at consistency for reoccurring ligands in the test set, differences in performance for the same ligands between DFT-NBD model and ECFP.

Figures 4.31 and 4.32, show that ECFP and DFT-NBD have different data points where each performs better than the other. Figure 4.31, for example, clearly shows that DFT-NBD has both better and worse predictions compared to ECFP, such as L73 (-7.6 -2.9) performing worse, or L27 (1.4, -0.5) better. Another example is Figure 4.32, L13 (2.5, -0.6) and L84 (-5.1, 1.6). This suggests that DFT-NBD and ECFP both use different correlations, which leads them to different predictions, with neither necessary better than the other.

Additionally, the reoccurring data points in the test split across different random states were analyzed to assess if models predict consistently for these data points, and results show that both models do not always produce consistent predictions. Among the 96 CEs, 24 were included in the test set more than once, of these, only 6 showed different prediction for at least one of the models. This highlights that the importance of the training data again.

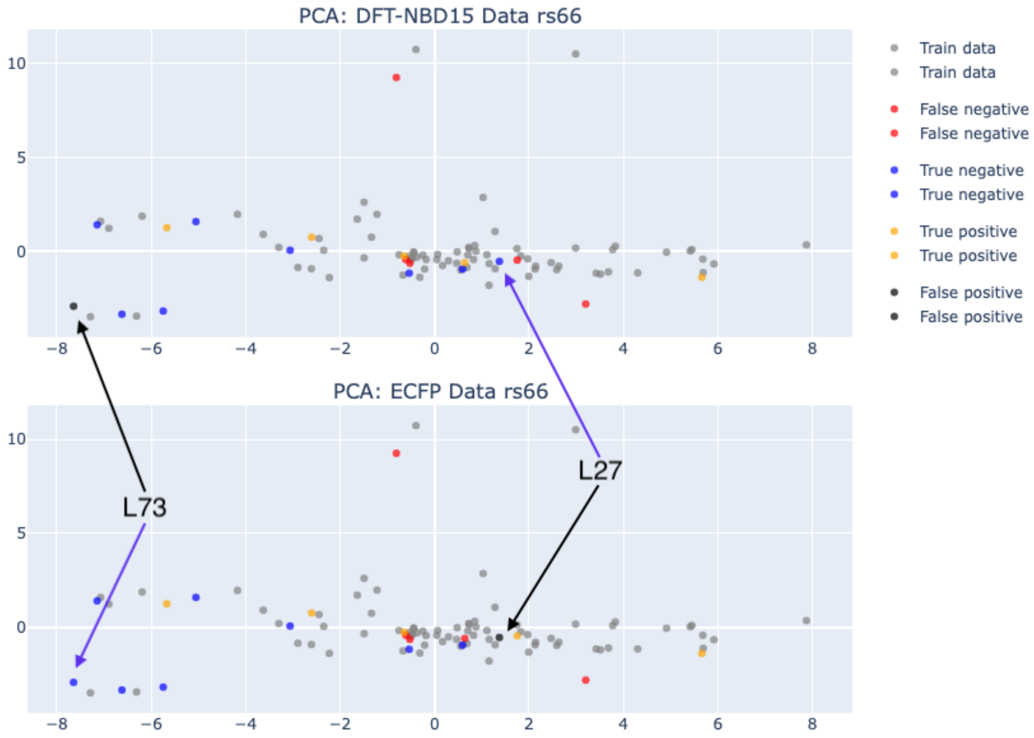


Figure 4.31: PCA of DFT-NBD (top) vs ECFP (bottom) for average 15 minute conversion dataset (random state 66). Both models have points where each outperforms the other.

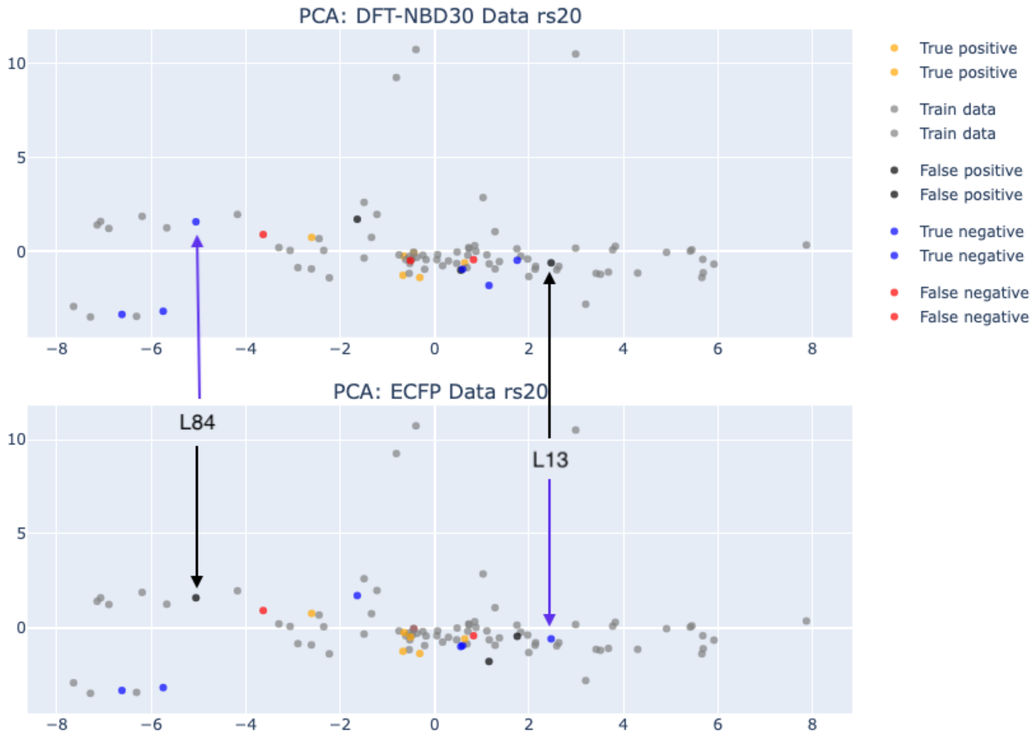


Figure 4.32: PCA of DFT-NBD (top) vs ECFP (bottom) for 30 minute conversion dataset (random state 66). Both models have points where each outperforms the other.

5

Conclusion & Outlook

5.1. Conclusion

The research question focused on investigating the effect of different types of representation on the performance of predictive Machine Learning (ML) models. The objective was to improve predictive models used for catalyst design of bidentate ligand coordinated transition metal (TM) complexes. These different representations were all DFT-based, with varying input structures and DFT calculations methods. Additionally, new experimental data created by High-Throughput Experimentation is used, which is closer to the simulated data.

The new reaction investigated was a hydrogenation reaction on a symmetrical substrate norbornadiene (NBD). Previous work investigated asymmetric hydrogenation reactions on different substrates. [16] The original representation (DFT-NBD), which is a handmade DFT Optimized structure, was combined with this new experimental data. This combination resulted in improved consistency of the models, but it did not necessarily improve their absolute performance. In comparison the models used in previous work [16], the current model clearly demonstrated better performance in DFT-based and ECFP representations compared to the random representations for regression. This suggests that the ML algorithms have captured useful patterns, but the performance $r^2 = 0.5$ is not sufficiently high. Additionally, this DFT-NBD representation, used a set of descriptors that included stereochemistry related descriptors. However, these descriptors were removed, as our new data used a symmetrical substrate. It was found that this removal had only negligible effect on the performance, which indicates that these descriptors do not influence the performance.

As mentioned, different representations were tested, adjustments were made to the input structure to more closely represent reality. This was achieved by using conformer ensembles (CEs) generated by CREST. These CEs account for the dynamic aspect of the complex by using all structures in the conformer ensemble (CE) and weighing them using Boltzmann averaging, or by selecting the most stable structure to represent the CE. However, CREST was not always able to find a more or equally stable conformer as DFT-NBD. This resulted in an energy increase in the lowest energy structure for approximately 25% of the CEs, leading to worse of equal performance compared to the static method. These higher energies were primarily caused by structures containing a ferrocene group, which CREST struggled with. Another hypothesis for the lack of improvement in these models could be the limitations of HTE . HTE is known to rapidly generate data, but also introduce inaccuracies. These are caused by using a general application for all samples, which can overlook the differences in catalysts, such as their activation times, sample variability and more. Depending on reaction, activity, and compound, these HTE can have very large deviations. [55] Our data, for instance, showed standard deviation values ranging from 0.9 to 39.5%, with an average of 11.8%. If experimental data contains significant errors, even perfect descriptors will fail to make accurate predictions.

Descriptors in both DFT-NBD and CREST representation were compared as part of the CREST analysis. The results showed that descriptors between DFT-NBD and CREST differed noticeably, specifically in steric descriptors. These descriptors are the most sensitive to changes in the input structure, as they are very geometry related. It was also noted that electronic descriptors changed the least. Between the Boltzmann averaged and minimum energy representations from CREST, it was found that these descriptors were almost identical, with the largest spread in again steric descriptors. Furthermore, we checked whether the similarity was due to minimum energy structures having dominant Boltzmann weights, but in over 70% of the CEs, this was not the case. Overall, the lowest energy conformer representations performed as good as the Boltzmann averaged representations.

Regardless of which DFT-based representation was used, the performance of all representations were similar or worse than ECFP, with for all models a good classification performance and a mediocre 0.3 to 0.7 R^2, with varying interquartile ranges (IQR). These large IQRs emphasize the importance of data splitting, as this has caused the large differences in performance.

Furthermore, none of the expensive DFT-based models has surpassed ECFP. Further analysis was done, but retrieving and learning from the patterns found is hard. This method works indeed the best and a possible explanation could be the consistency of how we described our data, coupled with the systematic method ECFP works, it can very nicely and effectively split topological features of these complexes. [56][57] These features might strongly correlate with the physical phenomena that drive the catalytic activity.

