

Featurization of chemical reactions for in silico catalyst screening

Andrian Mirza



Featurization of chemical reactions for in silico catalyst screening

by

Andrian Mirza

to obtain the degree of Master of Science
at the Delft University of Technology,
to be publicly defended on Tuesday, April 18, 2023 at 9:00 AM.

Performed at:

Inorganic Systems Engineering
Faculty of Applied Sciences

Under supervision of:

Prof. Dr. E. A. Pidko
MSc. A.V. Kalikadien

Student number: 5477514
Project duration: August 11, 2022 – April 11, 2023
Thesis committee: Prof. Dr. E. A. Pidko, TU Delft, AS, supervisor
Prof. Dr. F. Grozema, TU Delft, AS
Dr. G. A. Filonenko, TU Delft, 3ME

This manuscript is confidential and cannot be made public.



Abstract

Catalysts play an essential role in industry and for the general progress of mankind. With the parallel energy and technological transformations, it is important to create tools that aid in the development of better catalysts. To achieve this feat, it is firstly required to have a fully automated approach for *in silico* structure generation. Thus, in this study the *OBeLiX* workflow has been developed. The designed package includes a scaffold generation tool, a substituent placement tool, GFNN-xTB optimization and conformer search tools completed by a fully automated descriptor calculator. Even though descriptor databases can be found in literature, their reproducibility is limited. Consequently, the ability to reconstruct proposed approaches for new chemical reactions is hindered.

OBeLiX has been used to investigate a series of hydrogenation reactions catalyzed by rhodium phosphine complexes. The approach begins with the creation of a structure database for 192 such complexes. To simplify this process, it was opted to use a mechanistically relevant model catalyst structure. In the first step of the catalytic cycle, π -complexation occurs between the substrate and the metal center. Thus, a symmetric chelating norbornadiene molecule has been chosen to model the asymmetric substrates.

The generated database of model catalysts has been featurized through *OBeLiX*. The use of model structures underlined that the substrates have to be quantified as well. While for the complex model catalysts a series of chemically descriptive features have been created, the substrates were converted to two-dimensional fingerprints, and Sterimol parameters that describe the 3D size of the substrate around the double bond that is to be hydrogenated. Therefore, featurization of the chemical reaction has been achieved. Training machine learning algorithms on these features, yielded high correlations including out-of-sample binary reactivity classification for substrates outside the training set.

Contents

| | |
|---|------------|
| List of Figures | iv |
| List of Tables | vii |
| List of Code Listings | vii |
| 1 Introduction | 1 |
| 2 Theoretical background | 4 |
| 2.1 Density functional theory | 4 |
| 2.1.1 Exchange-Correlation Functionals | 5 |
| 2.1.2 Basis sets | 6 |
| 2.1.3 Potential energy surface | 6 |
| 2.1.4 Dispersion corrections | 7 |
| 2.2 GFN optimization tools | 7 |
| 2.2.1 Conformer search | 8 |
| 2.3 Phosphine ligands in homogeneous catalysis | 9 |
| 2.4 Descriptors | 9 |
| 2.4.1 Morfeus descriptors | 10 |
| 2.4.2 Descriptors from DFT calculations | 10 |
| 2.4.3 Subgraph search of transition metal complexes | 11 |
| 2.5 Machine learning models | 12 |
| 2.6 Computer-aided catalyst design | 13 |
| 3 Computational methods | 16 |
| 3.1 OBeLiX workflow | 16 |
| 3.1.1 Scaffold generation from SMILES representations | 16 |
| 3.1.2 Automated substituent placement | 17 |
| 3.1.3 Applied quantum-mechanical methods | 17 |
| 3.1.4 Descriptor calculation | 18 |
| 3.2 Machine learning pipeline | 20 |
| 3.2.1 Hierarchical clustering | 22 |
| 3.2.2 Evaluation of machine learning models | 22 |
| 4 Results & Discussion | 23 |
| 4.1 Substrate hydrogenation | 23 |
| 4.1.1 Experimental data analysis | 24 |
| 4.2 Ligand mapping | 25 |
| 4.2.1 Energy analysis | 27 |

| | | |
|----------|--|-----------|
| 4.3 | Conformer effects on descriptors | 28 |
| 4.4 | Reaction predictions | 30 |
| 4.4.1 | Feature selection | 30 |
| 4.4.2 | Hierarchical clustering | 31 |
| 4.4.3 | Full dataset regression models | 31 |
| 4.4.4 | General classification models | 33 |
| 4.4.5 | Hierarchical clustering classification models | 33 |
| 4.4.6 | Feature importance | 34 |
| 5 | Conclusion & Outlook | 35 |
| 5.1 | Conclusion | 35 |
| 5.2 | Outlook | 36 |
| 5.2.1 | New descriptors | 36 |
| 5.2.2 | Generation of unusual complexes | 36 |
| 5.2.3 | Reaction newtork explorers | 37 |
| | Acknowledgements | 38 |
| | Bibliography | 47 |
| | Appendices | 48 |
| A | Simplest model molecule | 48 |
| B | Conformer correlations | 49 |
| C | Protocol for 3D coordinates generation of sandwich TM complexes | 51 |
| D | GFN2-xTB preoptimized structure | 53 |
| E | Database contents | 54 |
| E.1 | Structures of publicly available substrates | 54 |
| E.2 | Ligands database | 55 |
| E.3 | Descriptor definitions | 62 |
| E.4 | Data availability | 63 |

List of Figures

| | | |
|-----|---|----|
| 1.1 | History of computation in chemistry with emphasis on the studies in homogeneous catalysis [2, 15–20]. | 2 |
| 1.2 | Reaction mechanism of a typical substrate used in this study. The first step indicates how the substrate attaches to the solvated Rh-complex. The second step of the mechanism involves the oxidation of the metal center from Rh(I) to Rh(III) through the oxidative addition of a hydrogen molecule. The migratory insertion step is skipped in this figure. The red highlight shows what bond is hydrogenated and the origin for the choice of a model substrate [26]. | 3 |
| 1.3 | Simplified structure of the taken approach. Step 1 includes structure and descriptor generation for both substrates and model catalysts. | 3 |
| 2.1 | Jacob’s ladder of exchange-correlation functional approximations which shows the increasing complexity of the different available functionals. PBE0 and B3LYP are by far the most used functionals across literature [35–37]. | 6 |
| 2.2 | GFN family of methods. In the center, the force field method is described and how the it is related to the GFNn-xTB methods. Z, fit - extra empirical parametrization of the elements. vdW - van der Waals interactions. EEQ - electronegativity-equilibrium for the description of pairwise interactions. Figure reproduced from [18]. | 7 |
| 2.3 | Workflows for difference computational packages for conformer search. CREST is the method of choice for this study (outlined in orange). Figure reproduced from [44]. | 8 |
| 2.4 | The main effects in phosphine ligands illustrated on a rhodium - biphosphine example. The bite angle is often considered when deriving structure-property relationships. Figure adapted from reference [50]. | 9 |
| 2.5 | Example of how the molecular graph functionality works. The full graph of the complex is split into subgraphs (ligands) by removing the metal center and performing a breadth first search on the independent subgraphs. | 11 |
| 2.6 | Classification of machine learning algorithms [61, 62]. In red are the algorithms applied in this work. | 12 |
| 2.7 | Computational workflows for catalyst design [12, 13, 22, 24, 28, 69]. Top (left to right) Dotson et al., Busch et al., Jover et al. ; Bottom (left to right): Ahneman et al., Gensch et al., Burrows et al. FF - force field, HT - high-throughput, LR/MLR - linear/multi-linear regression. | 14 |
| 3.1 | Local exploration of TM metals chemical space using MACE and ChemSpaX. | 17 |
| 3.2 | Model molecules used for machine learning purposes compared to Sigman et al [12]. | 19 |
| 3.3 | Calculation of the buried volume around the metal center at 3.5 Å accompanied by a steric map for Ligand 186 (see ligand database in Table E.1). The steric map and the buried volume calculations do not include NBD as per Fig 2.5. | 19 |

| | | |
|------|---|----|
| 3.4 | The general definition of the Sterimol parameters around a bond of interest. This illustration is taken from Miller et al [96]. | 20 |
| 3.5 | Compression of the Morgan fingerprint. The common bits are eliminated from the full 1024 digit long fingerprint [99]. | 20 |
| 3.6 | Machine learning pipeline used for reaction predictions. Both classification and regression models were tested in the ML model building phase. | 21 |
| 3.7 | The definition of the confusion matrix and the metrics that can be extracted from it to evaluate a binary classification model. | 22 |
| 4.1 | Representative ligand for each ligand class. The ligands with only one donor atom are taken twice when building the complexes. The top row contains bidentate ligand classes, while the bottom row contains monodentate ligand classes. | 23 |
| 4.2 | Heatmap of the experimental conversions for all substrates across all studied ligands and substrates. Horizontal axis - Substrates; Vertical axis - Ligands. | 25 |
| 4.3 | Deconstructed feature space for the first three principal components. The labels of the three plots indicate the ratio of explained variance for the principal components (PC). The colorbar represents the conversion of SM1. | 26 |
| 4.4 | Deconstructed feature space for the first three principal components. The labels of the three plots indicate the ratio of explained variance for the principal components (PC). The color mapping represent the ligand class. | 26 |
| 4.5 | Maps of donor properties for the ligand dataset colored by their respective chemical class. | 27 |
| 4.6 | Conversion distribution of SM1, SM2 and SM3 in the binding-interaction energy space. | 27 |
| 4.7 | Correlation matrix of the most relevant descriptors. Vertically the single structure properties are shown: bite angle, cone angle, rhodium NBO charge and buried volume at 4 Å. Horizontally, the same properties are shown, but conformer averaged. These descriptors have the abbreviation <i>avg</i> in front. | 28 |
| 4.8 | 3D geometries for Ligand 1 from Fig 4.9; On the right: DFT optimized structure; On the left: DFT optimized best conformer. | 29 |
| 4.9 | Energy difference between DFT optimized lowest energy conformer coming from CREST and DFT optimized single structure. The structures with a difference of more than 50 kJ/mol (≈ 12 kcal/mol) are shown. | 30 |
| 4.10 | Hierarchical classification of the experimental data. The reduced dataset does not contain SM1 and SM7. | 31 |
| 4.11 | Model performance for the same test set for the four applied models. Random Forest: $R^2 = 0.82$; XG Boost: $R^2 = 0.83$; Extra Trees: $R^2 = 0.81$; Gradient Boosting: $R^2 = 0.78$ | 32 |
| 4.12 | Sensitivity analysis of the number of estimators on the model performance for RF and GB models. | 33 |
| 4.13 | Feature importances of the Random Forest Regressor from Fig 4.11. Plot is shown in two parts. The compressed Morgan fingerprint is noted with 0-105. | 34 |
| 5.1 | General approach with possible packages to be used for each stage of this approach. The red highlighting in B indicates that Chem3D is not a high-throughput tool. CREST-ReNeGate connection indicates the tangency between OBeLiX and ReNeGate [44, 49, 91, 111–114]. | 37 |
| A.1 | Selection of model molecule based on the most likely square planar configuration | 48 |

| | | |
|-----|---|----|
| B.1 | Pairplots for the correlations shown in Fig 4.7 | 49 |
| B.2 | Full correlation matrix between conformer averaged properties and single structure properties | 50 |
| C.1 | Possible ferrocene representations according to Guzik et al. [115] | 51 |
| C.2 | Protocol schema for 3D coordinate generation for ferrocenyl metal complexes | 52 |
| D.1 | The difference between the handmade-xTB-DFT and handmade-DFT optimizations. | 53 |
| E.1 | Structures of publicly available substrates. SM4 and SM5 are confidential and will not be made available to the public. | 54 |

List of Tables

| | | |
|-----|---|----|
| 2.1 | Definition of a selection of descriptors highly relevant for homogeneous catalysts. The graphical representations are given below the definitions of the descriptors. . . | 10 |
| 4.1 | Experimental reaction dataset provided by the industrial partner used for training and validation of ML models. For SM7 and SM8 only half of the data is available. The chemical structures of the publicly available substrates can be seen in Fig E.1 . . | 24 |
| 4.2 | Modified descriptors for machine learning purposes. Maximum, minimum and standard deviation measures have been applied to a set of descriptors, that could convey different chemical information (relevant for PP and P ligands, where the difference in descriptors between the two phosphorus donors is minimal compared to the same difference in PN ligands). | 31 |
| 4.3 | Model performance across substrates. Table is split in three parts: model, performance per substrate, overall performance. | 32 |
| E.1 | Ligand database information | 55 |
| E.2 | Descriptor definitions. The package column indicates both the parsing and the calculator packages (e.g. Gaussian as NBO charge calculator; cclib as the parsing package). . . | 62 |

List of Code Listings

| | | |
|-----|---|----|
| 2.1 | Breadth first search algorithm implementation | 11 |
|-----|---|----|

Acronyms

| | |
|---------------|--|
| BV | Buried volume |
| CREST | Conformer-rotamer ensemble sampling tool |
| DFT | Density functional theory |
| DFTB | Density functional based tight-binding |
| ET | Extra Trees |
| FF | Force field |
| GB | Gradient boosting |
| GFN | Geometries, Frequencies, Noncovalent |
| GGA | Generalized gradient approximation |
| GPR | Gaussian process regression |
| HF | Hartree-Fock |
| HOMO | Highest occupied molecular orbital |
| HTE | High-throughput experimentation |
| IM | Intermediate |
| KRR | Kernel ridge regression |
| KS-DFT | Kohn-Sham density functional theory |
| LDA | Local density approximation |
| LUMO | Lowest unoccupied molecular orbital |
| ML | Machine learning |
| ML2 | Metal-Ligand Structure |
| MM | Molecular mechanics |
| NBD | Norbornadiene |
| NBO | Natural bonding orbital |
| OH | Octahedral |

| | |
|---------------|--|
| PBE | Perdew–Burke–Ernzerhof |
| PC | Principal component |
| PDE | Partial differential equation |
| PES | Potential energy surface |
| QSAR | Quantitative structure-activity relationship |
| QM | Quantum chemistry |
| RF | Random forest |
| RMSE | Root-mean-square error |
| SASA | Solvent available surface area |
| SMILES | Simplified molecular-input line-entry system |
| SP | Square planar |
| TM | Transition metal |
| TS | Transition state |
| XC | Exchange correlation |
| XG | Extreme Gradient Boosting |
| ZPE | Zero point energy |

Glossary

| | |
|------------------------------|---|
| Bayesian optimization | Approach that uses the Bayes Theorem [1] to direct the search in order to find the minimum or maximum of an objective function. |
| Cheminformatics | Use of computational and informational methods to understand chemistry [2]. |
| Decision tree | Flowchart-like structure in which individual internal nodes represent a test on a feature |
| Descriptor | Result of a logic or mathematical procedure which encodes multi-variate information about a molecule [3]. |
| Force field | Computational technique utilized for estimating the forces acting between atoms within a molecule, as well as the forces between different molecules. |
| Hyperparameter | Parameter whose value is used to control the learning process. |
| London dispersion | The weakest intermolecular force characterized by the formation of temporary dipoles between adjacent atoms. |
| Model catalyst | Global molecule used to model the behaviour of a real catalytic species for the purpose of simplifying data generation. |
| Molecular fingerprint | Encoded vector representation of a molecule that stores information about its structure [4]. |
| Molecular graph | Representation of the structural formula of a chemical compound in terms of graph theory. |
| Molecular mechanics | Computational method that computes the potential energy surface for atom arrangement using potentials that are derived from classical physics [5]. |
| One-hot encoding | Process of converting categorical data variables so they can be provided to machine learning algorithms to improve predictions. |
| Overfitting | Undesirable behaviour characterized by a model that is unstable to generalize on new data. Overfitted models perform well on training data but not on testing data. |
| Principal component | A linear combination of the variables in a dataset that captures the maximum amount of variation in the data. |

| | |
|-------------------------------------|--|
| Principal component analysis | Statistical method that involves linearly transforming a dataset into a new coordinate system, thereby reducing its dimensionality. In this new coordinate system, the majority of the variation in the data can be explained using fewer dimensions than in the original dataset. |
| Reinforcement learning | Machine learning training method based on rewarding desired behaviors and/or punishing undesired ones. |
| Supercomputer | Computer with a high performance level as compared to a general-purpose computer. |
| Supervised learning | Use of labeled datasets to train algorithms to classify data or predict outcomes accurately. |
| Unsupervised learning | Use of machine learning algorithms to analyze and cluster unlabeled datasets. |
| Virtual orbitals | Unoccupied molecular orbitals. |

1

Introduction

Catalysis is an interdisciplinary technology of high socio-economic importance. More than 95% of all chemical products use a catalyst in at least one step of their synthesis [6]. In basic terms, catalysis represents the acceleration of a chemical reaction through the means of a specialized chemical species that does not participate in the aggregate reaction and acts only as an intermediate. The intermediate processes to a reaction are known as mechanisms and are important in our understanding of catalysts. Catalysis comes in three varieties: homogeneous, heterogeneous and biocatalysis. The former has the catalyst and the reaction components in the same phase. By contrast, the latter implies that the catalyst is in a different phase from the reaction media [7]. Lastly, biocatalysis refers to the metabolic transformation of chemicals to produce new chemicals for industrial purposes [8]. Homogeneous catalysts aid a chemical reaction by reducing the energy barrier towards a transition state (very reactive species) [9]. Many complexes of precious transition metals (Ru, Os, Rh, Ir, Pd, and Pt) have demonstrated the ability to act as effective homogeneous catalysts for a variety of industrial reactions [10, 11]. As a result, the interest of the catalysis community shifted towards precious metals catalysts. Even though expensive, these catalysts provide slow rates of deactivation (i.e. turnover numbers).

The substitution of expensive, high-throughput experimental (HTE) campaigns with *in silico* techniques is a matter of active research, with a series of studies in the past modelling the reaction conversion and/or the enantiomeric ratio [12, 13]. A progress timeline of computation in the field of chemistry and homogeneous catalysis is shown in Fig 1.1. The initial theory for finding energy minima (i.e. the Hartree-Fock methods) of a molecule required a gargantuan number of calculations, which was not feasible at the time, nor is it feasible in the modern era. The theory was gradually optimized, yielding simplified methods, which is the base for the modern, chemically accurate density functional theory (DFT), where the energy is a function of the electron density. By the 1970s, the hardware and software revolution allowed the leap towards *in silico* use of DFT.

Optimizations with DFT are performed through supercomputers, the application of which dates back to the 1980s. The expectation that the outcome of catalytic reactions could be predicted with enough computational resources lead to a tremendous amount of research in this direction [14]. The first relevant computational study in homogeneous catalysis was the modelling of a full catalytic cycle for an alkene hydrogenation with Wilkinson's catalyst by Morokuma, Daniel and Koga [15]. The authors reported more than 200 hours of supercomputing time, having to simplify the structure by substituting phenyl rings with hydrogens, but the result was above expectations since the authors were able to predict a possible reactive intermediate with emphasis on the sensitivity

to the choice of ligand.

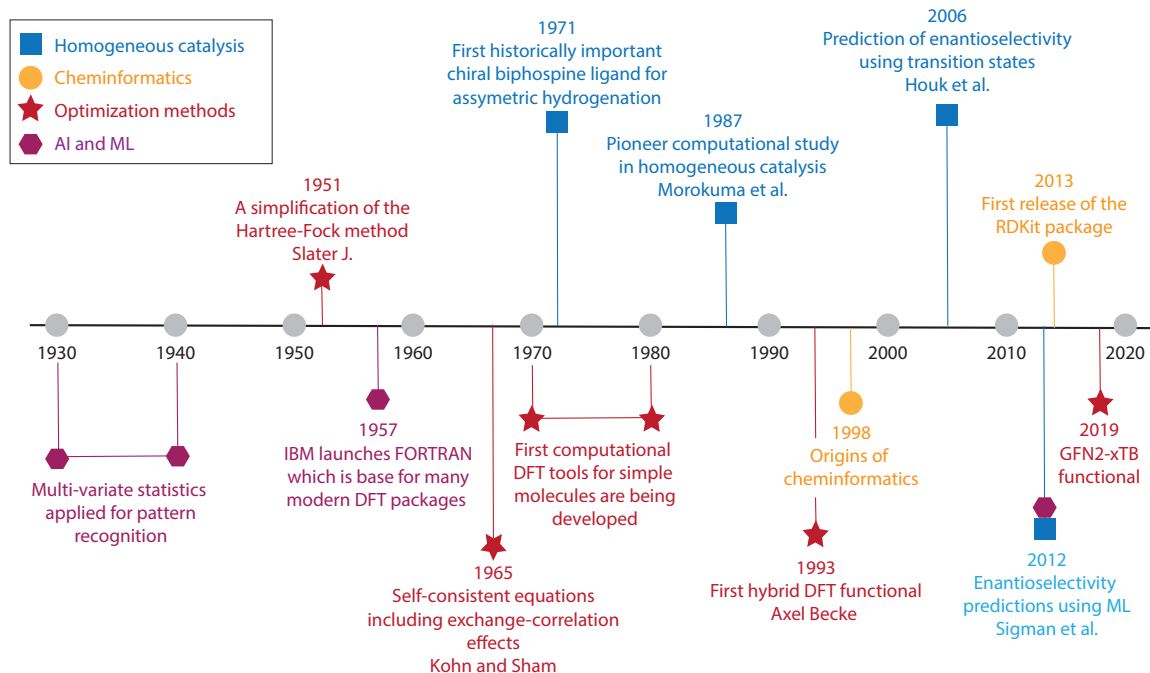


Figure 1.1: History of computation in chemistry with emphasis on the studies in homogeneous catalysis [2, 15–20].

Gradually, with the inception of cheminformatics, the desire to understand mechanisms converged towards a more data-driven approach, which is based on a set of physico-chemical parameters, commonly known as descriptors. Even though both statistical and quantum optimization methods have been created in the past century, complete data-driven studies did not occur in literature until the 2010s, with the study of enantioselectivity by Sigman et al (see Fig 1.1) [21]. Various approaches towards computational modelling of catalyst activity, turnover frequency and enantioselectivity are available in modern chemical literature. These state of the art approaches balance between the chemical accuracy provided by DFT and fully topology based studies [13, 17, 21–25].

In this study, these approaches are integrated into a new framework that allows the featurization of hydrogenation reactions for a number of substrates. Featurization of catalytic reactions through relevant descriptors can carry important mechanistic information. The underlying gap towards converting this concept into tangible predictions is the automation of data generation. Therefore, the scope of the research was to lay the foundations for an automated workflow for structure generation and descriptor calculation, which paired with machine learning can lead to *a priori* predictions of experimental results. This research focused on the catalytic hydrogenation of substrates containing C=C bonds, with rhodium phosphine complexes, with an example mechanism presented in Fig 1.2.

The catalyst featurization is done through descriptors of DFT optimized model molecules (the bonding origin of the model substrate is highlighted in red in Fig 1.2) and the substrates are generalized through a topological fingerprint and general 3D steric parameters around the C=C bond. The features, otherwise known as descriptors, are calculated through the *OBeLiX* computational workflow, designed in this study. Thus, the descriptor database can be reproduced and more importantly enhanced by other scientific groups. *OBeLiX* is a modular tool that allows structure gen-

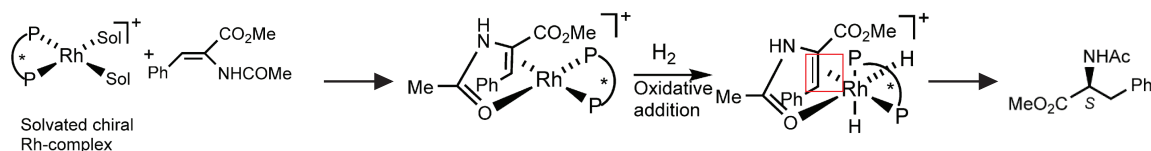


Figure 1.2: Reaction mechanism of a typical substrate used in this study. The first step indicates how the substrate attaches to the solvated Rh-complex. The second step of the mechanism involves the oxidation of the metal center from Rh(I) to Rh(III) through the oxidative addition of a hydrogen molecule. The migratory insertion step is skipped in this figure. The red highlight shows what bond is hydrogenated and the origin for the choice of a model substrate [26].

eration starting from string representations of molecules (i.e. SMILES [27]). If high-throughput exploration of the chemical space is the objective, then the scaffold generation and substituent placement tools can be used. After the generation of the structures, the descriptors are calculated in an automatic manner, where the indices of the donors and metal center are identified without user input, in contrast with other studies in literature where manual mapping is implemented [12, 28].

The modelling workflow designed in this research is summarized in Fig 1.3. Step 1 is achieved with the aid of the aforementioned *OBeLiX* package. Intermediate steps are implemented between the first two steps in form a descriptor analysis, where the ligands are mapped according to their features, and where it is identified whether the structures require conformer-averaged properties for experimental predictions. Conformer averaged properties have been successfully correlated with experimental data by Paton et al [29], Dotson et al [12], Gensch et al [28] and others. A part of these studies are discussed in section 2.6. The data analysis step is then followed by implementation and testing of machine learning models (Step 2 and 3). Step 3 is required to prove the validity of Step 2.

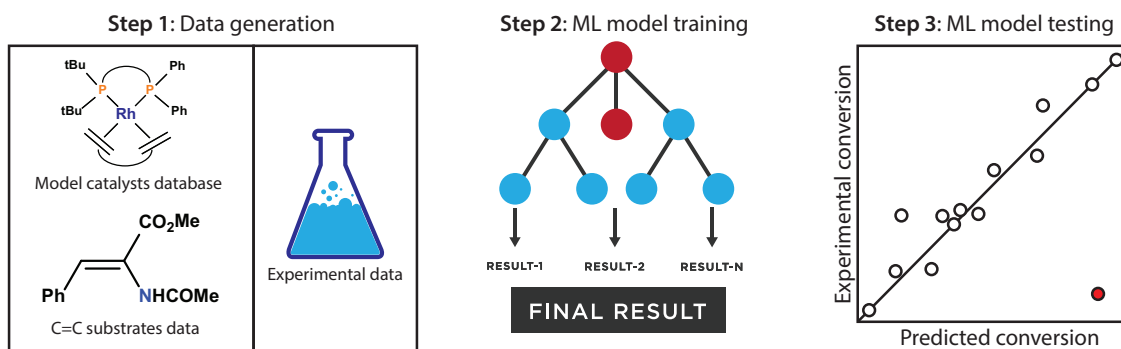


Figure 1.3: Simplified structure of the taken approach. Step 1 includes structure and descriptor generation for both substrates and model catalysts.

