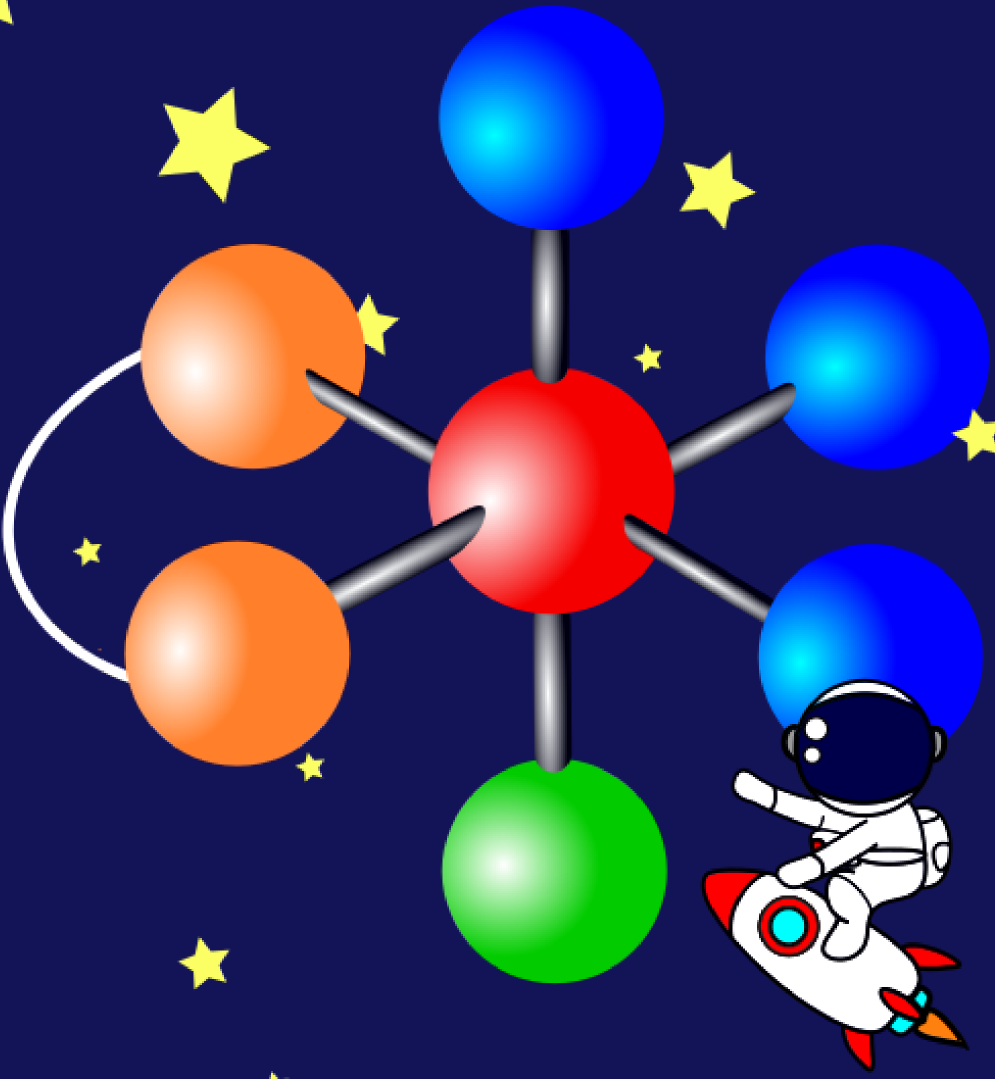
Exploring the configurational space of homogeneous catalysts

Joyce Sweere

Delft University of Technology



TU Delft

Exploring the configurational space of homogeneous catalysts

by

Joyce Sweere

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on Wednesday 26 July, 2024 at 09:00 AM.