5.2. Outlook

The highest level of theory has been reached in this thesis. Boltzmann averaging, combined with DFT Optimization, was used to obtain the most realistic representation of the complexes. However some methods could still be improved. Currently, CREST generated conformer ensembles with higher energies compared to manually generating a structure and optimizing it using DFT. Most of these increased energy structures were associated to ferrocene groups. If CREST is able to reliably generate correct conformers containing these ferrocene groups, it might be possible to actually obtain more accurate descriptors using geometries from CREST.

Improving reliability of High-Throughput Experimentation or finding alternative methods could improve the overall performance of the prediction models, as inconsistencies in experimental data may be the limiting factor in achieving a higher absolute performance.

Expanding the dataset could also significantly improve model performance, as results have shown strong dependency on the data split. Increasing the number data points would reduce the impact of the data split and make the models more consistent. Alternatively running more random states would give a more accurate box plots for the model performance.

In the future, different reactions can be analyzed, or different conformer ensemble generators can be used, to gain valuable insights.. These might identify where exactly the weaknesses are located and what can be changed to improve the current model.

Acknowledgments

Over the past six months, I have learned many new skills and gained experience in the computational aspect of chemistry, for which I am truly grateful.

First, I would to express my gratitude to my supervisor Professor Dr. Evgeny Pidko, for giving me this opportunity to do my thesis in the ISE group. I really appreciate the honest feedback and new ideas he shared during our progress meetings.

Above all, I would like to thank my daily supervisor, Adarsh Kalikadien, for his continuous support throughout this entire project. He was always available on Teams whenever I had questions, and his guidance in exploring new ideas, discussing results, explaining concepts, and improving visualizations in my thesis have been incredibly helpful.

I would also like to thank the other members of the ISE group, for giving their insights on topics related to my thesis. Additionally, I am grateful to TU Delft for providing me with two screens to work with and free coffee that kept me going.

Finally, I would like to thank my friends and family for providing moments of distraction and supporting me throughout this journey.

- *Tai Hong Chow*
Delft, February 2025

Bibliography

[1] Artificial Intelligence - Global | Statista Market Forecast, March 2024. URL `https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide#market-size`. [Online; accessed 19. Nov. 2024].

[2] Johannes G. de Vries and André H. M. de Vries. The Power of High-Throughput Experimentation in Homogeneous Catalysis Research for Fine Chemicals. *Eur. J. Org. Chem.*, 2003(5):799–811, March 2003. ISSN 1434-193X. doi: 10.1002/ejoc.200390122.

[3] Ligang Wang, Jiabin Wu, Shunwu Wang, Huan Liu, Yao Wang, and Dingsheng Wang. The reformation of catalyst: From a trial-and-error synthesis to rational design. *Nano Res.*, 17(4):3261–3301, April 2024. ISSN 1998-0000. doi: 10.1007/s12274-023-6037-8.

[4] Travis Williams, Katherine McCullough, and Jochen A. Lauterbach. Enabling Catalyst Discovery through Machine Learning and High-Throughput Experimentation. *Chem. Mater.*, 32(1):157–165, January 2020. ISSN 0897-4756. doi: 10.1021/acs.chemmater.9b03043.

[5] Gert-Jan M. Gruter, Alasdair Graham, Ben McKay, and François Gilardoni. R&D Intensification in Polymer Catalyst and Product Development by Using High-Throughput Experimentation and Simulation. *Macromol. Rapid Commun.*, 24(1):73–80, January 2003. ISSN 1022-1336. doi: 10.1002/marc.200390015.

[6] Bryan R. Goldsmith, Jacques Esterhuizen, Jin-Xun Liu, Christopher J. Bartel, and Christopher Sutton. Machine learning for heterogeneous catalyst design and discovery. *AIChE J.*, 64(7):2311–2323, July 2018. ISSN 0001-1541. doi: 10.1002/aic.16198.

[7] 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, and Alán Aspuru-Guzik. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.*, 144(3):1205–1217, January 2022. ISSN 0002-7863. doi: 10.1021/jacs.1c09718.

[8] Emil Roduner. Understanding catalysis. *Chem. Soc. Rev.*, 43(24):8226–8239, 2014. doi: 10.1039/C4CS00210E.

[9] Elizabeth L. Bell, William Finnigan, Scott P. France, Anthony P. Green, Martin A. Hayes, Lorna J. Hepworth, Sarah L. Lovelock, Haruka Niikura, Sílvia Osuna, Elvira Romero, Katherine S. Ryan, Nicholas J. Turner, and Sabine L. Flitsch. Biocatalysis. *Nat. Rev. Methods Primers*, 1(46):1–21, June 2021. ISSN 2662-8449. doi: 10.1038/s43586-021-00044-z.

[10] Piet W. N. M. Leeuwen. *Homogeneous Catalysis.* Springer Netherlands, 2006. ISBN 978-1-4020-2000-1.

[11] Serenella Medici, Massimiliano Peana, Alessio Pelucelli, and Maria Antonietta Zoroddu. Rh(I) Complexes in Catalysis: A Five-Year Trend. *Molecules*, 26(9):2553, April 2021. ISSN 1420-3049. doi: 10.3390/molecules26092553.

[12] Patrick J. Guiry and Cormac P. Saunders. The Development of Bidentate P,N Ligands for Asymmetric Catalysis. *Adv. Synth. Catal.*, 346(5):497–537, April 2004. ISSN 1615-4150. doi: 10.1002/adsc.200303138.

[13] J Caldwell. The Importance of Stereochemistry in Drug Action and Disposition. *J. Clin. Pharmacol.*, 32 (10):925–929, October 1992. ISSN 0091-2700. doi: 10.1002/j.1552-4604.1992.tb04640.x.

[14] Andrea Calcaterra and Ilaria D'Acquarica. The market of chiral drugs: Chiral switches versus de novo enantiomerically pure compounds. *J. Pharm. Biomed. Anal.*, 147:323–340, January 2018. ISSN 0731-7085. doi: 10.1016/j.jpba.2017.07.008.

[15] Mary McCarthy and Patrick J. Guiry. Axially chiral bidentate ligands in asymmetric catalysis. *Tetrahedron*, 57(18):3809–3844, April 2001. ISSN 0040-4020. doi: 10.1016/S0040-4020(01)00087-4.

[16] 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. *Chem. Sci.*, 15(34):13618–13630, August 2024. ISSN 2041-6520. doi: 10.1039/D4SC03647F.

[17] Oleg N. Temkin and P. P. Pozdeev. *Homogeneous Catalysis with Metal Complexes: Kinetic Aspects and Mechanisms.* Wiley, Hoboken, NJ, USA, April 2012. ISBN 978-0-470-66699-9.

[18] Christopher Masters. *Homogeneous Transition-metal Catalysis.* Springer Netherlands, Dordrecht, The Netherlands, 1981. ISBN 978-94-009-5880-7.

[19] James Keeler and Peter Wothers. *Chemical Structure and Reactivity: An Integrated Approach.* Oxford University Press, Oxford, England, UK, December 2013. ISBN 978-0-19960413-5.

[20] Gheorghe Duca. *Homogeneous Catalysis with Metal Complexes.* Springer, Berlin, Germany, 2012. ISBN 978-3-642-24629-6.

[21] Ilya D. Gridnev and Tsuneo Imamoto. Mechanism of enantioselection in Rh-catalyzed asymmetric hydrogenation. The origin of utmost catalytic performance. *Chem. Commun.*, (48):7447–7464, November 2009. ISSN 1359-7345. doi: 10.1039/B912440C.

[22] Miguel A. Esteruelas, Juana Herrero, Marta Martı'n, Luis A. Oro, and Vı'ctor M. Real. Mechanism of the hydrogenation of 2,5-norbornadiene catalyzed by [Rh(NBD)(PPh3)2]BF4 in dichloromethane: a kinetic and spectroscopic investigation. *J. Organomet. Chem.*, 599(2):178–184, April 2000. ISSN 0022-328X. doi: 10.1016/S0022-328X(99)00761-5.

[23] Paul C. J. Kamer and Piet W. N. M. van Leeuwen. *Phosphorus(III) Ligands in Homogeneous Catalysis: Design and Synthesis.* ohn Wiley Sons, May 2012. ISBN 978-0-47066627-2. doi: 10.1002/9781118299715.

[24] James H. Kim and Anthony R. Scialli. Thalidomide: the tragedy of birth defects and the effective treatment of disease. *Toxicol. Sci.*, 122(1):1–6, July 2011. ISSN 1096-0929. doi: 10.1093/toxsci/kfr088.

[25] Arthur A. Ciociola, Lawrence B. Cohen, Prasad Kulkarni, Costas Kefalas, Alan Buchman, Carol Burke, Tedd Cain, Jason Connor, Eli D. Ehrenpreis, John Fang, Ronnie Fass, Robyn Karlstadt, Dan Pambianco, Joseph Phillips, Mark Pochapin, Paul Pockros, Philip Schoenfeld, Raj Vuppalanchi, and The Fda-Related Matters Committee Of the American College of Gastroenterology. How Drugs are Developed and Approved by the FDA: Current Process and Future Directions. *Official journal of the American College of Gastroenterology | ACG*, 109(5):620, May 2014. ISSN 0002-9270. doi: 10.1038/ajg.2013.407.

[26] Jèssica Margalef, Maria Biosca, Pol de la Cruz Sánchez, Jorge Faiges, Oscar Pàmies, and Montserrat Diéguez. Evolution in heterodonor P-N, P-S and P-O chiral ligands for preparing efficient catalysts for asymmetric catalysis. From design to applications. *Coord. Chem. Rev.*, 446:214120, November 2021. ISSN 0010-8545. doi: 10.1016/j.ccr.2021.214120.

[27] Vivek Sinha. *The Molecular Basis of Clean Energy: Elucidating the Mechanism of Homogeneously Catalyzed Hydrogen Production from Methanol.* PhD thesis, Universiteit van Amsterdam, 2019.

[28] Wolfram Koch and Max C. Holthausen. *A Chemist's Guide to Density Functional Theory.* Wiley, July 2001. ISBN 978-3-52730372-4. doi: 10.1002/3527600043.

[29] Attila Szabo and Neil S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory (Dover Books on Chemistry).* Dover Publications, Mineola, NY, USA, July 1996. ISBN 978-0-48669186-2.