The main body of this report starts by introducing the theoretical background and methods necessary to comprehend the complexity of reaction predictions and machine learning models, as well as the tools for building and optimizing chemical structures. Then, the results of the research are presented with emphasis on the studied chemistry, the computational approaches and the correlation between the predicted and experimental conversion. Lastly, the conclusion, the outlook and the perspective of this work are given.

2

Theoretical background

This chapter describes general structure optimization methods such as density functional theory and semi-empirical methods, and discusses various conformer search algorithms with emphasis on Conformer-Rotamer-Ensemble-Sampling-Tool (CREST). The chapter ends with a discussion of machine learning algorithms followed by a literature review of common approaches involving machine learning and property/reaction predictions.

2.1 Density functional theory

Density functional theory (DFT) is a widely used computational method in the field of homogeneous catalysis. It is a method that can be used to predict the properties and behavior of catalytic systems. In the field of homogeneous catalysis, DFT has been used to study a wide range of systems, including enzymes, transition metal complexes, and organometallic compounds [30]. It has been used to study the mechanisms of catalytic reactions, to design new catalysts with improved properties, and to understand the factors that influence the reactivity and selectivity of catalytic systems [15, 24]. The base knowledge to understand DFT is presented step-by-step in the next paragraphs.

The electron is circa 1800 times smaller than the proton. As a result, the protons (being part of the nucleus) are not able to react as quickly to the changes in the surroundings. Thus, these two elementary particles can be separated into two distinct mathematical questions. This approximation is known as the Born-Oppenheimer approximation. If N nuclei are considered, we can express the ground state energy of the electrons as a function of the position of these nuclei $E(R_1, R_2, \dots, R_N)$. The Schrodinger equation represents a linear PDE that governs the wave function of a quantum system. In its time-independent, non-relativistic form, the Schrodinger equation takes the form [31]:

$$H\psi = E\psi \quad (2.1)$$

In Equation 2.1, H represents the Hamiltonian operator and ψ is the set of the eigenstates of the Hamiltonian. Depending on the described system, the Hamiltonian operator can take different shapes. However, the interest in computational chemistry lies mainly on multiple electrons-multiple nuclei systems, where a more intricate description of the Schrodinger equation is neces-

sary [5]:

$$\left[\frac{\hbar^2}{2m} \sum_{i=1}^N \nabla_i^2 + \sum_{i=1}^N V(r_i) + \sum_{i=1}^N \sum_{j<i}^N U(r_i, r_j) \right] \psi = E\psi \quad (2.2)$$

In Equation 2.2, the terms on the left-hand side represent in order: the kinetic energy of each electron, the energy of interaction between individual electrons and the collection of atomic nuclei, and the pairwise (in the formulation above) interaction between different electrons. The Schrodinger equation can be solved analytically only for small molecular systems. For systems with more atoms, approximations are needed. The electron wave function ψ is a function of the position of all electrons, so $\psi = \psi(r_1, r_2, \dots, r_N)$. It is possible to approximate ψ as a product of individual electron wave functions: $\psi = \psi_1(r)\psi_2(r)\dots\psi_N(r)$. With the current computational power, it is impossible to avoid this approximation, since the number of electrons in practical molecules gives impractical systems of equations (*i.e.* a large number of dimensions in the wave function) [5].

Moving from wave functions to electron density represented the entire theoretical ground for the density functional theory. The entire theory rests on two fundamental theorems proved by Hohenberg and Kohn, and a derivation of a set of equations by Kohn and Sham [32]. The first theorem states that the ground state energy of the Schrodinger equation is a unique functional of the electron density and there exists a mapping between the ground-state wave function and the ground state electron density. A functional, denoted by "[]", is similar to a function, but instead takes functions and returns single numbers. Restating the theory of Hohenberg and Kohn [33], the ground state energy functional E can now be defined as $E = E[\rho(r)]$, where $\rho(r)$ is the electron density function, dependent on the electron positions. The second Hohenberg-Kohn theorem states that the electron density that minimizes the energy of the functional is the true electron density that is corresponding to the full solution of the Schrodinger equation. In the Kohn-Sham formalism the ground state energy functional can be defined as:

$$E_{KS}[\rho(r)] = E_T[\rho(r)] + E_V[\rho(r)] + E_J[\rho(r)] + E_{XC}[\rho(r)] \quad (2.3)$$

Here, E_V is the electron interactions with nuclei and other electrons, E_T is the kinetic energy functional and E_{XC} is the exchange-correlation functional. E_J represents the electron repulsion term and is based on the Hartree-Fock assumption that electrons move in a potential created by other electrons and nuclei, thus the ground for a mean-field approximation for the repulsion term.

2.1.1 Exchange-Correlation Functionals

To solve the Kohn-Sham equations, the exchange-correlation functional ($E_{XC}[\rho(r)]$) present in Equation 2.3 has to be specified. The problem is that the true form of the exchange-correlation functional is not known, but its existence is intrinsic to the Hohenberg-Kohn theorems. However, in the case in which the electron density is constant at all points in space, the XC-functional can be derived exactly. This approximation looks like it has limited value, but it is the only way towards using the Kohn-Sham equations. Fig 2.1 shows the Jacob's ladder, with increasingly better approximations of the exchange-correlation functional. Each step of the ladder is introducing a new dependency which increases the accuracy of the calculation [34].

For the purpose of this study the hybrid PBE0 functional was used for geometry optimization. It uses the Hartree-Fock (HF) exchange energy and PBE exchange energy in a ratio of 3:1 respectively,

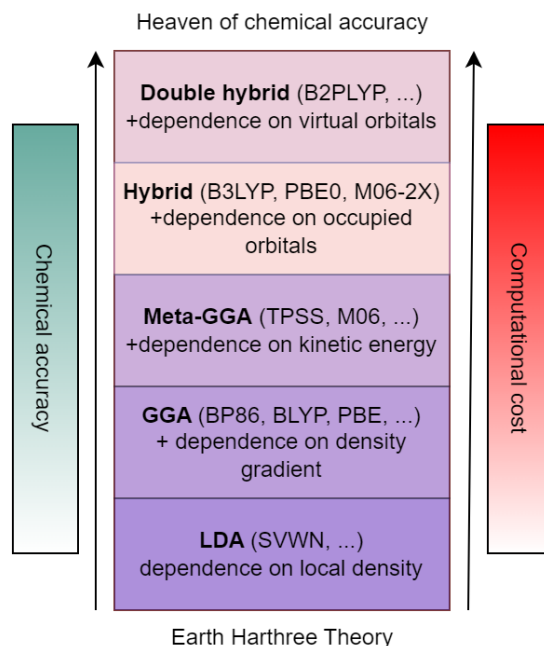


Figure 2.1: Jacob's ladder of exchange-correlation functional approximations which shows the increasing complexity of the different available functionals. PBE0 and B3LYP are by far the most used functionals across literature [35–37].

with a PBE correlation functional.

2.1.2 Basis sets

Basis sets represent a set of functions descriptive of the electronic wave function, that allow for the conversion of the HF differential equations into algebraic equations, which then can be solved with matrix methods, efficiently done by (super)computers [38]. For the research reported in this thesis, the double zeta basis set def2SVP is used. The def2SVP basis set consists of a split valence function with a polarization function on all atoms. By "split valence" it is meant that the valence electrons get a more complete description. As described by the name full name of the basis set def2SVP, the valence electrons are described by two basis sets instead of one for all the rest of the electrons in the atom. The polarization functions are allowing the electrons to get away from each other in order to minimize the electron-electron repulsion. The choice of the basis set and functional are done in a trial-and-error fashion, solely based on empirical evidence. The typical approach in science is to use a specific basis set for a specific type of chemistry, based on what literature reports and on personal experience [34].

2.1.3 Potential energy surface

Potential energy surface (PES) is a central concept in computational chemistry. A PES is the relationship between the energy and the structure of a molecule [39]. The energy minimization converges to the closest local minimum on the potential energy surface. Searching for global minima is still a matter of research. Current methods reduce to machine learning applications or using hundreds of intermediate DFT optimizations, because minima hopping is computationally demanding [40, 41].

Consequently, the starting structures supplied to DFT should be as accurate as possible. To understand the results generated by DFT calculations, one has to look at the vibrational frequencies which at local minima always have positive real values. If there are imaginary frequencies, the calculation should be restarted or changes should be made on the initial geometry.

2.1.4 Dispersion corrections

The Kohn-Sham description of the energy balance is incomplete due to the lack of the term accountable for London dispersion forces. The total corrected energy ($E_{\text{DFT-D3}}$) is given by:

$$E_{\text{DFT-D3}} = E_{\text{KS-DFT}} - E_{\text{disp}} \quad (2.4)$$

Here, $E_{\text{KS-DFT}}$ is the self-consistent Kohn-Sham energy as obtained from the chosen functional (e.g. PBE0) and E_{disp} is the dispersion correction as a sum of two and three body energies [42].

2.2 GFN optimization tools

Accurate and fast calculations for a large molecular system are still a great challenge in theoretical chemistry. While the Kohn-Sham theory can provide accurate representation for one individual large system in gas-phase, it cannot do so for large sets of molecules. The GFN methods described and developed by Bannwarth et al [43] are summarized in Fig 2.2. These methods are semi-empirical and parametrized for all elements with an atomic number lower than 86. Even though the recent progress of parallelized computer architectures has allowed for accurate DFT calculations, there is still need for electronic structure methods that are both simple and accurate, while also being efficient and applicable to large systems without requiring specialized hardware. The central method in Fig 2.2 is GFN-FF, which is also the simplest method from the GFN family. By adding different components to the energy functional of GFN-FF, GFNn-xTB (n = 0, 1, 2) methods were created.

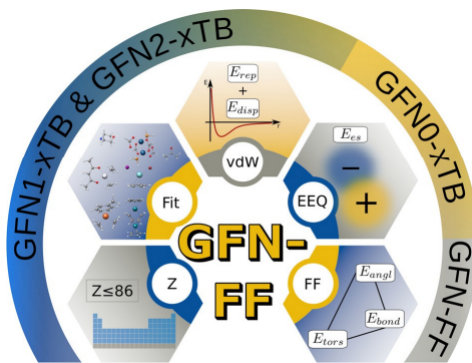


Figure 2.2: GFN family of methods. In the center, the force field method is described and how the it is related to the GFNn-xTB methods. Z, fit - extra empirical parametrization of the elements. vdW - van der Waals interactions. EEQ - electronegativity-equilibrium for the description of pairwise interactions. Figure reproduced from [18].

The total energy expression of the GFN2-xTB method is given by [18]:

$$E_{\text{GFN2-xTB}} = E_{\text{rep}} + E_{\text{disp}} + E_{\text{EHT}} + E_{\text{IES+IXC}} + E_{\text{AES}} + E_{\text{AXC}} + G_{\text{Fermi}} \quad (2.5)$$

The dispersion term E_{disp} was improved from D3 (introduced in section 2.1.4) to D4. D4 is a less empirical version of D3. The term E_{rep} describes the classical repulsion energy which represents a pairwise potential. The $E_{\text{IES+IXC}}$ term represents the isotropic electrostatic and XC-energy. The E_{AES} and E_{AXC} describe the anisotropic interactions and XC energy respectively. The E_{EHT} term is the extended Huckel contribution and represents the crucial contribution to the description of covalent bonds in these methods. The Fermi Gibbs energy terms describes the entropic contribution of the electronic free energy at finite temperature. Currently, the main application of this functional is to generate conformer ensembles in a relatively fast time, which cannot be achieved by DFT.

2.2.1 Conformer search

A conformer is a variation of a structure with the same chemical bonding, but different energy generated by the position of the atoms, bond rotation and repulsion. Multiple studies in literature reveal improvements in the correlation with experimental data (activity and enantioselectivity) when using a set of averaged properties instead of a single structure in a local minima coming from DFT [24, 29].

Many computational chemistry packages are available for searching conformers. According to a benchmarking study on phase-transfer catalysts performed by Trujillo and Iribarren [44], the methods of which are given in Fig 2.3, CREST ranked first in energy accuracy, structural accuracy, tunability and space exploration among the investigated methods. One of the main reasons methods like Balloon[45], RDKit conformer search [46] and Sterimol[29] have a lower performance than CREST lies in the underlying complexity of the applied optimization methods.

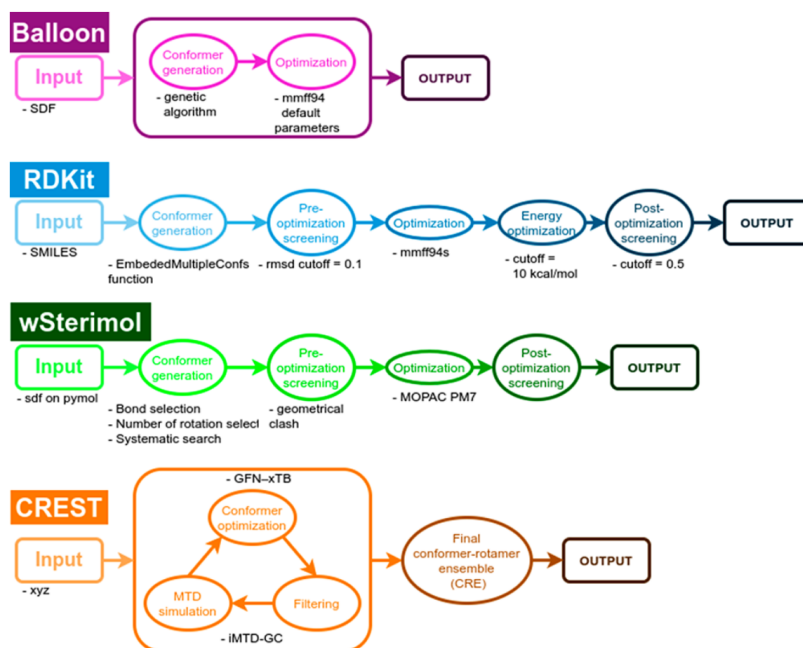


Figure 2.3: Workflows for different computational packages for conformer search. CREST is the method of choice for this study (outlined in orange). Figure reproduced from [44].

RDKit, Balloon and Sterimol are based on force field optimization of the generated conformer ensembles. Ebejer et al [47] identify these approaches as *methods for small organic molecules*. An alternative benchmarking study from Hutchinson et al [48] revealed that for a general molecular

system, the GFN methods rank well in energy computations, with a fair degree of correlation to DFT functionals. The *CREST* computation starts with a xTB optimization of the structure. For the creation of the CRE (conformer/rotamer ensemble), a composite algorithm made primarily of a pseudo-genetic procedure and long meta-dynamic simulations is used. The larger the molecular system, the larger the conformer ensemble, due to the increase in the number of rotatable bonds [49]. In this work, the conformer ensemble search was performed on a subsample of the studied phosphine ligands, the features of which are described in section 2.3 and section 2.4.

2.3 Phosphine ligands in homogeneous catalysis

A great variety of phosphine ligands are currently used in the field of homogeneous catalysis: ranging from monodentate (simple or phosphoramidites) and bidentate ligands to more complex ligands such as PNN or PNS pincers. One common feature of all these ligands is the presence of at least one phosphorus bond to the metal center. In the soft/hard acid-base theory phosphorus is known as a soft, strongly ligating atom for transition metals [50]. Even though most *d*-metals are able to undergo the elementary steps of a catalytic cycle, a number of catalytic reactions have been seen to be dominated by specific transition metals: e.g. Rh in the hydroformylation reaction [51]. Thus, the choice of the transition metal is of utmost importance. Correctly selected phosphine ligands are known to increase the reactivity of the metal center. This is achieved by two effects illustrated in Fig 2.4:

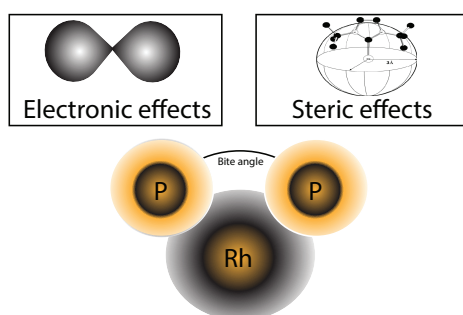


Figure 2.4: The main effects in phosphine ligands illustrated on a rhodium - biphosphine example. The bite angle is often considered when deriving structure-property relationships. Figure adapted from reference [50].

Normally, the selection is done by a combination of pre-existing knowledge with a high-throughput experimental campaign. As Fig 2.4 illustrates, steric and electronic effects affect the stability and activity of the catalysts. These effects can be quantified through descriptors for the further derivation of meaningful relationships (see section 2.4).

2.4 Descriptors

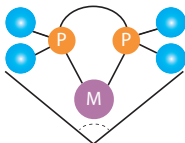
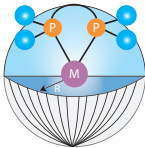
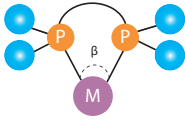
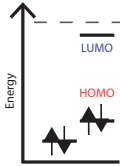
Descriptors are the result of a logic or mathematical procedure which encode multivariate information about a molecule [3]. They come in different variations, such as geometric, steric or electronic descriptors. Generally, geometric descriptors refer to various angles or bond lengths hypothesised to be the source of catalytic activity. On the other hand, steric descriptors are more complex, and they can in some contexts overlap with the geometric descriptors, but normally they are described

separately. The role of steric descriptors is to characterize a molecule's steric hindrance. Example of such descriptors include the buried volume or the SASA (solvent available surface area). The last descriptor class includes the electronic descriptors. For the purpose of this thesis, electronic descriptors will be mainly taken from DFT calculations. Occupancy of lone pairs in the donating atoms or charge at the donor and the metal center represent examples of DFT descriptors.

2.4.1 Morfeus descriptors

Morfeus [52] is a computational Python package that was designed for general use, but has primary applicability for TM complexes containing phosphine ligands. The package contains descriptors of electronic and steric origins. A selection of the most used descriptors in homogeneous catalysis are defined in Table 2.1. The rest of the descriptors are available in Appendix E.

Table 2.1: Definition of a selection of descriptors highly relevant for homogeneous catalysts. The graphical representations are given below the definitions of the descriptors.

| Descriptor | Class | Definition | Ref. |
|--|------------|--|------|
| Bite angle | Steric | The angle formed by the metal center with two donor atoms in a chelating complex. | [53] |
| Cone angle | Steric | The solid angle formed with the metal at the vertex and the outermost edge of the van der Waals spheres of the ligand atoms at the perimeter of the cone | [54] |
| Buried volume | Steric | Percentage of volume that is occupied by atoms in a molecule at a specified distance from an atom of interest. | [55] |
| HOMO-LUMO gap | Electronic | The difference in energy between the highest occupied molecular orbital and lowest unoccupied molecular orbital. | [56] |
| <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;"> <p>Cone angle</p>  </div> <div style="text-align: center;"> <p>Buried volume</p>  </div> <div style="text-align: center;"> <p>Bite angle</p>  </div> <div style="text-align: center;"> <p>HOMO-LUMO gap</p>  </div> </div> | | | |

2.4.2 Descriptors from DFT calculations

Natural bond orbital descriptors

Natural bond orbitals (NBOs) are localized orbitals that describe the molecular bonding pattern or electron pairs. NBOs describe residual resonance delocalization effects (departures from the idealized Lewis-type representation). The general objective of NBO methods in this study is to extract tangible chemical insights from DFT calculations, which are formulated in terms of well understood bonding concepts such as atomic charge, hybridization, bond order or charge transfer [57]. Electron density delocalization between occupied Lewis type NBOs and formally unoccupied non-Lewis NBOs correspond to a stabilizing donor-acceptor interaction.

Energetic descriptors

DFT calculations output several energetic parameters measured in Hartrees. Among them, the absolute energy E is the parameter used in this study [57]. This output is used for the calculation of binding and interaction energies which are described in section 3.1.3.

2.4.3 Subgraph search of transition metal complexes

The descriptors to be calculated with Morfeus require exact identification of the indices of the atoms forming the bite angle coming from the generated atomic coordinates files. For the purpose of identifying these indices, a graph based search was implemented. The first step of the search was to find the N connected atoms to the metal center which represent the starting points of the graph search. The metal is then completely excluded from the graph, which depending on the chosen geometry leaves N independent subgraphs, where the edges represent the bonds, and the nodes are atoms. Using a breadth-first search algorithm (see Code Listing 2.1) all visited nodes are determined and stored in a dictionary [58].

```

1 def bfs(visited, graph, node):
2     # visited -> list
3     # graph -> generated by the in-text mentioned procedure
4     # node - which node to start the search from -> ligand atoms connected to Rhodium
5     queue = [] # Next atom in list
6     visited.append(node) # append visited atoms
7     queue.append(node) # append neighbours of current atom in queue to queue
8     while queue:
9         s = queue.pop(0)
10        for neighbour in graph[s]:
11            if neighbour not in visited:
12                visited.append(neighbour)
13                queue.append(neighbour)
14    return visited

```

Code Listing 2.1: Breadth first search algorithm implementation

After determining all the visited nodes for each individual subgraph, the indices of the donors and metal have to be found. This can be done in two ways: either check if the visited nodes are identical (i.e. the two subgraphs are isomorphic) or if any of the atoms connected to the metal center are mapping each other [59]. The former approach is more consistent due to its unique output comparing to inconsistent results coming from the latter, where the method would fail if two monodentate ligands were used. Fig 2.5 illustrates the procedure introduced in this section.

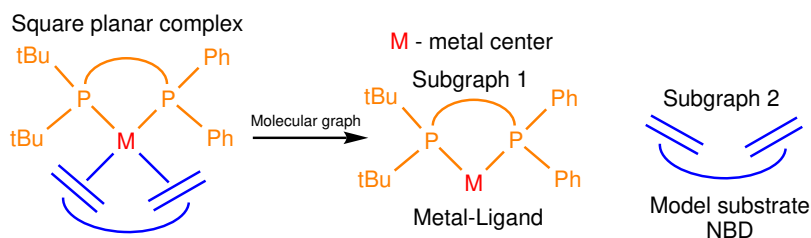


Figure 2.5: Example of how the molecular graph functionality works. The full graph of the complex is split into subgraphs (ligands) by removing the metal center and performing a breadth first search on the independent subgraphs.

The molecular graph functionality of the *OBeLiX* workflow will output the indices of the phosphine ligand(s) and the metal (further this (sub)structure will be denoted by ML2) and the indices forming the bite angle. The implementation of this approach for ligands of higher/lower hapticity is trivial. For monodentate ligands the subgraphs of maximum size are counted and extracted from the full graph, while for structures with more than two donor atoms the subgraph is identified based on the donating atoms that belong to the largest subgraph corresponding to the phosphine ligand.

2.5 Machine learning models

Correlating experimental data with molecular features with a high rate of accuracy is more favorable economically than setting up a full high-throughput experimental campaign. There are plenty of machine learning algorithms ranging from simple linear regression to complex regressors with hundreds of estimators [60]. Three main categories can be identified: unsupervised learning, supervised learning and reinforcement learning, each split into subcategories shown in Fig 2.6.

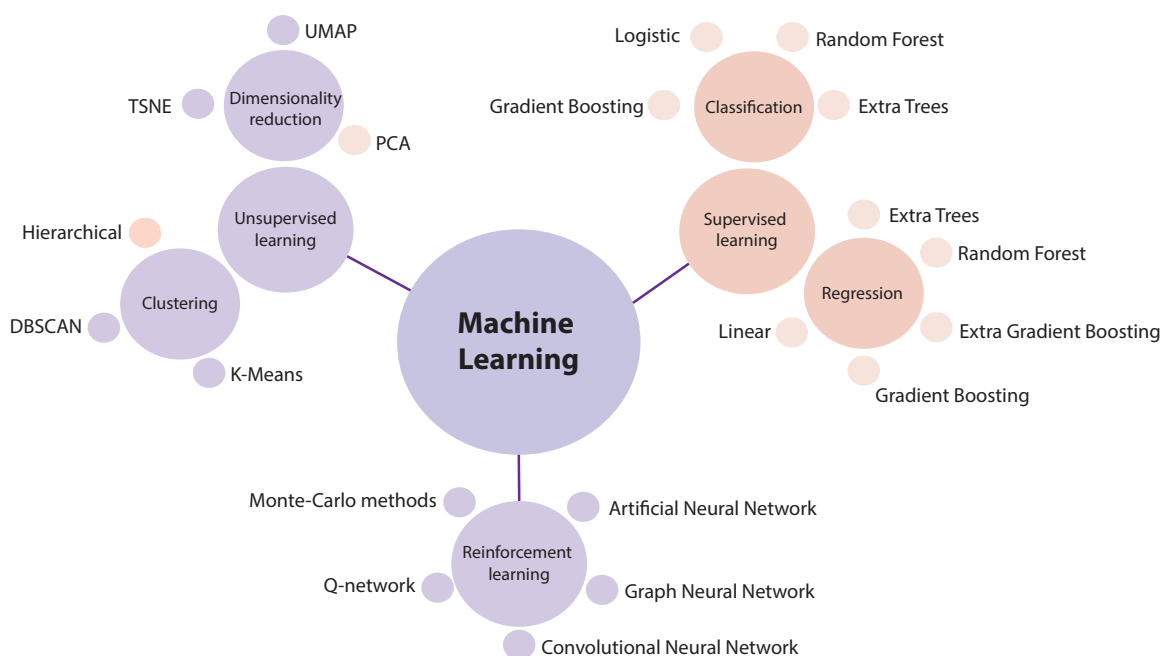


Figure 2.6: Classification of machine learning algorithms [61, 62]. In red are the algorithms applied in this work.

In chemistry, unsupervised learning can be used to analyze large datasets of molecular structures and identify patterns or similarities between them, which can help researchers discover new materials or predict the properties of existing ones. Supervised learning can be used to build predictive models that relate molecular structure to properties, such as solubility or reactivity [63]. Reinforcement learning can be used to design and optimize chemical processes, such as optimizing reaction conditions to maximize yield or selectivity [23]. The choice of learning method depends on the specific problem at hand and the type of data available, whether it is experimental measurements or computer simulations. As Fig 2.6 suggests, in this study several machine learning models have been used including hierarchical clustering, dimensionality reduction with principal component analysis (PCA), and supervised learning approaches in form of classification and regression analysis [64].

