

Optimisation of *in silico* techniques for homogeneous transition metal based catalysis research

Mark Heezen

 **TU**Delft

Optimisation of *in silico* techniques for homogeneous transition metal based catalysis research

by

Mark Heezen

to obtain the degree of Bachelor of Science
from Delft University of Technology & Leiden University,
to be defended publicly on July 13th, 2022 at 01:30 PM.

Performed at:

Inorganic Systems Engineering
Department of Chemical Engineering
Faculty of Applied Sciences
Delft University of Technology

Under supervision of:

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

Student number: 5188776/S2473089
Project duration: April 13, 2022 – July 13th, 2022
Thesis committee: Prof. Dr. E. A. Pidko, TU Delft, supervisor
Prof. Dr. F. C. Grozema, TU Delft

This thesis is confidential and cannot be made public until July 13th, 2022.

Abstract

Many drugs cannot be made without homogeneous catalysis. To increase the yield of drug synthesis, the search for new catalysts continues. Computational catalysis is becoming a more prominent tool since it enables to screen many catalysts without performing (m)any laboratory experiments. In this research a computational workflow has been created to calculate both electronic (e.g. dipole, ionisation potential, nucleophilicity, etc.) and steric (bite angle, buried volume, cone angle, etc.) molecular descriptors. The structures were automatically created starting from the metal centre, some bidentate phosphorus ligands, auxiliary ligands, and the substrate. An in-house computational workflow, MACE, is used for high-throughput generation of structures from the starting bidentate phosphorus ligands, by generating stereo-isomers around to metal centre. Afterwards small substituents (H, CH₃, Ph, etc.) are changed by ChemSpaX increasing the number of structures combinatorically. To reduce the computational cost of this workflow, it has been researched whether properties of the octahedral geometry could be predicted using properties from a simple model structure containing only the metal centre and the bidentate phosphorus ligand. This model structure did not show a correlation except for the electronic energy and the solvent accessible surface area, which are both primarily influenced by the number of electrons. Other correlations may be found if some other descriptors were calculated which were excluded now, like the HOMO-LUMO gap and the substrate binding energy. The workflow could be extended to machine learning and improved by including symmetry and optical isomerism.

Contents

List of Figures	iv
List of Tables	v
1 Introduction	1
1.1 Catalysis	1
1.2 Hydrogenation of imines	1
1.3 Computational catalysis	1
1.4 Research at Inorganic Systems Engineering (ISE)	2
2 Theory	3
2.1 Bidentate TM complexes	3
2.2 Structure optimisation	4
2.2.1 Density Functional Theory.	4
2.2.2 Extended Tight Binding (xTB) DFT.	5
2.2.3 Universal Force Field.	6
2.3 Descriptors	6
2.3.1 Electronic descriptors	6
2.3.2 Steric descriptors	7
2.3.3 Boltzmann average	8
3 Method	9
3.1 Reference and model structure	9
3.2 Structure generation	9
3.2.1 Backbones and substitutes.	10
3.2.2 MetAl Complexes Embedding (MACE)	10
3.2.3 Conversion of chemical structures	10
3.2.4 ChemSpaX.	11
3.3 Optimisation & conformer search.	11
3.4 Descriptor calculation	12
3.5 Descriptor comparison	12
4 Results & Discussion	13
4.1 Structure generation	13
4.1.1 Backbones and substitutes.	13
4.1.2 MACE	13
4.1.3 ChemSpaX.	14
4.2 Descriptors	14
5 Workflow discussion	16
5.1 Structure generation	16
5.1.1 Backbones and substitutes.	16
5.1.2 MACE	16
5.1.3 ChemSpaX.	17
5.2 Optimisation & conformer search.	17
5.3 Descriptor calculation	18
5.4 Descriptor comparison	18
5.5 Code & data availability.	18

6	Conclusion & Outlook	19
6.1	Conclusion	19
6.2	Outlook	19
6.2.1	Symmetry	19
6.2.2	Stereochemistry	19
6.2.3	Data storage	20
6.2.4	Machine Learning	20
6.2.5	Energy descriptors	20
	Acknowledgements	21
	Bibliography	22
A	CREST optimisation	27
B	Starting structures	28
C	Backbones	30
D	Substitutes	34
E	Descriptors correlation	35
F	Code & data availability	36

List of Figures

1.1	Hydrogenation of 2-Methyl-1-pyrroline.	2
2.1	OH, OH bidentate, SP, and SP bidentate geometry of an iridium complex.	3
2.2	A Jacob's ladder which ranks ways to solve for the exchange correlation energy from the hartree world to chemical heaven.	5
2.3	Bite angle in a bidentate complex.	7
2.4	Steric map from which the buried volume can be calculated.	8
2.5	Cone angle	8
3.1	Schematic overview of the computational workflow.	9
3.2	OH reference structure & SP model structure.	10
3.3	2 examples of backbones.	10
3.4	Example complexes after processing by MACE.	11
3.5	Example structures after processing by ChemSpaX.	11
4.1	Energy & SASA correlation between the OH- and BD-structure.	14
4.2	Lowest in energy CREST conformers from bb#6 with as substituents 2 H-atoms, CHCH ₃ CH ₃ and CH ₂ CH ₃ for all MACE conformers.	15
5.1	ChemSpaX optimisation error.	17
5.2	Duration of the CREST optimisation at a different number of cores.	18
A.1	Bonds inserted by GFN2- χ TB or UFF optimisation.	27

List of Tables

4.1	Number of isomers for OH and BD structure by bb#	13
4.2	r and r^2 values per descriptor for a part of the structures ($n = 147$) of bb# 6	14
B.1	Starting point structures.	28
C.1	Successfully generated structures.	30
C.2	Unsuccessfully generated structures.	32
D.1	Used substituents to enlarge the number of structures.	34
E.1	Cross correlated descriptors	35
F.1	Generated data and storage location	36
F.2	Generated code and storage location	37

Abbreviations and acronyms

ACN	Acetonitrile
bb#	Backbone number
BD	Bidentate
BFGS	Broyden–Fletcher–Goldfarb–Shanno algorithm
ChemE	Chemical Engineering
CREST	Conformer-rotamer ensemble sampling tool
csv	Comma separated values
DFT	Density functional theory
EA	Electron affinity
EF	Electrofugality
EP	Electrophilicity
GTO	Gaussian-type orbitals
HOMO	Highest occupied molecular orbital
IP	Ionisation potential
ISE	Inorganic Systems Engineering
LCAO	Linear combination of atomic orbitals
LUMO	Lowest unoccupied molecular orbital
MACE	Metal complexes embedding
MO	Molecular orbitals
Morfeus	Molecular features for machine learning
NF	Nucleofugality
NP	Nucleophilicity
OH	Octahedral
SASA	Solvent accessible surface area
SP	Square planar
TM	Transition metal
UFF	Universal force field
xTB	Extended tight binding

Introduction

Catalysis, in chemistry, is the modification of the rate of a chemical reaction, usually an acceleration, by addition of a substance not consumed during the reaction.

1.1. Catalysis

The above definition comes from ‘catalysis’ in Britannia Encyclopedia [1]. The referred modification of reaction rates can be over a broad range of systems, from industry to human bodies. For all these processes catalysts are available. Catalysts can be divided into three categories: biocatalysts, heterogeneous catalysts and homogeneous catalysts. Biocatalysis is the use of enzymes for chemical reactions. This field has much success currently, due to advanced tools for enzyme discovery with high-throughput laboratory environments [2]. Heterogeneous catalysts are defined as catalysts that are in a different aggregation state as the substrate. Most of the time, the catalyst is in the solid state. This type of catalysis is mostly used in (petro)chemical industry [3]. Homogeneous catalysis is defined as catalysis with the catalyst and the substrate in the same phase. Most of the time, this is the liquid phase [4]. In this research homogeneous catalysis is meant as homogeneous catalysis by transition metal (TM) complexes. Nevertheless, homogeneous catalysis is also possible by acids and bases [5]. Homogeneous catalysis has a few advantages over heterogeneous catalysis, namely its high selectivity, easy variability due to change of ligands and above all the ability to perform asymmetric catalysis. This means that a catalyst can guide the reaction to a specific enantiomer, resulting in an enantiomeric excess [4, 5]. Enantioselectivity is an important property since different enantiomers can have different properties in for example the human body by their rate of metabolism, potency and selectivity for receptors, and toxicity [6]. This results in enantioselectivity as a demand in drug synthesis.

1.2. Hydrogenation of imines

The reaction studied in this research is the asymmetric hydrogenation of the imine 2-Methyl-1-pyrroline to 2-Methylpyrrolidine (an amine), which is shown in Figure 1.1. The solvent is CH_2Cl_2 . In general, the hydrogenation of imines is interesting because stereospecific amines are commonly used for pharmaceuticals, agrochemicals, and fine chemicals [7, 8]. These stereospecific amines can be made by direct hydrogenation and by transfer hydrogenation. The difference between these methods is the donor for the H-atoms. In case of direct hydrogenation this is H_2 , as shown in Figure 1.1. In case of transfer hydrogenation, the donor is a small organic molecule like isopropanol or formic acid, which is less favourable due to the production of extra waste [9]. Both reactions are homogeneously catalysed. This is mostly done by an iridium-based catalyst [10], but catalysts based on ruthenium, rhodium and palladium are used as well [8]. This research focuses on an iridium based catalyst. The mechanism for both hydrogenation methods is still unknown and researched in for example [11].

1.3. Computational catalysis

The research into catalysts started in the nineteenth century. Until the 1950s, computational chemistry did not play any role in catalysis research [12]. New catalysts were discovered by the trial-and-error approach

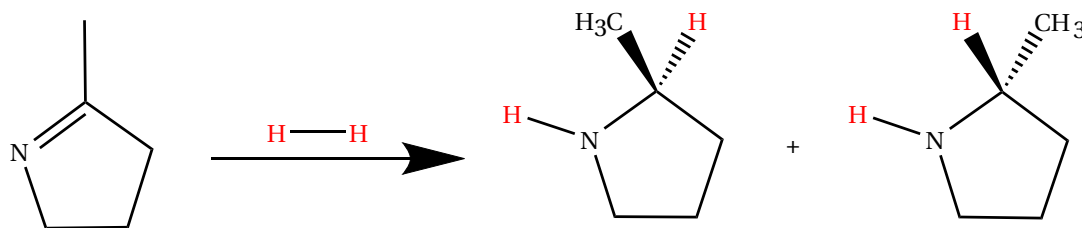


Figure 1.1: Hydrogenation of 2-Methyl-1-pyrroline.

in experiments [13]. From the 1950s, relationships, based on molecular descriptors started to dive up in research, from which Tolman's [14] are most well-known [15]. These descriptors can be calculated at different levels of theory corresponding with the quality of the descriptor. This opened the way for computational chemistry in the field of catalysis. From the 1990s this field became more prominent. Both for confirming hypotheses about reaction pathways as well as data approaches based on descriptor models [16]. This development was accelerated by the advancement of computational techniques like Density Functional Theory (DFT), which is one of the levels of theory used to calculate molecular descriptors. The 'holy grail' in computational catalysis engineering is the scenario that an entirely computational process leads to one proposed catalyst based on user-defined criteria like activity, availability, toxicity and cost by using data from the reaction mechanism and the use of machine learned properties about chemicals. This is not possible yet. Currently, breakthroughs in catalysis engineering are with the help of computational models, experimental procedures and dialogues between experts. This is partly due to the large computational cost involved to automate the entire process [17].

1.4. Research at Inorganic Systems Engineering (ISE)

The ISE-group, headed by Prof. Dr. E.A. Pidko, at Delft University of Technology, combines theoretical and practical knowledge of both hetero- and homogeneous catalysis to come up with new catalysts. Recently, the group started with a new road into data science & automation as part of the theoretical side of the group. This well-integrated approach is noticed by industry and ISE works together with multiple chemical and pharmaceutical companies.

The bigger picture of the research for this project is a question coming from a pharmaceutical company to use the new data science & automation approach together with the knowledge of mechanism research to suggest a new catalyst for the hydrogenation of 2-Methyl-1-pyrroline (section 1.2). The company possesses high-throughput laboratory equipment but is not able to process this data together with information from the mechanism and computational workflows. This is where ISE steps in, due to the strong integration between theory and practice in the group.

This research builds towards a computational workflow for catalyst discovery. In this research the goal was to create a workflow for chemical descriptor calculation and correlation. Furthermore, it aims to find a way to reduce the computational cost of this workflow by using a model structure to save computational cost. In chapter 2 the necessary theory to understand this thesis is explained. Afterwards, in chapter 3 the followed procedure is explained. This procedure led to the results which are explained and discussed in chapter 4. The computational workflow, which is explained in chapter 3, is discussed in chapter 5. The thesis ends with the conclusions and an outlook in chapter 6.

2

Theory

In this chapter relevant background information is given for a better understanding of this thesis. In section 2.1 a short introduction in inorganic chemistry is given. Section 2.2 explains the theoretical background on structure optimisation, specifically density functional theory and extended tight binding. This section assumes a basic knowledge of theoretical chemistry. This chapter ends in section 2.3 with an overview of molecular descriptors.