[30] David J. Wales and Tetyana V. Bogdan. Potential Energy and Free Energy Landscapes. *J. Phys. Chem. B*, 110(42):20765–20776, October 2006. ISSN 1520-6106. doi: 10.1021/jp0680544.

[31] Libretexts. 30.10: The Potential-Energy Surface Can Be Calculated Using Quantum Mechanics. *Chemistry LibreTexts*, April 2023. URL `https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Physical_Chemistry_%28LibreTexts%29/30%3A_Gas-Phase_Reaction_Dynamics/30.10%3A_The_Potential-Energy_Surface_Can_Be_Calculated_Using_Quantum_Mechanics`.

[32] Opt | Gaussian.com, January 2025. URL https://gaussian.com/opt. [Online; accessed 29. Jan. 2025].

[33] SP | Gaussian.com, January 2025. URL https://gaussian.com/sp. [Online; accessed 29. Jan. 2025].

[34] Adarsh V. Kalikadien, Evgeny A. Pidko, and Vivek Sinha. ChemSpaX: exploration of chemical space by automated functionalization of molecular scaffold. *Digital Discovery*, 1(1):8–25, February 2022. ISSN 2635-098X. doi: 10.1039/D1DD00017A.

[35] Vivek Sinha, Jochem J. Laan, and Evgeny A. Pidko. Accurate and rapid prediction of pKa of transition metal complexes: semiempirical quantum chemistry with a data-augmented approach. *Phys. Chem. Chem. Phys.*, 23(4):2557–2567, February 2021. ISSN 1463-9076. doi: 10.1039/D0CP05281G.

[36] 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. *J. Chem. Theory Comput.*, 15(3):1652–1671, March 2019. ISSN 1549-9618. doi: 10.1021/acs.jctc.8b01176.

[37] 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). *J. Chem. Theory Comput.*, 13(5):1989–2009, May 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00118.

[38] Stefan Grimme, Christoph Bannwarth, Eike Caldeweyher, Jana Pisarek, and Andreas Hansen. A general intermolecular force field based on tight-binding quantum chemical calculations. *J. Chem. Phys.*, 147 (16):161708, July 2017. ISSN 0021-9606. doi: 10.1063/1.4991798.

[39] Sebastian Spicher and Stefan Grimme. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angew. Chem. Int. Ed.*, 59(36):15665–15673, September 2020. ISSN 1433-7851. doi: 10.1002/anie.202004239.

[40] CREST, December 2024. URL https://crest-lab.github.io/crest-docs. [Online; accessed 29. Jan. 2025].

[41] Philipp Pracht, Fabian Bohle, and Stefan Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.*, 22(14):7169–7192, April 2020. ISSN 1463-9076. doi: 10.1039/C9CP06869D.

[42] EPiCs-group. obelix, January 2025. URL https://github.com/EPiCs-group/obelix. [Online; accessed 29. Jan. 2025].

[43] Hervé Clavier and Steven P. Nolan. Percent buried volume for phosphine and N-heterocyclic carbene ligands: steric properties in organometallic chemistry. *Chem. Commun.*, 46(6):841–861, January 2010. ISSN 1359-7345. doi: 10.1039/B922984A.

[44] Extended Connectivity Fingerprint ECFP | Chemaxon Docs, January 2025. URL https://docs.chemaxon.com/display/docs/fingerprints_extended-connectivity-fingerprint-ecfp.md. [Online; accessed 29. Jan. 2025].

[45] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, May 2010. ISSN 1549-9596. doi: 10.1021/ci100050t.

[46] Dave Berry. From Theory to Code: A Deep Dive into Molecular Extended-Connectivity Fingerprints (ECFPs) with Python. *Medium*, November 2024. URL https://medium.com/@musicalchemist/from-theory-to-code-a-deep-dive-into-molecular-extended-connectivity-fingerprints-ecfps-with-da1ed436925e.

[47] D. Bajusz, A. Rácz, and K. Héberger. *Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching*. December 2017. ISBN 978-0-12409547-2. doi: 10.1016/B978-0-12-409547-2.12345-5.

[48] Ivan Belcic and Cole Stryker. Supervised Learning, December 2024. URL https://www.ibm.com/think/topics/supervised-learning. [Online; accessed 11. Feb. 2025].

[49] Sára Finta. Research in high-throughput conformer search methods for homogeneous catalysis. Master's thesis, Delft University of Technology, 2024.

[50] Johnson & Johnson: Changing health for humanity, January 2025. URL `https://www.jnj.com`. [Online; accessed 30. Jan. 2025].

[51] Peter Dierkes and Piet W. N. M. van Leeuwen. The bite angle makes the difference: a practical ligand parameter for diphosphine ligands. *J. Chem. Soc., Dalton Trans.*, (10):1519–1530, January 1999. ISSN 0300-9246. doi: 10.1039/A807799A.

[52] Piet W. N. M. van Leeuwen, Paul C. J. Kamer, Joost N. H. Reek, and Peter Dierkes. Ligand Bite Angle Effects in Metal-catalyzed C-C Bond Formation. *Chem. Rev.*, 100(8):2741–2770, August 2000. ISSN 0009-2665. doi: 10.1021/cr9902704.

[53] Contributeurs aux projets Wikimedia. Ferrocène — Wikipédia, February 2025. URL `https://fr.wikipedia.org/w/index.php?title=Ferroc%C3%A8ne&oldid=222723213`. [Online; accessed 6. Feb. 2025].

[54] Simon Axelrod and Rafael Gómez-Bombarelli. Molecular machine learning with conformer ensembles. *Mach. Learn.: Sci. Technol.*, 4(3):035025, August 2023. ISSN 2632-2153. doi: 10.1088/2632-2153/acefa7.

[55] Quentin S. Hanley. The Distribution of Standard Deviations Applied to High Throughput Screening. *Sci. Rep.*, 9(1268):1–13, February 2019. ISSN 2045-2322. doi: 10.1038/s41598-018-36722-4.

[56] Hans Briem and Uta F. Lessel. *In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes*, volume 20. Springer Nature, December 2000. ISBN 978-0-7923-6633. doi: 10.1023/A:1008793325522.

[57] Nicholas Rhodes, David E. Clark, and Peter Willett. Similarity Searching in Databases of Flexible 3D Structures Using Autocorrelation Vectors Derived from Smoothed Bounded Distance Matrices. *J. Chem. Inf. Model.*, 46(2):615–619, March 2006. ISSN 1549-9596. doi: 10.1021/ci0503863.

Appendices contents

A Appendix A 1
- A.1 Methods . . . 2
- A.2 Results and Discussion . . . 18
  - A.2.1 Sensitivity analysis . . . 18
  - A.2.2 Removal stereochemistry . . . 19
  - A.2.3 ECFP . . . 19
- A.3 ChatGPT . . . 20
  - A.3.1 Examples . . . 21

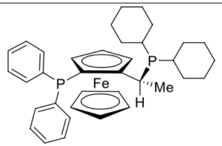
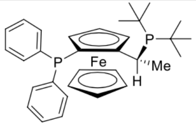
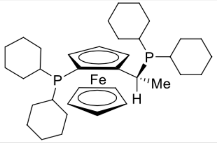
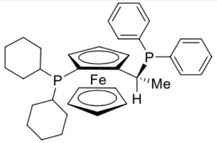
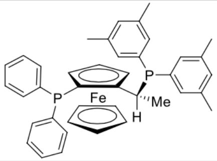
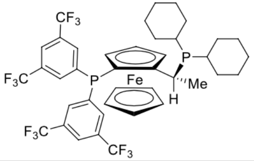
Appendix – 1

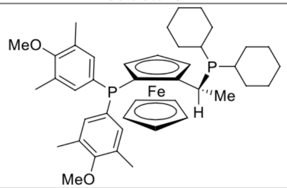
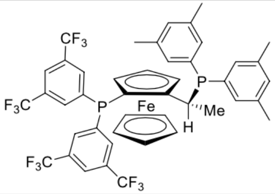
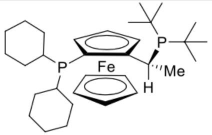
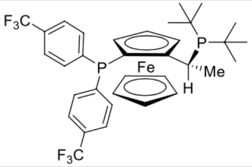
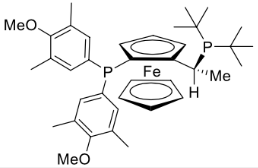
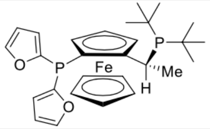
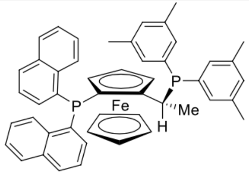
A

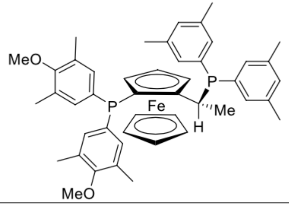
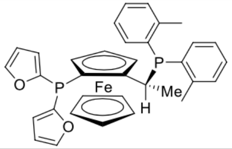
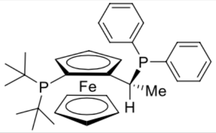
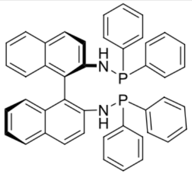
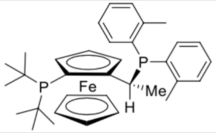
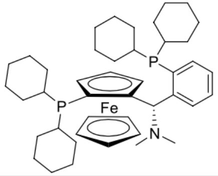
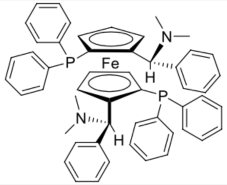
Appendix A