In the context of reaction predictions, the tree based regression and classification models can be used to predict the reactivity of a given chemical compound towards a specific reaction or set of reactions. The model takes as input a set of features that describe the structural, electronic, and physico-chemical properties of the molecule, such as its size, steric occupation, electronic properties at the donors, dipole moment and other.

During training, the model uses a subset of the available features and data to construct a decision tree, which splits the data into smaller subsets based on the input features. By using a large number of decision trees and randomly selecting a subset of the features and data for each tree, the tree based regression models can capture the non-linear relationships and interactions between the input features and the reactivity outcome, while avoiding overfitting and improving the generalization of the model. Once the model is trained, it can be used to predict the reactivity of new compounds. The predicted reactivity value can be used to guide the design and optimization of new chemical compounds with desired properties, potentially accelerating the discovery and development of new drugs, materials, and chemicals [46, 65–67].

2.6 Computer-aided catalyst design

Computational tools for catalyst design and reaction optimization have revolutionized the field of computational chemistry. These tools offer numerous advantages, including time and cost effectiveness, reproducible workflows, and minimized errors resulting from human bias or experimental limitations [68]. Fig 2.7 summarizes some state of the art approaches in computational catalyst design, the description of which is given below.

The success of data-driven methods hinges on the existence of vast databases. In a recent study, Gensch et al released Kraken - a publicly available virtual library for designing and refining catalytic processes that involve monodentate organophosphorus(III) ligands. This undertaking involved large-scale data generation, mapping of chemical space, as well as predicting properties and designing catalysts based on experimental data [70, 71]. In their approach, the authors first generated a dataset of monodentate phosphorus ligands comprising 1558 unique compounds, selected based on commercial availability and prevalence in literature. They then performed digital simulations of two versions of each compound: the free ligand and the ligand bound to the metal. Electronic, steric, thermodynamic, and molecular descriptors were used to represent the structures. To account for conformational dependencies, a complete ensemble of ligand conformations was calculated, as suggested in previous works [25, 72, 73]. These conformer are calculated through the Conformer-Rotamer-Ensemble-Sampling Tool (CREST) [49] and DFT geometry optimization is subsequently done on a representative set of the total conformers. Literature has shown that ligand properties on DFT level show similar trends to experimental properties, e.g. atomization energy, bond length, hydricity or activity. Therefore, DFT ligand properties can be used as target values for finding trends in lower levels of theory, such as semi-empirical DFT methods [74–82].

An example of a study that starts from the semi-empirical xTB methods is from Laplaza et al [24], which employed an automated workflow to investigate various reaction pathways in a Rh-catalyzed asymmetric C-H functionalization and predict their enantioselectivity. This approach utilizes the MolAssembler [83] library to create and sample transition states (TSs) at GFN2-xTB level of theory, which are then DFT optimized. In contrast to conventional 3D molecule visualization, MolAssembler models molecules as graphs, making it computationally more efficient to construct molecular graphs from text-based SMILES input rather than relying on (semi-empirical) DFT-based methods. Additionally, MolAssembler can model complexes that contain haptic bond-

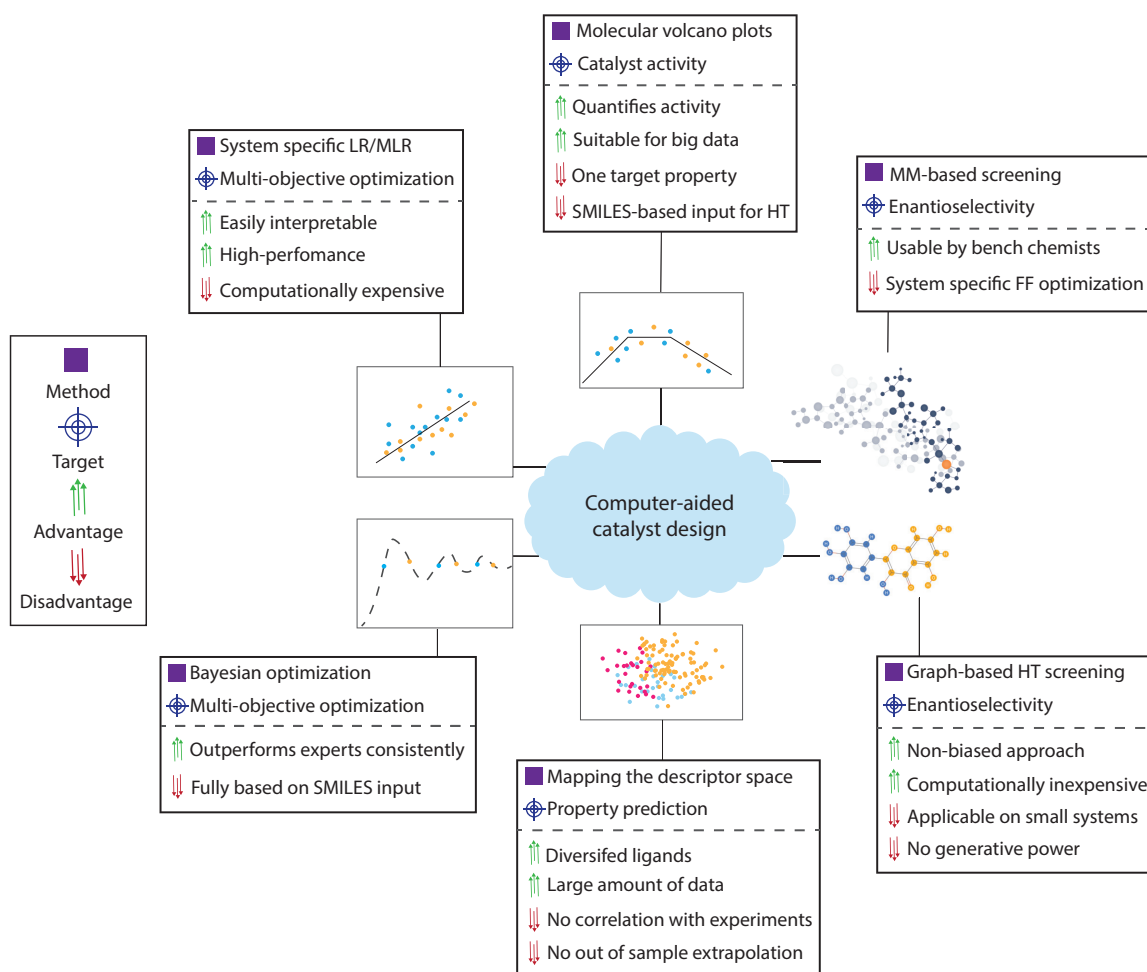


Figure 2.7: Computational workflows for catalyst design [12, 13, 22, 24, 28, 69]. Top (left to right) Dotson et al., Busch et al., Jover et al. ; Bottom (left to right): Ahneman et al., Gensch et al., Burrows et al. FF - force field, HT - high-throughput, LR/MLR - linear/multi-linear regression.

ing sites. These automated, graph-based workflows enable efficient screening of catalytic systems, provided the mechanism and a reasonable estimate of the transition state geometries are known.

In addition to enantioselectivity predictions, a challenging task is to identify catalytic activity. To determine the performance of a catalyst, Busch et al [84] utilized a commonly used technique in heterogeneous catalysis called the volcano plot. This plot is based on Sabatier's principle [85], which asserts that an ideal catalyst should have intermediate binding strength to a substrate. The energies of reaction intermediates that bind to the catalyst are interdependent through scaling relations, i.e. an empirical mathematical relationship exists between the energies of intermediates and transition states of a reaction or class of reactions across various catalysts. This relationship enables the expression of all reaction intermediates and transition states energies in terms of the energy of one or a few specific intermediates, generating linear free energy scaling relationships (LFSEs) based on the descriptor intermediate [86]. By expressing the reaction rate as a function of the energy of the intermediate, a characteristic volcano shape can be created.

Computationally intensive and intricate DFT methods can be replaced with cheminformatics, which is a relatively novel field. Shields et al [23] introduced a new method for optimizing chemical

reactions using Bayesian optimization, which was tested against expert chemists. The method was implemented in an open-source software tool that integrates with existing workflows. To develop the optimizer, data from two cross-coupling reactions were used with a Pd catalyst, and a combinatorial set of reaction conditions was tested. Three types of structure representation were used, including chemical-descriptor fingerprints, cheminformatics descriptors, and binary one-hot-encoded representations (a way to represent categorical data). The quantum chemical properties of reaction components were computed using DFT, and the Mordred package [87] was utilized to generate the one-hot-encoded representations [13, 23]. Benchmarking data-driven methods can be challenging, but this approach shows promise for optimizing chemical reactions.

As Fig 2.7 suggests, the methods described above come with several inherent disadvantages. These disadvantages are related to three key factors: structure representation, domain of applicability and computational cost. Thus, with the available tools, compromises have to be made. In this study, several of the techniques introduced in the original works have been used. Semi-empirical methods are implemented for understanding the effects of conformers on descriptors and general effects on bonding and stability. The mapping of the simplified descriptor space has been used to find the distribution of the experimental data and ligand classes across the feature space. A simplified version of the volcano plot methodology has been applied on the experimental data provided by the industrial partner. Finally, some of the machine learning algorithms used by Ahneman et al. have been applied in the context of reaction prediction in this study.

3

Computational methods

This chapter describes the general computational methods. The first part describes *OBeLiX* (Open Bidentate Ligand eXplorer), and more specifically the submodules of which it is composed. The second part introduces a structured approach for the building of machine learning models for general predictions of the experimental conversion for a set of substrates.

3.1 OBeLiX workflow

The OBeLiX Python package is a modular tool designed for the automated generation and featurization of transition metal complexes. The code of OBeLiX will be available at github.com/EPiCs-group. The core functions of the OBeLiX package (in order) are the following: scaffold generation, automated placement of functional groups, conformer ensemble search and descriptor calculation. DFT optimizations are part of the approach introduced in this thesis, but the DFT tools are to be accessed only through supercomputer interfaces.

3.1.1 Scaffold generation from SMILES representations

The generation of phosphorus ligands starting from SMILES [27] representations is not trivial and requires many manipulations. For this purpose a Python package built with the efforts of Chernyshov and Pidko is used. MACE is a tool designed to generate two types of transition metal chemical structures: octahedral and square planar complexes [88, 89]. The tool requires as input the metal and the SMILES representation of the ligands with precise indication of the donor atom(s) in the individual SMILES of the ligands. MACE can be also used to generate molecular scaffolds on which substituents are placed. MACE comes with the additional functionality of generating isomers for the complexes by changing the place of the ligands on the metal center (see the example of an octahedral complex with two stereoisomers in Fig 3.1). If two isomers are within a specified energy window from each other, they are removed from the output. MACE can regularly fail to provide the correct stereochemistry, thus a large part of the structures were corrected/made with the PerkinElmer software package Chem3D, where the stereochemistry of the chiral centers can be altered. To note, when the phosphine ligand SMILES was not available (ferrocenyl containing ligands), the scaffolds were manually constructed and functionalized with the approach described in the next section. The automated methods described in this and next section refer to the Metal-Ligand(s) structure without the substrate, which has been placed manually.

3.1.2 Automated substituent placement

The functionalization of a scaffold can be performed with ChemSpaX, a tool designed by Kalikadien et al [90]. The approach is combined with MACE to give fully automated generation of structures. An illustration of the combined approach is given in Fig 3.1. The MACE module of the structure generation requires only SMILES representations of the ligands, while for the substituent placement with ChemSpaX, 3D structures for the substituents are necessary. ChemSpaX comes with a database of substituents that was partly enhanced during the preparation of the catalyst database in this study.

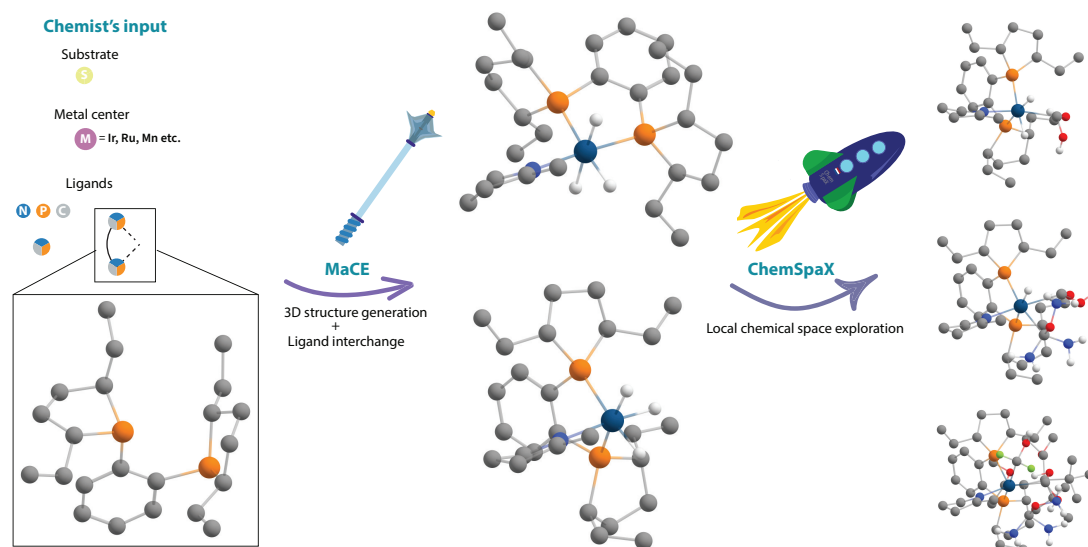


Figure 3.1: Local exploration of TM metals chemical space using MACE and ChemSpaX.

The ChemSpaX functionality was modified to use bromines as dummy atoms for substituent placement, instead of the default hydrogen atoms. Therefore, the indices of the atoms to be functionalized can be automatically identified. The core way in which the skeletons and substituents are paired has also been changed. In *OBeliX*, the user can supply any number of functionalization instructions for any number of skeletons. In the base *ChemSpaX* package, all skeletons are functionalized in the same way from a supplied list of functionalizations [90]. Other approaches for substituent placement are available in literature, such as MolSimplify [91], whose main framework revolves around a DFT-pretrained model that determines the skeleton structure, followed by a selective force field optimization. MolSimplify generates structures based on a neural network that could malfunction in certain scenarios (e.g. when there is high steric bulk or when the geometry of the scaffold is incorrectly identified).

The structures generated by Chem3D, MACE and/or ChemSpaX have been optimized with quantum mechanical methods. The electronic structure calculations methods referring to DFT and DFTB are further described in section 3.1.3.

3.1.3 Applied quantum-mechanical methods

Structure optimization and conformational sampling in the computational package *OBeliX* are performed using the tools from Bannwarth et al described in section 2.2. Both are preferably performed

at the GFN2-xTB level of theory, but the user of the package can choose any GFN level of theory, as illustrated in Fig 2.2. For DFT, Gaussian 16 C.01 [92] is the software of choice in this study. For structures functionalized with ChemSpaX, the OpenBabel [93] package was used to convert between different chemical formats.

Extended density functional tight-binding calculations

The conformer search was performed at the GFN2-xTB level of theory, with no solvation. Every metadynamics conformer search initiated with this tool is preceded by a simple GFN2-xTB optimization [49]. Depending on the size of the conformer ensemble, a number between 3 and 5 conformers were selected for further DFT optimization. These effects will be described in form of a descriptor comparison between conformer averaged and single DFT optimized structure properties.

Density functional theory calculations

All geometries were optimized using the PBE0 functional with a def2-SVP basis set in the gas phase coupled with natural bonding orbital analysis (NBO) [94]. This combination of functional and basis set were proven to give reliable geometry optimizations of TM complexes, even when compared to heavily parametrized functionals [81, 95]. For all norbornandiene complexes (introduced in Fig 3.2), the binding and interaction energies were calculated. The binding energy (ΔE_{bind}) is defined as the energy difference between the full complex and the sum of the DFT optimized Metal-Ligand and norbornandiene (NBD) energies. The interaction energy (ΔE_{int}) is calculated as the difference between the energy of the complex and the sum of the single point DFT energies of the Metal-Ligand and NBD derived from the optimized NBD-complex. The search of the NBD ligand was performed through a subgraph search method which is described in section 2.4.3.

$$\Delta E_{bind, NBD} = E_{DFT, opt, complex} - (E_{DFT, opt, ML2} + E_{DFT, opt, NBD}) \quad (3.1)$$

$$\Delta E_{int, NBD} = E_{DFT, opt, complex} - (E_{DFT, SP, ML2 \text{ from complex}} + E_{DFT, opt, NBD \text{ from complex}}) \quad (3.2)$$

The conformer ensemble generated by CREST (for ML2 structures) is arranged in energetic order, where the relative energy of 0 kJ/mol represents the lowest energy conformer. This specific conformer is always DFT optimized. The energy difference between this conformer and the single structure DFT optimization (denoted by ΔE_{++}) is calculated as below:

$$\Delta E_{++} = E_{DFT, opt, best \text{ conf. ML2}} - E_{DFT, opt, ML2} \quad (3.3)$$

The structures optimized with DFT have been used to generate a descriptor database. The same descriptor methods have been applied to the DFT optimized selection of conformers. The details of the descriptor calculator are presented in the next section.

3.1.4 Descriptor calculation

For reliable machine learning models, the training data has to contain both catalyst and substrate descriptors due to the different affinities of catalysts towards certain substrates [26]. For the catalysts, OBeLiX has a built-in descriptor calculator that computes steric and electronic properties through Morfeus and from DFT outputs. To simplify the data generation step, one has to use descriptors of a molecule that is binding similarly to the metal center, but is also representative for a number of substrates. The simplest model molecule for a homogeneous catalyst is discussed in

Appendix A.1, where just the ligand and the metal center are considered. Fig 3.2 shows the model molecule used by Sigman et al. in comparison with the model molecule used in this study. The main difference between these molecules is in the nature of the bonding between the chosen model ligands. As the bond of interest is C=C, using a similar π complex to that in the mechanism of the reaction is desirable. The steric properties of two chlorine atoms are also different, but the difference is less due to the size of the chlorine atoms compared to the carbon atoms in the norbornadiene.

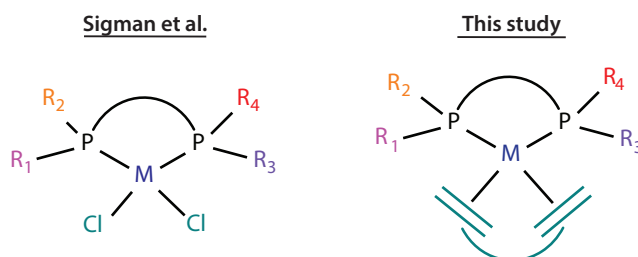


Figure 3.2: Model molecules used for machine learning purposes compared to Sigman et al [12].

For these TM catalysts, properties were calculated for the metal center and the donor atoms (P, N or S). All the descriptors that were further applied for machine learning are shown in Table E.2 and the values of all the descriptors are in the database accompanying this thesis. The donor properties and the quadrant/octant buried volume at the metal center depend on the definition of the two selected donor atoms for each ligand(s). The choice was made based on how strong the donor atoms are. Thus, the properties of the donor with maximum charge is assigned as *max* in the descriptor database, and the properties of the donor with minimum charge as *min*. As the workflow is designed to work at the GFNn-xTB level of theory, the charge is also calculated at this level using Morfeus [52].

Buried volume [55] calculations at the metal center are made by selecting the xz-direction and z-axis which determines the plane of the buried volume calculations. The illustration in Fig 3.3 depicts the taken approach. On the left an example of a steric map can be seen. It shows the steric

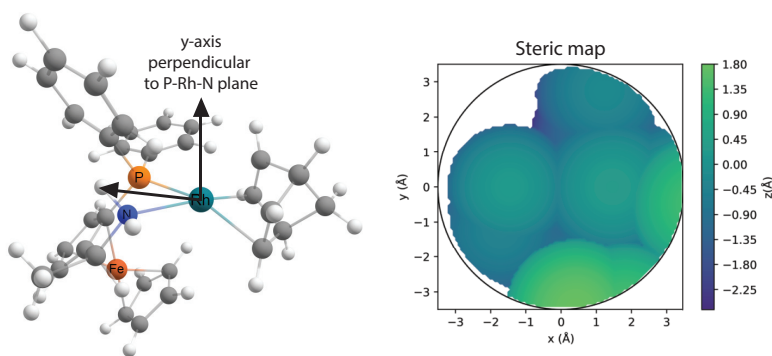


Figure 3.3: Calculation of the buried volume around the metal center at 3.5 Å accompanied by a steric map for Ligand 186 (see ligand database in Table E.1). The steric map and the buried volume calculations do not include NBD as per Fig 2.5.

occupancy of the ML2 structure after NBD has been removed. The left side of the steric map shows less occupancy corresponding to the two hydrogen atoms on the nitrogen donor, while the left side of the map corresponds to the steric occupation of the phenyl rings on the phosphorus donor.

For the substrates, it was opted to use 2D descriptors, and steric 3D descriptors provided by the

Sterimol package through Morfeus. The Sterimol parameters are illustrated in Fig 3.4 [29, 52]. B_1 , B_5 and L parameters sterically quantify the substrate in all its three dimensions, where L is the length of the substrate along the bond of interest, and B_1 and B_5 are the shortest and longest widths from the bond of interest which is the C=C bond for the studied reaction.

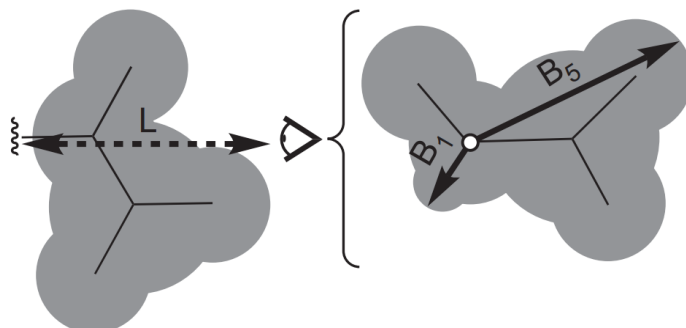


Figure 3.4: The general definition of the Sterimol parameters around a bond of interest. This illustration is taken from Miller et al [96].

Since the substrates are relatively simple, organic molecules, it can be assumed that this description is enough to quantify the steric inter-substrate differences. To account for the structural differences, 2D descriptors contained in the Morgan fingerprint have been used [97]. The calculation is done through the RDKit package [98]. Since there are common bits in the full Morgan fingerprint, these are eliminated. The compression of the Morgan fingerprint is shown in Fig 3.5. Morgan fingerprints are a reliable way to numerically describe the structural differences between different substrates. The bits are then used as features in the machine learning model training.

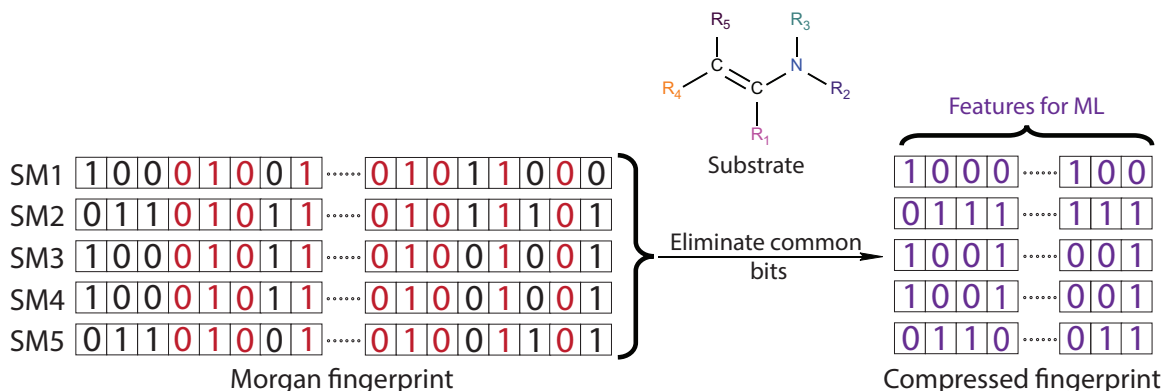


Figure 3.5: Compression of the Morgan fingerprint. The common bits are eliminated from the full 1024 digit long fingerprint [99].

The next section describes what machine learning algorithms and approaches have been applied on the generated database of descriptors for the catalysts and substrates.

3.2 Machine learning pipeline

The machine learning pipeline applied on part of the data generated during this research project is described in Fig 3.6. The stepwise approach starts with the preparation of the full dataset containing

the experimental target value, the substrate Sterimol descriptors and 2D compressed fingerprints, and descriptors of DFT optimized model catalyst structures. The second step involves the choice of the model. The training data has been passed through the *TPOT* package [100, 101], which outputs what machine learning models give the best performance on a randomly selected training set. Four models have been selected: gradient boosting (GB), random forest (RF), extra trees (ET) and extreme gradient boosting (XG). These models are further grid-search cross-validated with the *sklearn* Python library, where a set of hyperparameters are extensively tested aiming the choice of the best model. The last step involves the building of a predictive ML model, starting with the training of the four aforementioned models. The models are then tested on an arbitrary test set, and predicted results are stored. After predictions have been made, the model is assessed through the methods described in section 3.2.2.