2.1. Bidentate TM complexes

TM complexes consist of a metal atom or ion bound via dative bonds to so-called 'ligands'. Ligands are defined as any atom, ion or neutral molecule capable of donating an electron pair to bond to the central metal ion or atom through secondary valency [18]. The metal centre considered in this research is Ir^{3+} . Ir^{3+} has 6 valence electrons. Using the 18 electron rule by Langmuir [19], 6 ligands (each counting for 2 electrons) can bind to the iridium. This results in an octahedral (OH) structure in 3D which can be seen in Figure 2.1(a).

Complexes with a ligand bound twice to the metal centre by two different atoms are so-called bidentate complexes. The atoms that connect to the metal centre are called chelating atoms. This research was limited to phosphorus atoms as chelating atoms. An example of a bidentate complex with phosphorus atoms as chelating atoms is represented in Figure 2.1(b).

A geometry for metal centres with 4 ligands is square planar (SP). This geometry is shown in Figure 2.1(c). Metal centres that can be found in SP are metal centres with 8 valence electrons, such as Ir^+ . Metal centres with 8 valence electrons and 4 ligands, contributing 2 electrons each, result in 18 electrons which is preferred according to Langmuir's rule.

An SP structure can be generated from an OH structure by removing 2 auxiliary ligands. In the case of an bidentate ligand this results in a complex as shown in Figure 2.1(d).

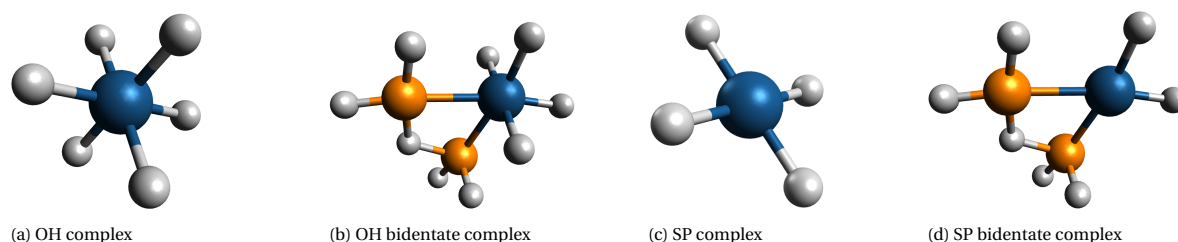


Figure 2.1: OH (2.1(a)), OH bidentate (2.1(b)), SP (2.1(c)), and SP bidentate (2.1(d)) geometry of an iridium complex. The blue spheres represent the iridium cores, the white spheres the (auxiliary) ligands and the orange sphere the phosphorus chelating atoms.

2.2. Structure optimisation

2.2.1. Density Functional Theory

Density Functional Theory (DFT) is one of the most popular quantum mechanical tools present. Its goal is the quantitative understanding of material properties from the fundamental laws of quantum mechanics [20]. Despite the original development for chemistry and material sciences, the applications from DFT range from geosciences until biology nowadays [21, 22]. Density functional theory is a method to approximate solutions to the Schrödinger equation, the fundamental equation in quantum chemistry which is shown in equation 2.1 [23].

$$\hat{H}\Psi = E\Psi \quad (2.1)$$

In this equation \hat{H} is the Hamiltonian, which is an operator working on the many-bodies wavefunction Ψ and E is the energy associated with this wavefunction. The Hamiltonian for a system with multiple electrons and nuclei is given by Equation 2.2.

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{V}_{nn} + \hat{V}_{ee} + \hat{V}_{en} \quad (2.2)$$

In this equation \hat{T}_e and \hat{T}_n are the sum over the kinetic energies of the electrons and the nuclei, respectively. \hat{V}_{nn} and \hat{V}_{ee} represent repulsion between the nuclei and between the electrons. At last, \hat{V}_{en} describes the attractive interaction between the electrons and the nuclei.

To approximate the solutions to the Schrödinger equation some approximations have to be made. First the Born-Oppenheimer approximation is used. The Born-Oppenheimer approximation states that the nuclei are fixed compared to the electrons due to their difference in mass. Since the mass of the proton is 1836 times higher than the electron's mass, the nuclei have a much smaller velocity. This approximation results in the ability to separate the many-bodies wavefunction Ψ in an electronic wavefunction ψ depending on the coordinates of both the electrons and nuclei and a nuclear wavefunction Φ depending on the coordinates of the nuclei only [24].

The Hamiltonian for the electronic wavefunction \hat{H}_{elec} is given in Equation 2.3.

$$\hat{H}_{elec} = \hat{T}_e + \hat{T}_n + \hat{V}_{ee} + \hat{V}_{ext} \quad (2.3)$$

In this equation \hat{V}_{ext} is a constant external potential given by the static nuclei and is given by the sum over the potential of all nuclei. Thomas and Fermi came up with the idea of using the electron density $\rho(\vec{r})$ as the central variable for quantum mechanical calculations instead of the many-electrons wavefunction. This concept leads to a significant decrease in variables from 3 variables per electron (x, y and z coordinate) to 3 variables with the use of a density. This also means that the number of variables does not depend on the number of electrons anymore, which makes this concept suitable for systems with a large number of electrons. Hohenberg and Kohn used this idea to formulate their theorems which are the basis for DFT [25].

The first Hohenberg-Kohn theorem states that the external potential \hat{V}_{ext} is a functional of the electron density only, resulting in the use of only the electron density to determine all properties. Their second theorem states that the variational principle applies [26]. The variational principle proves that every energy calculated by DFT is always larger than the energy of the ground state of the system. This means that the best energy of the ground state is the lowest energy found, and the best wavefunction is the wavefunction that corresponds to the lowest energy. The derivation of both postulates is beyond the scope of this research and can for example be found in [25].

These postulates are used in the Kohn-Sham approach which transforms a system of interacting electrons in a static external potential to a system of non-interacting electrons in an effective potential, named the Kohn-Sham potential $v_{KS}(\vec{r})$ [27]. This results in a set of single-particle equations instead of coupled Schrödinger equations, which are easier to solve [25]. Due to a difference in the exact kinetic energy $\hat{T}[\rho(\vec{r})]$ and the kinetic energy of a non-interaction electron gas (also called the Hartree energy) $\hat{T}_S[\rho(\vec{r})]$ and the difference between the exact electron-electron interaction energy and the classical electrostatic energy a new terms comes into play; the exchange-correlation energy $E_{XC}[\rho(\vec{r})]$ which is defined in Equation 2.4 [28].

$$E_{XC}[\rho(\vec{r})] = \hat{T}[\rho(\vec{r})] - \hat{T}_S[\rho(\vec{r})] + E_{ee}[\rho(\vec{r})] - E_H[\rho(\vec{r})] \quad (2.4)$$

Current research in theoretical chemistry aims to find the best functional to solve for the exchange-correlation energy [29]. The functionals to solve for the exchange-correlation energy are ranked in a Jacob's ladder as proposed by Perdew [30]. A figure of the Jacob's ladder can be seen in Figure 2.2. This ladder ranks

from the Hartree world (or sometimes called Hartree hell), which means no exchange-correlation energy, to chemical heaven. A method higher on the ladder means a higher chemical accuracy but an increased computational cost. It is the power of a theoretical or computational chemist to use the best functional to balance between chemical accuracy and computational cost for their purpose.

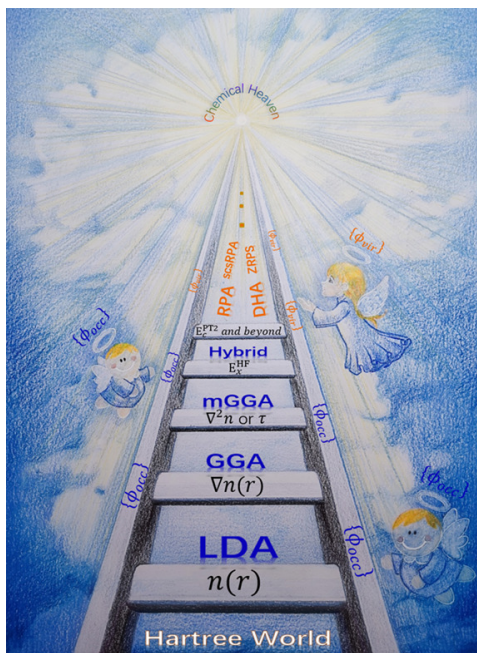


Figure 2.2: A Jacob's ladder which ranks ways to solve for the exchange correlation energy from the hartree world to chemical heaven. Higher on the ladder means a higher chemical accuracy but also increased computational cost. Image taken from [31].

Not only the operators have to be defined. A wavefunction has to be constructed as well. An electron wavefunction is constructed from a linear combination of numerical functions, so called basis functions [32]. These basis functions can be constructed in different ways. Gaussian-type orbitals (GTOs) are commonly used, due to their balance between accuracy and computational cost. GTOs are constructed through a linear combination of Slater atomic orbitals (LCAO). This means that the wavefunction for every Slater atomic orbital corresponds with a constant, which is determined with the DFT-calculation, to the GTO. The representation of atoms can be combined to the representation of molecules by multiplying these GTOs. The Gaussian Product Theorem predicts that two GTOs of different atoms multiplied yield a finite sum of Gaussian integrals centred on a point between the atoms. This reduces the number of integrals and therefore the computational cost [33].

Until now no interactions with other molecules have been taken into account. This is problematic since the reaction considered in this research takes place in a solvent. A solvent can be taken into account in two ways. The explicit solvent model models the solvent molecules explicitly. In this way the most realistic solvation is generated, but it is computationally very expensive. The implicit solvent model assumes a homogeneous polarizable medium for the solvent, thus taking effects of the charge distribution and the geometry coming from the solvent into account [25].

2.2.2. Extended Tight Binding (xTB) DFT

Grimme and coworkers developed GFN n -xTB with $n = 0, 1, 2$. GFN n -xTB are semiempirical DFT functionals meant for calculations on **G**eometries, **F**requencies and **N**oncovalent interactions [34]. Multiple versions of GFN n -xTB have been published. A semiempirical method is a method that uses systematic approximations to get to efficient computational workflows. These workflows are several orders of magnitude faster than *ab initio* calculations. These workflows, together with the use of modern supercomputers, lead to many applications for computational chemists. The downside of semiempirical DFT methods is their decrease in accuracy, which makes them not usable for highly quantitative computational studies [35].

The goal of GFN n -xTB is to describe large (1000 atoms or more) systems, specifically in chemistry and biology. Due to the computational cost it is impossible to do this with *ab initio* calculations yet. The first version

of GFN- x TB, later named GFN1- x TB was released in 2017 and started from DFTB3, which is a DFT-functional published in 2011. DFTB3 uses a third order approximation of the density $\rho(\vec{r})$ around a reference density to reduce computational cost. Corrections are made for Coulomb interactions between partial charges [36]. The changes made in GFN1- x TB compared to DFTB3 are the use of many element-specific parameters instead of the use of element-pair-specific parameters. These element-specific parameters have been found by fitting up to $Z = 86$ (Radon), which makes GFN1- x TB useful for many systems [37].

GFN1- x TB was improved to GFN2- x TB which was released in 2019. It gave more accurate and physically solid results than GFN1- x TB without a noticeable increase in computational demands [38]. The biggest flaw in GFN1- x TB was an inaccurate description of London-dispersion interactions. In 2019 a better model to describe London-dispersion interactions was released, called D4 [39]. This model is incorporated in GFN2- x TB. The second mentioned enhancement is a change in the description of electrostatic interactions between atoms. In GFN1- x TB all electrostatic interactions were accounted as spherical. In GFN2- x TB these interactions are modelled as quadrupole, which is a physically better description. The last improvement mentioned is the absence of the last element-pair-specific parameters and corrections, for example in H-H, N-H or halogen bonds. Only fitted element-specific parameters are used in GFN2- x TB. Besides this some minor improvements have been made which can be found in [38]. In the same article, the benchmarks of GFN2- x TB can be found which show a significant increase in performance compared with GFN1- x TB.

GFN n - x TB is still under development, with an attempt to reduce the computational resources even further in GFN0- x TB [34] and the recent introduction of a force field version called GFN-FF [40].

2.2.3. Universal Force Field

The last discussed method for structure optimisation is Universal Force Field (UFF). Force Field methods are computationally cheap methods to do rough optimisations of molecules based on information of the element, hybridisation and connectivity of atoms. UFF puts the atoms from a molecule in a force field that consists of the potential energies from the terms given in Equation 2.5.

$$E = E_R + E_\theta + E_\phi + E_\omega + E_{vdw} + E_{el} \quad (2.5)$$

In Equation 2.5, E_R represents the bond stretching according to a harmonic oscillator or a Morse potential. E_θ , E_ϕ and E_ω represent angular corrections for the angle bend, angle torsion and inversion respectively. E_{vdw} describes the Van der Waals interactions and E_{el} describes the electrostatic interactions [41].

2.3. Descriptors

Molecules can be numerically represented using descriptors. These descriptors give quantitative information over a certain property of a molecule. This numerical description is used for a computer to understand a molecule, since it can only do data analysis on numbers. Descriptors can be divided in two categories: 1) electronic properties (discussed in subsection 2.3.1), which represent electronic properties and 2) steric descriptors (discussed in subsection 2.3.2), which describe spatial properties.