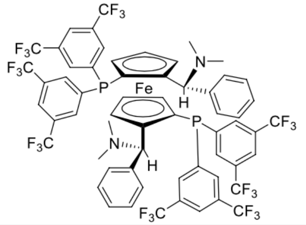
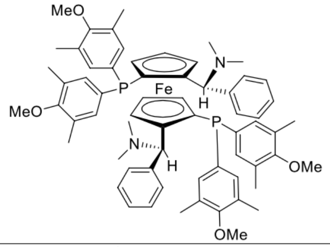
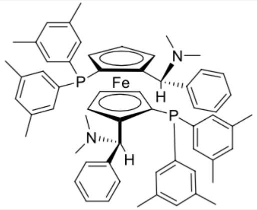
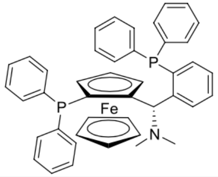
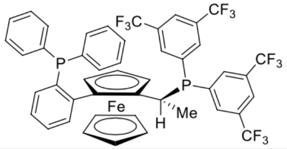
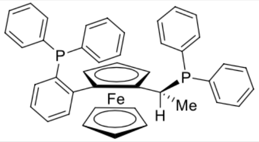
A.1. Methods

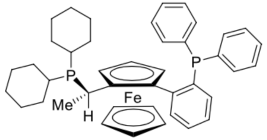
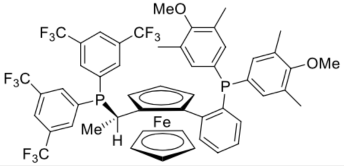
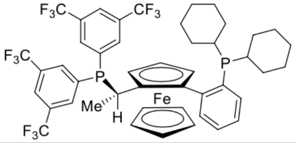
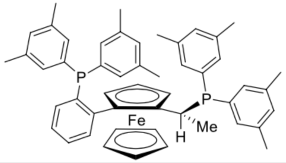
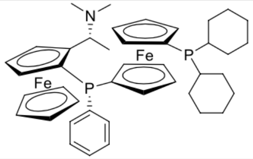
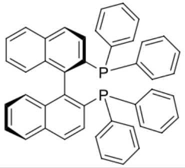
Definitive list of ligands (Oct ’23)

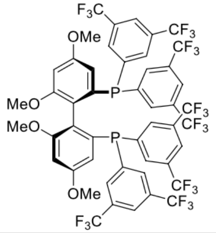
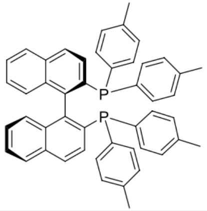
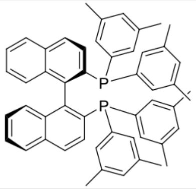
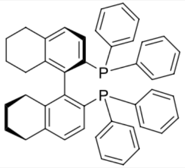
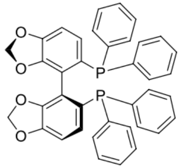
#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L1		SL-J001-1	155806-35-2	PP	$C_{36}H_{44}FeP_2$	1.01	594.54	Sigma-Aldrich	88717
L2		SL-J002-1	155830-69-6	PP	$C_{32}H_{40}FeP_2$	1.01	542.46	Sigma-Aldrich	88719
L3		SL-J003-1	167416-28-6	PP	$C_{36}H_{56}FeP_2$	1.04	606.64	Sigma-Aldrich	88721
L4		SL-J004-1	158923-09-2	PP	$C_{36}H_{44}FeP_2$	1.03	594.54	Sigma-Aldrich	88723
L5		SL-J005-1	184095-69-0	PP	$C_{40}H_{40}FeP_2$	1.03	638.55	Sigma-Aldrich	88725
L6		SL-J006-1	292638-88-1	PP	$C_{40}H_{40}F_{12}FeP_2$	1.06	866.53	Sigma-Aldrich	88727

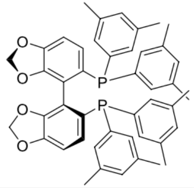
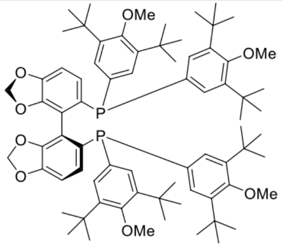
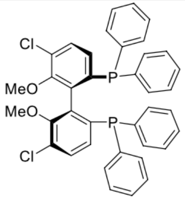
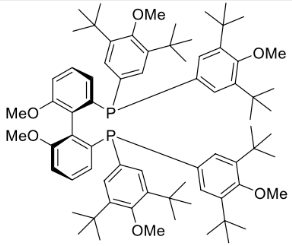
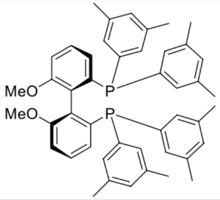
#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L7		SL-J007-1	360048-63-1	PP	$C_{42}H_{56}FeO_2P_2$	1.00	710.70	Sigma-Aldrich	88729
L8		SL-J008-1	166172-63-0	PP	$C_{44}H_{36}F_{12}FeP_2$	1.03	910.55	Sigma-Aldrich	88731
L9		SL-J009-1	158923-11-6	PP	$C_{32}H_{52}FeP_2$	1.03	554.56	Sigma-Aldrich	88733
L10		SL-J011-1	246231-79-8	PP	$C_{34}H_{38}F_6FeP_2$	1.04	678.46	Sigma-Aldrich	88735
L11		SL-J013-1	187733-50-2	PP	$C_{38}H_{52}FeO_2P_2$	1.03	658.62	Sigma-Aldrich	88737
L12		SL-J212-1	849924-41-0	PP	$C_{28}H_{36}FeO_2P_2$	1.08	522.39	abcr	AB426473
L13		SL-J404-1	851308-40-2	PP	$C_{48}H_{44}FeP_2$	1.03	738.67	STREM	26-1175

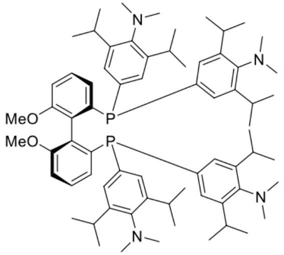
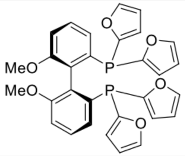
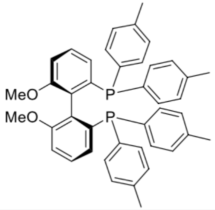
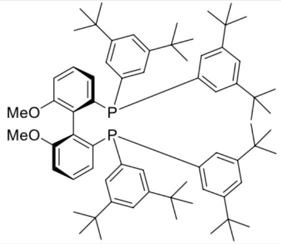
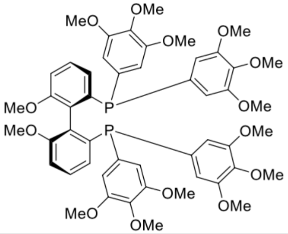
#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M$_w$ [g mol^{-1}]	Vendor	Catalog #
L14		SL-J418-1	849924-45-4	PP	$C_{46}H_{52}FeO_2P_2$	1.03	754.71	Sigma-Aldrich	88747
L15		SL-J452-1	849924-73-8	PP	$C_{34}H_{32}FeO_2P_2$	1.05	590.42	Sigma-Aldrich	88751
L16		SL-J502-1	223120-71-6	PP	$C_{32}H_{40}FeP_2$	1.06	542.46	Sigma-Aldrich	88753
L17		(*R*)-BINAM-P	74974-14-4	PP	$C_{44}H_{34}N_2P_2$	1.02	652.72	Sigma-Aldrich	708615
L18		SL-J505-1	849924-76-1	PP	$C_{34}H_{44}FeP_2$	1.07	570.52	Sigma-Aldrich	88755
L19		SL-T002-2	914089-00-2	PP	$C_{43}H_{63}FeNP_2$	1.01	711.78	Sigma-Aldrich	07542
L20		SL-M001-1	174467-31-3	PP	$C_{52}H_{50}FeN_2P_2$	1.03	820.78	Sigma-Aldrich	73463

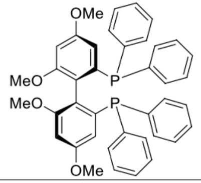
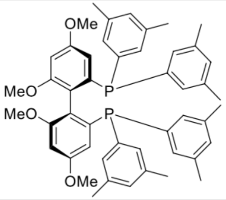
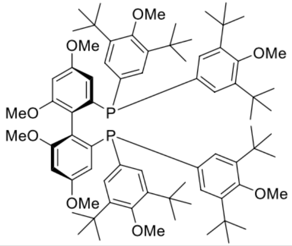
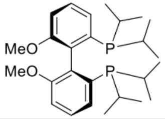
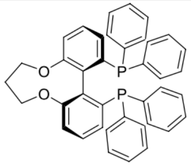
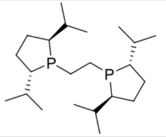
#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M$_w$ [g mol^{-1}]	Vendor	Catalog #
L21		SL-M003-1	494227-36-0	PP	$C_{60}H_{42}F_{24}FeN_2P_2$	1.02	1364.76	Sigma-Aldrich	73467
L22		SL-M004-1	494227-37-1	PP	$C_{64}H_{74}FeN_2O_4P_2$	1.04	1053.10	Sigma-Aldrich	73469
L23		SL-M009-1	793718-16-8	PP	$C_{60}H_{66}FeN_2P_2$	1.01	932.99	Sigma-Aldrich	73471
L24		SL-T001-2	850444-36-9	PP	$C_{43}H_{39}FeNP_2$	1.03	687.58	Sigma-Aldrich	73476
L25		SL-W001-1	565184-33-0 (AKA 387868-06-6)	PP	$C_{46}H_{32}F_{12}FeP_2$	1.05	930.54	Sigma-Aldrich	65671
L26		SL-W002-1	388079-58-1	PP	$C_{42}H_{36}FeP_2$	1.08	658.54	Sigma-Aldrich	65673