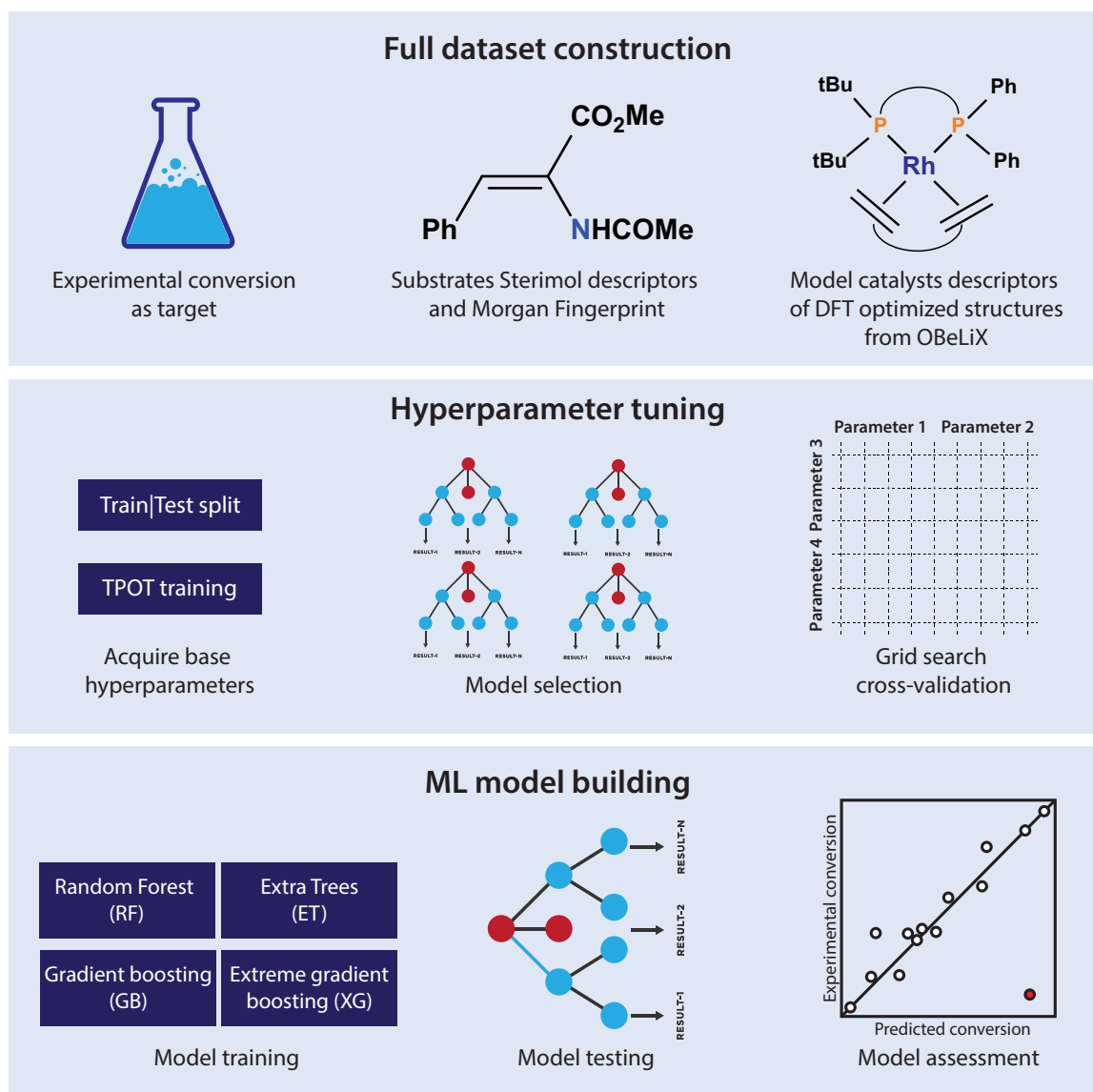


Figure 3.6: Machine learning pipeline used for reaction predictions. Both classification and regression models were tested in the ML model building phase.

3.2.1 Hierarchical clustering

Hierarchical clustering is a type of clustering algorithm used in unsupervised machine learning to group similar objects or data points together based on their characteristics or features. It is a popular technique used in data analysis, data mining, and pattern recognition applications. In hierarchical clustering, the data points are first assigned to individual clusters, and then these clusters are merged into larger clusters in a hierarchical manner, forming a tree-like structure called a dendrogram. The dendrogram represents the hierarchy of clusters, where the leaves of the tree correspond to individual data points and the branches correspond to the clusters formed at each level of the hierarchy [102]. Hierarchical clustering has been used to assert whether it is possible to predict the range of reactivity for substrates outside of the training set, based on chemical similarity.

3.2.2 Evaluation of machine learning models

The regression machine learning models shown in Fig 3.6 have been assessed with the traditional R^2 -score, which is the coefficient of determination of the regression model. The coefficient of determination is not necessarily a squared value, thus it can also be negative (i.e. the model is arbitrarily worse than a straight line). In that case the model has no predictive power. On the other side, the model cannot return an R^2 of more than 1, in which case the predictive values map the test set exactly. The root-mean-square error (RMSE) is calculated as in other similar investigations of catalyst activity [103]. RMSE describes how far on average is the predicted value from the observed experimental value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N ||y_{\text{obs}}(i) - y_{\text{pred}}(i)||^2}{N}} \quad (3.4)$$

For the classification models, a more detailed evaluation is necessary. The most common technique is to use a confusion matrix [104], which describes the general affinity of a binary classifier to make prediction for a certain category. Other measures can be extracted from the confusion matrix, which are given in Fig 3.7. These measures quantify the reliability of the model on an out-of-sample test set. That translates as the ability of the model to predict both of the chosen binary categories when training the model.

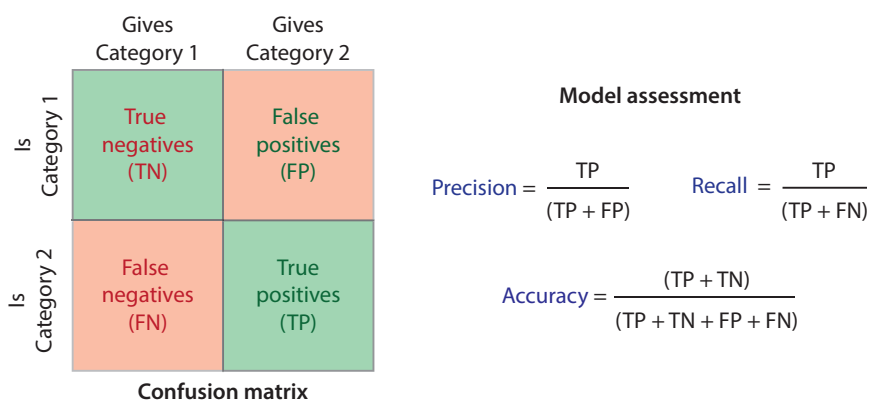


Figure 3.7: The definition of the confusion matrix and the metrics that can be extracted from it to evaluate a binary classification model.

4

Results & Discussion

The first part of the results describe the chemistry of hydrogenation reactions and summarizes the data generation step priorly introduced. The second part delves into the distribution of ligands across the feature space and identification of trends across this space. The third part is answering whether conformer averaged properties are needed for this study. The results are then concluded with the training and testing of machine learning models, built with the data from the descriptor database. These models are to answer the question whether it is possible or not to make predictions on out-of-sample inputs.

4.1 Substrate hydrogenation

This research focuses on hydrogenation reactions (see the mechanism in Fig 1.2), where the goal is to apply a novel approach for prediction of the conversion for this reaction. All used catalysts are chiral and have at least one donating phosphorus atom. Fig 4.1 shows 4 representative ligands from the 4 major ligand classes that have been studied.

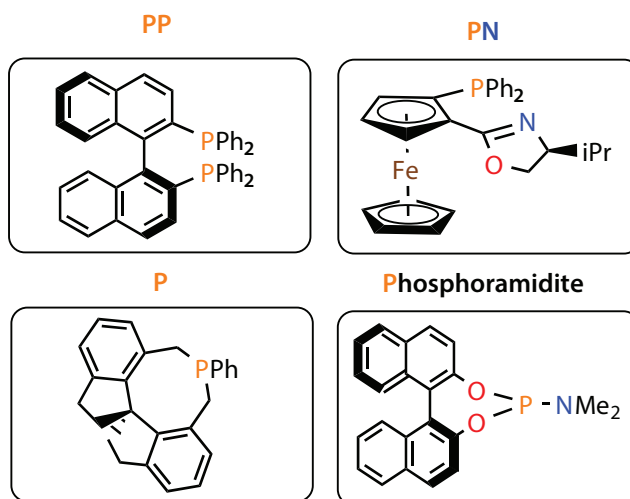


Figure 4.1: Representative ligand for each ligand class. The ligands with only one donor atom are taken twice when building the complexes. The top row contains bidentate ligand classes, while the bottom row contains monodentate ligand classes.

A total of 192 ligands have been investigated and geometries have been generated as part of this study. Besides the four groups in Fig 4.1, two pincers were also included in the dataset (one PNN and one PNS). The choice of the ligands was made by the industrial partner for a high-throughput experimental campaign based on some of the effects described in section 2.3 and commercial availability. The ligand names, formulas and CAS numbers are available in Table E.1. All the xyz geometries and DFT log files will be made available at github.com/EPiCs-group, as well as all CREST log and xyz files, along with the accompanying DFT optimizations of the conformer dataset.

4.1.1 Experimental data analysis

For this study the hydrogenation reaction of seven different substrates has been studied, labeled SM1-SM5 and SM7, SM8 (SM - starting material). The general structure of such a substrate can be seen in Fig 3.5. The experimental data have been provided by the industrial partner. For the machine learning models, the conversion after 16 hours (for all substrates) is the target property.

Table 4.1: Experimental reaction dataset provided by the industrial partner used for training and validation of ML models. For SM7 and SM8 only half of the data is available. The chemical structures of the publicly available substrates can be seen in Fig E.1

| Substrate | Solvent | Temperature | Pressure | Samples |
|--------------|----------|-------------|----------|-------------|
| SM1 | Methanol | 298 K | 5 bar | 192 |
| SM2 | Methanol | 298 K | 5 bar | 192 |
| SM3 | Methanol | 298 K | 5 bar | 192 |
| SM4 | Methanol | 323 K | 5 bar | 192 |
| SM5 | Methanol | 323 K | 20 bar | 192 |
| SM7 | Methanol | 323 K | 20 bar | 96 |
| SM8 | Methanol | 323 K | 20 bar | 96 |
| Total | | | | 1152 |

The distribution of the experimental data is a very important factor for training tree-based models. The more balanced the datasets, the better the algorithm comprehends the complex, non-linear relationships between the features and the experimental target. Fig 4.2 shows the conversion distribution across all substrates and all ligands in a detailed heat map, where the substrates can be found on the *x-axis* and the ligands on the *y-axis*. It can be seen that the conversion distribution is fairly uneven with very high performances for SM1 and SM2; average performances for SM3, SM4, SM7, and SM8, and overall poor catalyst performance for SM5. For SM5 in particular, only a small number of catalysts had a conversion rate higher than 0.5. It can be *a priori* predicted that the tree-based regressors will place all catalysts in the lower range of activity for SM5 since the model was trained on a few active catalysts for this specific reaction. Thus, the model cannot precisely learn what features determine the performance of SM5.

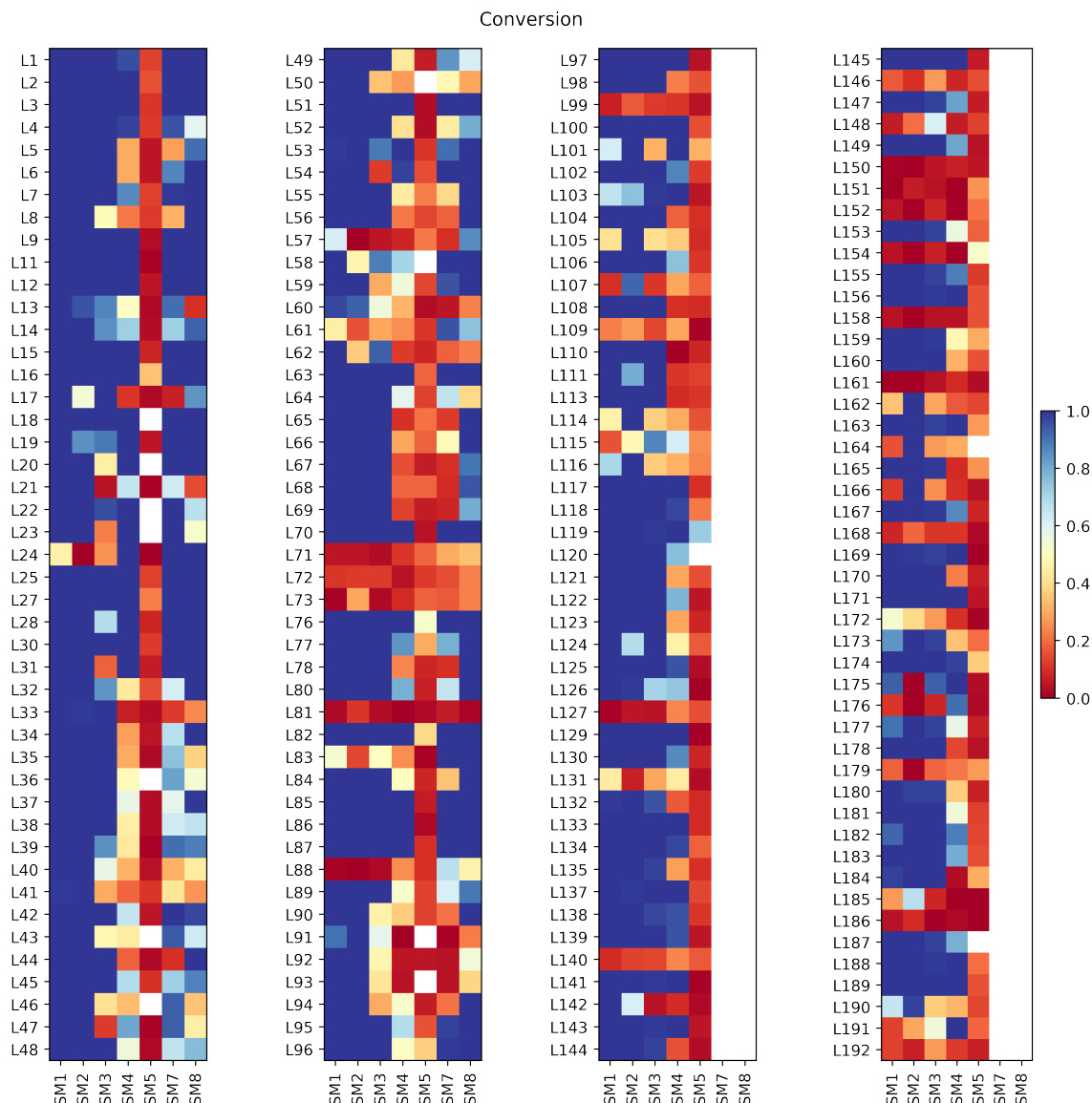


Figure 4.2: Heatmap of the experimental conversions for all substrates across all studied ligands and substrates. Horizontal axis - Substrates; Vertical axis - Ligands.

4.2 Ligand mapping

Ligand mapping is a concept that relates to the distribution of the model catalysts across the feature space. In literature [12, 28, 105], the most commonly used technique is the principal component analysis (PCA), where the first two or three components that maximize the amount of variance are considered, eliminating data collinearity. The general aim is to identify trends or cluster similar ligands across the feature space. For the purpose of this study, the general goal of a PCA map was enhanced in order to identify correlations with the experimental data. Fey et al [105] performed a similar analysis to identify trends in ligand binding energy. The decomposed maps of the three principal components are given in Fig 4.3, where the color represents the conversion for SM1. Sim-

ilar trends can be observed for SM2 and SM3, indicating their chemical similarity.

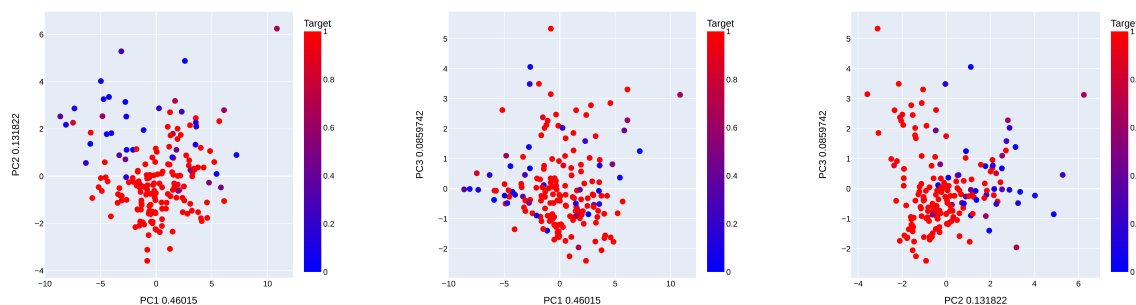


Figure 4.3: Deconstructed feature space for the first three principal components. The labels of the three plots indicate the ratio of explained variance for the principal components (PC). The colorbar represents the conversion of SM1.

It can be seen that the first component has the greatest percentage of explained variance ratio of 0.46, while PC2 and PC3 have lower weights of circa 0.13 and 0.08. The lowest conversions cluster around low values of PC1 and high values of PC2 (the map on the left in Fig 4.3). The same type of analysis has been performed again, but instead of coloring the feature space based on catalyst performance, the color now indicates the respective ligand class. This analysis is shown in Fig 4.4.

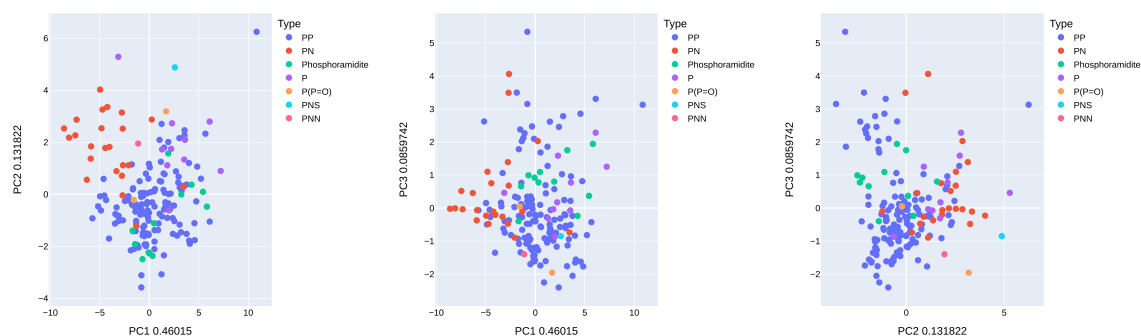


Figure 4.4: Deconstructed feature space for the first three principal components. The labels of the three plots indicate the ratio of explained variance for the principal components (PC). The color mapping represent the ligand class.

It can be concluded that the PN ligands give the lowest performance for the studied reactions. The PNS chelating ligand also gives low performance and one reason might be the strong binding to the metal center of nitrogen which is shown in Fig 4.6. Overall, the ligand class distribution seems to follow a clustering trend, where the main principal component is formed from the NBO charge of the Rh metal center. Nitrogen is an element with a higher electron affinity than phosphorus, thus the charge is more unevenly distributed in PN ligands than in the PP and P ligands. To further confirm this, the donor NBO charges, the donor bond distances and the donor buried volumes should be investigated. The maximum and minimum values for each measure were used as the axes in Fig 4.5.

Fig 4.5 confirms the statement about the PN ligands being electronically and sterically different from the ligands with two donor phosphorus atoms. The most obvious difference can be observed

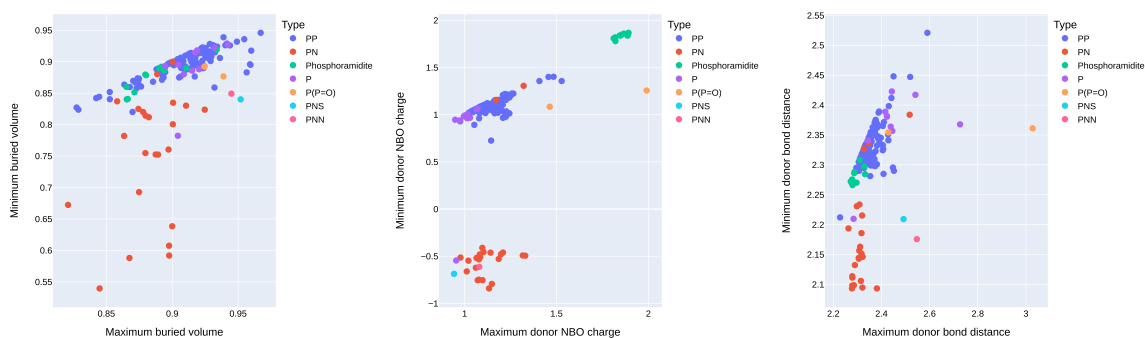


Figure 4.5: Maps of donor properties for the ligand dataset colored by their respective chemical class.

across the NBO charge of the donor and the bond distance, where the PN ligands form a separate cluster. The NBO charge of the N donor is highly negative comparing to the phosphorus donor in the same ligands. The phosphoramidites have the highest positive charge of all the ligands, because of the three highly electronegative atoms connected to the phosphorus (O, O, N). The PP ligands in the range of +1.5 charge for both donors have the motif of a phosphorus-containing ring being present. The next section presents a energy analysis associated with the analysis of the descriptors introduced in this section.

4.2.1 Energy analysis

The energy analysis has been performed to understand the degree of the binding and interaction energy of norbornadiene (NBD) and how the activity of the catalyst coming from experimental data is distributed across this energy space. According to previous studies (can be seen in Fig 2.7), it has been asserted that homogeneous catalysts follow the Sabatier principle in a similar way that heterogeneous catalysts do.

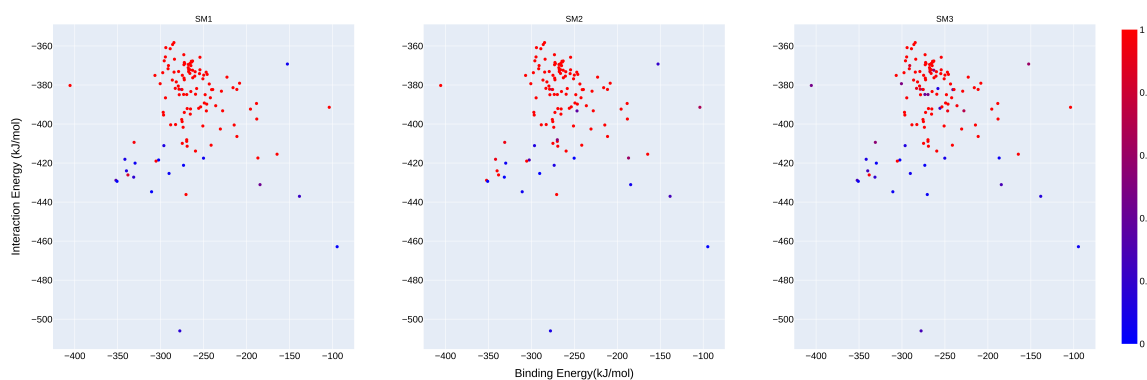


Figure 4.6: Conversion distribution of SM1, SM2 and SM3 in the binding-interaction energy space.

This means, that the active catalysts should have low interaction energies and intermediate binding energy. Performing this analysis for SM1 to SM3, it was observed that the inactive catalysts tend to appear in the same region for all three substrates, which suggests binding similarity of NBD

with the studied substrates. This region corresponds to a high interaction energy and low/high binding energies. Below $\Delta E_{\text{int,NBD}}$ of -420 kJ/mol, the catalysts seem to have a barrier towards binding to the substrate. Even though low-performing biphosphine catalysts are underrepresented for these three substrates, the low and high binding energies seem to be a point of inflection in catalyst performance. The magnitudes of the obtained binding energies were cross-validated with literature results [105, 106].

4.3 Conformer effects on descriptors

The conformer analysis has been performed to understand whether there is significant effect on the descriptors. As it can be seen from a part of the full correlation matrix in Fig 4.7, the descriptors for single structures correlate to a high extent with the conformer averaged descriptors and the R^2 spans the 0.7-0.8 range for the same descriptor correlations. Additionally, it can be seen that both the conformer averaged and single properties of buried volume at 4 Å correlate to a very high degree of 0.9 to the cone angle. This relationship is to be expected, since more space occupied around the metal center, means also an increase in the cone angle. The full correlation matrix is given in Appendix B. From Fig B.2 it can be discerned that only the dipole moment coming from DFT calculations is not correlating with its conformer averaged counterpart. These results suggest that the conformer ensemble is likely not required in this particular study. However, as mentioned in section 2.6, there have been reports of improved correlation with the experimental data when using conformer averaged properties.

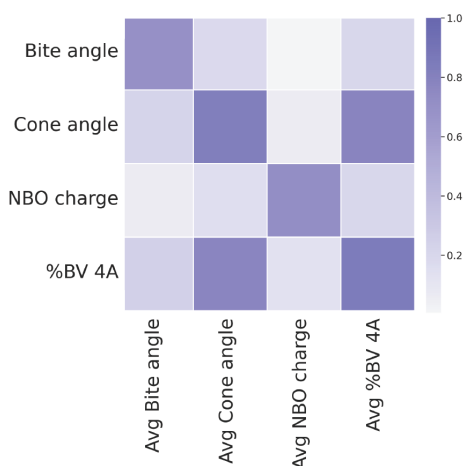


Figure 4.7: Correlation matrix of the most relevant descriptors. Vertically the single structure properties are shown: bite angle, cone angle, rhodium NBO charge and buried volume at 4 Å. Horizontally, the same properties are shown, but conformer averaged. These descriptors have the abbreviation *Avg* in front.

The energy change from DFT optimized single structure to the DFT optimized lowest energy conformer has been calculated. Fig 4.9 shows these energy differences. In the majority of cases the DFT optimized conformer structure and DFT optimized single structure converge to the same local energy minima. However, in several cases significant distances can be noticed, where the MD simulation with CREST changed the bonding of the metal-ligand structure.

The majority of ligands shown in Fig 4.9 show a common pattern and more specifically a penta-P-C-O ring attached to a phenyl ring (ligands 1, 3, 4, 5, 6). The DFT optimization on the best DFT

conformer converted the initially monodentate ligands into polydentate ligands. From Fig A it can be seen that these types of complexes, where a square planar configuration is formed from two bidentate ligands, are highly unlikely due to the excessive steric hindrance. The dihedral angle between the four donors (P, N, P, N) deviates by 23.5° from a planar configuration for ligand 1, which explains the largest energy difference across the dataset. The differences in coordination can be seen in Fig 4.8:

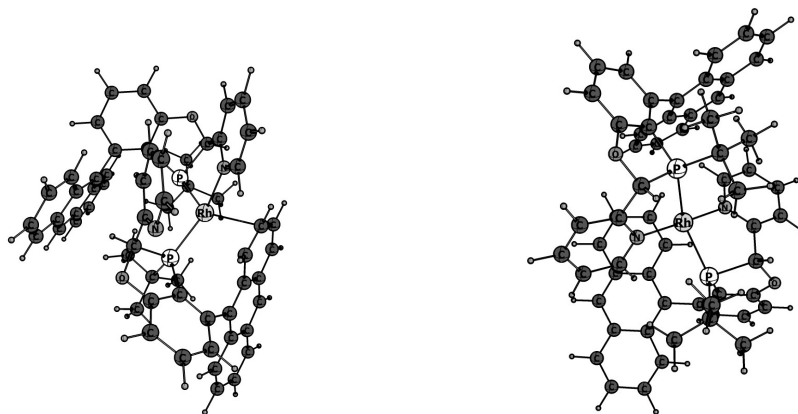


Figure 4.8: 3D geometries for Ligand 1 from Fig 4.9; On the right: DFT optimized structure; On the left: DFT optimized best conformer.