Performed at:
Inorganic Systems Engineering
Department of Chemical Engineering
Faculty of Applied Sciences

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

Student number:	5660173/s3134180	
Project duration:	22 April 2024 – 5 July 2024	
Thesis committee:	Prof. Dr. E. A. Pidko,	TU Delft, supervisor
	Dr. A. Somoza-Tornos,	TU Delft

Abstract

Transition metal complexes are important in homogeneous catalysis reactions, especially in asymmetric hydrogenation reactions. The coordinated ligands in the complexes provide stability and selectivity. With the right combination of ligands and metal centre a high selectivity can be reached. To find optimal metal-ligand combinations, the chemical space is explored with high throughput computational methods. In these methods descriptors are obtained, usually from a single structure with one specific ligand configuration, which might not represent reality very well. In earlier research the chemical space of iridium(III), ruthenium(II) and manganese(I) complexes has been explored. And in this research additional analysis was done to look at possible relations between ligand configurations and descriptors. This was done by using three unsupervised dimensionality reduction methods, i.e. PCA, t-SNE and UMAP. PCA showed that the ligand configuration could have an influence on mainly electronic descriptors, but failed to show clusters in terms of relative stability. t-SNE and UMAP showed some clusters for the stability, as well as overlapping between certain ligand configurations. However, no definitive relations have been found, thus optimising the analysis methods and performing other statistical analysis on the descriptors might give different outcomes.

Contents

List of Figures

Abbreviations and acronyms

BO	Born-Oppenheimer
DFT	Density functional theory
FF	Force field
GTO	Gaussian-type orbital
HF	Hatree-Fock
HTE	High-throughput experimentation
KS	Kohn-Sham
LCAO	Linear combination of atomic orbitals
PCA	Principal component analysis
STO	Slater-type orbital
t-SNE	t-distributed stochastic neighbour embedding
TM	Transition metal
UMAP	Uniform manifold approximation & projection

1
Introduction

Catalysts have a fundamental role in the chemical industry; they speed up reaction rates without being consumed, and alter reaction mechanisms in such a way that energy barriers become lower [1]. In the current chemical industry, catalysts are used in more than 90% of the processes [2]. Catalysts can be categorised in three groups, being biocatalysts, heterogeneous catalysts and homogeneous catalysts. In homogeneous catalysis the catalysts are in the same phase as the reactants, in heterogeneous catalysis the phases of the reactants and products differ and in biocatalysis enzymes are the catalysts [3].

Heterogeneous catalysis has a large advantage compared to homogeneous catalysis, namely the easy separation of reaction products and catalysts, due to the phase difference [4]. Although this separation is more difficult for homogeneous catalysis, it offers high selectivity. The interaction between reactant and catalyst is also potentially more efficient, since there is interaction with the whole catalyst while with a heterogeneous catalyst this interaction is only with the surface [5].

An example of a homogeneous catalysed reaction is asymmetric hydrogenation, which is used for pharmaceutical applications, agrochemicals and other fine chemical industries [6, 7]. With asymmetric hydrogenation a reaction can be steered towards a desired enantiomer of the product [3]. In the past, to obtain a specific enantiomer, racemic mixtures would be used. However, this can only lead to a maximum yield of 50% of one of the enantiomers. With asymmetric hydrogenation this yield could be, theoretically, 100%.

To reach this selectivity, transition metal (TM) complexes are used [8]. This is a transition metal atom surrounded by ligands [9]. With the right combination of ligands and metal centre in the TM complex, the selectivity towards a specific enantiomer can be reached [10]. Certain aspects of both the ligands and the transition metal have influence on the behaviour and stability of the complex. For the metal this includes, among other things, the oxidation state, the group and row of the periodic system [11]. Ligands can alter the electronic and steric environment of TM complexes, resulting in an adapted working of the catalyst [12]. Phosphines are the most important type of ligands used in asymmetric hydrogenation reactions [13]. Especially bisphosphine bidentate ligands excel in influencing the electronic and steric environment [8].

The focus in this research is on octahedral TM complexes. Here, the central atom is surrounded by six ligands [9]. An example of an octahedral ruthenium(II) complex with a bidentate ligand is given in Figure 1.1. The CO and NCC ligands are in the so-called axial positions.

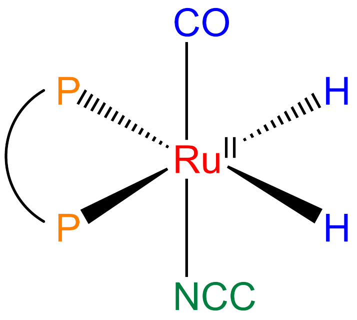


Figure 1.1: Schematic illustration of an octahedral TM-complex. Ru (red) is the metal centre, NCC (green) is a substrate ligand, PP (orange) is a bidentate ligand and CO and H (blue) are auxiliary ligands.