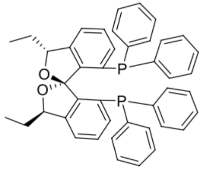
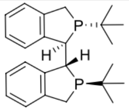
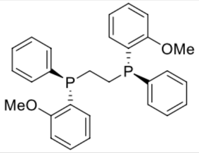
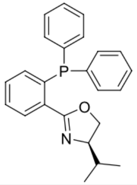
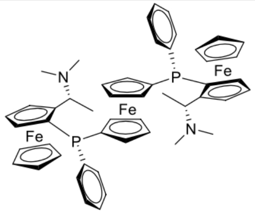
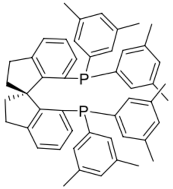
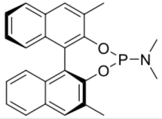
#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M$_w$ [g mol^{-1}]	Vendor	Catalog #
L27		SL-W003-2	849925-19-5	PP	$C_{42}H_{48}FeP_2$	1.06	670.64	Sigma-Aldrich	65676
L28		SL-W005-2	849925-20-8	PP	$C_{52}H_{44}F_{12}FeO_2P_2$	1.02	1046.70	Sigma-Aldrich	65678
L29		SL-W008-2	849925-22-0	PP	$C_{46}H_{44}F_{12}FeP_2$	1.07	942.63	Sigma-Aldrich	65682
L30		SL-W009-1	894771-28-9	PP	$C_{50}H_{52}FeP_2$	1.02	770.76	STREM	26-1555
L31		SL-F356-1	952586-19-5	PP	$C_{42}H_{53}Fe_2NP_2$	1.04	745.53	Sigma-Aldrich	779075
L32		(*R*)-BINAP	76189-55-4	PP	$C_{44}H_{32}P_2$	1.09	622.69	Sigma-Aldrich	693065

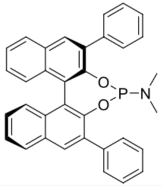
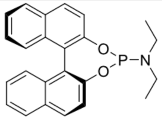
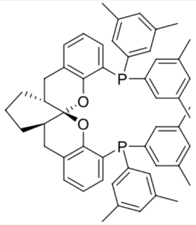
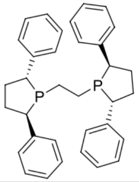
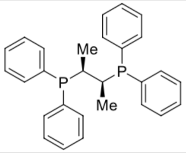
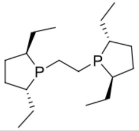
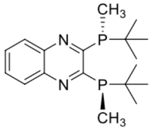
#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L33	 F3C OMe CF3 CF3 P MeO CF3 MeO CF3 P OMe CF3 CF3 F3C	(*R*)-BTFM-GarPhos	1365531-84-5	PP	$C_{48}H_{28}F_{24}O_4P_2$	1.07	1186.66	STREM	15-1663
L34		(*R*)-Tol-BINAP	99646-28-3	PP	$C_{48}H_{40}P_2$	1.04	678.80	Sigma-Aldrich	693049
L35		(*R*)-Xyl-BINAP (AKA (*R*)-DM-BINAP)	137219-86-4	PP	$C_{52}H_{48}P_2$	1.05	734.90	Sigma-Aldrich	692379
L36		(*R*)-H₈-BINAP	139139-86-9	PP	$C_{44}H_{40}P_2$	1.09	630.75	Sigma-Aldrich	692387
L37		(*S*)-SegPhos	210169-54-3	PP	$C_{38}H_{28}O_4P_2$	1.09	610.59	Sigma-Aldrich	693006

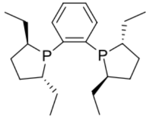
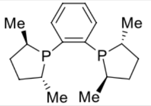
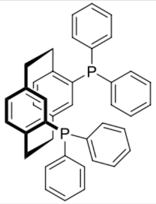
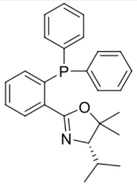
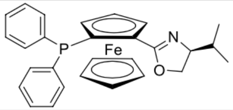
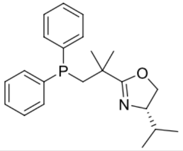
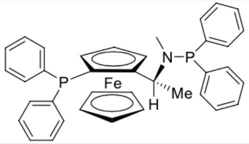
#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L38		(*S*)-Xyl-SegPhos (AKA (*S*)-DM-SegPhos)	210169-57-6	PP	$C_{46}H_{44}O_4P_2$	1.05	722.80	Sigma-Aldrich	692999
L39		(*S*)-DTBM-SegPhos	210169-40-7	PP	$C_{74}H_{100}O_8P_2$	1.09	1179.55	Sigma-Aldrich	692980
L40		(*R*)-Cl-MeO-BIPHEP	185913-97-7	PP	$C_{38}H_{30}Cl_2O_2P_2$	1.00	651.50	Sigma-Aldrich	96738
L41		SL-A109-1	352655-61-9	PP	$C_{74}H_{104}O_6P_2$	1.01	1151.59	Sigma-Aldrich	29512
L42		SL-A120-1	394248-45-4	PP	$C_{46}H_{48}O_2P_2$	1.07	694.84	Sigma-Aldrich	29516

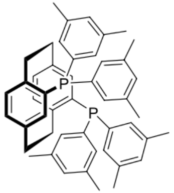
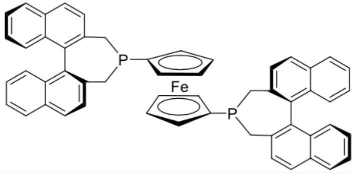
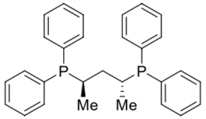
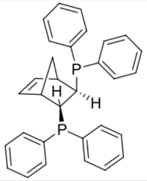
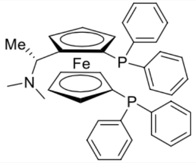
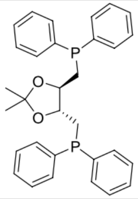
#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L43		SL-A107-1	352655-40-4	PP	$C_{70}H_{100}N_4O_2P_2$	1.02	1091.54	Sigma-Aldrich	29528
L44		SL-A108-2	145214-59-1	PP	$C_{30}H_{24}O_6P_2$	1.08	542.46	Sigma-Aldrich	29515
L45		SL-A102-2	133545-25-2	PP	$C_{42}H_{40}O_2P_2$	1.01	638.73	Sigma-Aldrich	29521
L46		SL-A121-1	192138-05-9	PP	$C_{70}H_{96}O_2P_2$	1.10	1031.48	Sigma-Aldrich	29524
L47		SL-A104-1	256390-47-3	PP	$C_{50}H_{56}O_{14}P_2$	1.05	942.93	Sigma-Aldrich	29526

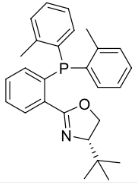
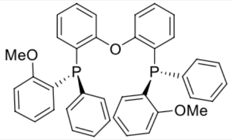
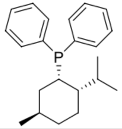
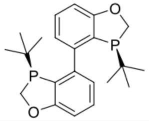
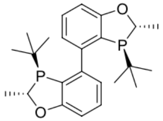
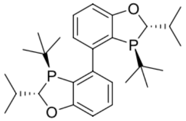
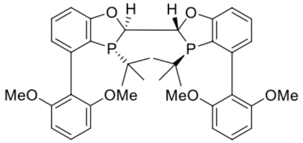
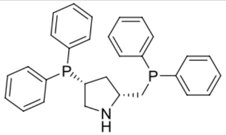
#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L48		(*R*)-GarPhos	1365531-75-4	PP	$C_{40}H_{36}O_4P_2$	1.06	642.67	Sigma-Aldrich	754927
L49		(*R*)-Xyl-GarPhos	1365531-89-0	PP	$C_{48}H_{52}O_4P_2$	1.00	754.89	Sigma-Aldrich	761389
L50		(*R*)-DTBM-GarPhos	1365531-98-1	PP	$C_{76}H_{108}O_8P_2$	1.03	1211.64	Sigma-Aldrich	761419
L51		(*S*)-^{i}Pr-BIPHEP	150971-43-0	PP	$C_{26}H_{40}O_2P_2$	1.03	446.55	STREM	15-0655
L52		(*R*)-C$_3$-TunePhos	301847-89-2	PP	$C_{39}H_{32}O_2P_2$	1.05	594.63	STREM	15-0175
L53		(*S*,*S*)-^{i}Pr-BPE	528854-34-4	PP	$C_{22}H_{44}P_2$	1.07	370.54	Sigma-Aldrich	668435

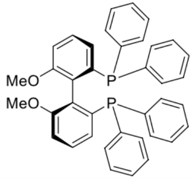
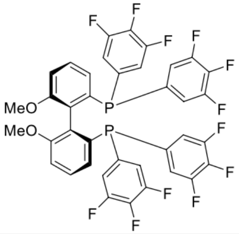
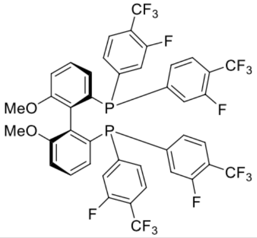
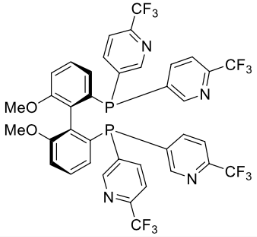
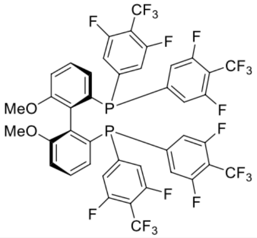
#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L54		(*R*,*R*,*R*)-SPIRAP (AKA (*R*)-CrabPhos)	NA	PP	$C_{43}H_{38}O_2P_2$	1.03	648.72	Sigma-Aldrich	905240
L55		(*R*,*R*,*S*,*S*)-DuanPhos	528814-26-8	PP	$C_{24}H_{32}P_2$	1.09	382.46	Sigma-Aldrich	657697
L56		(*R*,*R*)-DIPAMP	55739-58-7	PP	$C_{28}H_{28}O_2P_2$	1.06	458.48	Sigma-Aldrich	697761
L57		(*R*)-^{i}Pr-PHOX	164858-78-0	PN	$C_{24}H_{24}NOP$	1.09	373.44	Sigma-Aldrich	72575
L58		SL-F131-1 (AKA Trifer)	899811-43-9	PP	$C_{50}H_{54}Fe_3N_2P_2$	1.03	912.48	STREM	26-1260
L59		(*R*)-Xyl-SDP	917377-75-4	PP	$C_{49}H_{50}P_2$	1.04	700.89	STREM	15-5168
L60		(*S*)-DM-MonoPhos	185449-86-9	Phosphoramidite	$C_{24}H_{22}NO_2P$	2.00	387.42	STREM	15-1255