Similarly, for ligand 5 the deviation from the SP configuration is 27.3° . For ligands 3 and 7 a SP configuration was formed with one of the oxygens. Oxygen coordination to transition metals is relatively weak. Thus, the energy is lower than the ionic form of the ML2 structure. Ligand 4 opened up to a 180° angle between the donor phosphorus. Even though in terms of CREST energy this structure is the most stable, the more accurate PBE0 functional determined the opposite. Ligand 2, which is a PNS pincer creates a stabilizing configuration when DFT optimized, where an agostic interaction is formed with the hydrogen from one of the methyl groups on the tBu radicals, whereas the CREST-DFT optimized structure does not have this stabilizing bond. For ligand 6, one of the two P-monodentate ligands formed a tridentate configuration with the nitrogen and one of the oxygens connected to the sulphur. Ligand 8 forms stabilizing bonds through an agostic interaction and formation of a π -complexation with one of the phenyls. Even though the metal forms the correct coordination, the dihedral angles formed by the donor atoms is 11.2° and 35.2° respectively for each of the carbons in the C=C bond. These eight molecules clearly underline the issues with the CREST algorithm, where a minimal valence basis set with a polarization function on elements with the atomic number $Z > 9$ is used [18, 49]. This can partially explain the tendency of under-coordinated complexes to form π -bonds with adjacent electron clouds, even though they are energetically not favorable from a DFT standpoint. An interesting case of how GFN2-xTB pre-optimization affected the following DFT optimization is given in Appendix D.

The energy landscape for the studied ligands is quite bleak considering that extra computational resources have been used to perform these calculations. Overall, only 8 structures have proven to improve after the metadynamics simulations. This can be correlated with the type of structure that has been used, which is unstable due to the undercoordinated metal center. However, this type of optimization and analysis can lead to the discovery of transition states, which is very desirable in homogeneous catalysis.

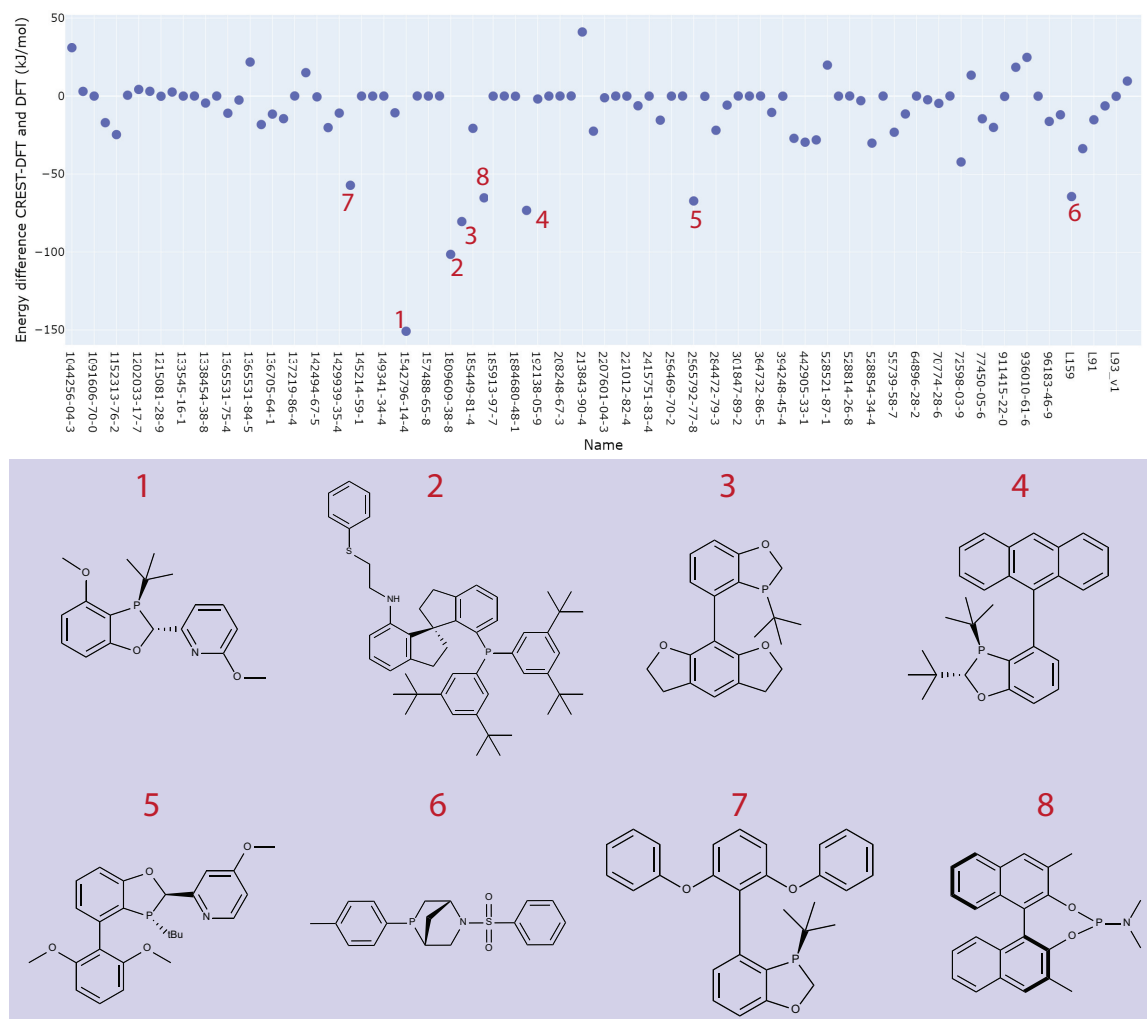


Figure 4.9: Energy difference between DFT optimized lowest energy conformer coming from CREST and DFT optimized single structure. The structures with a difference of more than 50 kJ/mol (≈ 12 kcal/mol) are shown.

4.4 Reaction predictions

The reaction prediction was performed using machine learning algorithms and the multitude of descriptors that were calculated. The method entailed using chemically informed descriptors for the model catalyst structures, and 2D and 3D descriptors for the substrate.

4.4.1 Feature selection

The features selection for the training of the available machine learning models was based on the capability of the features to convey the same type of information across ligand classes. Thus, the donor related properties have been modified as follows. A maximum, minimum function was applied on pairs of donor descriptors, as shown in Table 4.2 and maximum, minimum and standard deviation has been applied on the quadrant and octant buried volumes.

Table 4.2: Modified descriptors for machine learning purposes. Maximum, minimum and standard deviation measures have been applied to a set of descriptors, that could convey different chemical information (relevant for PP and P ligands, where the difference in descriptors between the two phosphorus donors is minimal compared to the same difference in PN ligands).

| Descriptors | Applied functions |
|-------------------------|--------------------------------------|
| Quadrant buried volumes | maximum, minimum, standard deviation |
| Octant buried volumes | maximum, minimum, standard deviation |
| Donors buried volumes | maximum, minimum |
| Donors NBO charges | maximum, minimum |
| Lone pair occupancies | maximum, minimum |

4.4.2 Hierarchical clustering

The hierarchical clustering was performed to show what ligands have similar chemical properties. The clustering was performed on Sterimol and the compressed Morgan fingerprint descriptors. The dendrogram in Fig 4.10 shows that SM1 and SM3 have similar structural motifs, as well as SM7 and SM8. Thus, a model where the similar substrates (according to this approach) are removed from training set has also been investigated. SM2 is similar to SM1 and SM3, but above a specific threshold, meaning that it is advisable to keep two out of these three substrates for predictions.

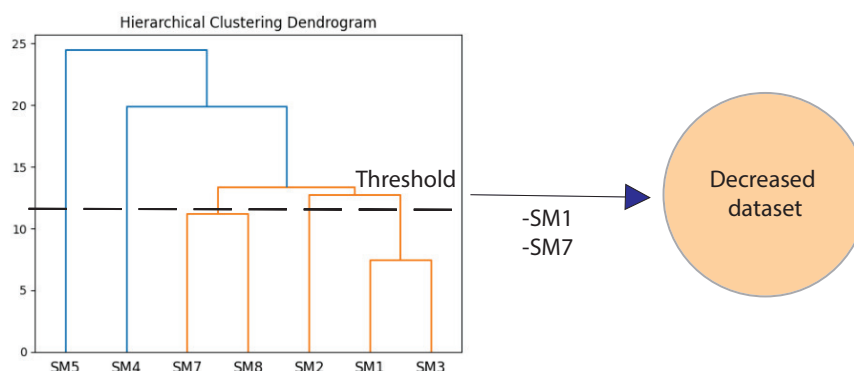


Figure 4.10: Hierarchical classification of the experimental data. The reduced dataset does not contain SM1 and SM7.

4.4.3 Full dataset regression models

The size of the test data is important when training the model. As recommended in literature [107], a 70-30 or 80-20 split should be normally applied to any type of data, independent of its origin, to avoid overfitting the data to a specific, smaller test set. Thus, a 75-25 training-test split has been chosen for the data at hand.

The regression models include the data presented in Table 4.1. The models, suggested at different runs of the TPOT-Regressor, had their hyperparameters optimized with a grid-cross search validation. The accuracy of all the used machine learning models is shown in Table 4.3 and Fig 4.11. The best performing model is the Extreme Gradient Boosting model with an R^2 of 0.83, while the worst performing model is the Gradient Boosting model. The performance of the model varies substrate

by substrate, with the SM5 having a very poor performance ranging between 0.01 and 0.03. As previously mentioned this was expected due to how skewed the data for the SM5 substrates is towards low catalyst activity. However, if training the same model in the absence of SM5 data (even with the auto-ML tools) gives a much worse performance ranging between 0.6 and 0.7. This means that despite SM5 performance being so low, it has a role in describing the features that contribute to the lower performance of a general substrate. The performance of all the models per substrate is given in Table 4.3.

Table 4.3: Model performance across substrates. Table is split in three parts: model, performance per substrate, overall performance.

| Model | SM1 | SM2 | SM3 | SM4 | SM5 | SM7 | SM8 | Train | Test | RMSE |
|-------------------|------|------|------|------|------|------|------|-------|------|-------|
| Random Forest | 0.81 | 0.67 | 0.69 | 0.69 | 0.03 | 0.66 | 0.64 | 0.93 | 0.82 | 0.180 |
| XG Boost | 0.79 | 0.69 | 0.72 | 0.74 | 0.03 | 0.69 | 0.85 | 0.93 | 0.83 | 0.177 |
| Extra Trees | 0.76 | 0.72 | 0.72 | 0.69 | 0.01 | 0.71 | 0.79 | 0.93 | 0.81 | 0.19 |
| Gradient Boosting | 0.58 | 0.63 | 0.67 | 0.74 | 0.03 | 0.56 | 0.71 | 0.83 | 0.78 | 0.21 |

The visualization of the information in Table 4.3 is given in Fig 4.11, where four machine learning models are shown. The distribution of the test data is skewed towards low and high performance. All four model show similar performance for an individual reaction. The root-mean-squared error for the four models is ranging between 0.177 and 0.21. This deviation can be partially explained by the degree of the experimental error which is around 30%, according to an experimental reproducibility study performed by the industrial partner and partially by the internal error of tree-based models like the ones shown in Table 4.3 [107].

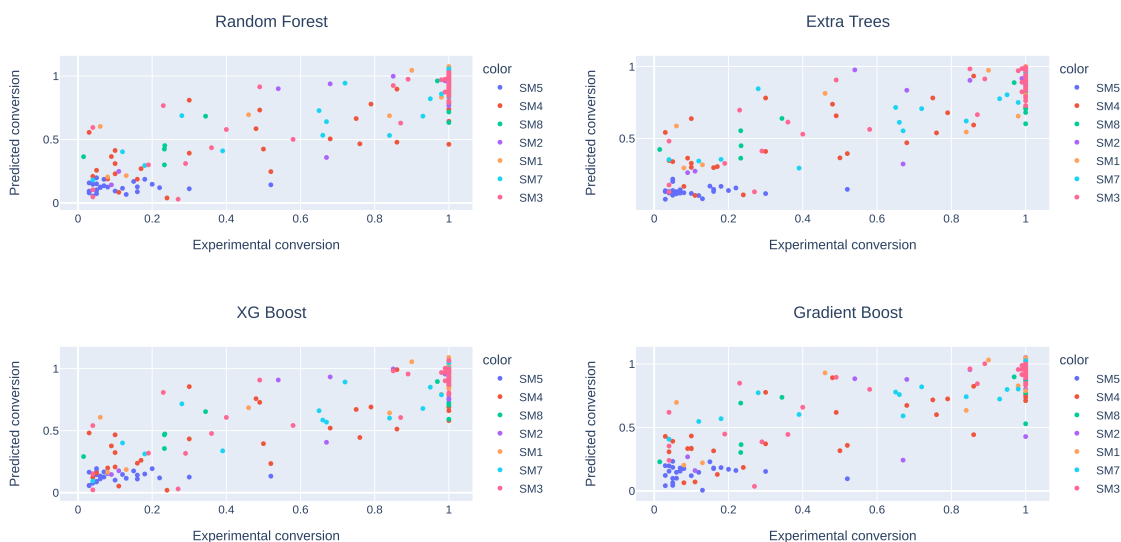


Figure 4.11: Model performance for the same test set for the four applied models. Random Forest: $R^2 = 0.82$; XG Boost: $R^2 = 0.83$; Extra Trees: $R^2 = 0.81$; Gradient Boosting: $R^2 = 0.78$.

Sensitivity study of the number of estimators on the model performance

Hyperparameters are vital when training a machine-learning model, because they directly dictate

the behaviour and performance of the chosen model, especially for tree-based regressor like Random Forest and Gradient Boosting regressors. In the following analysis, the effect of the number of estimators for each model has been assessed. This analysis is relevant in the context of computational time. The higher the number of estimators, the higher the computation time, since there is additional trees that need to be evaluated. TPOT suggested a number of 100 estimators for all the studied models, but this parameter can be optimized. This analysis is necessary for medium to large size datasets, because the general time complexity of tree based classifiers is of order $O(d * n * \log(n))$, where d is the number of features and n is the number of samples [107].

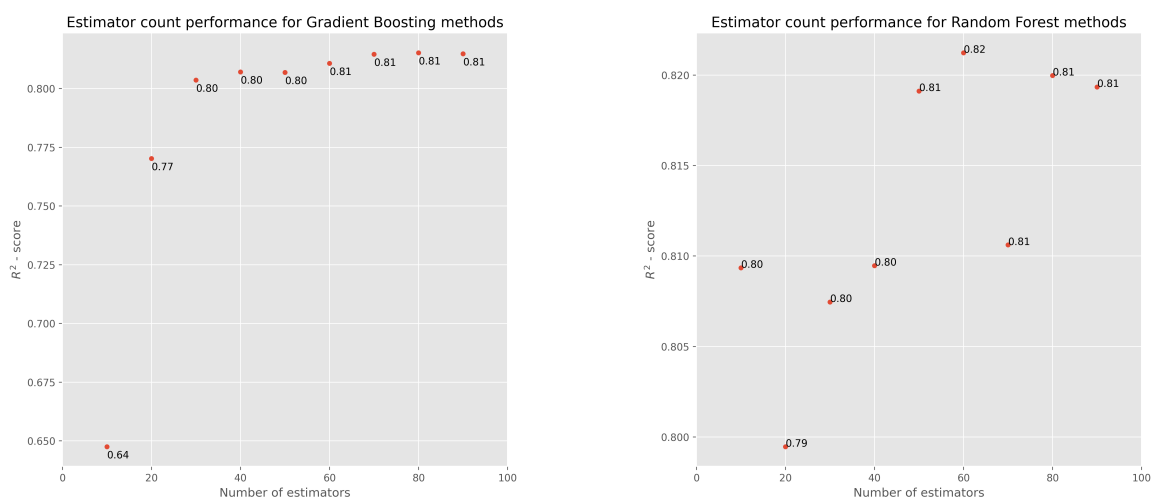


Figure 4.12: Sensitivity analysis of the number of estimators on the model performance for RF and GB models.

Fig 4.12 shows the reason why Random Forest is the model of choice if results were to be extrapolated on a new dataset. The steep rise in performance when moving from 10 to 20 estimators characterizes the overfitting nature of gradient boosting models, whereas random forest is not susceptible to overfitting describing the random nature of the RF algorithm.

4.4.4 General classification models

For the general binary classification of reactive and non-reactive ligands a transformation approach has been used, where the previously discussed regressors have been converted to classifiers, to assess if the general range of catalytic activity is maintained. This approach starts by converting the continuous predicted values from the regressors into discrete binary values. The choice of the threshold is assessed by the data distribution in the test set and set at a 0.7 conversion rate. The results of the classification model yielded a 0.92, 0.92 and 0.95 accuracy, recall and precision scores.

4.4.5 Hierarchical clustering classification models

The random forest model has been chosen as the best model in this study. Thus, the regression on the hierarchical clustering approach presented in section 4.4.2 was performed using this model and then transformed into a classifier. Making binary predictions for SM1 and SM7 as per Fig 4.10 yielded high value of 0.8 accuracy for SM1, since the chosen threshold was set at 0.75 conversion. For SM7 an accuracy score of 0.75 was obtained. The recall and precision scores for SM1 were 0.99

catalysts, as it can be seen in Fig 4.6.

4.4.6 Feature importance

performance.

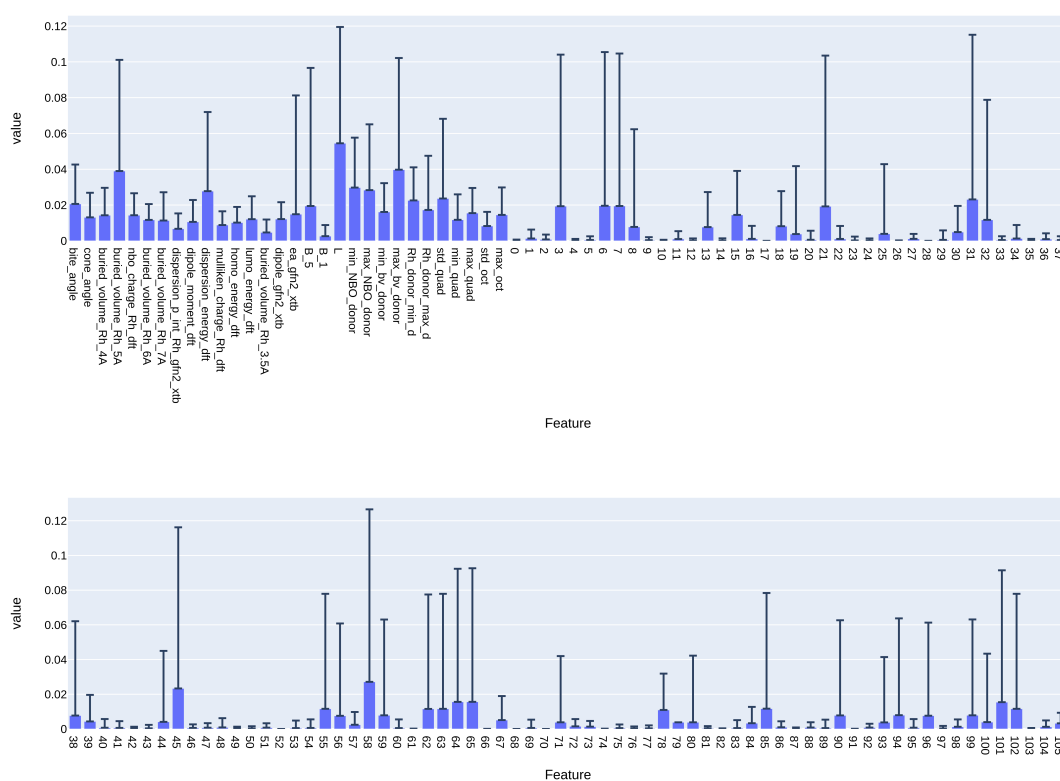


Figure 4.13: Feature importances of the Random Forest Regressor from Fig 4.11. Plot is shown in two parts. The compressed Morgan fingerprint is noted with 0-105.

suitability with the catalytic species in the first step of the mechanism shown in Fig 1.2.

5

Conclusion & Outlook

5.1 Conclusion

In this research project, an automated computational workflow (*OBeLiX*) for generation of 3D structures for transition metal complexes and descriptor calculation has been introduced. In *OBeLiX*, the structure generation starts from the simplest string representation for ligands, assembling TM complexes. In this work, the *OBeLiX* workflow has been applied to investigate a hydrogenation reaction from a data science perspective. Experimental data for 192 ligands have been offered by the industrial partner. To reduce the number of needed computations, a model molecule has been proposed, where the binding substrates are substituted by a model substrate, norbornadiene. Using *OBeLiX*, an additional database of descriptors has been generated for these model catalysts. The dataset was then enhanced with 3D steric and 2D topological descriptors for all the substrates. In this way, a large dataset of 1152 featurized reactions has been created.

Before applying any machine learning algorithm on the gathered descriptors, an analysis of the data has been performed. Among the 192 ligands, the PN ligands stood out as chemically different from the ligands with two donor phosphorus atoms. This is largely due to different steric and electronic properties of nitrogen comparing to phosphorus. The experimental data provided by the industrial partner showed that PN ligands are overall the worst performing. The reason for this might be that PN ligands have high interaction energies with the substrate, as demonstrated in this study.

Moreover, it was discovered that there is a high correlation between conformer averaged and single structure descriptors. The conformer ensemble found with CREST, has shown an increase in the local energy minima for some metal-ligand(s) structures. This proved that CREST and semi-empirical methods have a relatively high probability to distort molecules. Thus, conformer averaged properties of a 192 ligand dataset would be highly error prone. However, the semi-empirical level of theory could be used in high-throughput computational campaigns, where the rate of error is less significant than in the small scale study performed in this work.

The final step was to train a machine learning model with predictive capabilities. Two approaches have been tested. *In the first approach*, all the substrates have been considered in both training and testing sets. A regression coefficient as high as 0.83 has been obtained on the test set, and between 0.93 and 0.96 on the training set. The regression model was converted into a classification model to assess whether the general range of reactivity is correctly predicted for a featurized reaction. The

classification model yielded an accuracy score of 0.92.

In the second approach, the test set was composed of substrates that were not present in the training dataset, following an hierarchical clustering, which determined the substrates that are chemically similar. As a result, a classifier model has been trained to make predictions on substrates outside of the test set. This method yielded accuracy scores in the 0.75-0.8 range. The approach is thus promising, because less experimental data is necessary to make these predictions, even though the overall accuracy is slightly lower than using the full dataset.

Performing a feature importance analysis on the full dataset, it was concluded that both catalyst and substrate descriptors have high importance for making *in silico* predictions. This study has proven the power of using chemical and topological descriptors in combination for reaction predictions and confirmed the potential of reaction featurization. As a final remark, the application of the OBeLiX workflow has shown the importance of automation and data science in the world of automated catalyst design.

5.2 Outlook

The perspective of automated catalyst design is bright, the humanity being just at the inception of this field. There is no clear and direct path towards achieving autonomous catalyst discovery, as there are numerous decisions to be made regarding design and implementation. While progress in machine learning is advancing rapidly, it can take some time for the latest developments to be adopted by other fields [108]. The most important developments to be made are in the adoption of deep reinforcement learning approaches, which were earlier introduced in Fig 2.6. Several developments and enhancements of the research presented in this thesis can be made, and are introduced in the next paragraphs.

5.2.1 New descriptors

The advancement of machine learning can be used to the advantage of *in silico* discovery in homogeneous catalysis. The modelling approach introduced in this study can be enhanced by increasing the number of descriptors in the realm of deep-learning. For instance, a number of graph-based descriptors have been made available, that can also be applied in the context of the predictions made during this thesis [109]. These topology-based descriptors can be very powerful in describing certain molecular interactions pertaining to certain values of target properties such as conversion or enantioselectivity. The reason why molecular design is so cumbersome for chiral catalysts is the conformational complexity of these structures. Descriptors can be used to describe this aspect. For instance, the number of rotatable bonds, bond distances at active sites and other important parameters such as local buried volumes. Conformer averaged descriptors also have their place in the study of activity and selectivity. The full descriptor matrix can only be compiled when all structures can be equally generated by the computer, rather than the chemist. Some of these difficulties and possible solutions are presented in the next section.

5.2.2 Generation of unusual complexes

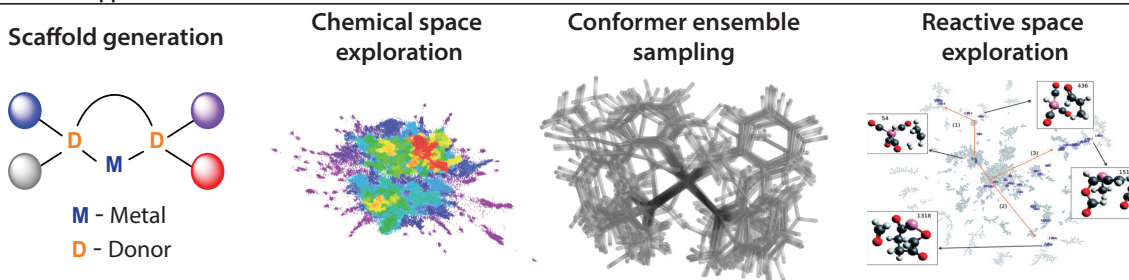
A problem encountered during the preparation of the catalyst database, is the generation of π -complexes. In this research, the ferrocene containing scaffolds were made using expert intuition rather than automation. However, during the later stages of the project it was discovered that there

are possible work-arounds. One such solution is presented in Appendix C, where a combination of semi-empirical methods are used, to generate a ferrocene based Metal-Ligand structure. The structure can then be further refined with DFT based methods. Once the generation of all structures is achieved, one can start wondering about other aspects of fully automated explorations, such as providing automated mechanistic insights.