The ligand configuration within a complex also has influence on the catalytic properties. In Figure 1.2 is an example given of possible ligand configurations in octahedral complexes. An example of how the configuration influences the catalytic properties is the trans effect, which describes the effect the influence of a ligand on the substitution rate of the ligand trans to itself [14, 15].

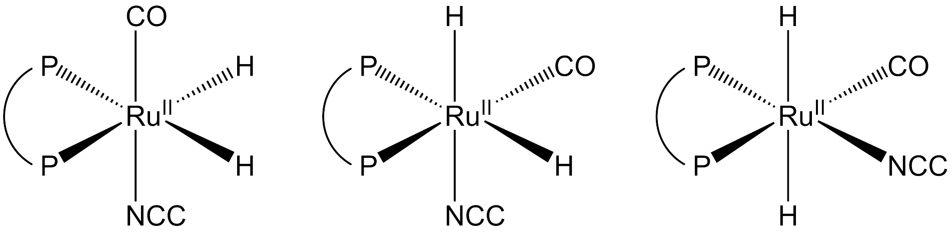


Figure 1.2: Illustration of possible ligand configurations for a octahedral TM-complex featuring a metal centre (Ru) and various ligands (H, CO, NCC, P)

To research these metal-ligand combinations and ligand configurations, high throughput experimentation (HTE) methods have proven to be very helpful in navigating the chemical space of TM-complexes [16]. In HTE many chemical experiments are performed in parallel. This saves times compared to classic catalyst testing. However, digitally navigating through the large generated datasets is still a matter of ongoing research. HT computational methods, such as quantitative structure-activity/property relationships (QSAR/QSPR), are used to enhance HTE. The general workflow for these methods is as follows: first, a virtual structure of the catalyst is generated, then this structure is optimised using a quantum chemistry tool and at last descriptors are extracted and used to model the experimental property of interest [17].

The obtained descriptors are usually extracted from a single generated structure and used for calculations/statistical models of different TM complexes. This might not be an accurate representation of the catalytic active species in experiment though and thus it is important to find a way to include these species in a representation used for statistical models.

In earlier research the chemical space of 919 octahedral TM complexes, i.e. iridium(III), ruthenium(II) and manganese(I) complexes featuring 88 bisphosphine ligands, was explored by using density functional theory (DFT) for structure optimisation. For these complexes, descriptors have been generated and used to find linear relations between ligand configuration and stability. The main findings include the preference of iridium complexes for the H-N axial ligand configuration, the absence of the descriptors showing a linear trend for the most stable configuration and the coexistence of various ligand configurations under reaction conditions.

In this thesis the focus is on dimensionality reduction methods to see if and what relations can be found between descriptors and (non-stable) ligand configurations. There are a lot of descriptors, which generated a data set with many dimensions. With dimensionality reduction methods the amount of dimensions can be reduced. The applied methods are principal component analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE) and uniform manifold approximation & projection (UMAP). The analysis is split into four parts. The first part will be using PCA to look at the variance in descriptors. The second part is inspecting whether different ligand families can be distinguished by a computer, instead of a chemist, based on these descriptors. The third part consists of dividing the complexes into energy categories based on relative stability. This is to see if the descriptors can be linked to stability. And for the fourth part an energy threshold is applied on the complexes. This threshold was used in the previous research to indicate if multiple ligand configurations coexist. The purpose of this part is similar to the third, but now focused on linking descriptors to whether ligand configurations coexist under reaction conditions or not.

The outline of this thesis is as follows: first, a theoretical background is provided, followed by a short explanation of the analysis methods used. Afterwards, the results are discussed, and at last a conclusion and the outlook is given.

2 Theory

The described theory in this chapter is a summary of the background theory used for the generated data from the previous research. Methods based on DFT and force fields were used for structure optimisation. Besides an explanation on this, descriptors used in this thesis are discussed.

2.1. Density functional theory

Density functional theory (DFT) is a very useful tool in quantum mechanics. It allows the simplification of a many-electron problem to describe many electronic properties of molecules [18]. The time-independent, non-relativistic form of the Schrödinger equation is shown in Equation 2.1 [19].