2.3.1. Electronic descriptors

Dipole

The dipole moment is a vector which represents the distribution of charges over a molecule due to a difference of electronegativity between atoms. The dipole is the length of this vector [42].

Ionisation potential and electron affinity

The ionisation potential is defined as the change in energy between the molecule and the molecule as cation, so after removing an electron. This energy barrier gives information about the likelihood for the molecule to lose an electron. The electron affinity is the opposite change. It is the difference between the molecule and the molecule as anion, i.e. after adding an electron [43].

Nucleophilicity and electrophilicity

Nucleophilicity and electrophilicity describe the eagerness of molecules to donate or accept electrons respectively, and therefore make chemical bonds [42]. The electrophilicity and nucleophilicity are calculated with IP the ionisation potential and EA the electron affinity using the equations in Equation 2.6 and 2.7 [44].

$$EP = \frac{(IP + EA)^2}{IP - EA} \quad (2.6)$$

$$NP = -IP \quad (2.7)$$

Nucleofugality and electrofugality

Nucleofugality (NF) and electrofugality (EF) are used to describe chemical groups eager to leave with or without an electron pair, to complete their octet [45]. The electrofugality and nucleofugality are calculated as in Equation 2.8 and 2.9, with IP and EA still the ionisation potential and electron affinity respectively.

$$EF = \frac{(3IP - EA)^2}{8(IP - EA)} \quad (2.8)$$

$$NF = \frac{(IP - 3EA)^2}{8(IP - EA)} \quad (2.9)$$

Highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO)

The Linear Combination of Atomic Orbitals (LCAO) - Molecular Orbitals (MO) theory, as introduced by Lennard-Jones [46] states that by combining atoms into molecules electrons are combined in molecular orbitals, originating from their atomic orbitals. By arranging all possible MO's from the lowest to the highest, these orbitals get filled until some point. The orbital that is (partially) filled is called HOMO and the orbital that is just not filled is called LUMO. The energies of these orbitals are used as the descriptor.

2.3.2. Steric descriptors

Bite angle

The bite angle is defined as the angle between the two chelating ligand atoms and the metal atom and is assigned with θ in Figure 2.3.

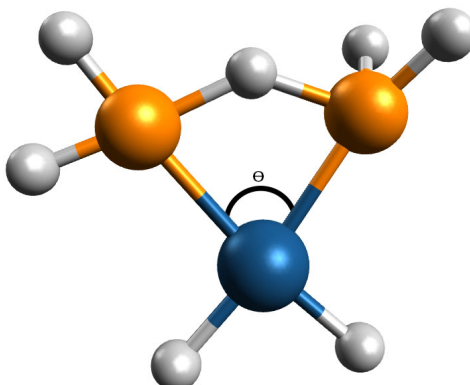


Figure 2.3: Bite angle in a bidentate complex denoted with θ .

The bite angle is a much used descriptor in inorganic chemistry since it is both influenced by steric and by electronic properties [16]. Since the descriptor itself is spatial it is categorised as a steric descriptor in this thesis.

Buried volume

For the cone angle multiple definitions can be found. In this research the definition from Cavallo and coworkers is used [47]. The buried volume is a relative value calculated from a steric map. A steric map can be seen in Figure 2.4. This figure is made by drawing a sphere of 3.5 Å around the metal centre. The first two dimensions are shown on the x and y axes. The third dimension is indicated by colour. Afterwards, it is measured which part of this sphere is occupied and thereafter divided by the total volume. The volume occupation of the metal centre and hydrogen atoms are excluded by default.

Cone angle

For the cone angle multiple definitions can be found. In this research the definition from Allen and coworkers is used [48]. They describe a systematic approach to computationally calculate the cone angle. The cone angle is the angle between two lines starting at the metal centre and tangent to the furthest atom of the ligand as can be seen in Figure 2.5. The angle is indicated with θ .

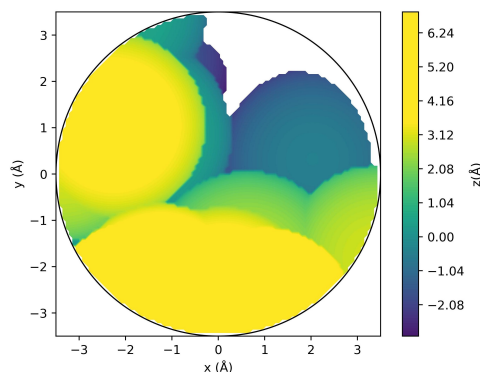


Figure 2.4: Steric map from which the buried volume can be calculated.

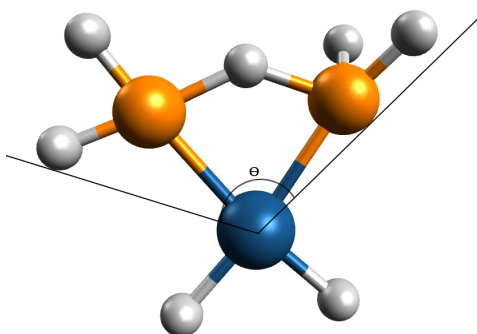


Figure 2.5: Cone angle indicated with θ between the tangent lines in black.

Dispersion

The London dispersion force is an attractive force originating from a transient dipole in atoms that induce transient dipoles in nearby atoms [49]. This dipole can be measured in all atoms separately. Pollice and Chen introduced a method to calculate a quantitative descriptor of this interaction potential [50]. This method is used with one adaption. Instead of the electron density isosurface, a surface is created from the electron densities since this saves computational cost [51]. The dispersion is always seen as the dispersion of the metal centre.

Solvent Accessible Surface Area (SASA)

The solvent accessible surface area gives information about how much area of the molecule can be reached by a solvent. A computational workflow to calculate the SASA is described in [52].

2.3.3. Boltzmann average

The Boltzmann average is a weighted average over a sample of chemical structures based on their energies with a temperature high enough that quantum mechanical effects may be neglected. The Boltzmann average for an observable A is equal to the expectation value, denoted as $\langle A \rangle$, at a certain temperature. The Boltzmann average for the observable A is defined in Equation 2.10, with E_i the energy of the chemical structure i , k_B Boltzmann's constant and T the absolute temperature [53, 54].

$$\langle A \rangle = \frac{\sum_i A e^{-E_i/k_B T}}{\sum_i e^{-E_i/k_B T}} \quad (2.10)$$

3

Method

At the start of this chapter the choice for the used reference structure and model structure is discussed. Next, the computational workflow shown in Figure 3.1, is described. First, the way of high-throughput generation of structures is explained. Afterwards, the optimisation and conformer search is discussed. Lastly, the calculation of the descriptors is explained.

3.1. Reference and model structure

It is chosen to build the reference structure as shown in Figure 3.2(a). This is an OH geometry with Ir^{3+} as the metal centre. The ligands are the chosen bidentate phosphorus ligand (shown as two connected phosphorus atoms) and 3 hydrides (H^-) as auxiliary ligands. For the substrate it is chosen to not include 2-Methyl-1-pyrroline but acetonitrile (ACN). This structure is chosen to decrease the computational power and avoid stereochemical effects. Computational cost is reduced since ACN does not contain any rotatable bonds and 2-Methyl-1-pyrroline contains one rotatable bond, which decreases the computational cost for the conformer search. After attaching an hydrogen atom in the hydrogenation reaction 2-Methyl-1-pyrroline possesses a chiral centre, resulting in two enantiomers. Since the two enantiomeric products can differ in reaction energy, calculations regarding the reaction energy get more complicated. To avoid this, ACN, as a flat linear molecule is chosen over 2-Methyl-1-pyrroline.

The model structure is shown in Figure 3.2(b). For the model structure, a flat geometry is chosen. This flat geometry might result in a decrease in computational cost over the three dimensional OH geometry when optimising the structure and calculating the descriptors. This structure is modelled with Ir^{3+} as the metal centre and the chosen bidentate phosphorus ligand. All auxiliary ligands and the substrate are not included. This means that the structure is formally not a square planar structure since it does not have four ligands at the metal structure. Therefore this structure will be named as bidentate (BD). This model structure has a charge of +3e.

3.2. Structure generation

This section consists of a part about separation of the starting structures in backbones and substitutes. After that, the packages MACE and ChemSpaX are explained for high-throughput generation of structures.

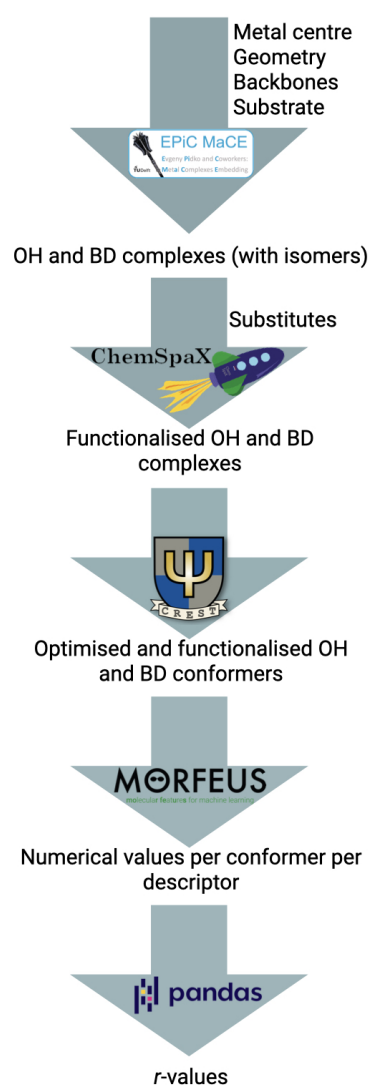


Figure 3.1: Schematic overview of the computational workflow.



Figure 3.2: Structure templates for OH (3.2(a)) and SP (3.2(b)).

3.2.1. Backbones and substitutes

At the start of this research, the industrial partner provided ninety ligands they tested in an iridium based complex for the hydrogenation of 2-Methyl-1-pyrroline. To be able to fulfill this research in the designated amount of time, the number of structures had to be limited to about twenty structures. First, structures with non-covalent bonds, like ferrocenes, were omitted, since MACE cannot process these. Second, only structures with two chelating phosphorus atoms were selected for two reasons. The first reason was that with two phosphorus chelating atoms the researchers were always able to distinguish the chelating atoms from the other atoms. The second reason was that a trend in the descriptors might easier be spotted with the same chelating atoms. From the remaining structures, 18 were chosen based on their variety in size and heteroatoms. Next, the structures were split in backbones and substitutes. The base for the backbone is everything between the two phosphorus atoms. This part was examined further and small organic groups like H, CH₃, isopropyl and phenyl were taken off and listed as substitutes. For purposes which will be explained later, all substitutes were replaced by a monovalent atom not occurring anywhere else in the structures. For this purpose bromine was chosen. Two examples of backbones with bromines as substitutes are shown in Figure 3.3.

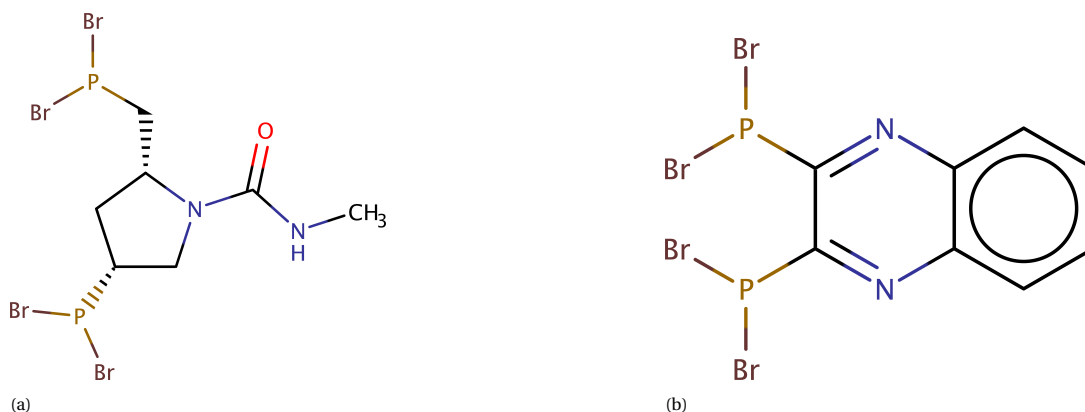


Figure 3.3: 2 examples of backbones in 3.3(a) and 3.3(b).

3.2.2. MetAl Complexes Embedding (MACE)

The backbones were fed to MACE commit 68 [55] as ligands, together with the metal centre, auxiliary ligands, and ACN for both the BD and the OH structure as specified in section 3.1. MACE is able to construct 3D coordinates for a SP or OH complex from this information in the requested geometry. MACE searches for all different stereoisomers, originating from different positions of the ligands around the metal centre. The results of MACE for the backbone presented in Figure 3.3(a) are shown in Figure 3.4. Figure 3.4(a) until 3.4(d) show four stereoisomers for the OH complex. For the BD model only one isomer could be found by MACE which is shown in Figure 3.4(e).