5.2.3 Reaction newtork explorers

The power of artificial intelligence can be also be applied in the context of mechanistic exploration, which can explain phenomena that a black-box machine learning model could never be able to explain [108]. Several tools have been made available earlier by Blau et al [110], Hashemi et al [111] and Maeda et al [112], as well as a set of commercial packages. As a conceptual example, the ReNeGate graph based approach will be used to demonstrate further applications of the OBeLiX package. The full approach is summarized in Fig 5.1. ReNeGate can generate new features for the studied ligands. For instance, a general catalyst deactivation coefficient could be applied as a filtering parameter to the overall experimental predictions. In other words, ReNeGate can indirectly describe the reactivity of a given transition state through reaction network exploration. Since ReNeGate uses the conformers coming from the CREST MD simulations, the integration with OBeLiX is straightforward.

A. General approach



B. Alternative tools for each step



C. In-house approach using in-house designed packages

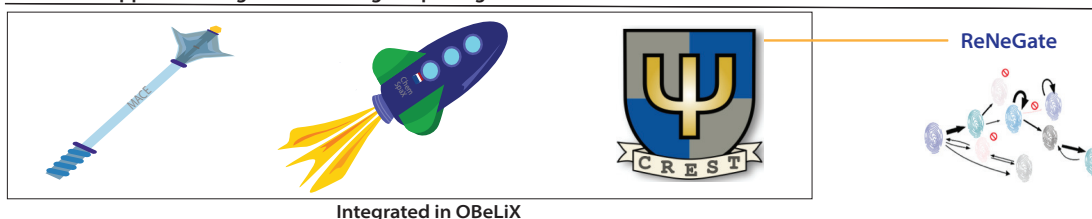


Figure 5.1: General approach with possible packages to be used for each stage of this approach. The red highlighting in B indicates that Chem3D is not a high-throughput tool. CREST-ReNeGate connection indicates the tangency between OBeLiX and ReNeGate [44, 49, 91, 111–114].

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to everyone who has supported me throughout my master's thesis journey. First, I would like to thank my supervisor **Evgeny Pidko**. Thanks for your valuable lessons on how to do proper science and for the little chemistry lessons you would give every time there was an opportunity. I learned a lot from you and thanks for all the support throughout these 8 months.

Special thanks go to my daily supervisor, **Adarsh Kalikadien**, for his guidance throughout this journey. Thank you for teaching me how to write good and structured code and for all the debates we had and for fixing my git every time I had major issues with it. I wish you the very best of luck with your PhD, and I hope we can have collaborations in the future.

I would love to extend my appreciation to the ISE staff for the quality time spent together and to all other postdocs, PhD and master students I had the opportunity to exchange ideas with. Moreover, I would love to thank our collaborators from Janssen: Laurent Lefort, Robert van Putten and Cecile Valsechi for the highest quality of scientific debates, it was a lot of fun meeting with you every week and exchanging ideas.

I would love to thank all the people in Delft that made it enjoyable and fun, to my colleagues and friends who shared my struggles. Furthermore, I would love to thank my family. Even in these difficult times for our little country, I got all the unconditional help and support from them which I really appreciate. I hope to see you all again soon!

Bibliography

- [1] James Joyce. *Bayes Theorem*. Ed. by Edward N. Zalta. Spring 2019. Metaphysics Research Lab, Stanford University, 2019.
- [2] Wendy L. Williams, Lingyu Zeng, Tobias Gensch, Matthew S. Sigman, Abigail G. Doyle, and Eric V. Anslyn. "The Evolution of Data-Driven Modeling in Organic Chemistry". In: *ACS Central Science* 7 (10 Oct. 2021), pp. 1622–1637. ISSN: 23747951. DOI: [10.1021/acscentsci.1c00535](https://doi.org/10.1021/acscentsci.1c00535).
- [3] Mati Karelson, Victor S Lobanov, and Alan R Katritzky. "Quantum-Chemical Descriptors in QSAR/QSPR Studies". In: *Chemical Reviews* (1996), pp. 1027–1043.
- [4] H. L. Morgan. "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service." In: *Journal of Chemical Documentation* 5.2 (1965), pp. 107–113. DOI: [10.1021/c160017a018](https://doi.org/10.1021/c160017a018).
- [5] Herbert Goldstein, Charles Poole, and John Safko. *Classical Mechanics*. 3rd ed. Addison-Wesley, 2000.
- [6] GECATS - German Catalysis Society. *The Digitalization of Catalysis-Related Sciences*. Mar. 2019.
- [7] Taylor Hugh S. *Encyclopedia Britannica: Catalysis*. Date Accessed: [15/08/2022]. July 2018.
- [8] Elizabeth L. Bell, William Finnigan, Scott P. France, Anthony P. Green, Martin A. Hayes, Lorna J. Hepworth, Sarah L. Lovelock, Haruka Niikura, Silvia Osuna, Elvira Romero, Katherine S. Ryan, Nicholas J. Turner, and Sabine L. Flitsch. "Biocatalysis". In: *Nature Reviews Methods Primers* 1 (1 Dec. 2021). ISSN: 26628449. DOI: [10.1038/s43586-021-00044-z](https://doi.org/10.1038/s43586-021-00044-z).
- [9] Racha Abed Ali Abdine, Gaspard Hedouin, Françoise Colobert, and Joanna Wencel-Delord. "Metal-Catalyzed Asymmetric Hydrogenation of C=N Bonds". In: *ACS Catalysis* 11 (1 Jan. 2021), pp. 215–247. ISSN: 21555435. DOI: [10.1021/acscatal.0c03353](https://doi.org/10.1021/acscatal.0c03353).
- [10] J A Davies' and F R Hartley. "Complexes of the Platinum Metals Containing Weak Donor Ligands". In: *Chem. Rev* 81 (1981), pp. 79–90.
- [11] Brian R. James. "Hydrogenation Reactions Catalyzed by Transition Metal Complexes". In: *Advances in Organometallic Chemistry* 17 (1979). Ed. by F.G.A. Stone and Robert West, pp. 319–405. ISSN: 0065-3055. DOI: [doi.org/10.1016/S0065-3055\(08\)60327-5](https://doi.org/10.1016/S0065-3055(08)60327-5).
- [12] Jordan J. Dotson, Lucy van Dijk, Jacob C. Timmerman, Samantha Grosslight, Richard C. Walroth, Francis Gosselin, Kurt Püntener, Kyle A. Mack, and Matthew S. Sigman. "Data-Driven Multi-Objective Optimization Tactics for Catalytic Asymmetric Reactions Using Bisphosphine Ligands". In: *Journal of the American Chemical Society* 145 (1 Jan. 2022), pp. 110–121. ISSN: 0002-7863. DOI: [10.1021/JACS.2C08513](https://doi.org/10.1021/JACS.2C08513).

- [13] Derek T. Ahneman, Jesús G. Estrada, Shishi Lin, Spencer D. Dreher, and Abigail G. Doyle. "Predicting reaction performance in C–N cross-coupling using machine learning". In: *Science* 360 (6385 Apr. 2018), pp. 186–190. ISSN: 0036-8075. DOI: [10.1126/science.aar5169](https://doi.org/10.1126/science.aar5169).
- [14] Klavs F Jensen and Donald G Truhlar. *Supercomputer Research in Chemistry and Chemical Engineering An Introduction*. 1984.
- [15] N. Koga, J. Han C. Daniel, X. Y. Fu, and K. Morokuma. "Potential energy profile of a full catalytic cycle of olefin hydrogenation by the Wilkinson catalyst". In: *Am. Chem. Soc.* (1987), pp. 3455–3456. DOI: doi.org/10.1021/ja00245a044.
- [16] Axel D. Becke. "A new mixing of Hartree-Fock and local density-functional theories". In: *The Journal of Chemical Physics* 98 (2 1993), pp. 1372–1377. ISSN: 00219606. DOI: [10.1063/1.464304](https://doi.org/10.1063/1.464304).
- [17] Daniel H. Ess, Steven Wheeler, Robert G. Iafe, Lai Xu, Nihan Celebi-Olcum, and Kendall N. Houk. "Bifurcations on Potential Energy Surfaces of Organic Reactions". In: *Angew Chem* (2008).
- [18] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. "GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multiple Electrostatics and Density-Dependent Dispersion Contributions". In: *Journal of Chemical Theory and Computation* 15.3 (2019). PMID: 30741547, pp. 1652–1671.
- [19] Jolene P. Reid and Matthew S. Sigman. "Holistic prediction of enantioselectivity in asymmetric catalysis". In: *Nature* 571 (7765 July 2019), pp. 343–348. ISSN: 14764687. DOI: [10.1038/s41586-019-1384-z](https://doi.org/10.1038/s41586-019-1384-z).
- [20] Robin Haunschild, Andreas Barth, and Bernie French. "A comprehensive analysis of the history of DFT based on the bibliometric method RPYS". In: *Journal of Cheminformatics* 11 (1 Nov. 2019). ISSN: 17582946. DOI: [10.1186/s13321-019-0395-y](https://doi.org/10.1186/s13321-019-0395-y).
- [21] Kaid C. Harper, Elizabeth N. Bess, and Matthew S. Sigman. "Multidimensional steric parameters in the analysis of asymmetric catalytic reactions". In: *Nature Chemistry* (2012).
- [22] Lauren C. Burrows and Luke T. Jesikiewicz and Gang Lu and Steven J. Geib and Peng Liu and Kay M. Brummond. "Computationally Guided Catalyst Design in the Type I Dynamic Kinetic Asymmetric Pauson-Khand Reaction of Allenyl Acetates". In: *Journal of the American Chemical Society* 139 (42 Oct. 2017), pp. 15022–15032. ISSN: 15205126.
- [23] Benjamin J Shields and Jason Stevens and Jun Li and Marvin Parasram and Farhan Damani and Jesus I Martinez Alvarado and Jacob M Janey and Ryan P Adams and Abigail G Doyle. "Bayesian reaction optimization as a tool for chemical synthesis". In: *Nature* 590 (7844 2021), pp. 89–96. ISSN: 1476-4687. DOI: [10.1038/s41586-021-03213-y](https://doi.org/10.1038/s41586-021-03213-y).
- [24] Rubén Laplaza, Jan Grimo Sobez, Matthew D. Wodrich, Markus Reiher, and Clémence Corminboeuf. "The (not so) simple prediction of enantioselectivity - a pipeline for high-fidelity computations". In: *Chemical Science* 13 (23 May 2022), pp. 6858–6864. ISSN: 20416539. DOI: [10.1039/d2sc01714h](https://doi.org/10.1039/d2sc01714h).
- [25] R. Angharad Baber, Mairi F. Haddow, Ann J. Middleton, A. Guy Orpen, Paul G. Pringle, Anthony Haynes, Gary L. Williams, and Rainer Papp. "Ligand Stereoelectronic Effects in Complexes of Phospholanes, Phosphinanes, and Phosphapanes and Their Implications for Hydroformylation Catalysis". In: *Organometallics* 26 (3 Jan. 2007), pp. 713–725. ISSN: 0276-7333. DOI: [10.1021/om060912v](https://doi.org/10.1021/om060912v).

- [26] Dipak Kumar Mandal. *Stereochemistry and Organic Reactions Conformation, Configuration, Stereoelectronic Effects and Asymmetric Synthesis*. Elsevier, 2021, pp. 475–490. ISBN: 978-0-12-824092-2.
- [27] David Weininger. “**SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules**”. In: *Journal of Chemical Information and Computer Sciences* 28.1 (1988), pp. 31–36. DOI: [10.1021/ci00057a005](https://doi.org/10.1021/ci00057a005).
- [28] 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**”. In: *Journal of the American Chemical Society* 144 (3 Jan. 2022), pp. 1205–1217. ISSN: 15205126. DOI: [10.1021/jacs.1c09718](https://doi.org/10.1021/jacs.1c09718).
- [29] Alexandre V. Brethomé, Stephen P. Fletcher, and Robert S. Paton. “**Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters**”. In: *ACS Catalysis* 9 (3 Mar. 2019), pp. 2313–2323. ISSN: 21555435. DOI: [10.1021/acscatal.8b04043](https://doi.org/10.1021/acscatal.8b04043).
- [30] J.M. Seminario. *Recent Developments and Applications of Modern Density Functional Theory*. Elsevier, 1996, pp. 288–295. ISBN: 0-444-82404-9.
- [31] Schrodinger E. “An Undulatory Theory of the Mechanics of Atoms and Molecules.” In: *Annalen der Physik* 81 (1926), p. 109.
- [32] W Kohn and L J Sham. “Self-Consistent Equations Including Exchange and Correlation Effects”. In: 140 (1965), A1133–A1138.
- [33] P. Hohenberg and W. Kohn. “**Inhomogeneous Electron Gas**”. In: *Phys. Rev.* 136 (3B Nov. 1964), B864–B871. DOI: [10.1103/PhysRev.136.B864](https://doi.org/10.1103/PhysRev.136.B864).
- [34] Dmitrij Rappoport, Nathan R M Crawford, Filipp Furche, and Kieron Burke. *Which functional should I choose?* 2008.
- [35] Sergio Sousa, Pedro Fernandes, and Maria Ramos. “**General Performance of Density Functionals**”. In: *The journal of physical chemistry. A* 111 (Nov. 2007), pp. 10439–52.
- [36] Car R. “**Fixing Jacob’s ladder**”. In: *Nature Chemistry* 8 (2016), pp. 820–821.
- [37] Konstantinos D. Vogiatzis, Mikhail V. Polynski, Justin K. Kirkland, Jacob Townsend, Ali Hashemi, Chong Liu, and Evgeny A. Pidko. “**Computational Approach to Molecular Catalysis by 3d Transition Metals: Challenges and Opportunities**”. In: *Chemical Reviews* 119 (4 Feb. 2019), pp. 2453–2523. ISSN: 15206890. DOI: [10.1021/acs.chemrev.8b00361](https://doi.org/10.1021/acs.chemrev.8b00361).
- [38] Vivek Sinha. “The Molecular Basis of Clean Energy Elucidating the Mechanism of Homogeneously Catalyzed Hydrogen Production from Methanol The Molecular Basis of Clean Energy: Elucidating the Mechanism of Homogeneously Catalyzed Hydrogen Production from Methanol.” 2019.
- [39] Errol G. Lewars. *The Concept of the Potential Energy Surface*. Cham: Springer International Publishing, 2016, pp. 9–49. ISBN: 978-3-319-30916-3. DOI: [10.1007/978-3-319-30916-3_2](https://doi.org/10.1007/978-3-319-30916-3_2).
- [40] Carr S. F., R. Garnett, and C. S. Lo. “**Accelerating the search for global minima on potential energy surfaces using machine learning**”. In: *The Journal of Chemical Physics* 145.15 (2016), p. 154106.
- [41] Goedecker S., Hellmann W., and Lenosky T. “**Global minimum determination of the Born-Oppenheimer surface within density functional theory**.” In: *Physical Rev Lett.* (2005).

- [42] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. "A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu". In: *The Journal of Chemical Physics* 132.15 (2010), p. 154104.
- [43] Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. "Extended tight-binding quantum chemistry methods". In: *WIREs Computational Molecular Science* 11.2 (2021), e1493. DOI: 10.1002/wcms.1493.
- [44] Iñigo Iribarren and Cristina Trujillo. "Efficiency and Suitability when Exploring the Conformational Space of Phase-Transfer Catalysts". In: *Journal of Chemical Information and Modeling* (Nov. 2022). ISSN: 1549960X. DOI: 10.1021/acs.jcim.2c00934.
- [45] Mikko J. Vainio and Mark S. Johnson. "Generating conformer ensembles using a multiobjective genetic algorithm". In: *Journal of Chemical Information and Modeling* 47 (6 2007), pp. 2462–2474. ISSN: 1549960X. DOI: 10.1021/Ci6005646.
- [46] Sereina Riniker and Gregory A. Landrum. "Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation". In: *Journal of Chemical Information and Modeling* 55 (12 Dec. 2015), pp. 2562–2574. ISSN: 1549960X. DOI: 10.1021/acs.jcim.5b00654.
- [47] Jean Paul Ebejer, Garrett M. Morris, and Charlotte M. Deane. "Freely available conformer generation methods: How good are they?" In: *Journal of Chemical Information and Modeling* 52 (5 May 2012), pp. 1146–1158. ISSN: 1549960X. DOI: 10.1021/ci2004658.
- [48] Dakota Folmsbee and Geoffrey Hutchison. "Assessing conformer energies using electronic structure and machine learning methods". In: *International Journal of Quantum Chemistry* 121 (1 Jan. 2021). ISSN: 1097461X. DOI: 10.1002/qua.26381.
- [49] Stefan Grimme. "Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations". In: *Journal of Chemical Theory and Computation* 15.5 (2019), pp. 2847–2862.
- [50] Jason A. Gillespie, Erik Zuidema, Piet van Leeuwen, and Paul C. J. Kamer. *Phosphorus (III) Ligands in Homogeneous Catalysis: Design and Synthesis. Chap. 1: Phosphorus Ligand Effects in Homogeneous Catalysis and Rational Catalyst Design*. John Wiley & Sons, 2012.
- [51] Raffaello Lazzaroni, Roberta Settambolo, and Aldo Caiazzo. *Rhodium catalyzed hydroformylation. Chap. 2: Hydroformylation with unmodified rhodium catalysts*. Kluwer Academic Publishers, 2002.
- [52] Kjell Jorner, Tobias Gensch, Pascal Friedrich, and Gabriel dos Passos Gomes. *Morfeus: Molecular Features for Machine Learning, version 0.7.2*. 2022.
- [53] "Ligand bite angle effects in metal-catalyzed C-C bond formation". In: *Chemical Reviews* 100 (8 Aug. 2000), pp. 2741–2769. ISSN: 00092665. DOI: 10.1021/cr9902704.
- [54] Jenna A. Bilbrey, Arianna H. Kazez, Jason Locklin, and Wesley D. Allen. "Exact ligand cone angles". In: *Journal of Computational Chemistry* 34 (14 May 2013), pp. 1189–1197. ISSN: 01928651. DOI: 10.1002/jcc.23217.
- [55] Albert Poater, Biagio Cosenza, Andrea Correa, Simona Giudice, Francesco Ragone, Vittorio Scarano, and Luigi Cavallo. "SambVca: A web application for the calculation of the buried volume of N-heterocyclic carbene ligands". In: *European Journal of Inorganic Chemistry* (13 SPEC. ISS. May 2009), pp. 1759–1766. ISSN: 14341948. DOI: 10.1002/ejic.200801160.
- [56] Jean Luc Bredas. "Mind the gap!" In: *Materials Horizons* 1 (1 Jan. 2014), pp. 17–19. ISSN: 20516355. DOI: 10.1039/c3mh00098b.

- [57] Eric D. Glendening, Clark R. Landis, and Frank Weinhold. "Natural bond orbital methods". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2 (1 Jan. 2012), pp. 1–42. ISSN: 17590876. DOI: [10.1002/wcms.51](https://doi.org/10.1002/wcms.51).
- [58] Adrian Bondy and U.S.R. Murty. *Graph Theory*. Springer London, 2011.
- [59] Nenad Trinajstić. *Chemical graph theory*. 2nd ed. CRC Press, 2019.
- [60] Heather J. Kulik. *What's Left for a Computational Chemist To Do in the Age of Machine Learning?* Office of Scientific and Technical Information (OSTI), 2013.
- [61] Akinsola Jet and Hinmikaiye J O. "Supervised Machine Learning Algorithms: Classification and Comparison". In: *International Journal of Computer Trends and Technology* 48 (2017). ISSN: 2231-2803. DOI: [10.14445/22312803/IJCTT-V48P126](https://doi.org/10.14445/22312803/IJCTT-V48P126).
- [62] Tian Lan and Qi An. "Discovering Catalytic Reaction Networks Using Deep Reinforcement Learning from First-Principles". In: *Journal of the American Chemical Society* 143.40 (2021), pp. 16804–16812.
- [63] Yehia Amar, Artur M. Schweidtmann, Paul Deutsch, Liwei Cao, and Alexei Lapkin. "Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis". In: *Chemical Science* 10 (27 2019), pp. 6697–6706. ISSN: 20416539. DOI: [10.1039/c9sc01844a](https://doi.org/10.1039/c9sc01844a).
- [64] Octavio Loyola-Gonzalez. "Black-box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view". In: *IEEE Access* 7 (2019), pp. 154096–154113. ISSN: 21693536. DOI: [10.1109/ACCESS.2019.2949286](https://doi.org/10.1109/ACCESS.2019.2949286).
- [65] Seongyong Kim and Sunghwan Kim. "Prediction of the rate constant of phenol hydroxylation using random forest regression". In: *Chemosphere* 252 (2020), p. 126511.
- [66] Nathalie Sturm, Matthias A Meier, and Manfred Kansy. "Machine learning models for solubility prediction: Overview and outlook". In: *Expert opinion on drug discovery* 14.11 (2019), pp. 1135–1149.
- [67] Robin Gautier and Jean-Louis Reymond. "Chemical space navigation: a toolkit for reaction-based de novo design of drug-like molecules". In: *Journal of chemical information and modeling* 58.7 (2018), pp. 1364–1385.
- [68] Agustí Lledós. "Computational Organometallic Catalysis: Where We Are, Where We Are Going". In: *European Journal of Inorganic Chemistry* 2021 (26 July 2021), pp. 2547–2555. ISSN: 1099-0682. DOI: [10.1002/EJIC.202100330](https://doi.org/10.1002/EJIC.202100330).
- [69] Jesús Jover, Natalie Fey, Jeremy N. Harvey, Guy C. Lloyd-Jones, A. Guy Orpen, Gareth J. J. Owen-Smith, Paul Murray, David R. J. Hose, Robert Osborne, and Mark Purdie. "Expansion of the Ligand Knowledge Base for Chelating P,P-Donor Ligands (LKB-PP)". In: *Organometallics* 31 (2012), pp. 5302–5306. ISSN: 0276-7333. DOI: [10.1021/om300312t](https://doi.org/10.1021/om300312t).
- [70] Christoph Kuhn and David N Beratan. "Inverse Strategies for Molecular Design". In: *Journal of Physical Chemistry* (1996).
- [71] Carl Poree and Franziska Schoenebeck. "A holy grail in chemistry: Computational catalyst design: Feasible or fiction?" In: *Accounts of Chemical Research* 50 (3 Mar. 2017), pp. 605–608. ISSN: 15204898. DOI: [10.1021/acs.accounts.6b00606](https://doi.org/10.1021/acs.accounts.6b00606).
- [72] Natalie Fey, James A. S. Howell, Jonathan D. Lovatt, Paul C. Yates, Desmond Cunningham, Patrick McArdle, Hugo E. Gottlieb, and Simon J. Coles. "A molecular mechanics approach to mapping the conformational space of diaryl and triarylphosphines". In: *Dalton Transactions* (46 2006), p. 5464. ISSN: 1477-9226. DOI: [10.1039/b610123b](https://doi.org/10.1039/b610123b).

- [73] Jennifer Crawford and Matthew Sigman. "Conformational Dynamics in Asymmetric Catalysis: Is Catalyst Flexibility a Design Element?" In: *Synthesis* 51 (05 Mar. 2019), pp. 1021–1036. ISSN: 0039-7881. DOI: [10.1055/s-0037-1611636](https://doi.org/10.1055/s-0037-1611636).
- [74] Sawsan Dacrory and Asmaa M. Fahim. "Synthesis, anti-proliferative activity, computational studies of tetrazole cellulose utilizing different homogenous catalyst". In: *Carbohydrate Polymers* 229 (Feb. 2020), p. 115537. ISSN: 01448617. DOI: [10.1016/j.carbpol.2019.115537](https://doi.org/10.1016/j.carbpol.2019.115537).
- [75] Alister S. Goodfellow and Michael Bühl. "Hydricity of 3d Transition Metal Complexes from Density Functional Theory: A Benchmarking Study". In: *Molecules* 26 (13 July 2021), p. 4072. ISSN: 1420-3049. DOI: [10.3390/molecules26134072](https://doi.org/10.3390/molecules26134072).
- [76] Theresa Sperger, Italo A. Sanhueza, Indrek Kalvet, and Franziska Schoenebeck. "Computational Studies of Synthetically Relevant Homogeneous Organometallic Catalysis Involving Ni, Pd, Ir, and Rh: An Overview of Commonly Employed DFT Methods and Mechanistic Insights". In: *Chemical Reviews* 115 (17 Sept. 2015), pp. 9532–9586. ISSN: 0009-2665. DOI: [10.1021/acs.chemrev.5b00163](https://doi.org/10.1021/acs.chemrev.5b00163).
- [77] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. "Generalized Gradient Approximation Made Simple". In: *Physical Review Letters* 77 (18 Oct. 1996), pp. 3865–3868. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.77.3865](https://doi.org/10.1103/PhysRevLett.77.3865).
- [78] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. "A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu". In: *The Journal of Chemical Physics* 132 (15 Apr. 2010), p. 154104. ISSN: 0021-9606. DOI: [10.1063/1.3382344](https://doi.org/10.1063/1.3382344).
- [79] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. "Effect of the damping function in dispersion corrected density functional theory". In: *Journal of Computational Chemistry* 32 (7 May 2011), pp. 1456–1465. ISSN: 01928651. DOI: [10.1002/jcc.21759](https://doi.org/10.1002/jcc.21759).
- [80] R. Ditchfield, W. J. Hehre, and J. A. Pople. "Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules". In: *The Journal of Chemical Physics* 54 (2 Jan. 1971), pp. 724–728. ISSN: 0021-9606. DOI: [10.1063/1.1674902](https://doi.org/10.1063/1.1674902).
- [81] Carlo Adamo and Vincenzo Barone. "Toward reliable density functional methods without adjustable parameters: The PBE0 model". In: *Journal of Chemical Physics* 110 (13 Apr. 1999), pp. 6158–6170. ISSN: 00219606. DOI: [10.1063/1.478522](https://doi.org/10.1063/1.478522).
- [82] Florian Weigend and Reinhart Ahlrichs. "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy". In: *Physical Chemistry Chemical Physics* 7 (18 2005), p. 3297. ISSN: 1463-9076. DOI: [10.1039/b508541a](https://doi.org/10.1039/b508541a).
- [83] Jan Grimo Sobez and Markus Reiher. "Molassembler: Molecular Graph Construction, Modification, and Conformer Generation for Inorganic and Organic Molecules". In: *Journal of Chemical Information and Modeling* 60 (8 Aug. 2020), pp. 3884–3900. ISSN: 1549960X. DOI: [10.1021/acs.jcim.0c00503](https://doi.org/10.1021/acs.jcim.0c00503).
- [84] Michael Busch, Matthew D. Wodrich, and Clémence Corminboeuf. "A Generalized Picture of C–C Cross-Coupling". In: *ACS Catalysis* 7 (9 July 2017), pp. 5643–5653. ISSN: 21555435. DOI: [10.1021/acscatal.7b01415](https://doi.org/10.1021/acscatal.7b01415).
- [85] P Sabatier. "La catalyse en chimie organique, Librairie Polytechnique". In: (1913).