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

Where $\hat{H}$ is the Hamiltionian, Ψ the wave function and $\hat{E}$ the energy of the system. To find solutions for the Schrödinger equation, the Hamiltonian has to be described. For a system with N electrons and M nuclei, this will look like Equation 2.2 [18]:

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

with Z_A and Z_B representing the mass of the nuclei. R_{AB} is the distance between nuclei, r_{iA} is the distance between a nucleus and an electron and r_{ij} is the distance between electrons.

This equation can be rewritten as:

$$\hat{H} = T_e + T_n + V_{en} + V_{nn} + V_{ee} \tag{2.3}$$

where T_e and T_n stand for the kinetic energies of electrons and nuclei, respectively. V_{en}, V_{nn} and V_{ee} stand for electron-nuclei Coulombic interaction, nuclei-nuclei Coulombic repulsion and electron-electron Coulombic repulsion, respectively.

To actually solve the Schrödinger equation, approximations are needed. The first one is called the Born-Oppenheimer (BO) approximation. This states that nuclei are much larger than electrons, and thus have more mass which makes them significantly slower [20]. The kinetic energy for the nuclei becomes zero and the nuclei-nuclei Coulombic repulsion term becomes a constant. With that the Hamiltonian can be written as Equation 2.4.

$$\hat{H} = \frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A\neq 1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} = T_e + V_{en} + V_{ee} \tag{2.4}$$

The second approximation is the Hatree-Fock (HF) approximation, this assumes that electrons move independently of each other. This results in the wave function being written in the form of the Slater determinant [21]. However, with the Hatree-Fock approximation electron correlation is neglected. To correct for this, DFT can be used. This uses the electron density instead of the wave function [22]. With the wave function an N electron system leads to 3N variables, but the electron density only depends on three variables. DFT is based on two theorems, described by Hohenberg and Kohn. The first one states that all electronic properties are described by the properties of the ground-state. The second theorem establishes that the electron distribution energy is described as a functional of the electron density, with this functional being a minimum for the ground-state density.

A common approach for DFT is the Kohn-Sham (KS) approach [22]. In this approach the energy is split into parts [23], shown in Equation 2.5.

$$E = E^T + E^V + E^J + E^{XC} \tag{2.5}$$

E^T is the kinetic interaction energy, E^V the electron-nuclei interaction energy, E^J the Coulomb self-interaction of the density ρ, and E^{XC} the exchange correlation functional.

2.1.1. Exchange correlation

The exchange correlation functional is important in KS DFT. This functional cannot be exactly known, so approximations are made [24]. These approximations are arranged on the so called 'Jacob's ladder'. A requirement of this ladder is that each rung of the ladder should have a better approximation in terms of performance [25]. The first rung on the ladder contains the local-density approximation (LDA). This approximation is based on the principle that the energy is solely based on the density in the point evaluated [22]. The second step is the generalised gradient approximation (GGA), which adds a gradient correction to LDA [25]. On top of that, meta-GGA functionals also take the kinetic energy density into account. The last two rungs are the hybrid and double-hybrid functionals. Hybrid functionals are based on occupied orbitals, and double-hybrid functionals are based on both occupied and virtual orbitals. An illustration of the Jacob's ladder in given in Figure 2.1. For the data used in this research the PBE0 functional has been used. This is a hybrid variant of the PBE (Perdew-Burke-Emzerhof) functional, with a contribution of 25% from HF [26, 27].

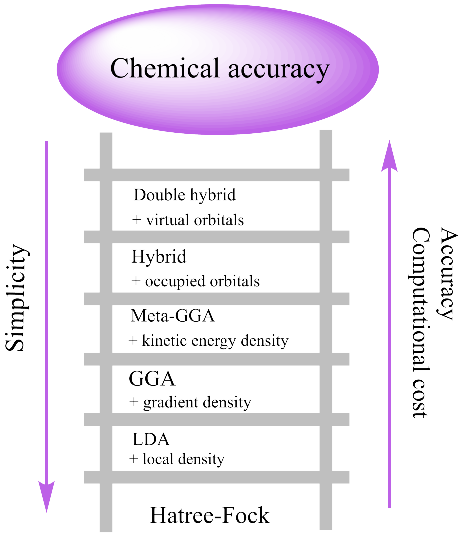


Figure 2.1: Jacob's ladder with DFT exchange correlation functional approximations. Going along the ladder decreases simplicity, increases accuracy and computational cost [25].