3.2.3. Conversion of chemical structures

MACE yields as output only a XYZ file. An XYZ file only contains the element symbol and the coordinates for each atom. For ChemSpaX both a XYZ file and an MDL Molfile are necessary. An MDL Molfile contains not only the coordinates, but includes bonding information. This information includes the atoms bonded and

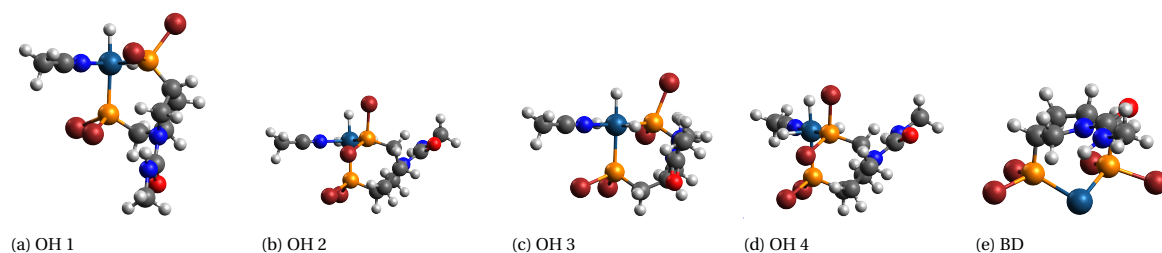


Figure 3.4: Example structures with the backbone presented in Figure 3.3(a). 3.4(a) until 3.4(d) show the four OH structures and 3.4(e) shows the only BD structure.

the bond type. The XYZ files had to be converted to MDL Molfiles. This is done with the python package openbabel version 3.1.1 [56].

3.2.4. ChemSpaX

The separation of the structures in separate backbones and substitutes opened possibilities for highthroughput generation of structures, since all substitutes can be placed on any open space on the backbone. This is done with a modified version of the package ChemSpaX commit 117 [57]. ChemSpaX replaces a certain substituent, bromine in this case, by one of the previously defined substitutes. A few example structures of the structure from MACE given in Figure 3.4(a) are given in Figure 3.5, with the substitutes indicated under the subfigures. ChemSpaX has been modified to support parallel use and take substitutes from a previously defined file which contained all options of substitutes based on combinatorics. Different substituent files were made for a different number of substitution sites. After the modification, ChemSpaX was able to increase the amount of structures combinatorially. The modification to ChemSpaX will be available on GitHub shortly [58]. After the use of ChemSpaX a conversion was necessary again. The modified version of ChemSpaX gives as output both a XYZ file and an MDL Molfile, but the XYZ file is not an optimised version of last structure. Therefore the MDL Molfiles were converted to XYZ files for the next step. This is done in the same way as indicated in subsection 3.2.3.

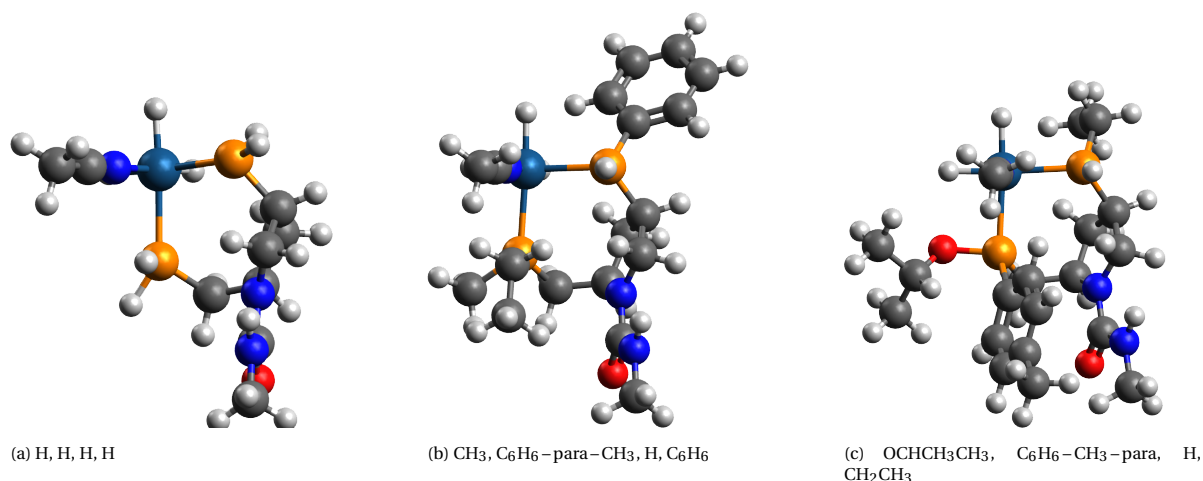


Figure 3.5: Example structures from ChemSpaX with the MACE structure presented in Figure 3.4(a). Under the subfigures the placed substitutes are indicated.

3.3. Optimisation & conformer search

Before descriptor calculation could take place, all ChemSpaX generated structures had to be optimised such that the descriptors would be calculated at a sufficient level of theory. Together with the optimisation, a conformer search is performed. This is done with the Conformer-Rotamer Ensemble Sampling Tool (CREST) developed by Grimme and coworkers [59], which uses the different versions of GFN n -xTB version 6.5.2 for optimisation [38]. First, the structures were optimised on the GFN2-xTB level. Then GFN2-xTB was used

for the OH structures in CREST. Optimising the BD structures using GFN2-*x*TB led to newly created bonds between atoms from the substrate (O, N, C and H) to the iridium core, instead of their original position. Therefore it is chosen to constrain all of these atoms during a separate GFN2-*x*TB optimisation and use a GFN-FF optimisation in CREST. This resulted in an optimisation of the Ir and P-atoms only. An example of this optimisation problem is shown in Appendix A.

3.4. Descriptor calculation

The next step was the calculation of the descriptors described in section 2.3. This is done with the python package MOleculaR FEatureS for machine learning (morfeus) version 0.6.0 [60]. This package is able to calculate the electronic descriptors, as described in subsection 2.3.1, directly over the conformer ensemble originating from CREST. This is not possible for the steric descriptors, as described in subsection 2.3.2. Therefore a python script is written that saves all conformers from the CREST conformers file in separate XYZ files. This way, the descriptors could be calculated for all descriptors. All calculated descriptors from each conformer were put in a pandas dataframe [61]. The energies found by CREST had to be converted from relative to absolute energies. The energies reported by CREST are relative to the lowest conformer. This means that the energy of the lowest conformer is set to zero. The absolute electronic energy of this conformer is found in the CREST logfile and added to all conformers. This way, all conformers have the same reference and different structures can be compared. Afterwards all properties are boltzmann averaged with a copy of the function from morfeus. This resulted in one row of descriptors per chemical structure. To be able to automate the steric descriptor calculation, the metal centre and the chelating atoms should be found in the conformer xyz-file in the same spot for every conformer. Therefore a script was written to automatically place the Ir atom as the first atom in the list and the phosphorus atoms on places 2 and 3.

3.5. Descriptor comparison

Since the goal of this research is to compare the quality of the descriptors from the BD structures with the OH structures it is chosen to take the boltzmann average for each descriptor over the structures generated by MACE. This results in two values per descriptor for every combination of backbone and substitutes. The first value is the boltzmann average over all MACE-structures for the OH structure. The second value is the boltzmann average over all MACE-structures for the BD structure. All these values were written, together with the structure information, to a pandas DataFrame. Pandas has a built in workflow to produce a correlation matrix. In this matrix all correlation coefficients, also called *r*-values, are shown between the different variables. The main diagonal shows the correlation coefficients for the same descriptor coming from the BD and OH structure.

4

Results & Discussion

This chapter builds up to the descriptor correlation, which are the main results of this research. This chapter is built up in the same structure as the Method chapter. At the start, the way of high-throughput generation of structures is discussed. Next, the optimisation and conformer search is discussed. At last, the results of the descriptor calculation are shown.

4.1. Structure generation

The industrial partner provided ninety ligands structures. From these structures 32 contained a non-covalent bond, like a ferrocene, and were therefore discarded directly. Due to the other criteria explained in subsection 3.2.1 this was reduced to 18 structures. These structures are shown Table B.1 in Appendix B.

4.1.1. Backbones and substitutes

The found backbones from the structures in Table B.1 that were successfully processed by MACE can be found in Table C.1 in Appendix C, together with the bb# (backbone number), which was used to keep track of the structures, and the number of substituent sites. The backbones that did not successfully produced isomers can be found in Table C.2 with the same properties. In both tables the places where the substituents will come are indicated by R. The number of backbones in Table C.1 and C.2 is smaller than the number of structures shown in Table B.1, due to overlapping backbones. The structures with the CAS-numbers 55739-58-7 and 64896-28-2 both have the backbone with bb# 2. The same is the case for structures with the CAS-numbers 136705-64-1 and 136705-65-2, namely bb# 8. Structures with the CAS-numbers 1202033-19-9, 1228758-57-3 and 1884680-45-8 all have bb# 13. The backbones that failed while being processed by MACE are discussed in subsection 4.1.2. The chosen substitutes are listed in Table D.1 in Appendix D.

4.1.2. MACE

MACE generated the complexes as explained in subsection 3.2.2. The number of conformers found for the OH and BD structures are listed in Table 4.1. Some backbones gave errors while being processed by MACE. MACE could not construct a single complex for bb#'s 3 and 4. After consultation of Wenjun Yang MSc, it was found that these structures have too much tension to form a bidentate complex. Therefore these structures were discarded. MACE could not produce any BD complex for bb# 10. Since both structures are needed to make the comparison between the descriptors this structure was discarded as well. Bb# 8 was discarded since it took MACE 24 hours to construct 511 OH complexes with this backbone and 24 hours to create 127 BD complexes as well. Both jobs were terminated due to the time limit which was present at DelftBlue during the post beta test phase. The most likely explanation for the termination is a non-convergence in MACE for this structure.

Bb#	OH	BD
1	5	3
2	2	1
5	2	1
6	4	1
7	2	1
9	2	1
11	8	2
12	2	1
13	8	3
14	2	1

Table 4.1: Number of isomers for OH and BD structure by bb#

4.1.3. ChemSpaX

Initially, ChemSpaX was made for manual input of the substituent in a bash file and a manual input of the substituent site in the input file. Afterwards one batch of input files could be processed in series with the same substitutes. Some changes were implemented to allow for automation. First, lists of all possible combinations of substitutes were made by combinatorics. This was done by writing every combination of substitutes to a new line of a text file. ChemSpaX was altered in such a way that the file with the right number of substitutes was read. Then a single job was submitted with every line of the text file as an argument for that job. For this to be possible, ChemSpaX was changed to accept the substitutes as input on the command line. Another update was the removal of identically named temporary files. In all temporary files the substitute names are written, so these can be distinguished. With these improvements ChemSpaX could be used in parallel to generate the structures in bulk corresponding with the backbones in Table C.1 and substitutes in Table D.1. The structures for all substitutes for bb# 6, 7, 12, 13, and 14 have been created. The structures corresponding to bb# 10 and 11 are partially generated. The order of generation was chosen based on the number of substituent sites, starting with the least substituent sites.

4.2. Descriptors

The descriptors that are calculated can be found in section 2.3. The descriptors have been calculated for 147 structures originating from bb# 6. Due to time limitations only structures with the substituent H on spot 1 (for the OH backbone) or spot 3 (for the BD backbone) have been taken into account. Explanations of the spots can be found in Table C.1. The r and r^2 -values for the relations between the OH and BD structures can be found for every descriptor in Table 4.2. All descriptor correlation values can be found in Appendix E. The only good agreement for both structures is found between the **energy** and **SASA** calculations for the OH and BD structure with an r^2 -value of 0.989 and 0.938 respectively. This is in agreement with the plots shown in Figure 4.1. The good agreement is the case for the cross references as well. These have both an r^2 value of 0.947. This can be explained by the strong correlation of these properties with the number of electrons present in the structure. A linear correlation is present between the electrons in the OH and BD structures, since they only differ with a fixed amount due to the difference in auxiliary ligands and the substrate. It is seen that for high values of the SASA the error from the trend line is substantially larger than at low SASA values. This can be explained since at high SASA values more atoms are likely to be in the molecule resulting in a larger molecule. Larger molecules can have more configurations which influence the SASA, resulting in the SASA not depending on the number of electrons only.