- [86] Megha Anand and Jens K. Nørskov. "Scaling Relations in Homogeneous Catalysis: Analyzing the Buchwald–Hartwig Amination Reaction". In: *ACS Catalysis* 10 (1 Jan. 2020), pp. 336–345. ISSN: 2155-5435. DOI: [10.1021/acscatal.9b04323](https://doi.org/10.1021/acscatal.9b04323).
- [87] Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. "**Mordred: a molecular descriptor calculator**". In: *Journal of Cheminformatics* 10 (1 2018), p. 4. ISSN: 1758-2946. DOI: [10.1186/s13321-018-0258-y](https://doi.org/10.1186/s13321-018-0258-y).
- [88] Wenjun Yang, Ivan Yu Chernyshov, Manuela Weber, Evgeny A Pidko, and Georgy A Filonenko. "Switching Between Hydrogenation and Olefin Transposition Catalysis via Silencing NH Cooperativity in Mn(I) pincer complexes". In: *ACS Catalysis* (2022). DOI: [10.4121/19704391](https://doi.org/10.4121/19704391).
- [89] Ivan Chernyshov. *Metal Complex Embedding*. 2020.
- [90] Adarsh V. Kalikadien, Evgeny A. Pidko, and Vivek Sinha. "**ChemSpaX: exploration of chemical space by automated functionalization of molecular scaffold**". In: *Digital Discovery* 1 (1 2022), pp. 8–25.
- [91] Efthymios I. Ioannidis, Terry Z.H. Gani, and Heather J. Kulik. "molSimplify: A toolkit for automating discovery in inorganic chemistry". In: *Journal of Computational Chemistry* (Aug. 2016), pp. 2106–2117. ISSN: 1096987X. DOI: [10.1002/jcc.24437](https://doi.org/10.1002/jcc.24437).
- [92] M. J. Frisch et. al. *Gaussian 16 Revision C.01*. Gaussian Inc. Wallingford CT. 2016.
- [93] "Open Babel: An Open chemical toolbox". In: *Journal of Cheminformatics* 3 (10 Oct. 2011). ISSN: 17582946. DOI: [10.1186/1758-2946-3-33](https://doi.org/10.1186/1758-2946-3-33).
- [94] Eric D. Glendening, Clark R. Landis, and Frank Weinhold. "NBO 6.0: Natural bond orbital analysis program". In: *Journal of Computational Chemistry* 34 (16 June 2013), pp. 1429–1437. ISSN: 1096987X. DOI: [10.1002/jcc.23266](https://doi.org/10.1002/jcc.23266).
- [95] Carlo Adamo and Vincenzo Barone. "**Inexpensive and accurate predictions of optical excitations in transition-metal complexes: The TDDFT/PBE0 route**". In: *Theoretical Chemistry Accounts* 105 (2 2000), pp. 169–172. ISSN: 1432881X. DOI: [10.1007/s002140000202](https://doi.org/10.1007/s002140000202).
- [96] Scott J. Miller. "Asymmetric catalysis: Correlating sterics in catalysis". In: *Nature Chemistry* 4 (5 May 2012), pp. 344–345. ISSN: 17554330. DOI: [10.1038/nchem.1339](https://doi.org/10.1038/nchem.1339).
- [97] Zhonghua Wang, Lu Liang, Zheng Yin, and Jianping Lin. "Improving chemical similarity ensemble approach in target prediction". In: *Journal of Cheminformatics* 8 (1 2016). ISSN: 17582946. DOI: [10.1186/s13321-016-0130-x](https://doi.org/10.1186/s13321-016-0130-x).
- [98] Greg Landrum. *RDKit: Open-Source Cheminformatics*. version 2022.09.1.
- [99] Lagnajit Pattanaik and Connor W. Coley. "**Molecular Representation: Going Long on Fingerprints**". In: *Chem* 6.6 (2020), pp. 1204–1207. ISSN: 2451-9294. DOI: doi.org/10.1016/j.chempr.2020.05.002.
- [100] Trang T Le, Weixuan Fu, and Jason H Moore. "Scaling tree-based automated machine learning to biomedical big data with a feature set selector". In: *Bioinformatics* 36.1 (2020), pp. 250–256.
- [101] Randal S. Olson, Ryan J. Urbanowicz, Peter C. Andrews, Nicole A. Lavender, La Creis Kidd, and Jason H. Moore. *Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 – April 1, 2016, Proceedings, Part I*. Ed. by Giovanni Squillero and Paolo Burelli. Springer International Publishing, 2016. Chap. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization, pp. 123–137. ISBN: 978-3-319-31204-0. DOI: [10.1007/978-3-319-31204-0_9](https://doi.org/10.1007/978-3-319-31204-0_9).

- [102] Fionn Murtagh and Pedro Contreras. "Algorithms for hierarchical clustering: An overview". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2 (1 Jan. 2012), pp. 86–97. ISSN: 19424795. DOI: [10.1002/widm.53](https://doi.org/10.1002/widm.53).
- [103] Olaf Wiest, Mandana Saebi, Bozhao Nan, John E Herr, Jessica Wahlers, Zhichun Guo, Andrzej Zuranski, Thierry Kogej, Per-Ola Norrby, Abigail G Doyle, and Nitesh V. Chawla. "On the Use of Real-World Datasets for Reaction Yield Prediction". In: *Chemical Science* (2023). ISSN: 2041-6520. DOI: [10.1039/D2SC06041H](https://doi.org/10.1039/D2SC06041H).
- [104] Olivier Caelen. "A Bayesian interpretation of the confusion matrix". In: *Annals of Mathematics and Artificial Intelligence* 81 (3-4 Dec. 2017), pp. 429–450. ISSN: 15737470. DOI: [10.1007/s10472-017-9564-8](https://doi.org/10.1007/s10472-017-9564-8).
- [105] Natalie Fey, Alexander Koumi, Andrei V. Malkov, Jonathan D. Moseley, Bao N. Nguyen, Simon N.G. Tyler, and Charlotte E. Willans. "Mapping the properties of bidentate ligands with calculated descriptors (LKB-bid)". In: *Dalton Transactions* 49 (24 June 2020), pp. 8169–8178. ISSN: 14779234. DOI: [10.1039/d0dt01694b](https://doi.org/10.1039/d0dt01694b).
- [106] Jaya Mehara, Anurag Noonikara-Poyil, Adway O. Zacharias, Jana Roithova, and H. V. Rasika Dias. "Binding Interactions in Copper, Silver and Gold π -Complexes". In: *Chemistry - A European Journal* 28 (13 Mar. 2022). ISSN: 15213765. DOI: [10.1002/chem.202103984](https://doi.org/10.1002/chem.202103984).
- [107] Gilles Louppe. "Understanding Random Forests: From Theory to Practice". July 2014.
- [108] Gabriel dos Passos Gomes, Robert Pollice, and Alán Aspuru-Guzik. "Navigating through the Maze of Homogeneous Catalyst Design with Machine Learning". In: *Trends in Chemistry* 3 (2 2021), pp. 96–110. ISSN: 2589-5974. DOI: doi.org/10.1016/j.trechm.2020.12.006.
- [109] Dejun Jiang, Zhenxing Wu, Chang Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models". In: *Journal of Cheminformatics* 13 (1 Dec. 2021). ISSN: 17582946. DOI: [10.1186/s13321-020-00479-8](https://doi.org/10.1186/s13321-020-00479-8).
- [110] Xiaowei Xie, Evan Walter Clark Spotte-Smith, Mingjian Wen, Hetal D. Patel, Samuel M. Blau, and Kristin A. Persson. "Data-Driven Prediction of Formation Mechanisms of Lithium Ethylene Monocarbonate with an Automated Reaction Network". In: *Journal of the American Chemical Society* 143 (33 Aug. 2021), pp. 13245–13258. ISSN: 15205126. DOI: [10.1021/jacs.1c05807](https://doi.org/10.1021/jacs.1c05807).
- [111] Ali Hashemi, Sana Bougueroua, Marie-Pierre Gageot, and Evgeny A. Pidko. "ReNeGate: A Reaction Network Graph-Theoretical Tool for Automated Mechanistic Studies in Computational Homogeneous Catalysis". In: *Journal of Chemical Theory and Computation* (Dec. 2022). ISSN: 1549-9618. DOI: [10.1021/acs.jctc.2c00404](https://doi.org/10.1021/acs.jctc.2c00404).
- [112] Takahashi Keisuke and Satoshi Maeda. "Mining hydroformylation in complex reaction network via graph theory". In: *RSC Advances* 11 (38 2021), pp. 23235–23240. DOI: [10.1039/D1RA03395F](https://doi.org/10.1039/D1RA03395F).
- [113] Yanfei Guan, Victoria M. Ingman, Benjamin J. Rooks, and Steven E. Wheeler. "AARON: An Automated Reaction Optimizer for New Catalysts". In: *Journal of Chemical Theory and Computation* 14 (10 Oct. 2018), pp. 5249–5261. ISSN: 15499626. DOI: [10.1021/ACS.JCTC.8B00578](https://doi.org/10.1021/ACS.JCTC.8B00578).

- [114] Miguel Steiner and Markus Reiher. “Autonomous Reaction Network Exploration in Homogeneous and Heterogeneous Catalysis”. In: *Topics in Catalysis* 65 (1-4 Feb. 2022), pp. 6–39. ISSN: 15729028. DOI: [10.1007/s11244-021-01543-9](https://doi.org/10.1007/s11244-021-01543-9).
- [115] Alan Aspuru Guzik et. al. “SELFIES and the future of molecular string representations”. In: *Patterns* 3.10 (2022), p. 100588. ISSN: 2666-3899. DOI: doi.org/10.1016/j.patter.2022.100588.



Simplest model molecule

Rhodium (I) complexes have square planar configuration. Considering this geometry in a solution in the presence of bidentate phosphorus ligands, there are two possible molecular arrangements. They are shown in Figure A.1.

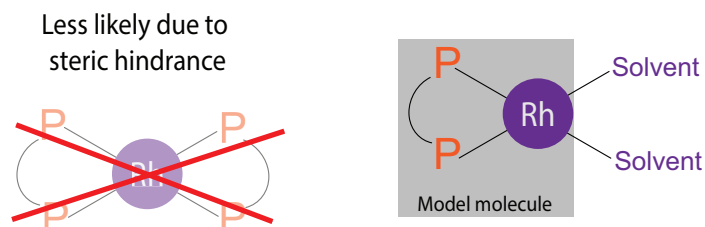


Figure A.1: Selection of model molecule based on the most likely square planar configuration

The model molecule can be interpreted in two ways: as a conceptual molecule and as a proper molecular intermediate. In the former case, the Rhodium metal center would have a charge of 0 and multiplicity of 2, since both bonds with the phosphorus would come from the lone pairs present on the phosphorus atoms, while in the latter the charge and multiplicity would be +1 and 1 respectively, similarly to a real square planar complex.

B

Conformer correlations

The plots of the conformer correlations presented in Fig 4.7 are given in Fig B.1:

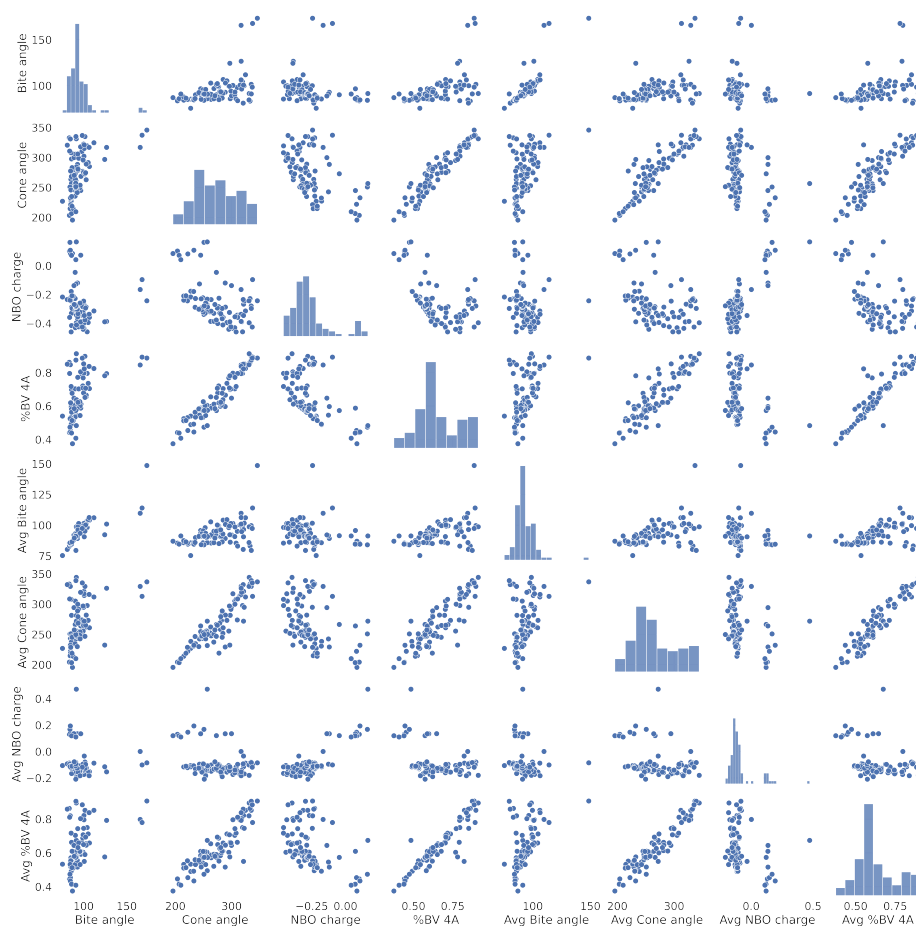


Figure B.1: Pairplots for the correlations shown in Fig 4.7

The full correlation matrix is given in Fig B.2:

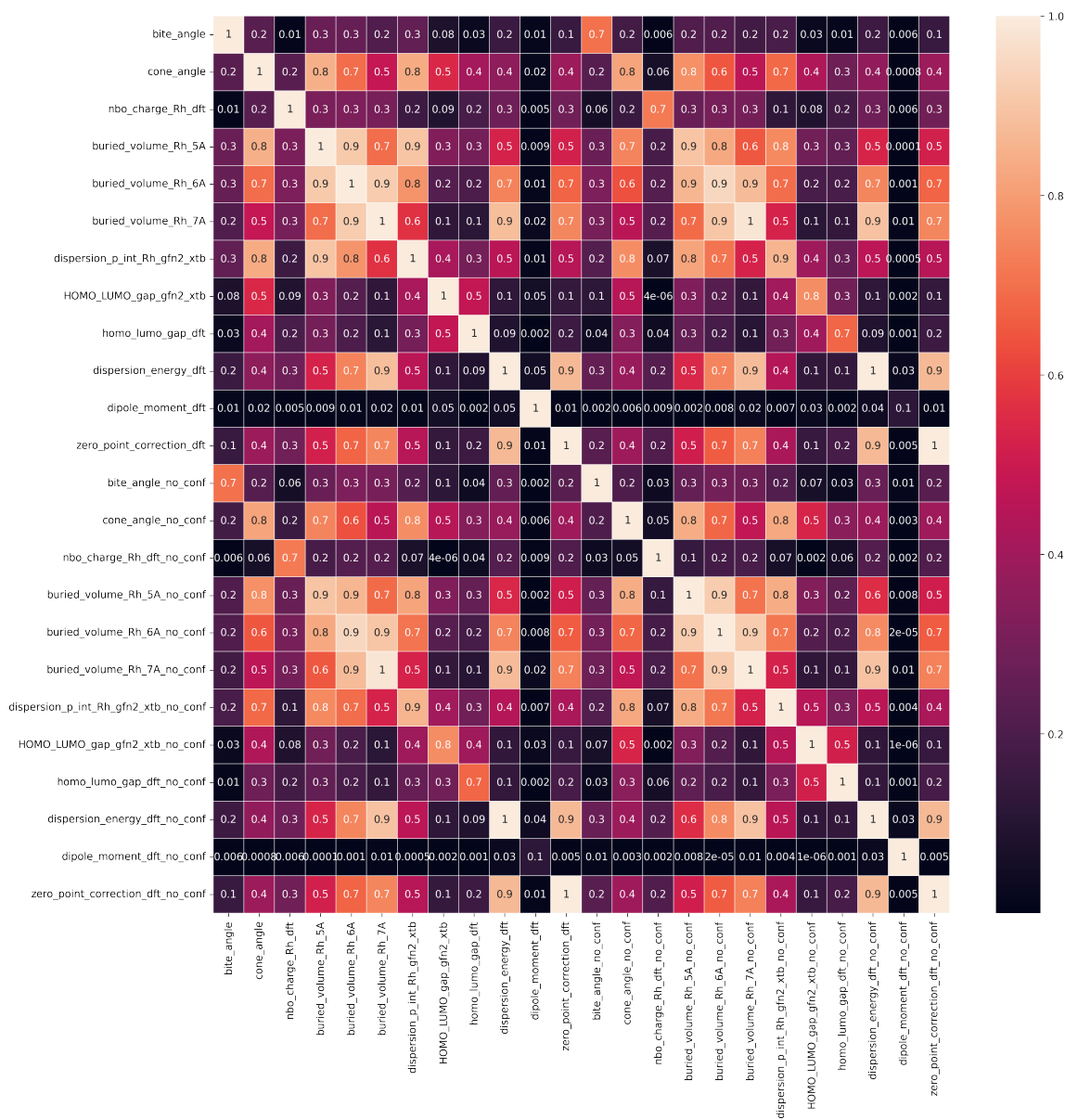


Figure B.2: Full correlation matrix between conformer averaged properties and single structure properties



Protocol for 3D coordinates generation of sandwich TM complexes

Generation of complexes with high π -electron interaction from string representation is a complicated matter and an active research topic. However, several generation sequences can be tried to reach the final goal. Figure C.1 displays different representations of ferrocenes [115]. Representation (1) is not interpretable as a SMILES. (2) is a flat 2D representation of a ferrocene which is chemically inaccurate. (3) and (4) do not convey the correct chemical information, but are the only alternatives worth investigating. The number of bonds on (3) and (4) convey the correct chemical message, but are still incorrect since in SMILES representation a covalent bond represents a full bond while a more accurate explanation is the presence of partial, delocalized bonds.

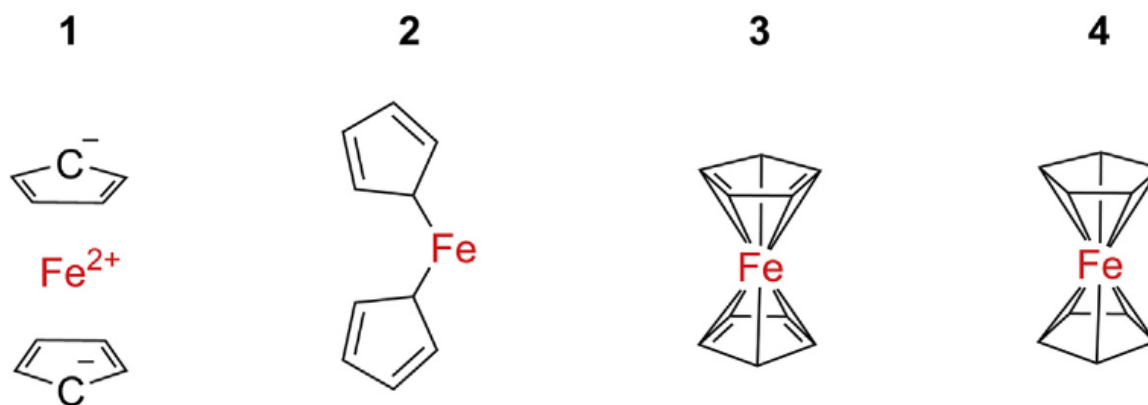


Figure C.1: Possible ferrocene representations according to Guzik et al. [115]

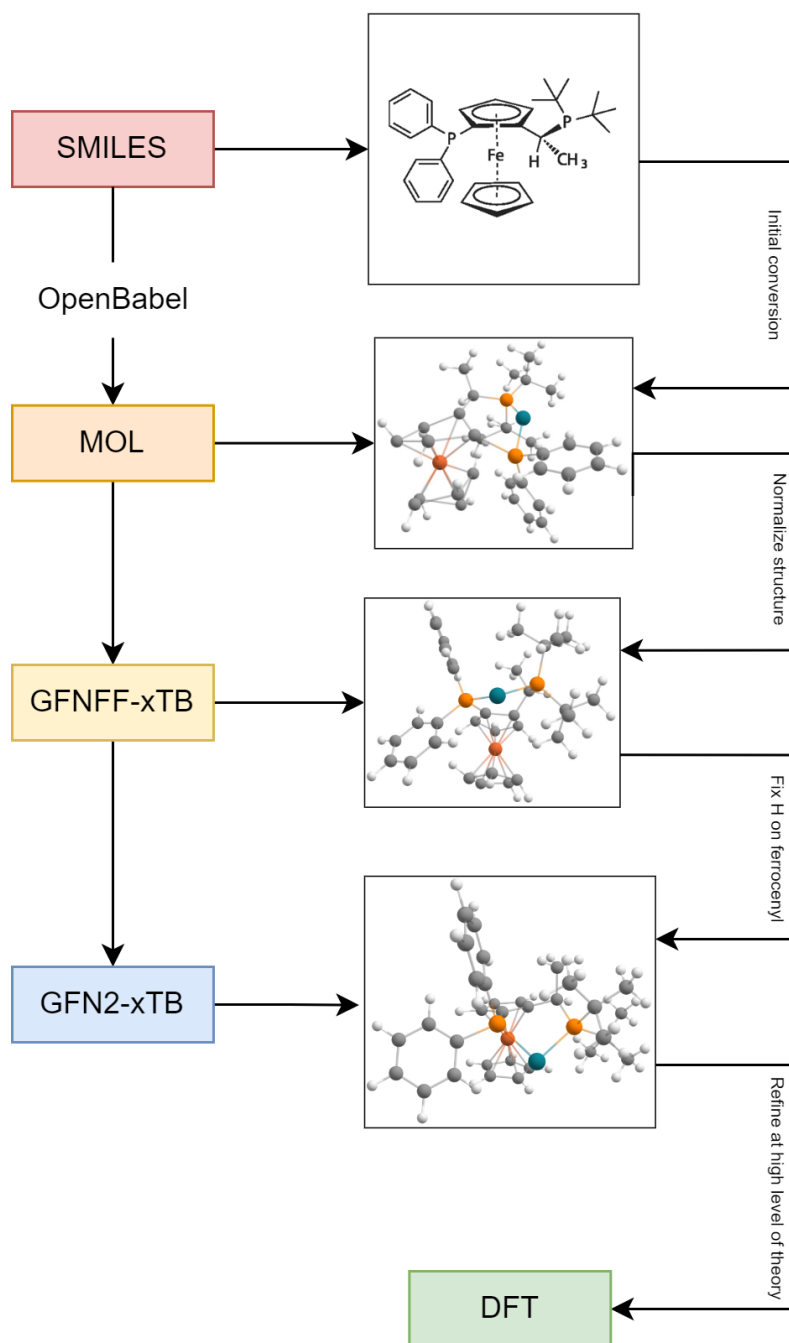


Figure C.2: Protocol schema for 3D coordinate generation for ferrocenyl metal complexes

D

GFN2-xTB preoptimized structure

The second pathway is to apply the semi-empirical GFN2-xTB optimization, which then followed by DFT.

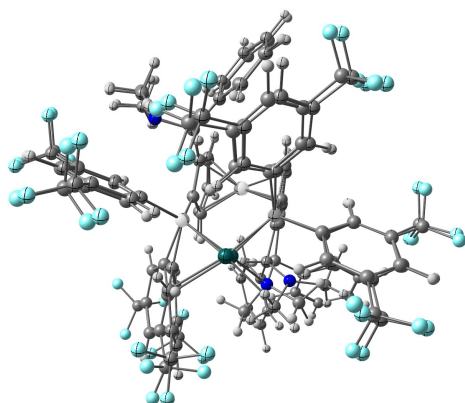


Figure D.1: The difference between the handmade-xTB-DFT and handmade-DFT optimizations.

Since xTB optimization underperformed for 494227-36-0, two optimization strategies were compared. The handmade complex was optimized in two ways: xTB \rightarrow DFT and direct DFT optimization. The obtained result is of high interest, since the two obtained complexes have different ligating atoms, but at the same time, the RMSD of 1.61 is relatively low (see Fig. D.1). The xTB optimized structure generated a pincer-like structure, adding a nitrogen coordination to the under-coordinated Rhodium, while the directly DFT optimized structure has a C-H agostic interaction and a π - interaction with one of the adjacent phenyl rings, completing the square planar configuration. This results in a 40 kJ/mol energy difference, with the pincer being the more stable configuration.

The initial xTB geometry was incorrect due to the proximity of the two metals present in the molecule. The xTB single point calculation showed the structure was ca. 300 kJ/mol more stable than the DFT-xTB optimized structures. However, DFT corrected the mistake, resulting in a more stable geometry.