2.1.2. Basis sets

For calculations with DFT, a choice for a basis set has to be determined. This is a mathematical description for the orbitals in a system [28]. For most calculations a linear combination of atomic orbitals (LCAO) is used [29]. Slater-type orbitals (STOs) represent the electron density quite good, since they are similar to the atomic orbitals in hydrogen atoms [28].

Another type of orbitals used is the Gaussian-type orbitals (GTOs). These are much easier to calculate than STOs, but are not a very good representation of the electron density [28]. To overcome this, multiple GTOs are used together to give a higher accuracy.

For the data on which the analysis in this report is based, the def2-SVPP and def2-TZVPP basis sets were used. The def2 sets are second generation default basis sets [30]. SV stands for split valence (also called valence double zeta), and TZV stands for valence triple zeta. PP indicates extra polarisation functions. The def2-SVPP set is used for geometry optimisation and the def2-TZVPP for further refining of energies.

2.1.3. Dispersion correction

The KS-approach for DFT calculations does not properly take the London-dispersion interactions into account [31]. These interactions are described by Equation 2.6.

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

To correct for this, numerous methods have been developed. Most are the so-called DFT-D methods. In this research the DFT-D3 method is used.

2.1.4. Geometry optimisation

Geometry optimisation is used to define molecule structures with a minimum total energy on the potential energy surface (PES). This surface can be seen as a landscape with peaks and valleys, which represent transition states, reactants, intermediates and products [32]. The first and second derivative of the energy with respect to position are used for the optimisation. To characterise a minimum on the PES, the gradient of the potential energy must be zero and the eigenvalues of the Hessian must be positive. The Hessian is the matrix with the second derivatives of the energy.

2.2. Force fields

Another tool in exploring the chemical space of TM-complexes is the application of force field (FF) methods. These FF methods have simple energy functions that allow predictions of structural properties to be at a relatively low computational cost [33]. There are several FF methods available, each made for a specific chemical domain. For TM complexes, the universal force field (UFF) method is best [34]. This is a more general FF method, which can describe the large number of TM elements and the possible geometries and oxidation states.

2.3. Descriptors

For all the studied complexes, descriptors have been obtained. These are representations of physical-chemical properties, with several factors influencing these properties. In this work, the descriptors have previously been generated using the Open bidentate ligand explorer (OBeLiX) workflow. These descriptors are categorised in three categories: geometric, steric and electronic.

2.3.1. Geometric descriptors

Geometric descriptors are used to describe angles and lengths between atoms, for example the bond length. Three well-studied angles for bidentate ligands are the Tolman cone angle, the exact cone angle and the bite angle.

The Tolman cone angle was one of the first described properties for phosphine ligands, and was originally described for symmetric monodentate ligands [35]. A cone with its origin at the metal centre and spreading edges along the Van der Waals spheres of the outermost atoms is defined by this parameter, see Figure 2.2b as an example. While the Tolman cone angle works for monodentate ligands, an additional angle is defined for phosphine ligands [36]. This angle, the exact cone angle, is better in describing asymmetric ligands. For this research the exact cone angle is used.

Next to the exact cone angle, the bite angle is used in this research. This angle, see Figure 2.2a, is described by the angle on the central metal between the two donor atoms of the ligand.

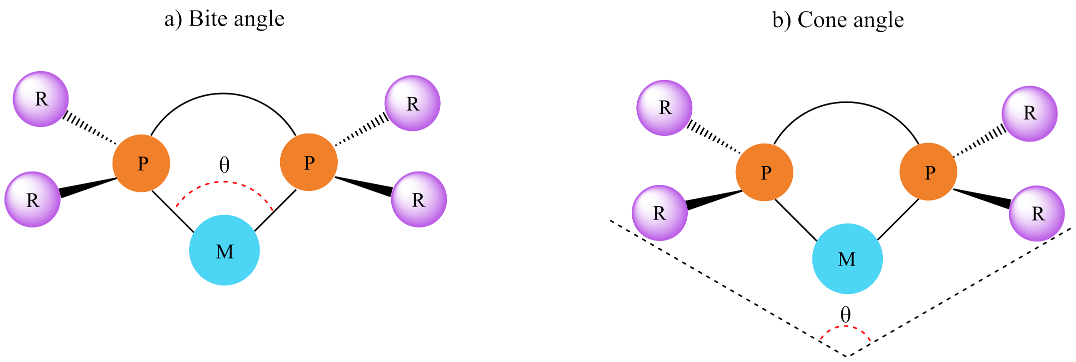


Figure 2.2: Illustration of a) the bite angle and b) the cone angle

2.3.2. Steric descriptors

Steric effects lead to enantioselectivity in asymmetric catalysis. Overlapping electron clouds in molecules have interactions, resulting in change in reactivity and selectivity [37].