Table 4.2: r and r^2 values per descriptor for a part of the structures ($n = 147$) of bb# 6

Backbone number	r	r^2
Energy	0.994	0.989
Dipole	-0.148	0.022
Electron affinity	0.644	0.415
Electrophilicity	0.491	0.241
Nucleophilicity	-0.504	0.254
Electrofugality	0.341	0.116
Nucleofugality	-0.111	0.012
HOMO	-0.089	0.008
Ionisation Potential	-0.504	0.254
LUMO	0.555	0.308
Bite angle	-0.070	0.005
Burried volume	0.533	0.284
Cone angle	-0.069	0.005
Dispersion	0.554	0.307
SASA	0.968	0.938

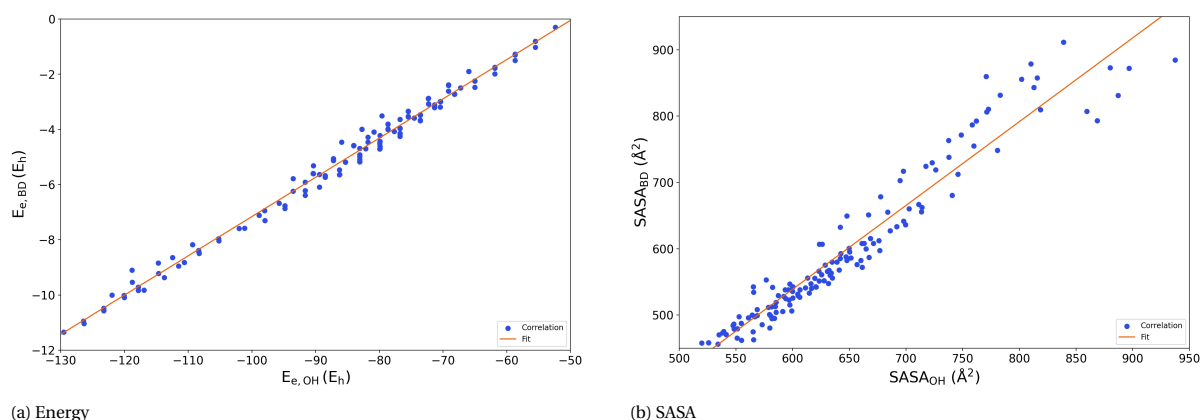


Figure 4.1: Correlation for the Energy (4.1(a)) and SASA (4.1(b)) between the OH structure (x-axis) and the BD-structure (y-axis)

Between the other electronic descriptors no (strong) correlation is found. For the **dipole** this makes sense since the dipole is influenced by many atoms and definitely by the substrate, which is only present in the OH structure and contains a nitrile. A nitrile is a very electron withdrawing group and therefore influences the dipole [62]. For the **HOMO** and **LUMO** it makes sense that these are not strongly correlated since those values are system specific. It would have been better to take a look at the HOMO-LUMO gap. Possibly a correlation can be found for this descriptor. The **electron affinity** and **ionisation potential** are most likely effected by the choice for the model structure. The model structure cannot exist due to a lack of electrons around the iridium centre. This results in an unreal urge for electrons at the iridium centre. This has the consequence of a very low electron affinity and a very high ionisation potential. The **Electrophilicity**, **nucleophilicity**, **electrofugality** and **nucleofugality** are all calculated directly from the ionisation potential and the electron affinity as shown in subsection 2.3.1. Therefore the error in the electron affinity and the ionisation potential results in an error in these descriptors. The same error results probably in the missing correlation for the **dispersion** of the iridium metal centre. The dispersion is also calculated based on the ionisation potential and the polarizabilities [50].

An absent correlation between the steric descriptors was unexpected, since some of these properties primarily depend on the metal centre and the chelating atoms only. The most likely explanation for an absence of the correlation is the lower-level optimisation for the BD structure. As discussed in section 3.3, only force field optimisation was used. UFF optimisation is not very accurate [38]. Another contribution to the error may come from the different MACE-structures. For the sake of the explanation below, all lowest in energy CREST conformers from bb# 6 with as substitutes two H-atoms, CH_2CH_3 and CHCH_3CH_3 are shown in Figure 4.2. All these molecules are visualised in the same way according to two criteria. First, the phosphorus atom with the two H-atoms was always visualised on the right side. Second, the ring with the nitrogen atoms was always visualised on top. A possible absence of correlation for the **cone angle** can easily be explained. The cone angle is, by default, calculated over all ligands. These ligands include the hydrides and the substrate in the OH geometry (as can be seen in Figure 4.2(b) until 4.2(e)) and do not in the BD model (as can be seen in Figure 4.2(a)). A relationship may be found if the hydrides and the substrate were not considered when calculating the cone angle over the OH structures. This can be done by removing the hydrides and the substrate through a substructure search. Afterwards the cone angle can be calculated without optimising the structure, so the cone angle over the leftover part does not change. The difference in ligands cannot be true for the **buried volume**, since for the buried volume calculation the hydrides, like all H-atoms, and the substrate were not included. The lack of relationship between the BD structure and the OH structures could be explained by the different MACE-structures, which are present for the OH structures but not for the BD structure. In this example a difference is seen between the spatial structure of the BD structure (Figure 4.2(a)) and OH0 (Figure 4.2(b)) resulting in a difference in buried volume. The lack of correlation between the buried volume questions the applicability of the model structure. The absence of a correlation from the **bite angle** was the most surprising since the iridium and phosphorus atoms are the only atoms optimised in the BD structure. Nevertheless the space for optimisation was probably too limited since all other atoms in the molecule were constrained.

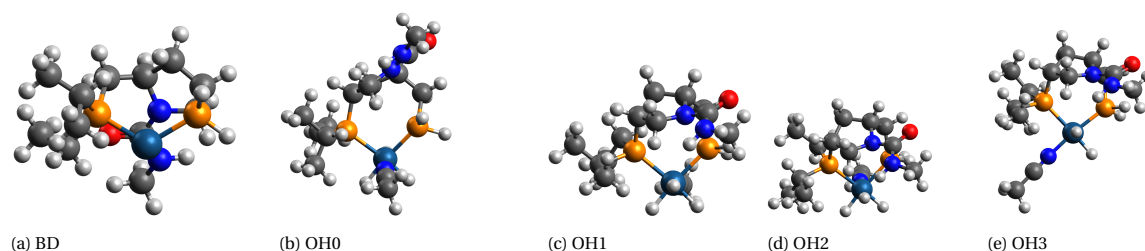


Figure 4.2: The lowest in energy CREST conformers from bb# 6 with as substituents 2 H-atoms, CHCH_3CH_3 and CH_2CH_3 are given in 4.2a until e for the different MACE isomers

5

Workflow discussion

This chapter starts with a general discussion about setting up a computational workflow. After that, it follows the same structure as in the results & discussion chapter to discuss the challenges encountered during setting up a computational workflow for high-throughput research. First, the way of high-throughput generation of structures is discussed, followed by the optimisation and conformer search. Next, the process of the descriptors calculation and comparison is discussed. This chapter ends with a paragraph concerning code & data availability.

As shown in chapter 4 the intended computational workflow has been set up during the project. The big downside of doing computational research is the need for computing hours at supercomputer facilities. During this research Snellius [63] (7,407,526), Tetralith [64] (945,710) and DelftBlue [65] (1,116,063) were used. After the supercomputer name the number of CPU-hours requested for the project is indicated in parenthesis. This number of CPU-hours is sadly much higher than it should be, since the supercomputers were initially not used efficiently. During the project it was learned how to properly request the right number of cpu hours for a job. Before that time all jobs requested the same amount of resources as an example job which used 32 cpus. For most jobs 1 or 2 cpus would have been cheaper and speed up the job, since communicating between the different cpus was the slowest step. Therefore, it is recommended to teach students how to efficiently work with a supercomputer and not let them figure out how to work with a supercomputer by trial-and-error.

5.1. Structure generation

5.1.1. Backbones and substitutes

The separation of the structures into backbones and substitutes was not part of the computational workflow. This is done by hand of a chemist, Prof. dr. E.A. Pidko, in case of this research. This is still a step done by a chemist, which could be subjective. It could be investigated whether this step can be replaced by a computational workflow to reach a workflow which is as unbiased as possible.

5.1.2. MACE

MACE is not published yet since it is not stable enough. This became clear during this research. First, some structures could not be processed by MACE. This was known for ferrocenes so these were excluded beforehand. Other structures yielded a very large amount of isomers or no structures at all, as discussed in subsection 4.1.2. The last issue encountered with MACE was stereochemistry. The stereochemistry from bb#'s 1 and 12 (as indicated in Table C.1) had to be removed for MACE to be able to process the structure. Bb#'s 1 and 12 with stereochemistry kept giving the warning 'Unrecognized atom type'. A clue for that has not been found, since bb#'s 6 and 13 also have stereochemistry and did not give this error. The ideal solution for this would be to delete all optical isomerism from the structures and let MACE or another program introduce this into the structures. In this way all structures with different optical isomerism can be screened.

5.1.3. ChemSpaX

ChemSpaX was not meant to be used in a computational workflow as in this project. Therefore this program is altered to make it able to run in parallel by using command line input as substitutes and renaming all temporary files to an identical name. A fundamental problem in ChemSpaX that could not be solved within this project was the issue of overlapping substitutes. If large substitutes were used, specifically cyclohexane and 2,4,6-triisopropylbenzene, the substitutes overlapped with other atoms resulting in an optimisation problem. This leads to unbonded atoms after the next optimisation. An example of this problem is shown in Figure 5.1. For this reason cyclohexane was excluded from the substitutes list. Currently, the developer of ChemSpaX is in the process of rewriting the program and removing the use of the UFF optimisation from openbabel but including the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) from the Atomic Simulation Environment, which leads to better optimisations with the same computational cost [66].

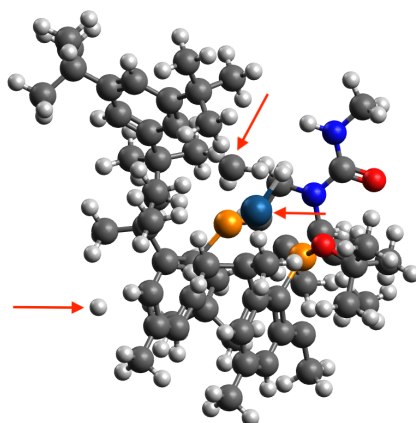


Figure 5.1: ChemSpaX optimisation error. The unbounded atoms are highlighted by an arrow.

Furthermore, the indexing in ChemSpaX is not ideal. To be able to functionalise certain groups, the indices of the atoms to be replaced and the indices of the atoms to which they are connected have to be given by hand. It would be easier to indicate just the element sort which has to be replaced. ChemSpaX should then be able to know which atoms to detach and where to attach the new substitutes. The developer of ChemSpaX is currently coding a function to be able to functionalise all atoms from a certain element. In this research, the complication was encountered that the OH and BD structures were not labelled in the same order. The order of the XYZ files was used since the atom number had to be counted. This led to difficulties when comparing the descriptors.

A small change made to ChemSpaX was the naming of the substitutes. The substitutes folder used a combination of `_` and `-` in the names of substitutes. Since a single character was needed to differentiate the different substitutes in the last step to compare structures with the same substitutes all naming has been replaced to `-` for ChemSpaX substitutes. For structures that were already generated a name processing script was written that looked for specific substitutes which had a `_` in the name and replaced it to `-`, e.g. `C6H6-CH3_para` got renamed to `C6H6-CH3-para`.

5.2. Optimisation & conformer search

CREST was the computationally most demanding part of the project. During the time CREST calculations were done the researchers had a meeting with Dr. D. Palagin and Dr. J. Thies about optimising CREST for the DelftBlue supercomputer. The program itself is written in Fortran which can perform parallel tasks to optimise efficiency. It is not useful to submit a CREST job for multiple tasks since tasks have to be started separately which is not done by CREST but multiple cores could be helpful. This is tested and the results for CREST-optimisations with a different number of cpu's on the OH structure of bb# 6 with as substituents 2 hydrogens and 2 benzenes can be seen in Figure 5.2. The data point for 9 cores is missing, since this job was not done within 24 hours and therefore cancelled automatically because the DelftBlue supercomputer had a 24-hour time limit in their post-beta phase. The trend seen in this figure is a decrease in computation time until 5 cores and afterwards an increase in computation time. This makes sense since initially adding more cores decreases the calculation time but later on the time it takes to communicate between the different cores

becomes the limiting factor, resulting in an increase in calculation time. For CREST this point is at 5 cores so all CREST calculations have been run on 5 cores. To ensure this statement, it can be validated by taking a larger sample. The outliers at 3 and 9 cores are probably caused by a simultaneously running job on the same node using a lot of resources. This can be validated by reserving a node exclusively to do this performance test. This has not been done since the optimal point was already found.

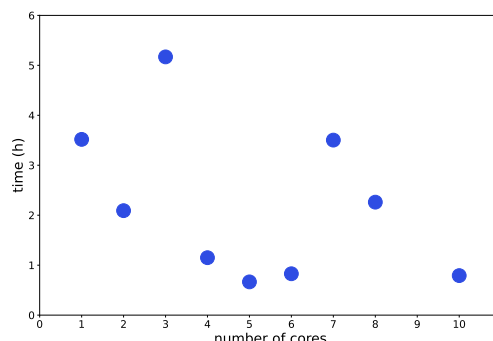


Figure 5.2: The duration of the CREST optimisation for bb# 6 with as substituents 2 hydrogens and 2 benzenes for a different number of cores.

5.3. Descriptor calculation

The descriptor calculation is done by Morfeus only. Other descriptor calculators such as Mordred [67] could be used to enlarge the numbers of descriptors calculated. This could quite easily be added to the workflow but has not been done due to a lack of time. Special attention was paid to the descriptor calculation of the energy. The energy is not calculated like the steric descriptors but the energy is used as given by CREST. CREST uses the conformer with the lowest energy as a reference. This way structures could not be compared to each other since the reference energies were not similar. This is solved by adding the energy of the lowest conformer from the CREST log file to the relative energies for every conformer.