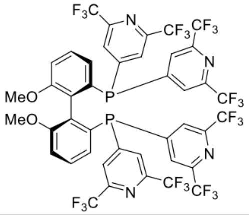
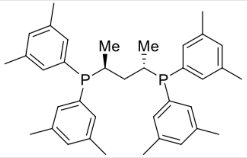
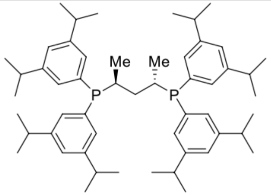
#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L61		(*R*)-Ph-MonoPhos	936010-61-6	Phosphoramidite	$C_{34}H_{26}NO_2P$	2.00	511.56	STREM	15-5620
L62		(*S*)-NEt$_2$-MonoPhos	252288-04-3	Phosphoramidite	$C_{24}H_{22}NO_2P$	2.00	387.42	STREM	15-1231
L63		(*R*,*R*,*R*)-Xyl-SKP	1429939-35-4	PP	$C_{52}H_{54}O_2P_2$	1.03	772.95	STREM	15-4320
L64		(*R*,*R*)-Ph-BPE	528565-79-9	PP	$C_{34}H_{36}P_2$	1.09	506.61	Sigma-Aldrich	667811
L65		(*S*,*S*)-ChiraPhos	64896-28-2	PP	$C_{28}H_{28}P_2$	1.04	426.48	Sigma-Aldrich	259098
L66		(*R*,*R*)-Et-BPE	136705-62-9	PP	$C_{18}H_{36}P_2$	1.06	314.43	Sigma-Aldrich	668478
L67		(*R*)-QuinoxP	866081-62-1	PP	$C_{18}H_{28}N_2P_2$	1.06	334.38	Sigma-Aldrich	676403

#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L68		(*R*,*R*)-Et-DuPhos	136705-64-1	PP	$C_{22}H_{36}P_2$	1.04	362.48	Sigma-Aldrich	668494
L69		(*R*,*R*)-Me-DuPhos	147253-67-6	PP	$C_{18}H_{28}P_2$	1.08	306.37	Sigma-Aldrich	665258
L70		(*S*)-PhanePhos	192463-40-4	PP	$C_{40}H_{34}P_2$	1.09	576.66	Sigma-Aldrich	682136
L71		(*S*)-Me-^{i}Pr-PHOX	1152313-76-2	PN	$C_{26}H_{28}NOP$	1.05	401.49	Sigma-Aldrich	719641
L72		SL-N003-2	163169-29-7	PN	$C_{28}H_{28}FeNOP$	1.09	481.36	Sigma-Aldrich	717398
L73		(*S*)-NeoPHOX	1199225-38-1	PN	$C_{22}H_{28}NOP$	1.05	353.45	Sigma-Aldrich	729264
L74		(*R*,*S*)-Me-BoPhoz	406680-94-2	PP	$C_{37}H_{35}FeNP_2$	1.06	611.49	Sigma-Aldrich	682322

#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L75		(*R*)-Xyl-PhanePhos	325168-89-6	PP	$C_{48}H_{50}P_2$	1.08	688.88	Sigma-Aldrich	682306
L76		(*S*,*S*)-f-Binaphane	544461-38-3	PP	$C_{54}H_{40}FeP_2$	1.09	806.71	STREM	26-0243
L77		(*R*,*R*)-BDPP	96183-46-9	PP	$C_{29}H_{30}P_2$	1.09	440.51	STREM	15-0432
L78		(*R*,*R*)-NorPhos	71042-55-2	PP	$C_{31}H_{28}P_2$	1.09	462.51	STREM	15-0140
L79		(*R*,*S*)-BPPFA	74311-56-1	PP	$C_{38}H_{37}FeNP_2$	1.06	625.51	Sigma-Aldrich	344087
L80		(*R*,*R*)-DIOP	32305-98-9	PP	$C_{31}H_{32}O_2P_2$	1.07	498.54	Sigma-Aldrich	237655

#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L81		(*S*)-Tol-^{t}Bu-PHOX	218460-00-5	PN	C27H30NOP	1.10	415.52	Sigma-Aldrich	688533
L82		(*S*,*S*)-DPE-Phos	2119686-55-2	PP	$C_{38}H_{32}O_3P_2$	1.06	598.62	STREM	15-1279
L83		(*S*)-NMDPP	43077-29-8	P	$C_{22}H_{29}P$	2.00	324.45	STREM	15-3490
L84		(*S*,*S*)-BABIBOP	2207601-04-3	PP	$C_{22}H_{28}O_2P_2$	1.03	386.41	STREM	15-6410
L85		(*S*,*S*,*S*,*S*)-Me-BABIBOP	2207601-10-1	PP	$C_{24}H_{32}O_2P_2$	1.06	414.47	STREM	15-6420
L86		(*S*,*S*,*S*,*S*)-^{i}Pr-BABIBOP	2207601-12-3	PP	$C_{28}H_{40}O_2P_2$	1.08	470.57	STREM	15-6430
L87		(*R*,*R*,*R*,*R*)-Bis-BIDIME	1884680-48-1	PP	$C_{38}H_{44}O_6P_2$	1.03	658.71	STREM	15-6240
L88		(*R*,*R*)-PPM	77450-05-6	PP	$C_{29}H_{29}NP_2$	1.05	453.51	STREM	15-7210

#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L89		SL-A101-1	133545-16-1	PP	$C_{38}H_{32}O_2P_2$	1.03	582.62	Sigma-Aldrich	29510
L90		(*S*)-MeO-F$_{12}$-BIPHEP	116008-37-6	PP	$C_{38}H_{20}F_{12}O_2P_2$	1.06	798.50	Kanto	25971-95
L91		(*R*)-MeO-F$_{16}$-BIPHEP	NA	PP	$C_{42}H_{24}F_{16}O_2P_2$	1.03	926.57	Kanto	2022-008K
L92		(*R*)-MeO-Py-F$_{12}$-BIPHEP	NA	PP	$C_{38}H_{24}F_{12}N_4O_2P_2$	1.06	858.56	Kanto	2022-009K
L93		(*R*)-MeO-F$_{20}$-BIPHEP	NA	PP	$C_{42}H_{20}F_{20}O_2P_2$	1.02	998.54	Kanto	2022-010K

#	Structure	Alias or name	CAS#	Type	Formula	Eq. to Rh	M_w [g mol^{-1}]	Vendor	Catalog #
L94		(*R*)-MeO-BFPy-BIPHEP	NA	PP	$C_{42}H_{20}F_{24}N_4O_2P_2$	1.03	1130.56	Kanto	2022-011K
L95		(*S*,*S*)-XylSKEWPhos	551950-92-6	PP	$C_{37}H_{46}P_2$	1.08	552.72	Kanto	05843-68
L96		(*S*,*S*)-DIPSKEWPhos	NA	PP	$C_{53}H_{78}P_2$	1.10	777.15	Kanto	05845-55

A.2. Results and Discussion

A.2.1. Sensitivity analysis

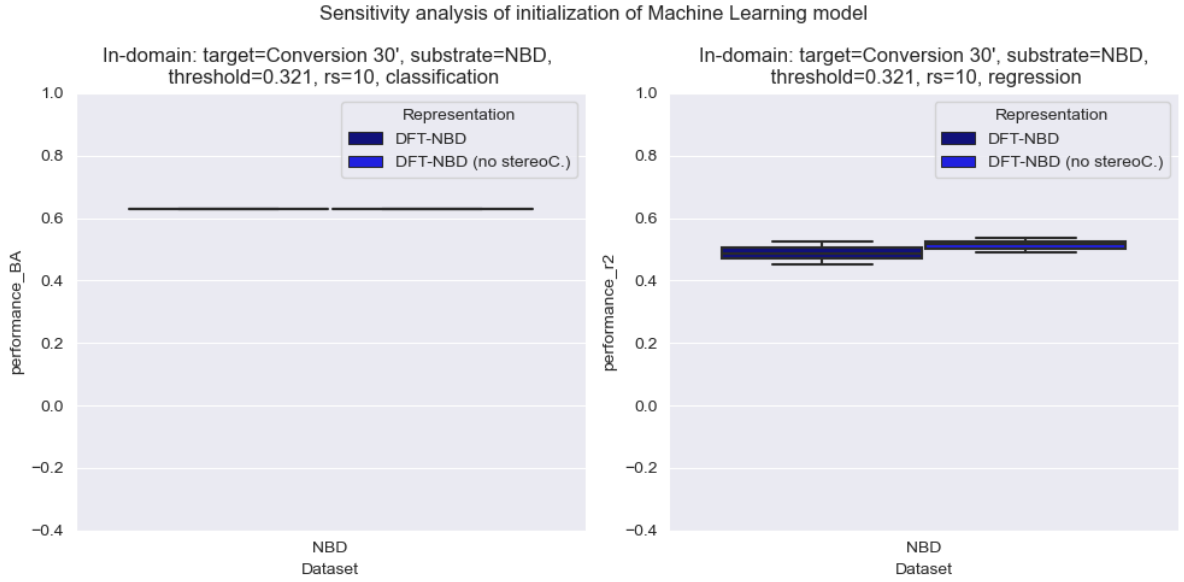


Figure A.1: Sensitivity analysis of initialization of the Machine Learning model. Box plots are shown of the performance of the models using classification (left) and regression (right). Classification show no change after different initializations and regression show small changes.