E

Database contents

E.1 Structures of publicly available substrates

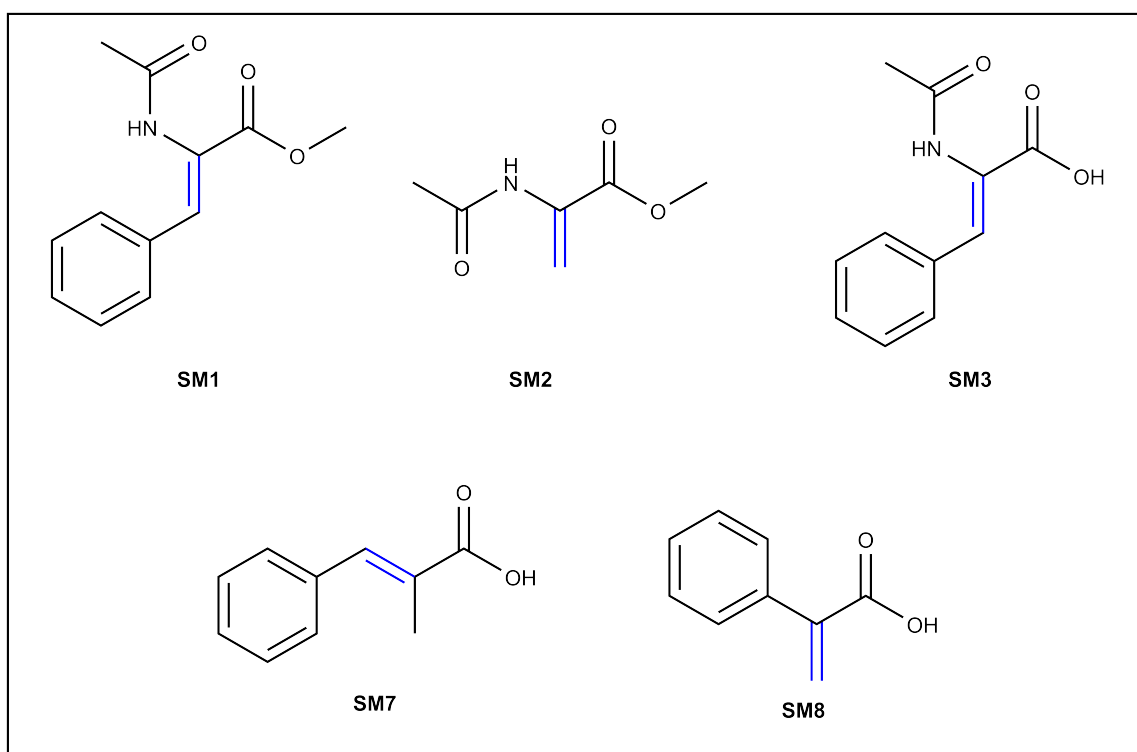


Figure E.1: Structures of publicly available substrates. SM4 and SM5 are confidential and will not be made available to the public.

E.2 Ligands database

Table E.1: Ligand database information

| Nr. | Commercial name | CAS | Formula |
|-----|-----------------|-------------|---|
| L1 | SL-J001-1 | 155806-35-2 | C ₃₆ H ₄₄ FeP ₂ |
| L2 | SL-J002-1 | 155830-69-6 | C ₃₂ H ₄₀ FeP ₂ |
| L3 | SL-J003-1 | 167416-28-6 | C ₃₆ H ₅₆ FeP ₂ |
| L4 | SL-J004-1 | 158923-09-2 | C ₃₆ H ₄₄ FeP ₂ |
| L5 | SL-J005-1 | 184095-69-0 | C ₄₀ H ₄₀ FeP ₂ |
| L6 | SL-J006-1 | 292638-88-1 | C ₄₀ H ₄₀ F ₁₂ FeP ₂ |
| L7 | SL-J007-1 | 360048-63-1 | C ₄₂ H ₅₆ FeO ₂ P ₂ |
| L8 | SL-J008-1 | 166172-63-0 | C ₄₄ H ₃₆ F ₁₂ FeP ₂ |
| L9 | SL-J009-1 | 158923-11-6 | C ₃₂ H ₅₂ FeP ₂ |
| L10 | SL-J011-1 | 246231-79-8 | C ₃₄ H ₃₈ F ₆ FeP ₂ |
| L11 | SL-J013-1 | 187733-50-2 | C ₃₈ H ₅₂ FeO ₂ P ₂ |
| L12 | SL-J212-1 | 849924-41-0 | C ₂₈ H ₃₆ FeO ₂ P ₂ |
| L13 | SL-J404-1 | 851308-40-2 | C ₄₈ H ₄₄ FeP ₂ |
| L14 | SL-J418-1 | 849924-45-4 | C ₄₆ H ₅₂ FeO ₂ P ₂ |
| L15 | SL-J452-1 | 849924-73-8 | C ₃₄ H ₃₂ FeO ₂ P ₂ |
| L16 | SL-J502-1 | 223120-71-6 | C ₃₂ H ₄₀ FeP ₂ |
| L17 | (R)-BINAM-P | 74974-14-4 | C ₄₄ H ₃₄ N ₂ P ₂ |
| L18 | SL-J505-1 | 849924-76-1 | C ₃₄ H ₄₄ FeP ₂ |
| L19 | SL-T002-2 | 914089-00-2 | C ₄₃ H ₆₃ FeNP ₂ |
| L20 | SL-M001-1 | 174467-31-3 | C ₅₂ H ₅₀ FeN ₂ P ₂ |
| L21 | SL-M003-1 | 494227-36-0 | C ₆₀ H ₄₂ F ₂₄ FeN ₂ P ₂ |
| L22 | SL-M004-1 | 494227-37-1 | C ₆₄ H ₇₄ FeN ₂ O ₄ P ₂ |
| L23 | SL-M009-1 | 793718-16-8 | C ₆₀ H ₆₆ FeN ₂ P ₂ |
| L24 | SL-T001-2 | 850444-36-9 | C ₄₃ H ₃₉ FeNP ₂ |
| L25 | SL-W001-1 | 387868-06-6 | C ₄₆ H ₃₂ F ₁₂ FeP ₂ |
| L26 | SL-W002-1 | 388079-58-1 | C ₄₂ H ₃₆ FeP ₂ |
| L27 | SL-W003-2 | 849925-19-5 | C ₄₂ H ₄₈ FeP ₂ |

| | | | |
|-----|--------------------|--------------|-----------------|
| L28 | SL-W005-2 | 849925-20-8 | C52H44F12FeO2P2 |
| L29 | SL-W008-2 | 849925-22-0 | C46H44F12FeP2 |
| L30 | SL-W009-1 | 894771-28-9 | C50H52FeP2 |
| L31 | SL-F356-1 | 952586-19-5 | C42H53Fe2NP2 |
| L32 | (R)-BINAP | 76189-55-4 | C44H32P2 |
| L33 | (R)-BTfM-GarPhos | 1365531-84-5 | C48H28F24O4P2 |
| L34 | (R)-Tol-BINAP | 99646-28-3 | C48H40P2 |
| L35 | (R)-Xyl-BINAP | 137219-86-4 | C52H48P2 |
| L36 | (R)-H8-BINAP | 139139-86-9 | C44H40P2 |
| L37 | (S)-SegPhos | 210169-54-3 | C38H28O4P2 |
| L38 | (S)-Xyl-SegPhos | 210169-57-6 | C46H44O4P2 |
| L39 | (S)-DTBM-SegPhos | 210169-40-7 | C74H100O8P2 |
| L40 | (R)-Cl-MeO-BIPHEP | 185913-97-7 | C38H30Cl2O2P2 |
| L41 | SL-A109-1 | 352655-61-9 | C74H104O6P2 |
| L42 | SL-A120-1 | 394248-45-4 | C46H48O2P2 |
| L43 | SL-A107-1 | 352655-40-4 | C70H100N4O2P2 |
| L44 | SL-A108-2 | 145214-59-1 | C30H24O6P2 |
| L45 | SL-A102-2 | 133545-25-2 | C42H40O2P2 |
| L46 | SL-A121-1 | 192138-05-9 | C70H96O2P2 |
| L47 | SL-A104-1 | 256390-47-3 | C50H56O14P2 |
| L48 | (R)-GarPhos | 1365531-75-4 | C40H36O4P2 |
| L49 | (R)-Xyl-GarPhos | 1365531-89-0 | C48H52O4P2 |
| L50 | (R)-DTBM-GarPhos | 1365531-98-1 | C76H108O8P2 |
| L51 | (S)-iPr-BIPHEP | 150971-43-0 | C26H40O2P2 |
| L52 | (R)-C3-TunePhos | 301847-89-2 | C39H32O2P2 |
| L53 | (S,S)-iPr-BPE | 528854-34-4 | C22H44P2 |
| L54 | (R,R,R)-SPIRAP | NA | C43H38O2P2 |
| L55 | (R,R,S,S)-DuanPhos | 528814-26-8 | C24H32P2 |
| L56 | (R,R)-DiPamp | 55739-58-7 | C28H28O2P2 |
| L57 | (R)-iPr-PHOX | 164858-78-0 | C24H24NOP |
| L58 | SL-F131-1 | 899811-43-9 | C50H54Fe3N2P2 |

| | | | |
|-----|-----------------------|--------------|-------------|
| L59 | (R)-Xyl-SDP | 917377-75-4 | C49H50P2 |
| L60 | (S)-DM-MonoPhos | 185449-86-9 | C24H22NO2P |
| L61 | (R)-Ph-Monophos | 936010-61-6 | C34H26NO2P |
| L62 | (S)-NEt2-MonoPhos | 252288-04-3 | C24H22NO2P |
| L63 | (R,R,R)-Xyl-SKP | 1429939-35-4 | C52H54O2P2 |
| L64 | (R,R)-Ph-BPE | 528565-79-9 | C34H36P2 |
| L65 | (S,S)-ChiraPhos | 64896-28-2 | C28H28P2 |
| L66 | (R,R)-Et-BPE | 136705-62-9 | C18H36P2 |
| L67 | (R)-QuinoxP | 866081-62-1 | C18H28N2P2 |
| L68 | (R,R)-Et-DuPhos | 136705-64-1 | C22H36P2 |
| L69 | (R,R)-Me-DuPhos | 147253-67-6 | C18H28P2 |
| L70 | (S)-PhanePhos | 192463-40-4 | C40H34P2 |
| L71 | (S)-Me-iPr-PHOX | 1152313-76-2 | C26H28NOP |
| L72 | SL-N003-2 | 163169-29-7 | C28H28FeNOP |
| L73 | (S)-NeoPHOX | 1199225-38-1 | C22H28NOP |
| L74 | (R,R)-Me-BoPhoz | 406680-94-2 | C37H35FeNP2 |
| L75 | (R)-Xyl-PhanePhos | 325168-89-6 | C48H50P2 |
| L76 | (S,S)-f-Binaphane | 544461-38-3 | C54H40FeP2 |
| L77 | (R,R)-BDPP | 96183-46-9 | C29H30P2 |
| L78 | (R,R)-NorPhos | 71042-55-2 | C31H28P2 |
| L79 | (R,S)-BPPFA | 74311-56-1 | C38H37FeNP2 |
| L80 | (R,R)-DIOP | 32305-98-9 | C31H32O2P2 |
| L81 | (S)-Tol-tBu-PHOX | 218460-00-5 | C27H30NOP |
| L82 | (S,S)-DPE-Phos | 2119686-55-2 | C38H32O3P2 |
| L83 | (S)-NMDPP | 43077-29-8 | C22H29P |
| L84 | (S,S)-BABIBOP | 2207601-04-3 | C22H28O2P2 |
| L85 | (S,S,S,S)-Me-BABIBOP | 2207601-10-1 | C24H32O2P2 |
| L86 | (S,S,S,S)-iPr-BABIBOP | 2207601-12-3 | C28H40O2P2 |
| L87 | (R,R,R,R)-Me-BIBOP | 1884680-48-1 | C38H44O6P2 |
| L88 | (R,R)-PPM | 77450-05-6 | C29H29NP2 |
| L89 | SL-A101-2 | 133545-16-1 | C38H32O2P2 |

| | | | |
|------|--|--------------|-----------------|
| L90 | (S)-MeO-F12-BIPHEP | 116008-37-6 | C38H20F12O2P2 |
| L91 | (R)-MeO-F16-BIPHEP | NA | C42H24F16O2P2 |
| L92 | (R)-MeO-py-F12-BIPHEP | NA | C38H24F12N4O2P2 |
| L93 | (R)-MeO-F20-BIPHEP | NA | C42H20F20O2P2 |
| L94 | (R)-MeO-BFPy-BIPHEP | NA | C42H20F24N4O2P2 |
| L95 | (S,S)-XylSKEWPhos | 551950-92-6 | C37H46P2 |
| L96 | (S,S)-DIPSKEWPhos | NA | C53H78P2 |
| L97 | SL-W022-1 | 849925-29-7 | C44H48FeP2 |
| L98 | catASium D(R) | 99135-95-2 | C35H33NP2 |
| L99 | (2R)-1-[(1S)-1-Aminoethyl]-2-(diphenylphosphino)ferrocene | 607389-84-4 | C24H24FeNP |
| L100 | SL-W012-1 | 565184-30-7 | C38H44FeP2 |
| L101 | SL-W030-1 | 1854067-62-1 | C34H52FeP2 |
| L102 | (S,S)-Et-FerroTANE | 290347-66-9 | C24H36FeP2 |
| L103 | SL-W029-1 | 1854067-50-7 | C38H56FeP2 |
| L104 | (S)-NMe2-MonoPhos | 157488-65-8 | C22H18NO2P |
| L105 | SL-F103-1 | 55700-44-2 | C26H28FeNP |
| L106 | (R)-Xyl-P-Phos | 442905-33-1 | C46H50N2O4P2 |
| L107 | (S)-2-(Diphenylphosphinomethyl)pyrrolidine | 60261-46-3 | C17H20NP |
| L108 | (R)-ProPhos | 67884-32-6 | C27H26P2 |
| L109 | (3R)-3-(1,1-Dimethylethyl)-2,3-dihydro-4-(2-methoxyphenyl)-1,3-benzoxaphosphole | 1338454-28-6 | C18H21O2P |
| L110 | (2S,3R)-2-[Bis(1,1-dimethylethyl)phosphino]-3-(1,1-dimethylethyl)-2,3-dihydro-4-methoxy-1,3-benzoxaphosphole | 1215081-28-9 | C20H34O2P2 |
| L111 | (R,R)-BenzP* | 919778-41-9 | C16H28P2 |
| L112 | SL-J216-1 | 849924-43-2 | C40H44FeP2 |
| L113 | (S,S)-1-Naphthyl-DiPamp | 256469-70-2 | C34H28P2 |
| L114 | (S,R)-PPFA | 55650-58-3 | C26H28FeNP |
| L115 | SL-F173-1 | 166172-70-9 | C30H24F12FeNP |
| L116 | (R)-Xyl-SDP Oxide | 1462321-89-6 | C49H50OP2 |
| L117 | (R)-SITCP | 856407-37-9 | C25H23P |

| | | | |
|------|---|--------------|--------------|
| L118 | (R,R,R)-Tol-SKP | 1429939-32-1 | C48H46O2P2 |
| L119 | (R,R)-BCPM | 114751-47-2 | C34H49NO2P2 |
| L120 | (R)-DiFluorPhos | 503538-69-0 | C38H24F4O4P2 |
| L121 | (R,R)-Me-BPE | 129648-07-3 | C14H28P2 |
| L122 | (R)-SynPhos | 445467-61-8 | C40H32O4P2 |
| L123 | (R)-SIPhos | 443965-14-8 | C19H20NO2P |
| L124 | (3R,8R)-Tetrahydro-N,N,2,2-tetramethyl-4,4,8,8-tetraphenyl-1,3-dioxolo[4,5-e][1,3,2]dioxaphosphepin-6-amine | 213843-90-4 | C33H34NO4P |
| L125 | (R)-SDP | 917377-74-3 | C41H34P2 |
| L126 | (R,R,R,R)-Ph-BIBOP | 2301856-53-9 | C34H36O2P2 |
| L127 | (R,S)-Ph-Bn-SIPHOX | 2074610-05-0 | C39H34NOP |
| L128 | (R,R)-iPr-BPF | 849950-54-5 | C30H48FeP2 |
| L129 | (R)-Tol-SDP | 528521-87-1 | C45H42P2 |
| L130 | (R)-DMM-GarPhos | 1365531-93-6 | C52H60O8P2 |
| L131 | 8-[(3R)-3-(1,1-Dimethylethyl)-2,3-dihydro-1,3-benzoxaphosphol-4-yl]benzo[1,2-b:5,4-b']difuran | 1835717-07-1 | C21H23O3P |
| L132 | (S)-PipPhos | 284472-79-3 | C25H22NO2P |
| L133 | (R)-An-PhanePhos | 364732-86-5 | C44H42O4P2 |
| L134 | (S)-BINAPINE | 528854-26-4 | C52H48P2 |
| L135 | (S)-H8-MonoPhos | 389130-06-7 | C22H26NO2P |
| L136 | (R,R)-Me-Ferrocene | 540475-45-4 | C22H32FeP2 |
| L137 | (R,R)-Et-Ferrocene | 147762-89-8 | C26H40FeP2 |
| L138 | (S,S,S,S)-MeO-BIBOP | 1202033-19-9 | C24H32O4P2 |
| L139 | (R)-CTH-BINAM | 208248-67-3 | C44H42N2P2 |
| L140 | (2R)-1-[(R)-Aminophenylmethyl]-2-(diphenylphosphino)ferrocene | 498580-48-6 | C29H26FeNP |
| L141 | (1R,2S)-TaniaPhos-OH | 851308-43-5 | C41H34FeOP2 |
| L142 | 2-[2-[(2R,5R)-2,5-Dimethyl-1-phospholanyl]phenyl]-1,3-dioxolane | 1044256-04-3 | C15H21O2P |
| L143 | (R,R)-BPPM | 72598-03-9 | C34H37NO2P2 |
| L144 | (S)-MorfPhos | 185449-81-4 | C24H20NO3P |

| | | | |
|------|---|--------------|----------------|
| L145 | (R,R,R)-Ph-SKP | 1360823-43-3 | C44H38O2P2 |
| L146 | (S,R)-N-PINAP | 1173836-08-2 | C38H30N3P |
| L147 | (R)-CTH-P-Phos | 221012-82-4 | C38H34N2O4P2 |
| L148 | (R)-SIPHOS-PE | 500997-69-3 | C33H32NO2P |
| L149 | (R)-Tol-GarPhos | 1365531-81-2 | C44H44O4P2 |
| L150 | (R)-DTB-SpiroSAP-Ph | 1809609-38-8 | C53H66NPS |
| L151 | SL-N004-1 | 1226898-27-6 | C29H30FeNOP |
| L152 | SL-N011-2 | 950201-43-1 | C36H32FeNOP |
| L153 | (S,S,S,S)-BIBOP | 1202033-17-7 | C22H28O2P2 |
| L154 | SL-N009-2 | 706814-27-9 | C32H24F12FeNOP |
| L155 | SL-J408-1 | 950982-69-1 | C44H48FeP2 |
| L156 | (2R,2R)-2,2-bis(diphenylphosphino)-1,1-biferrocene | 136274-57-2 | C44H36Fe2P2 |
| L157 | (R)-Cy-GarPhos | 2829282-18-8 | C40H60O4P2 |
| L158 | (R)-DTB-SpiroPAP-6-Me | 1298133-26-2 | C52H65N2P |
| L159 | Exo-4-Methoxyphenyl Kwon [2.2.1] Bicyclic Phosphine | 1975180-37-0 | C19H22NO3PS |
| L160 | Endo-4-Methoxyphenyl Kwon [2.2.1] Bicyclic Phosphine | 1883493-01-3 | C19H22NO3PS |
| L161 | (R,R)-(Diphenylphosphino)-phenylbenzeneethanamine | 1091606-68-6 | C26H24NP |
| L162 | (1R,2R)-2-(Diphenylphosphino)-2,3-dihydro-1H-inden-1-amine | 1091606-70-0 | C21H20NP |
| L163 | (S,S)-tBuPh-SKEWPhos | 911415-22-0 | C45H62P2 |
| L164 | (R,R)-(S,S)-PhTRAP | 137096-37-8 | C48H44Fe2P2 |
| L165 | (R)-BINAPhane | 253311-88-5 | C50H36P2 |
| L166 | (1R)-8-(Diphenylphosphino)-1,2,3,4-tetrahydro-1-naphthalenamine | 960128-64-7 | C22H22NP |
| L167 | (R,R)-iPr-DuPhos | 136705-65-2 | C26H44P2 |
| L168 | (3R)-4-[2,6-Bis(1-methylethoxy)phenyl]-3-(1,1-dimethylethyl)-2,3-dihydro-1,3-benzoxaphosphole | 1338454-38-8 | C23H31O3P |
| L169 | SL-M002-1 | 494227-35-9 | C52H74FeN2P2 |
| L170 | (S)-DTBM-BINAP | 541502-07-2 | C80H104O4P2 |

| | | | |
|------|--|--------------|--|
| L171 | (S,S,S,S)-Et-BABIBOP | 2415751-83-4 | C ₂₆ H ₃₆ O ₂ P ₂ |
| L172 | (R,R,R,R)-WingPhos | 1884680-45-8 | C ₅₀ H ₄₄ O ₂ P ₂ |
| L173 | 2-[(2S,3S)-3-(1,1-Dimethylethyl)-2,3-dihydro-4-methoxy-1,3-benzoxaphosphol-2-yl]pyridine | 2565792-52-9 | C ₁₇ H ₂₀ NO ₂ P |
| L174 | SL-J681-1 | 1221745-90-9 | C ₂₈ H ₃₂ FeOP ₂ |
| L175 | (S,Sp)-p-Tol-TaniaPhos | NA | C ₄₇ H ₄₇ FeNP ₂ |
| L176 | (R,Rp)-2-Furyl-TaniaPhos | NA | C ₃₅ H ₃₁ FeNO ₄ P ₂ |
| L177 | (R)-DM-MorfPhos | 864529-90-8 | C ₂₇ H ₂₆ NO ₂ P |
| L178 | (R)-C2-TunePhos | 301847-88-1 | C ₃₈ H ₃₀ O ₂ P ₂ |
| L179 | (R)-QUINAP | 149341-34-4 | C ₃₁ H ₂₂ NP |
| L180 | SL-J015-1 | 649559-65-9 | C ₃₆ H ₃₆ FeO ₂ P ₂ |
| L181 | SL-J403-1 | 166172-60-7 | C ₄₀ H ₂₈ F ₁₂ FeP ₂ |
| L182 | SL-J425-1 | 849924-49-8 | C ₄₄ H ₄₈ FeO ₂ P ₂ |
| L183 | (R,R)-CyPP | 70774-28-6 | C ₃₂ H ₃₄ P ₂ |
| L184 | (R,R)-MeO-BoQPhos | 1542796-16-6 | C ₁₈ H ₂₂ NO ₃ P |
| L185 | 2-[(2R,3R)-4-(2,6-Dimethoxyphenyl)-3-(1,1-dimethylethyl)-2,3-dihydro-1,3-benzoxaphosphol-2-yl]-6-methoxypyridine | 2565792-77-8 | C ₂₅ H ₂₈ NO ₄ P |
| L186 | 2-[(2R,3R)-4-(9-Anthracenyl)-3-(1,1-dimethylethyl)-2,3-dihydro-1,3-benzoxaphosphol-2-yl]pyridine | 1542796-14-4 | C ₃₀ H ₂₆ NOP |
| L187 | (S)-SunPhos | 765312-54-7 | C ₄₂ H ₃₆ O ₄ P ₂ |
| L188 | (1R)-1-[Bis[3,5-bis(1,1-dimethylethyl)-4-methoxyphenyl]phosphino]-2-[(1R)-1-(dicyclohexylphosphino)ethyl]ferrocene | 1453803-83-2 | C ₅₄ H ₈₀ FeO ₂ P ₂ |
| L189 | (1R,4R)-1,4-dimethyl-1,4-butanediylbis(diphenylphosphine) | 142494-67-5 | C ₃₀ H ₃₂ P ₂ |
| L190 | (2R,3R)-4-(9-Anthracenyl)-3-(1,1-dimethylethyl)-2,3-dihydro-2-(1-methylethyl)-1,3-benzoxaphosphole | 1891002-60-0 | C ₂₈ H ₂₉ OP |
| L191 | (S,S)-XantPhos | 2119686-35-8 | C ₄₁ H ₃₆ O ₃ P ₂ |
| L192 | (3R)-3-(1,1-Dimethylethyl)-4-(2,6-diphenoxyphenyl)-2,3-dihydro-1,3-benzoxaphosphole | 1441830-74-5 | C ₂₉ H ₂₇ O ₃ P |

E.3 Descriptor definitions

Table E.2: Descriptor definitions. The package column indicates both the parsing and the calculator packages (e.g. Gaussian as NBO charge calculator; cclib as the parsing package).

| Descriptor | Definition | Package |
|---------------------------|---|---------------------|
| Bite angle | See definition in Table 2.1 | Morfeus |
| Cone angle | See definition in Table 2.1 | Morfeus |
| Buried volume | See definition in Table 2.1 | Morfeus |
| Quadrant BV (4 quadrants) | North-West, North-East, South-West, South-East quadrant buried volumes. | Morfeus |
| Octant BV (8 octants) | See definitions in Morfeus' documentation [52]. | Morfeus |
| nbo_charge_Rh_dft | NBO charge at the metal center. | Gaussian 16, cclib |
| min_NBO_donor | Minimum NBO charge at the donors. | Gaussian 16, cclib |
| max_NBO_donor | Maximum NBO charge at the donors. | Gaussian 16, cclib |
| min_bv_donor | Minimum buried volume at the donors. | Morfeus |
| max_bv_donor | Maximum NBO charge at the donors. | Morfeus |
| Lone pair occupancy | Electron occupancy of the lone pair between the metal center and the donors | Gaussian 16, OBeLiX |
| L | See definition in Fig 3.4 | Morfeus, Sterimol |
| B1 | See definition in Fig 3.4 | Morfeus, Sterimol |
| B5 | See definition in Fig 3.4 | Morfeus, Sterimol |
| std_quad | Standard deviation of quadrant buried volumes. | Morfeus |
| std_oct | Standard deviation of octant buried volumes. | Morfeus |
| homo_energy_dft | See definition in Table 2.1 | Gaussian 16, cclib |
| lumo_energy_dft | See definition in Table 2.1 | Gaussian 16, cclib |
| HOMO.LUMO_gap | See definition in Table 2.1 | Morfeus/Gaussian 16 |
| Fingerprint (0-N) | Fingerprints of the substrates are given as the difference in fingerprint bits among the investigated substrates. | RDKit |

E.4 Data availability

The data will be available with the publication of the OBeLiX workflow and the full results of this study.