5.4. Descriptor comparison

The descriptor comparison is done by the Pearson (or standard) correlation coefficient. This coefficient only takes linear relationships into account. With this coefficient a quick selection of interesting (possibly) linear relations could be made. Other types of relationships (exponential, quadratic, etc.) are not found by this method. It would be wise to implement possibilities to find these relationships in the computational workflow. Non-linear regression models are available, such as [68], but this has not yet been implemented in the workflow.

5.5. Code & data availability

All data and code generated for this project will be stored on the ISE server, accompanied by a README file, shortly. An overview of the data and code generated for this project is shown in Table E1 and E2 in Appendix F, sorted by process step and including their location of storage. All gen-scripts are used for *generation* of a certain program on computer clusters. These files provide information about the needed resources and combine certain programs.

6

Conclusion & Outlook

6.1. Conclusion

This research set a base for a workflow for automated screening of catalysts within the ISE-group at Delft University of Technology.

The first goal of this research was to work towards a workflow for comparison of molecular descriptors for screening catalysts. From some ligands, complexes have been created by MACE. These complexes have been altered by a modified version of ChemSpaX which is able to work in an automatised mode now. Conformers have been found using CREST with 5 cpu-cores. This seemed to be the most efficient number of cpu-cores on DelftBlue. Descriptors have been calculated and compared via an automated process. Thus a workflow for computational screening of catalysts has been setup. Improvements to this workflow can definitely be made. These are discussed in section 6.2.

The second goal was to compare the BD model structure with the OH reference structure. It is likely that the BD model is not a good model due to the non correlating bite angle. Since this descriptor only depends on the iridium metal centre and the phosphorus chelating atoms a correlation was expected. The absence of some other correlations could be explained by a non proper choice of the descriptors. This is the case for the HOMO, LUMO and cone angle. Changes to the workflow should be made to calculate these descriptors or small variations on them, like the gap between the HOMO and the LUMO. This model structure does not seem to perform well due to not being able to chemically exist, resulting in optimisation problems. This led to unrealistic values for electronic properties around the metal centre. A slightly more complicated model structure could be tried, namely addition of one hydride and a change of the metal centre from Ir^{3+} to Ir^+ resulting in a flat neutral structure which could exist.

6.2. Outlook

A basis for the computational workflow is set during this project. Somethings that were encountered but have not been solved yet are discussed below.

6.2.1. Symmetry

In the current workflow symmetry is not included. This leads to an inefficient use of resources since the same structure gets created by ChemSpaX multiple times by adding the same substitutes in for ChemSpaX different places, but resulting in the same structure due to symmetry. This becomes computationally even more demanding if CREST is used on both structures, yielding the same results. This can be solved manually by taking the symmetry into account before using ChemSpaX but a computational method would be the preferred way.

6.2.2. Stereochemistry

Stereochemistry was not the focus of this project. The stereochemistry of the ligands were originally copied from the starting dataset. Due to errors arising in MACE sometimes the stereochemistry had to be deleted. It would be a big improvement if stereochemistry could be added to the workflow, especially optical isomerism, since enantiomerism is important in drug synthesis.

6.2.3. Data storage

Storage is always something that should be taken care of when working with (big) data. This workflow produces different results. First, the structures generated by ChemSpaX can be used for a lot of optimisations and other research, so a database of these structure is helpful to save computational time. Next, the optimised conformers with CREST are useful to store, since these are needed to calculate additional descriptors in the future, without going to the entire optimisation process again. Furthermore, the calculated descriptors are helpful to be stored to prove certain correlations. And last, the final step of the research (the correlations) should obviously be stored. For the structures no database format has been found yet. Currently, the structures are stored in compressed folders per bb# and geometry with the substitutes in the file name. A database format with an efficient search engine could be helpful. The descriptors are currently stored in comma separated values (csv) files. This data storage could significantly be improved by for example an online portal as presented in [69].

6.2.4. Machine Learning

The entire workflow from generating the structures until descriptor comparison is computationally very intensive. It could therefore be worthy to feed all data and structures to a machine learning model and see whether such a model can predict the descriptors accurately for new structures. After that, a new level of complexity could be added by asking the model to recommend a new structure based on certain descriptors. This will be much more computationally efficient since countless chemical structures exists and it is undoable to calculate the descriptors for all of them.

6.2.5. Energy descriptors

An improvement that can be made is the calculation of the energy descriptor. The current energy descriptor is the electronic energy of the entire complex. It would be more interesting to see what the binding energy of the substrate is, defined as the energy of the entire complex minus the energy of the free substrate. This energy is known to influence reaction mechanisms [4] and could therefore be a good starting point to bridge between the theoretical model and the experiments.

Acknowledgements

This project was performed as the final thesis for the bachelor Molecular Science & Technology at Delft University of Technology and Leiden University. The research was performed at Delft University of Technology in the Chemical Engineering (ChemE) department at the research section Inorganic Systems Engineering. I would like to start thanking the entire section for welcoming me into their diverse group and made me feel part of it. The picture at the bottom of this page was taken at one of the group activities I joined.

I want to especially thank my daily supervisor Adarsh Kalikadien. Since I was also a teaching assistant in a course over the first six weeks in the fourth quarter I arranged my thesis spot quite early. Adarsh started only a month as a PhD student when we had our first meeting about my spot in the ISE-group. He was directly very enthusiastic and flexible to have me as one of his first bachelor students. During the project time we had many meetings about the project, the bigger research picture and more (non) science related stuff. Furthermore, I would like to thank Evgeny Pidko for having me in his group and the spot-on direct comments which highlighted the questionable spots in my project.

Last, I would like to thank my direct colleague Britt van Dongen. During the nine weeks we worked together on our projects most days we could be found side-by-side at our desks in the office. We worked very hard but always made time to catch up when the last one came in. During the days, we celebrated break troughs, motivated each other during setbacks (of which I had a lot during my project) inspired each other and exchanged ideas to continue our research.

Thank you all!

*- Mark Heezen
Delft, July 2022*



Bibliography

- [1] Hugh S. Taylor. catalysis, 6 2018. URL <https://www.britannica.com/science/catalysis>. Date accessed: 26-04-2022.
- [2] Elizabeth L Bell, William Finnigan, Scott P France, Anthony P Green, Martin A Hayes, Lorna J Hepworth, Sarah L Lovelock, Haruka Niikura, Sílvia Osuna, Elvira Romero, Katherine S Ryan, Nicholas J Turner, and Sabine L Flitsch. Biocatalysis. *Nature Reviews Methods Primers*, 1(1):46, 2021. doi: 10.1038/s43586-021-00044-z.
- [3] Leon Lefferts, Emiel Hensen, and Hans Niemantsverdriet. Heterogeneous Catalysis. In Ulf Hanefeld and Leon Lefferts, editors, *Catalysis An integrated textbook for students*, chapter 2, pages 15–71. Wiley-VCH Verlag GmbH & Co., Weinheim, 1 edition, 2018.
- [4] Elisabeth Bouwman, Martin C. Feiters, and Robertus J.M. Klein Gebbink. Homogeneous Catalysis. In Ulf Hanefeld and Leon Lefferts, editors, *Catalysis An integrated textbook for students*, chapter 3, pages 73–125. Wiley-VCH Verlag GmbH & Co., Weinheim, 1 edition, 2018.
- [5] Ram Karan, Rohit Bhatia, and Ravindra K Rawal. Chapter 9 - Applications of homogeneous catalysis in organic synthesis. In Inamuddin, Rajender Boddula, Abdullah M Asiri, and Mohammed Muzibur Rahman, editors, *Green Sustainable Process for Chemical and Environmental Engineering and Science*, pages 159–188. Elsevier, 2021. ISBN 978-0-12-819720-2. doi: <https://doi.org/10.1016/B978-0-12-819720-2.00010-2>.
- [6] Jonathan McConathy and Michael J. Owens. Stereochemistry in Drug Action. *Primary Care Companion to The Journal of Clinical Psychiatry*, 5(2):70, 6 2003. doi: 10.4088/PCC.V05N0202.
- [7] Allyn M Kaufmann and Jeffrey P Krise. Lysosomal Sequestration of Amine-Containing Drugs: Analysis and Therapeutic Implications. *Journal of Pharmaceutical Sciences*, 96(4):729–746, 2007. doi: <https://doi.org/10.1002/jps.20792>.
- [8] Nicolas Fleury-brégeot, De Fuente, Sergio Castellón, and Carmen Claver. Highlights of Transition Metal-Catalyzed Asymmetric Hydrogenation of Imines. pages 1346–1371, 2010. doi: 10.1002/cctc.201000078.
- [9] Chao Wang, Barbara Villa-Marcos, and Jianliang Xiao. Hydrogenation of imino bonds with half-sandwich metal catalysts. *Chemical Communications*, 47(35):9773, 2011. doi: 10.1039/c1cc12326b.
- [10] Kathrin Helen and Annette Bayer. Enantioselective imine hydrogenation with iridium-catalysts : Reactions , mechanisms and stereocontrol. *Coordination Chemistry Reviews*, 268:59–82, 2014. doi: 10.1016/j.ccr.2014.01.023.
- [11] Britt J.M. van Dongen. Study of the imine hydrogenation mechanism over an Ir-based catalyst using DFT. Technical report, Delft University of Technology, Delft, 6 2022.
- [12] Walter Thiel. Computational Catalysis-Past, Present, and Future. *Angewandte Chemie International Edition*, 53(33):8605–8613, 8 2014. doi: 10.1002/anie.201402118.
- [13] Felix Studt. Grand Challenges in Computational Catalysis. *Frontiers in Catalysis*, 1, 4 2021. doi: 10.3389/fctls.2021.658965.
- [14] Chadwick A. Tolman. Steric Effects of Phosphorus Ligands in Organometallic Chemistry and Homogeneous Catalysis. *Chemical Reviews*, 77(3):313–348, 1977.
- [15] Natalie Fey, A. Guy Orpen, and Jeremy N. Harvey. Building ligand knowledge bases for organometallic chemistry: Computational description of phosphorus(III)-donor ligands and the metal-phosphorus bond. *Coordination Chemistry Reviews*, 253(5-6):704–722, 3 2009. doi: 10.1016/j.ccr.2008.04.017.

- [16] Derek J. Durand and Natalie Fey. Computational Ligand Descriptors for Catalyst Design. *Chemical Reviews*, 119(11):6561–6594, 6 2019. doi: 10.1021/ACS.CHEMREV.8B00588.
- [17] Jesús Jover and Natalie Fey. The Computational Road to Better Catalysts. *Chemistry - An Asian Journal*, 9(7):1714–1723, 7 2014. doi: 10.1002/asia.201301696.
- [18] Vasishta Bhatt. Basic Coordination Chemistry. In *Essentials of Coordination Chemistry*, pages 1–35. Elsevier, 2016. doi: 10.1016/b978-0-12-803895-6.00001-x.
- [19] Irving Langmuir. Types of valence. *Science*, 54(1386):59–67, 7 1921. doi: 10.1126/SCIENCE.54.1386.59.
- [20] S. Kurth, M.A.L. Marques, and E.K.U. Gross. Density-Functional Theory. In *Encyclopedia of Condensed Matter Physics*, pages 395–402. Elsevier, 2005. doi: 10.1016/B0-12-369401-9/00445-9.
- [21] Preet Sharma. Density Functional Theory in Biology. *Annals of Chemical Science Research*, 2(1), 2020. doi: 10.31031/acsr.2020.02.000530.
- [22] John R. Brodholt and L. Vocablo. Applications of density functional theory in the geosciences. *MRS Bulletin*, 31(9), 2006. doi: 10.1557/mrs2006.176.
- [23] E. Schrödinger. Quantisierung als Eigenwertproblem. *Annalen der Physik*, 385(13):437–490, 1 1926. doi: 10.1002/ANDP.19263851302.
- [24] M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389(20):457–484, 1 1927. doi: 10.1002/ANDP.19273892002.
- [25] Anoop Kumar Kushwaha. A Brief Review of Density Functional Theory and Solvation Model. 3 2022. doi: 10.26434/CHEMRXIV-2022-VLHM0.
- [26] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864, 11 1964. doi: 10.1103/PHYSREV.136.B864.
- [27] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A), 1965. doi: 10.1103/PhysRev.140.A1133.
- [28] O. Gunnarsson, M. Jonson, and B. I. Lundqvist. Exchange and correlation in inhomogeneous electron systems. *Solid State Communications*, 24(11):765–768, 12 1977. doi: 10.1016/0038-1098(77)91185-1.
- [29] Pragya Verma and Donald G. Truhlar. Status and Challenges of Density Functional Theory. *Trends in Chemistry*, 2(4):302–318, 4 2020. doi: 10.1016/J.TRECHM.2020.02.005.
- [30] John P. Perdew and Karla Schmidt. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conference Proceedings*, 577(1):1, 8 2001. doi: 10.1063/1.1390175.
- [31] Igor Ying Zhang and Xin Xu. On the top rung of Jacob's ladder of density functional theory: Toward resolving the dilemma of SIE and NCE. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 11(1):e1490, 1 2021. doi: 10.1002/WCMS.1490.
- [32] Andrew Leach. *Molecular modelling: principles and applications*. Prentice Hall, Harlow England ;New York, 2nd ed. edition, 2001. ISBN 9780582382107.
- [33] S.F. Boys. A general method of calculation for the stationary states of any molecular system. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 200(1063):542–554, 2 1950. doi: 10.1098/RSPA.1950.0036.
- [34] Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. Extended tight-binding quantum chemistry methods. 11(2), 3 2021. doi: 10.1002/wcms.1493.
- [35] Anders S. Christensen, Tomáš Kubař, Qiang Cui, and Marcus Elstner. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chemical Reviews*, 116(9):5301–5337, 5 2016. doi: 10.1021/ACS.CHEMREV.5B00584.