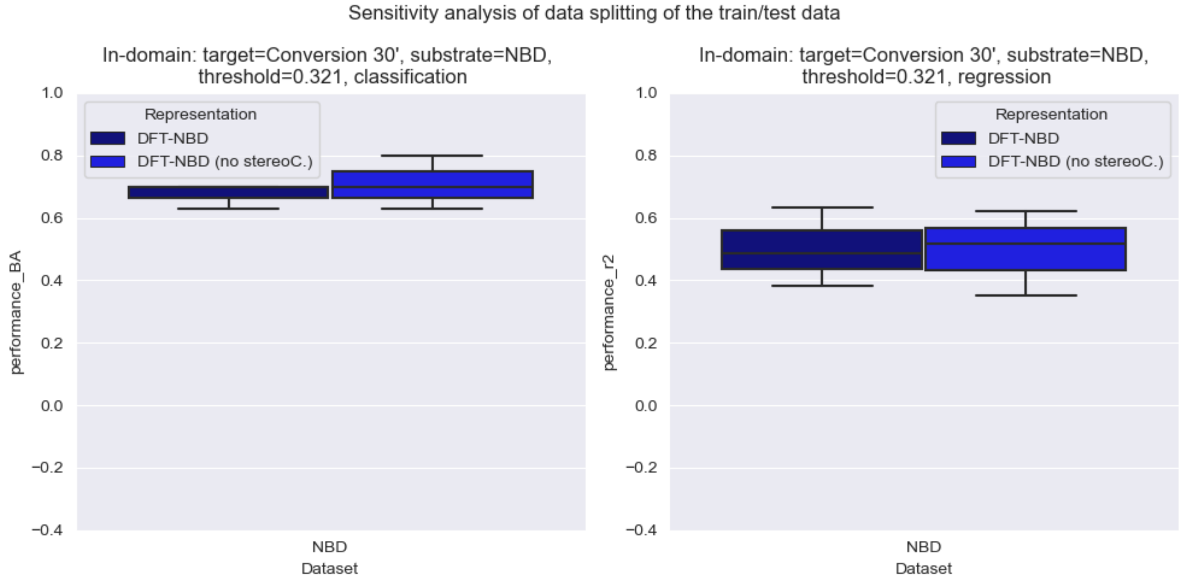


Figure A.2: Sensitivity analysis of data splitting of the train/test data. Boxplots are shown of the performance of the ML models using classification (left) and regression (right). While classification show a small change in performance, regression show a much larger change.

A.2.2. Removal stereochemistry

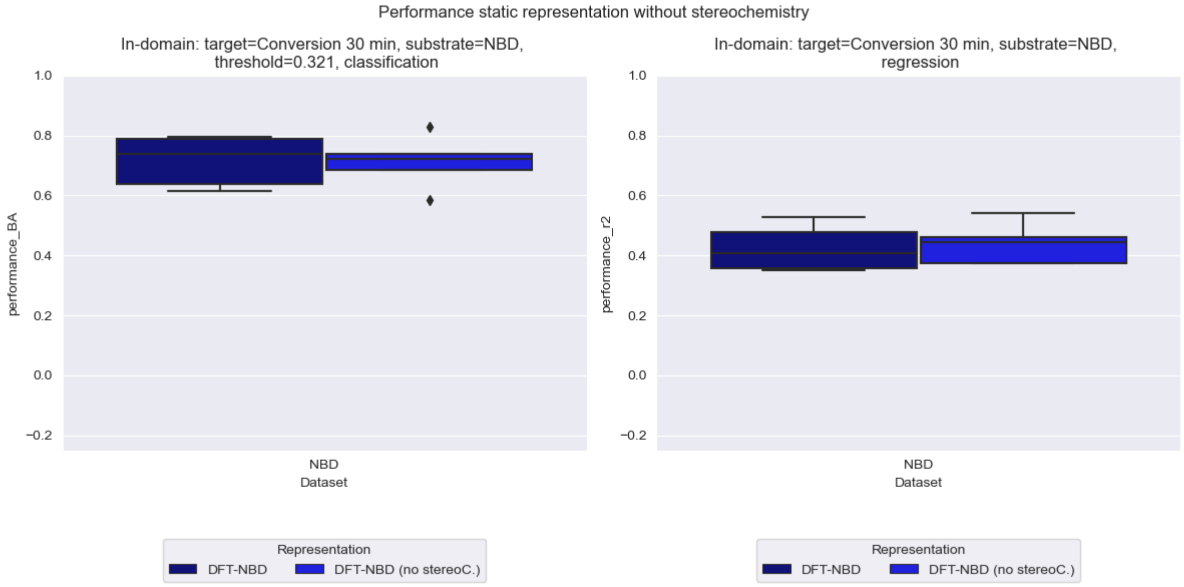


Figure A.3: Effect of removing the stereochemistry descriptors on the NBD 30 min dataset. On the left the classification performances are shown for DFT-NBD with stereochemistry and DFT-NBD without stereochemistry. Performances are similar

A.2.3. ECFP

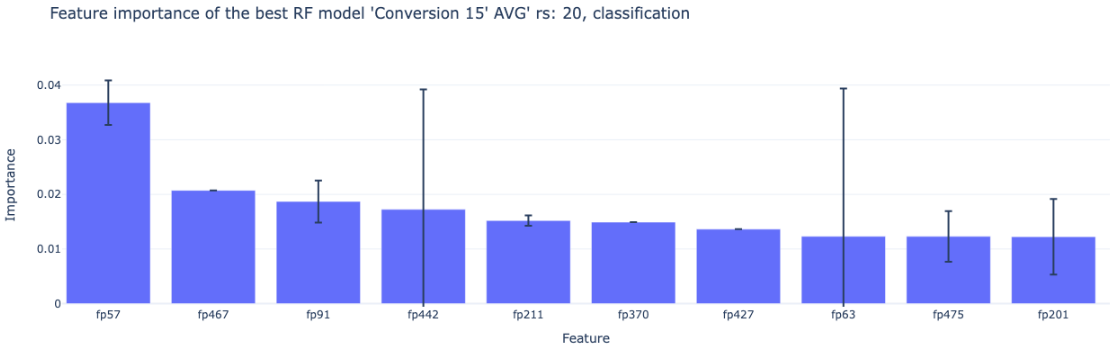


Figure A.4: ECFP Figure Importance plot of random state 20.

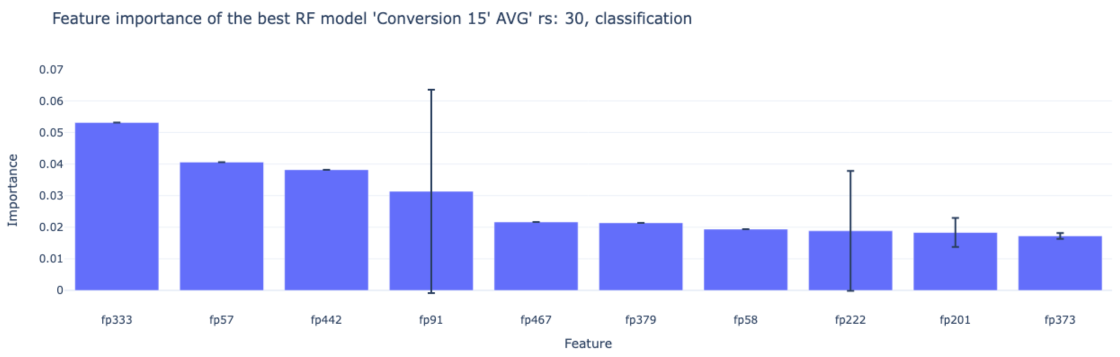


Figure A.5: ECFP Figure Importance plot of random state 30.

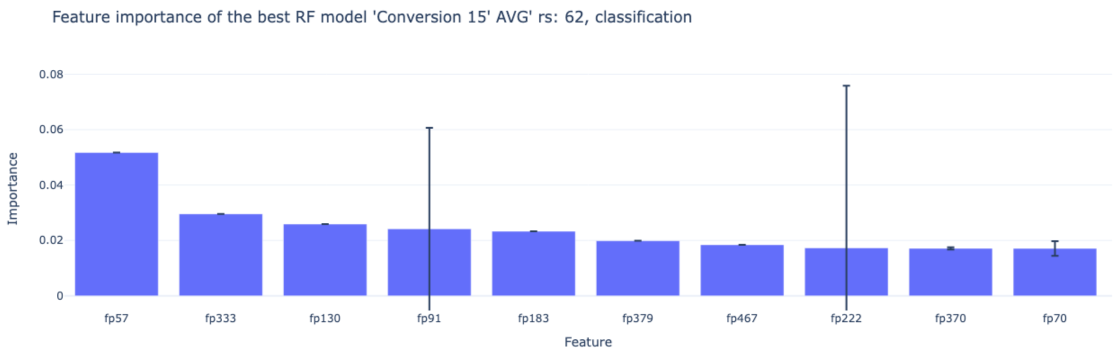


Figure A.6: ECFP Figure Importance plot of random state 62.