For more complicated, asymmetric ligands, such as bisphosphine bidentate ligands, the buried volume (%V_{bur}) is used as a steric descriptor. %V_{bur} calculates how much of the sphere volume of the TM-complex is occupied by ligands [38]. Figure 2.3 shows an example of the buried volume with the metal "M" as the centre of the sphere. The advantage of the buried volume is that the ligands do not have to be symmetrical [39]. Descriptor calculations using the Tolman cone angle have shown to be difficult and even meaningless sometimes for such ligands [40]. For this research the buried volume has been measured for several (partial) spheres with varying radii, and with various atoms as the centre.

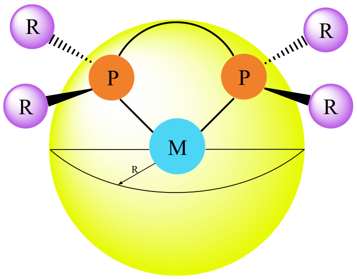


Figure 2.3: Illustration of the buried volume of a ligand in a transition metal complex

2.3.3. Electronic descriptors

Electronic descriptors are used to describe electronic interactions. [38]. In general, there are many electronic descriptors, such as bond energies, proton/electron affinities, molecular orbital energies and more. The electronic descriptors generated for this thesis can be divided into global and local descriptors.

Global

The global electronic descriptors used in this research are the HOMO-LUMO gap, the dipole moment and the dispersion energy. The HOMO-LUMO gap describes the gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). The dipole moment is defined as the product of the charge, at the ends of a molecule, and the distance between the charges [41]. The dispersion energy stands for the London dispersion force. This force is created when a temporary dipole arises and causes a temporary dipole in nearby atoms [42].

Local

The lone pair occupancy and the NBO charge are the local electronic descriptors that are used. The lone pair occupancy is a measure of the filling of the maximum/minimum bidentate ligand donor by a lone electron pair. And the NBO charge stands for the natural bonding orbital charge. This has been measured on a number of specific atoms.

3
Methods

Three methods have been used to analyse the data, namely principal component analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE) and uniform manifold approximation and projection (UMAP). In this chapter a short explanation for each method is given.

3.1. PCA

PCA is the first analysis methods that was used to see if any relations can be found between the descriptors and ligand configurations. The goal of PCA is to reduce the dimensions in a multi-dimensional data set, and see if any clusters can be found in the data [43]. PCA is an unsupervised method, which focuses on finding underlying relations in the data [44]. The outcome of interest in not taken into account, whereas with supervised methods this outcome of interest influences the dimensionality reduction.

With a large dataset it can be difficult to see correlations between the variables. PCA provides a solution by generating new variables that capture as much as possible of the variation in the data [45]. These newly obtained variables are called principal components [45]. These are linear combinations of all the original dimensions of which the first (PC1) explains the most of the variation in the data, PC2 explains the second most variation etc. [43]. By plotting PC2 vs PC1, called a score plot, it could be possible to see any clusters.

The main downside of PCA is that it is a linear method, meaning only linear relations could be found with PCA. However, it is quite interesting to also explore any possible non-linear relations. Thus two other, non-linear method besides PCA were used to analyse the data.

3.2. t-SNE

t-SNE is a very popular analysis method as alternative for PCA. To reduce the number of dimensions, it projects the data from the high-dimensional space onto a low-dimensional space. To construct clusters in the low-dimensional space a student-t distribution is used, unlike a Gaussian distribution which is used by stochastic neighbourhood embedding (SNE) [46]. To do this, the data points are 'allowed' to interact through repulsion and attraction [47]. All points show repulsion to each other, but there is attraction to a points nearest neighbour. The perplexity defines how many points are considered as neighbours. Changing this parameter has influence on the outcome. t-SNE does not return variables like the principal components in PCA, but what is returned can best be described as the dimensions from the low-dimensional space. So, in the constructed graphs two dimensions, called t-SNE 1 and t-SNE 2, are plotted against each other.