- [36] Michael Gaus, Qiang Cui, and Marcus Elstner. DFTB3: Extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB). *Journal of Chemical Theory and Computation*, 7(4):931–948, 4 2011. doi: 10.1021/CT100684S.
- [37] Stefan Grimme, Christoph Bannwarth, and Philip Shushkov. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements ($Z = 1-86$). *Journal of Chemical Theory and Computation*, 13(5):1989–2009, 5 2017. doi: 10.1021/ACS.JCTC.7B00118.
- [38] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, 3 2019. doi: 10.1021/acs.jctc.8b01176.
- [39] Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Hagen Neugebauer, Sebastian Spicher, Christoph Bannwarth, and Stefan Grimme. A generally applicable atomic-charge dependent London dispersion correction. *The Journal of Chemical Physics*, 150(15):154122, 4 2019. doi: 10.1063/1.5090222.
- [40] Sebastian Spicher and Stefan Grimme. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angewandte Chemie International Edition*, 59(36):15665–15673, 9 2020. doi: 10.1002/ANIE.202004239.
- [41] A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, 12 1992. doi: 10.1021/JA00051A040.
- [42] K L Kapoor. *Quantum Chemistry and Molecular Spectroscopy*, volume 4. 6 edition, 2020.
- [43] Ralph H. Petrucci, F. Geoffrey Herring, Jeffry D. Madura, and Carey Bissonnette. *GENERAL CHEMISTRY: Principles and Modern Applications*. Pearson, 10 edition, 2011. ISBN 978-0132064521.
- [44] Kjell Jorner. Source code morfeus, 2022. URL https://kjelljorner.github.io/morfeus/_modules/morfeus/xtb.html#XTB. Date accessed: 03-07-2022.
- [45] Thomas H. Lowry and Kathleen Schueller Richardson. *Mechanism and Theory in Organic Chemistry*. Harper & Row, Michigan, 1987. ISBN 9780063504288.
- [46] J. E. Lennard-Jones. The electronic structure of some diatomic molecules. *Transactions of the Faraday Society*, 25(0):668–686, 1 1929. doi: 10.1039/TF9292500668.
- [47] Laura Falivene, Raffaele Credendino, Albert Poater, Andrea Petta, Luigi Serra, Romina Oliva, Vittorio Scarano, and Luigi Cavallo. SambVca 2. A Web Tool for Analyzing Catalytic Pockets with Topographic Steric Maps. *Organometallics*, 35(13):2286–2293, 7 2016. doi: 10.1021/ACS.ORGANOMET.6B00371.
- [48] Jenna A. Bilbrey, Arianna H. Kazez, Jason Locklin, and Wesley D. Allen. Exact ligand cone angles. *Journal of Computational Chemistry*, 34(14):1189–1197, 5 2013. doi: 10.1002/JCC.23217.
- [49] Joseph Feher. Chemical Foundations of Physiology I: Chemical Energy and Intermolecular Forces. *Quantitative Human Physiology*, pages 46–58, 1 2017. doi: 10.1016/B978-0-12-800883-6.00005-7.
- [50] Robert Pollice and Peter Chen. A Universal Quantitative Descriptor of the Dispersion Interaction Potential. *Angewandte Chemie International Edition*, 58(29):9758–9769, 7 2019. doi: 10.1002/ANIE.201905439.
- [51] Kjell Jorner, Gabriel dos Passos Gomes, Pascal Friedrich, and Tobias Gensch. Dispersion descriptor - Morfeus, 2021. URL <https://kjelljorner.github.io/morfeus/dispersion.html>. Date accessed: 15-06-2022.
- [52] A. Shrake and J. A. Rupley. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*, 79(2):351–371, 9 1973. doi: 10.1016/0022-2836(73)90011-9.

- [53] Kim Sharp and Franz Matschinsky. Translation of Ludwig Boltzmann's Paper "On the Relationship between the Second Fundamental Theorem of the Mechanical Theory of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium" *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissen Classe. Abt. II*, LXXVI 1877, pp 373-435 (Wien. Ber. 1877, 76:373-435). Reprinted in *Wiss. Abhandlungen*, Vol. II, reprint 42, p. 164-223, Barth, Leipzig, 1909. *Entropy* 2015, Vol. 17, Pages 1971-2009, 17(4):1971-2009, 4 2015. doi: 10.3390/E17041971.
- [54] Daniel V. Schroeder. *An introduction to thermal physics*. Addison-Wesley, San Francisco, 1999. ISBN 9780192895547.
- [55] Ivan Chernyshov. Mace: Metal complexes embedding, 2021. URL <https://pypi.org/project/epic-mace/>.
- [56] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An Open chemical toolbox. *Journal of Cheminformatics*, 3(10):1-14, 10 2011. doi: 10.1186/1758-2946-3-33/TABLES/2.
- [57] Adarsh V. Kalikadien, Evgeny A. Pidko, and Vivek Sinha. ChemSpaX : exploration of chemical space by automated functionalization of molecular scaffold . *Digital Discovery*, 1(1):8-25, 2022. doi: 10.1039/d1dd00017a.
- [58] Adarsh Kalikadien, Evgeny Pidko, and Vivek Sinha. EPiCs-group/chemspax: A Python tool for local chemical space exploration of any structure based on their 3D geometry. URL <https://github.com/EPiCs-group/chemspax>. Date accessed: 01-07-2022.
- [59] Philipp Pracht, Fabian Bohle, and Stefan Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics*, 22(14):7169-7192, 4 2020. doi: 10.1039/c9cp06869d.
- [60] Kjell Jorner, Gabriel dos Passos Gomes, Pascal Friedrich, and Tobias Gensch. morfeus, 2022. URL <https://github.com/kjelljorner/morfeus>.
- [61] The pandas development team. pandas-dev/pandas: Pandas, 2022.
- [62] F. Delbecq. The cyano group as a model for an electron-withdrawing substituent : Two examples. *Journal of Molecular Structure*, 93:353-357, 1 1983. doi: 10.1016/0022-2860(83)90421-0.
- [63] SURE Snellius: de nationale supercomputer. URL <https://www.surf.nl/snellius-de-nationale-supercomputer>. Date accessed: 29-06-2022.
- [64] NSC. Tetralith. URL <https://www.nsc.liu.se/systems/tetralith/>. Date accessed: 29-06-2022.
- [65] Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase 1). <https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1>, 2022.
- [66] Ask Hjorth Larsen, Jens JØrgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, Bjørk Hammer, Cory Hargus, Eric D. Hermes, Paul C. Jennings, Peter Bjerre Jensen, James Kermode, John R. Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob SchiØtz, Ole Schütt, Mikkel Strange, Kristian S. Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W. Jacobsen. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 6 2017. doi: 10.1088/1361-648X/AA680E.
- [67] Hirotomo Moriwaki, Yu Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, 10(1):1-14, 2 2018. doi: 10.1186/S13321-018-0258-Y/FIGURES/6.
- [68] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert

- Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.
- [69] 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. *Journal of the American Chemical Society*, 144(3):1205–1217, 1 2022. doi: 10.1021/JACS.1C09718.

A

CREST optimisation

This appendix elaborates on the problems with the optimisation of the BD structures, as mentioned in section 3.3. A GFN2-*x*TB optimisation led for the conformers originating from bb# 6 to newly created bonds with the nitrogen atom and the carbon atom involved. The output of one of the structures is shown in Figure A.1(a). Therefore a constrainfile was added which forced the positions of the O, N and C atoms. This led to the same structure as shown in Figure A.1(a). Then, it was chosen to use UFF optimisation instead of *x*TB optimisation. Using the same constrain file the structure in Figure A.1(b) was found. In this figure a hydrogen transferred to the iridium centre. To prevent atoms from binding to the iridium centre a UFF optimisation was used, together with a constrain file which constrained all C, N, O and H atoms. In other words only the Iridium and phosphorus atoms were optimised.

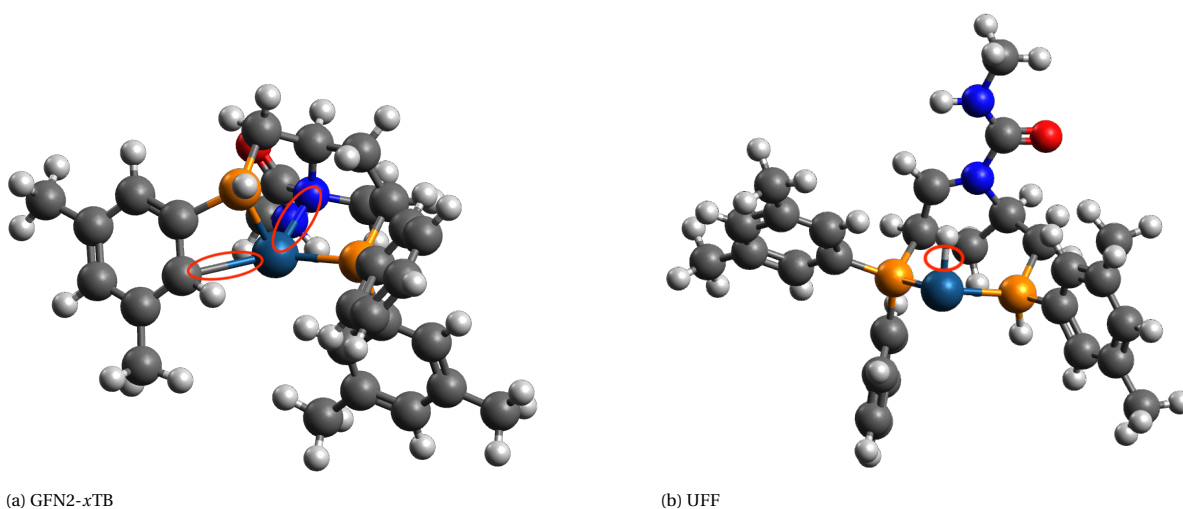


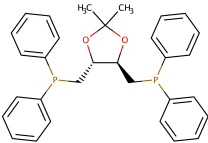
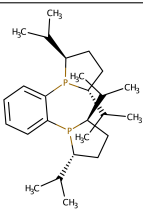
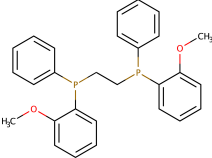
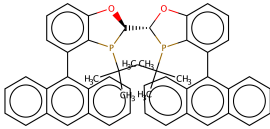
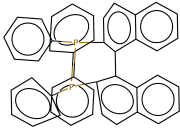
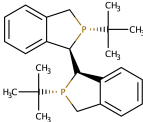
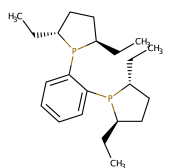
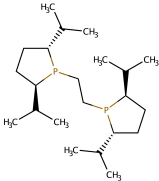
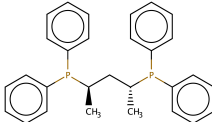
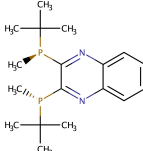
Figure A.1: The two circled bonds were not present in the model structure and inserted by the GFN2-*x*TB (a) or UFF (b) optimisation.

B

Starting structures

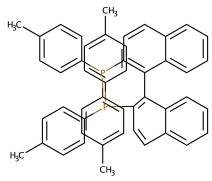
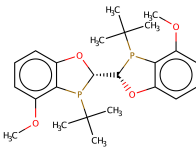
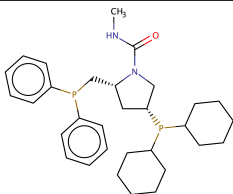
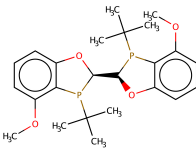
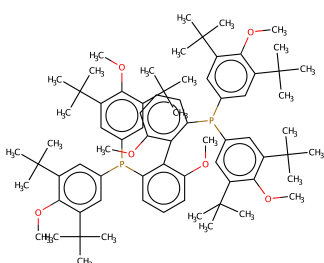
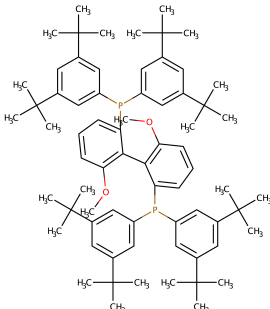
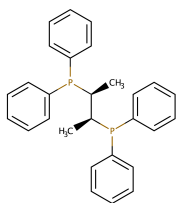
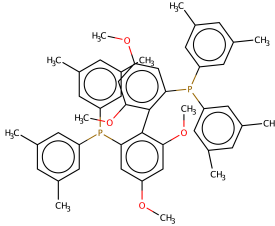
This appendix shows the chosen starting structures in Table B.1 elaborating on section 4.1. The criteria on which these structures have been chosen are discussed in subsection 3.2.1 and 4.1.