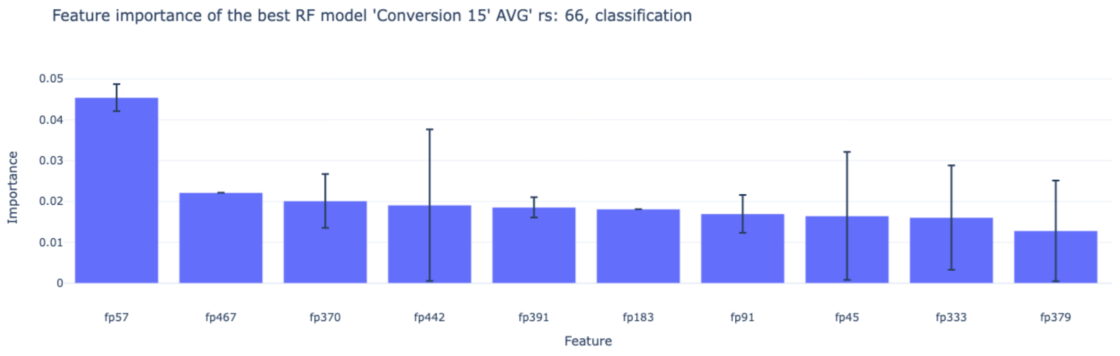
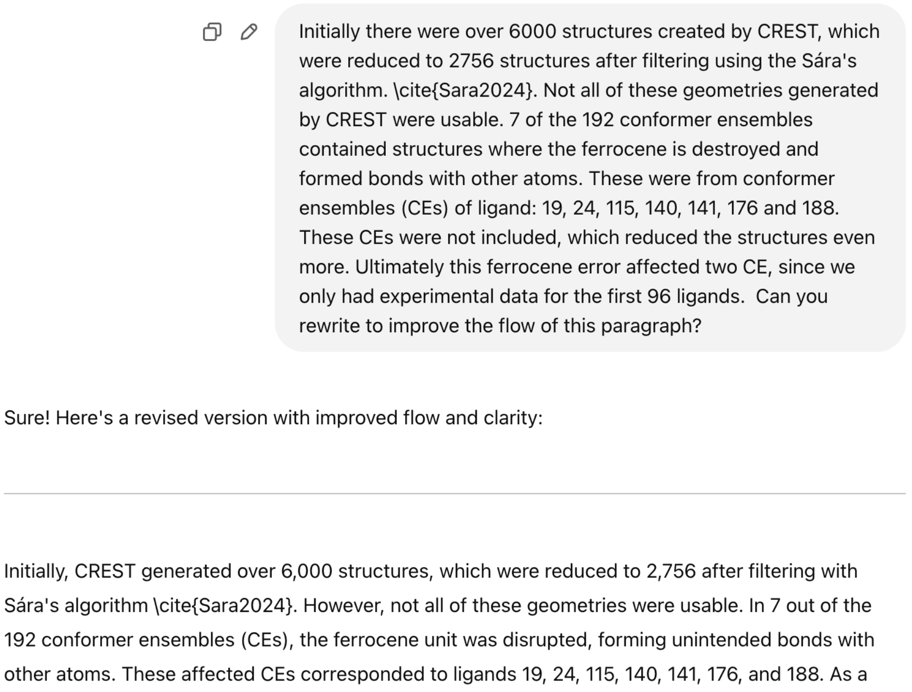


Figure A.7: ECFP Figure Importance plot of random state 66.

A.3. ChatGPT

ChatGPT was used during this thesis to assist and improve my writing and coding. It was primarily used for tedious tasks, which I otherwise had googled. Things such as how to retrieve an entire row or column when a specific value is in a DataFrame, and suggesting synonyms or word to refine my flow. I always made sure I checked everything I asked ChatGPT, to prevent repeating sentences or having inconsistencies. For the coding part, I tried to avoid copy pasting entire codes, as it could alter the calculations or data if I needed to continue on those. I usually only copy pasted code if I ask to display data, for example how to make a Seaborn barplot, or how to change the colors in a plot. When I use ChatGPT to rewrite my text, I try to only use parts that improved the flow or avoid lengthy sentences.

A.3.1. Examples

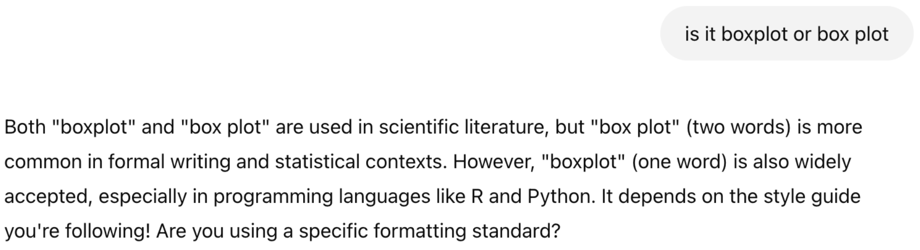


Initially there were over 6000 structures created by CREST, which were reduced to 2756 structures after filtering using the Sára's algorithm. \cite{Sara2024}. Not all of these geometries generated by CREST were usable. 7 of the 192 conformer ensembles contained structures where the ferrocene is destroyed and formed bonds with other atoms. These were from conformer ensembles (CEs) of ligand: 19, 24, 115, 140, 141, 176 and 188. These CEs were not included, which reduced the structures even more. Ultimately this ferrocene error affected two CE, since we only had experimental data for the first 96 ligands. Can you rewrite to improve the flow of this paragraph?

Sure! Here's a revised version with improved flow and clarity:

Initially, CREST generated over 6,000 structures, which were reduced to 2,756 after filtering with Sára's algorithm \cite{Sara2024}. However, not all of these geometries were usable. In 7 out of the 192 conformer ensembles (CEs), the ferrocene unit was disrupted, forming unintended bonds with other atoms. These affected CEs corresponded to ligands 19, 24, 115, 140, 141, 176, and 188. As a

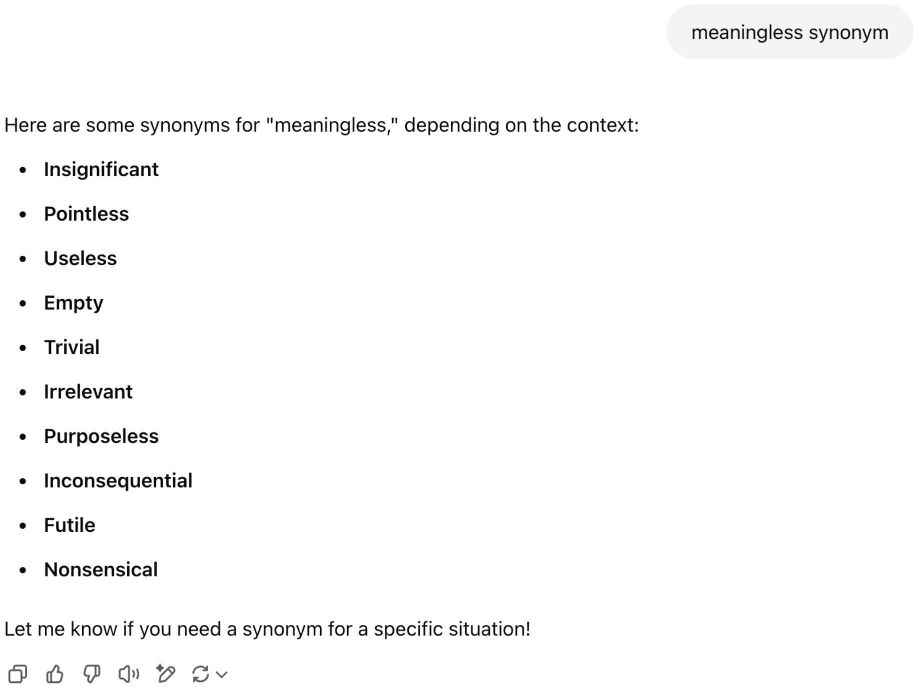
Figure A.8: Example where I inputted a paragraph to see if I missed punctuation and changed structure if it is better.



is it boxplot or box plot

Both "boxplot" and "box plot" are used in scientific literature, but "box plot" (two words) is more common in formal writing and statistical contexts. However, "boxplot" (one word) is also widely accepted, especially in programming languages like R and Python. It depends on the style guide you're following! Are you using a specific formatting standard?

Figure A.9: Example where I checked spelling of words.



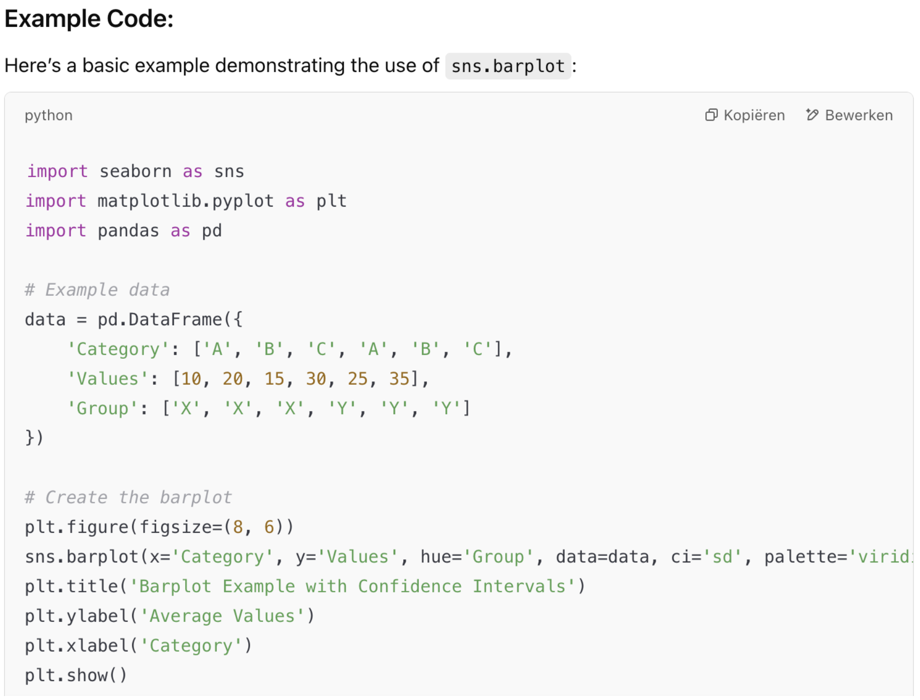
meaningless synonym

Here are some synonyms for "meaningless," depending on the context:

- **Insignificant**
- **Pointless**
- **Useless**
- **Empty**
- **Trivial**
- **Irrelevant**
- **Purposeless**
- **Inconsequential**
- **Futile**
- **Nonsensical**

Let me know if you need a synonym for a specific situation!

Figure A.10: Example where I used ChatGPT for finding synonyms.



Example Code:

Here's a basic example demonstrating the use of `sns.barplot`:

```python
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Example data
data = pd.DataFrame({
    'Category': ['A', 'B', 'C', 'A', 'B', 'C'],
    'Values': [10, 20, 15, 30, 25, 35],
    'Group': ['X', 'X', 'X', 'Y', 'Y', 'Y']
})

# Create the barplot
plt.figure(figsize=(8, 6))
sns.barplot(x='Category', y='Values', hue='Group', data=data, ci='sd', palette='virid
plt.title('Barplot Example with Confidence Intervals')
plt.ylabel('Average Values')
plt.xlabel('Category')
plt.show()
```

Figure A.11: Example where I asked how to create a sns.barplot, when I have a pd.DataFrame.