3.3. UMAP

UMAP is another non-linear dimension reduction method. With UMAP a fuzzy representation of the high-dimensional data and low-dimensional data gets constructed. Then the low-dimensional representation gets adjusted to get as close as possible to similarity with the high-dimensional representation [48]. The main advantage of UMAP over t-SNE is the ability to perform better with very large datasets [49]. Additionally, the global structure is better preserved using UMAP, in contrast to t-SNE which preserves the local structure better [50]. Similar to t-SNE, in the graphs two dimensions are plotted, but now called UMAP 1 and UMAP 2.

4

Results and discussion

This chapter discusses the results of the research. The data set that was analysed consist of all the TM-complexes together with certain properties of the complexes and the values for the descriptors. The most important properties are the ligand configuration, the ligand itself and the relative energy difference. The ligand configuration tells which ligands are in the axial positions. The possible configurations for these complexes are H-H, H-N, C-H and C-N. The relative energy difference is the difference in energy for each complex compared to the H-N configuration. This configuration was chosen as the reference since it proved to be the most stable in the previous research.

The results are given according to the four parts in which the analysis was done. First, the results of the descriptors variance are given in section 4.1. Next, the results from inspecting whether ligand families can be distinguished are presented in section 4.2. Then, the results from the energy categories are shown in section 4.3. At last, the results from the energy threshold are given in section 4.4.

4.1. Descriptor variance

The first analysis that was done is PCA. The principle components all have an explained variance. This tells how much of the variation in the data is explained by this principle component. In the appendix Figure A.1 is given which shows the explained variance. PC1 and PC2 have an explained variance of 36% and 14%, respectively.

The loadings plot of the PCs tells what is represented by the principle components. In Figure 4.1 it is visible that steric descriptors, the cone angle and the bite angle have a negative correlation with PC1. This would suggest that a distinction between different ligands can be made based on the values for PC1. The contribution to PC2 comes mostly from electronic descriptors, of which the NBO charges have the strongest correlation with PC2. In contrast to PC1, based on the values for PC2 a distinction between the different metal centres could be made.