Table B.1: Starting point structures.

CAS-number	Chemical structure	CAS-number	Chemical structure
32305-98-9		136705-65-2	
55739-58-7		1884680-45-8	
76189-56-5		528814-26-8	
136705-64-1		528854-34-4	
96183-46-9		866081-62-1	

Continued on next page

Table B.1: Starting point structures. (Continued)

CAS-number	Chemical structure	CAS-number	Chemical structure
99646-28-3		1202033-19-9	
122709-72-2		1228758-57-3	
133545-25-2		192138-05-9	
64896-28-2		1365531-89-0	

C

Backbones

This appendix shows the chosen backbones based on the procedure explained in subsection 3.2.1 and 4.1.1 elaborating on subsection 4.1.1. Table C.1 shows the backbones that were successfully processed by MACE. Table C.2 shows the backbones that were unsuccessfully processed by MACE. These are discussed in more detail in subsection 4.1.2.

Table C.1: Successfully generated structures. The substituent number for the OH and BD geometry is indicated in dark and light grey respectively.

Bb#	Number of substituents	Chemical structure
1	6	
2	8	

Continued on next page

Table C.1: Successfully generated structures. The substituent number for the OH and BD geometry is indicated in dark and light grey respectively. (Continued)

Bb#	Number of substituents	Chemical structure
6	4	
7	4	
9	8	
11	4	

Continued on next page

Table C.1: Successfully generated structures. The substituent number for the OH and BD geometry is indicated in dark and light grey respectively. (Continued)

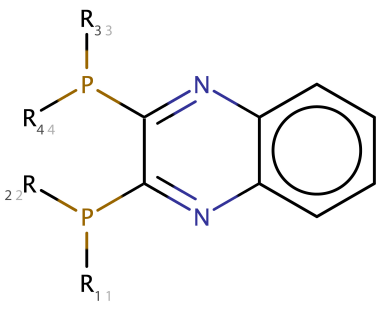
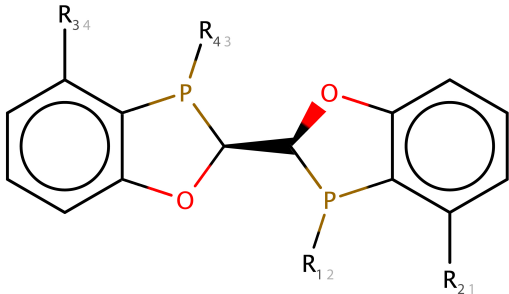
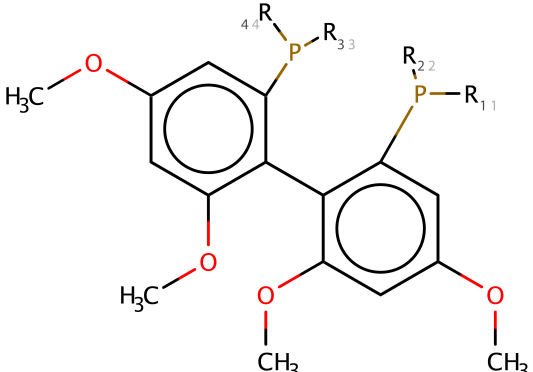
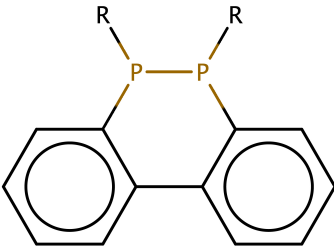
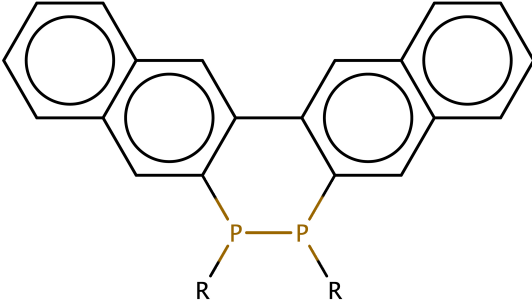
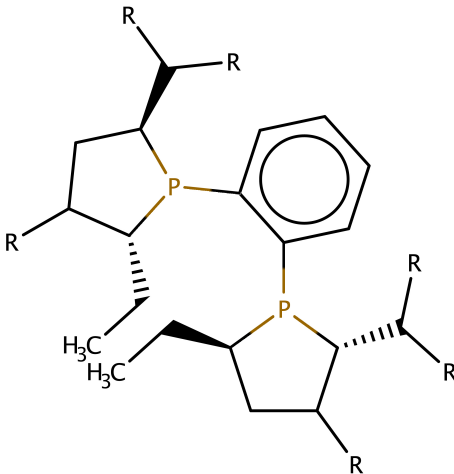
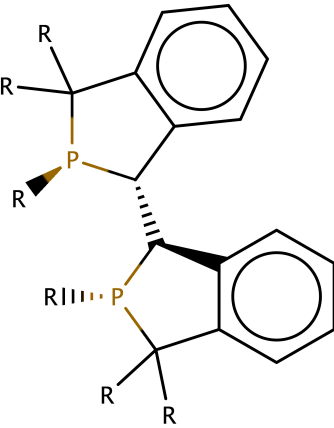
Bb#	Number of substituents	Chemical structure
12	4	
13	4	
14	4	

Table C.2: Unsuccessfully generated structures

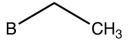
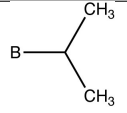
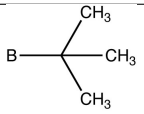
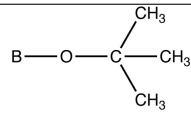
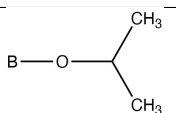
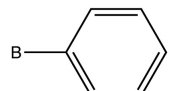
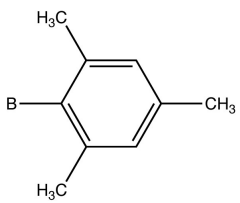
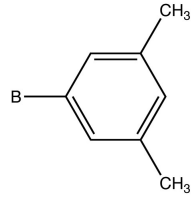
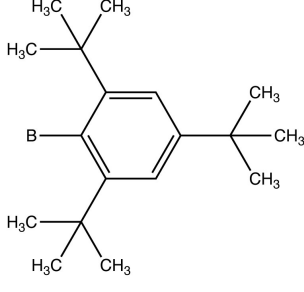
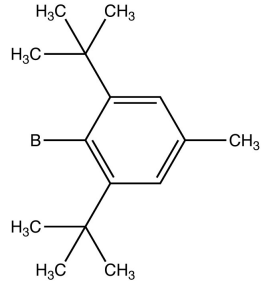
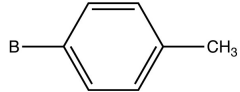
Bb#	Number of substituents	Chemical structure
3	2	
4	2	
5	8	
10	6	

D

Substitutes

This appendix shows the chosen substitutes in Table D.1 elaborating subsection 4.1.1. These substitutes are found based on the procedure explained in subsection 3.2.1.

Table D.1: Used substituents to enlarge the number of structures (B is the abbreviation for backbone).

Name in ChemSpaX	Chemical structure	Name in ChemSpaX	Chemical structure
H	B—H	CH ₃	B—CH ₃
CH ₂ CH ₃		CHCH ₃ CH ₃	
CCH ₃ CH ₃ CH ₃		OCCH ₃ CH ₃ CH ₃	
OCHCH ₃ CH ₃		C ₆ H ₆	
C ₆ H ₆ —CH ₃ —ortho—1-2-para		C ₆ H ₆ —CH ₃ —meta—1-2	
C ₆ H ₆ —iPr—ortho—1-2-para		C ₆ H ₆ —iPr—ortho—1-2-CH ₃ —para	
C ₆ H ₆ —CH ₃ —para			

E

Descriptors correlation

This appendix elaborates on Table 4.2 in chapter 4. In Table E.1 all r -values can be found for the cross references between the OH and BD structure for all descriptors. The squared values from the main diagonal are mentioned in Table E.1.

Table E.1: Cross correlated descriptors

OH															
	Energy	Dipole	Electron affinity	Electrophilicity	Nucleophilicity	Electroflugety	Nucleofugality	HOMO	Ionisation potential	LUMO	Bite angle	Burried volume	Cone angle	Dispersion	SASA
BD	Energy	0.994	-0.702	-0.523	-0.580	-0.358	0.361	-0.110	0.580	0.660	0.258	-0.586	0.050	-0.583	-0.973
	Dipol	-0.617	-0.148	0.495	0.441	0.323	-0.090	-0.027	-0.413	-0.462	-0.060	0.364	-0.108	0.540	0.598
	Electron affinity	-0.803	-0.045	0.644	0.497	0.346	-0.293	0.094	-0.530	-0.596	-0.212	0.382	-0.110	0.521	0.789
	Electrophilicity	-0.793	-0.003	0.586	0.491	0.343	-0.179	0.054	-0.520	-0.553	-0.174	0.329	-0.129	0.537	0.768
	Nucleophilicity	0.755	0.049	-0.569	-0.446	-0.303	0.241	-0.085	0.504	0.518	0.206	-0.367	0.096	-0.506	-0.736
	Electrofugety	-0.800	-0.013	0.594	0.491	0.341	-0.195	0.061	-0.526	-0.556	-0.184	0.343	-0.124	0.541	0.776
	Nucleofugality	-0.738	0.017	0.521	0.456	0.319	-0.111	0.030	-0.481	-0.497	-0.145	0.282	-0.130	0.511	0.709
	HOMO	0.750	0.055	-0.586	-0.454	-0.312	0.262	-0.089	0.501	0.535	0.212	-0.344	0.109	-0.501	-0.737
	Ionisation potential	-0.755	-0.049	0.569	0.446	0.303	-0.241	0.085	-0.504	-0.504	-0.206	0.367	-0.096	0.506	0.736
	LUMO	0.782	0.050	-0.605	-0.470	-0.321	0.265	-0.098	0.525	0.555	0.212	-0.380	0.104	-0.516	-0.764
	Bite angle	-0.201	0.034	0.137	0.137	0.105	0.019	-0.038	-0.111	-0.120	-0.070	0.084	-0.037	0.107	0.195
	Burried volume	-0.764	-0.075	0.522	0.431	0.263	-0.173	0.149	-0.594	-0.489	-0.273	0.533	-0.042	0.511	0.753
	Cone angle	-0.763	-0.050	0.502	0.438	0.275	-0.113	0.119	-0.577	-0.471	-0.253	0.461	-0.069	0.518	0.735
	Dispersion	-0.817	-0.059	0.592	0.486	0.320	-0.202	0.072	-0.584	-0.558	-0.307	0.427	-0.070	0.554	0.816
	SASA	-0.973	-0.071	0.712	0.537	0.381	-0.351	0.064	-0.549	-0.663	-0.233	0.520	-0.080	0.615	0.968

F

Code & data availability

This appendix elaborates on section 5.5. Table F.1 shows the created data for this project, their file format and the location of storage, sometimes accompanied by a short description. The same categories can be found for the code in Table F.2.

Table F.1: Generated data and storage location

Step in the process	Name	Format	Short description	Location
Backbones & substitutes	Chosen ligands	.mrz, .mol		ISE storage
	Generated complexes	.xyz, .mol		ISE storage
MACE	Generated substituted complexes	.mol		ISE storage
	Logfiles	.out	Printed output for every ChemSpax step	ISE storage
CREST	CREST-folders	multiple	The entire folder which results from a CREST calculation by structure.	ISE storage
	Conformers	.xyz	All CREST-conformers in a separate file by structure.	ISE storage
	Constrain	.inp	Constrain details for the BD complexes.	ISE storage
	Descriptors per conformer	.csv		ISE storage
Descriptor calculation	Descriptors per structure	.csv		ISE storage
Descriptor comparison	Descriptors per substituent combination	.csv		ISE storage
	r-values	.csv		ISE storage & Appendix E

Table E2: Generated code and storage location

Step in the process	Name	Format	Short description	Location
MACE	Use MACE	.py	Specifying the input for MACE	ISE storage
	Gen MACE	.sh		ISE storage
ChemSpaX	New substitutes	.xyz & .mol	Some substitutes needed were not present in the program. The new ones are added to the program.	ISE storage &
CREST	Modified program	multiple	The entire program including the changes made as described in subsection 3.2.4 and 4.1.3.	GitHub [58]
	Gen CREST	.sh	A check whether CREST jobs successfully terminated.	ISE storage
Descriptor calculation	Check CREST	.sh		ISE storage
	Gen descriptors	.sh		ISE storage
	Name change	.py	Change of _ to - for substitutes	ISE storage
	Atom order	.py	The ability to let the XYZ files start with the metal centre and the phosphorus atoms after it.	ISE storage
	Descriptor per conformer	.py	The calculation of all descriptors per conformer.	ISE storage
Descriptor comparison	AbsE	.py	The change from relative energies to absolute conformer energies.	ISE storage
	Preperation	.py	The search for descriptors over different isomers and calculating the boltzmann average.	ISE storage
	Comparison	.py	Checking for matching structures and calculation of the correlation matrix.	ISE storage