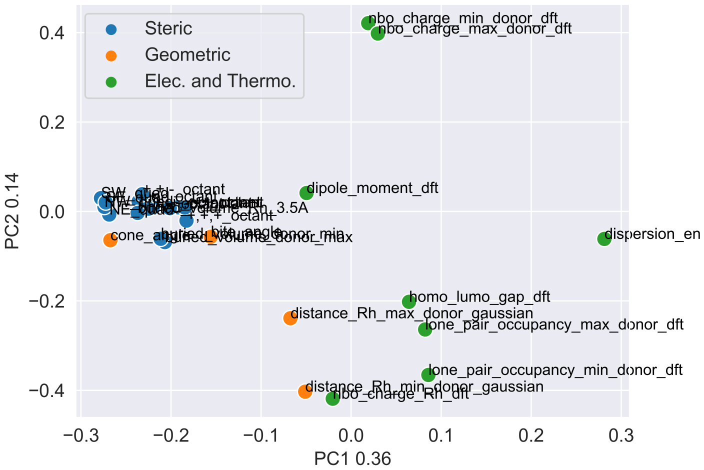


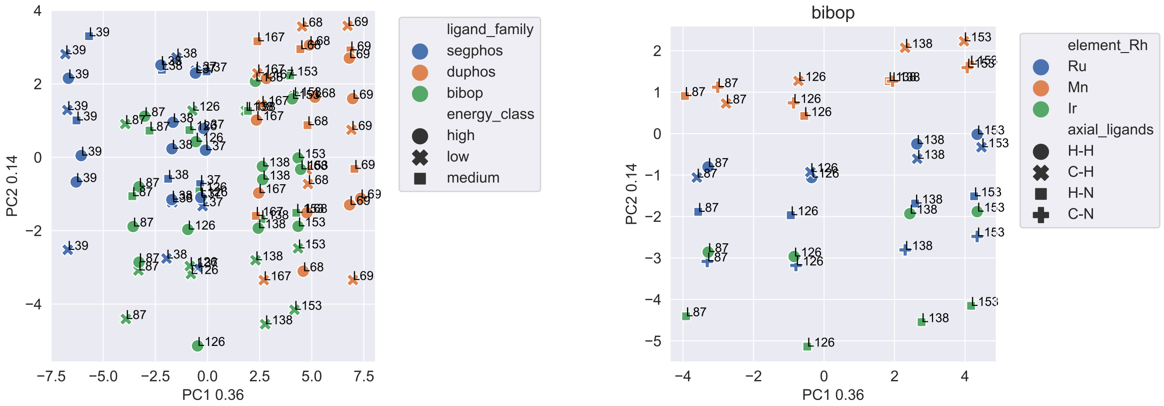
Figure 4.1: Scatter plot of the loadings for each descriptor. Colours indicate descriptor category.

4.2. Ligand families

These findings can be confirmed by looking at Figure 4.2. In this graph score plots are given for PCA. For this report PC2 is plotted against PC1 for all of the score plots. If there is any clustering in the data, this can be seen in a score plot.

The 88 bisphosphine ligands have been categorised in families, which represent ligands with similar backbones. In Figure 4.2a three ligand families have been plotted, namely segphos, duphos and bibop. An example of what the ligands look like from these families is given in the appendix (Figure A.2). In Figure 4.2a there is a clear distinction visible between the different metal centres. Manganese has only positive values for PC2, at least for these families. Ruthenium is spread around 0 for PC2, and iridium has only negative values. Additionally, each ligand has roughly the same values for PC1, even for the different metal centres.

Figure 4.2b shows only the bibop family, containing four different ligands. Similar to Figure 4.2a the metal centres are well distinguishable. What is also remarkable is the ligand configuration. Per type of metal centre one configuration has approximately the same values for PC2. This would imply that the ligand configuration could have some influence on the descriptors contributing to PC2.



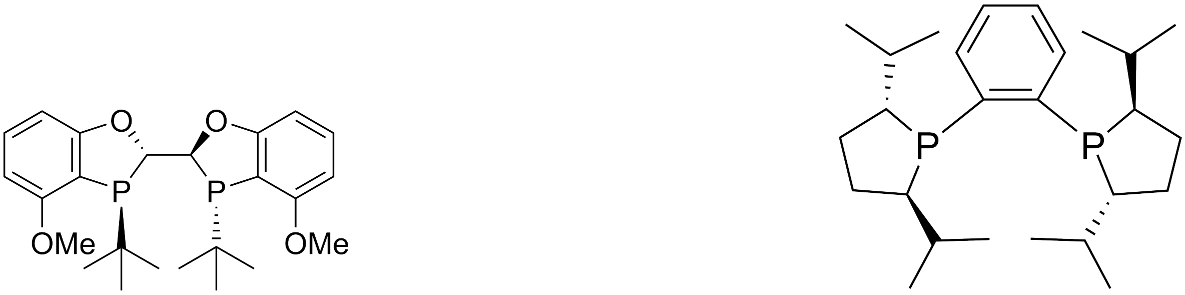
(a)

(b)

Figure 4.2: a) PCA score plot of the first two PCs for ligand families segphos, duphos and bibop. Colours indicate the family and marker style the metal centre. For each point it is indicated which ligand that complex contains, shown by 'L' followed by the number of the ligand. b) PCA score plot of the first two PCs for ligand family bibop. Colours indicate the type of metal centre and marker style the ligand configuration.

With t-SNE and UMAP these distinctions are less clear. The graphs for the segphos, duphos and bibop families and the bibop alone are given in the appendix (Figure A.3 and Figure A.4 respectively). Although there is some grouping visible between the families, the differences between metal centres are less distinguishable. The different ligands and configurations are also not well distinguished.

Furthermore, something else is remarkable about Figure 4.2a. Different ligand families would be expected to have different sterics, because of the different backbones. Since PC1 mainly consists of steric descriptors and PC2 of electronic descriptors, we would expect to see a difference in values for PC1 between different families. However, L138 and L167 have roughly the same values for PC1 despite of being from different families. The structures of these ligands are shown in Figure 4.3. The ligands do not have similar groups in the structure, so it remains remarkable that they have similar values for PC1.



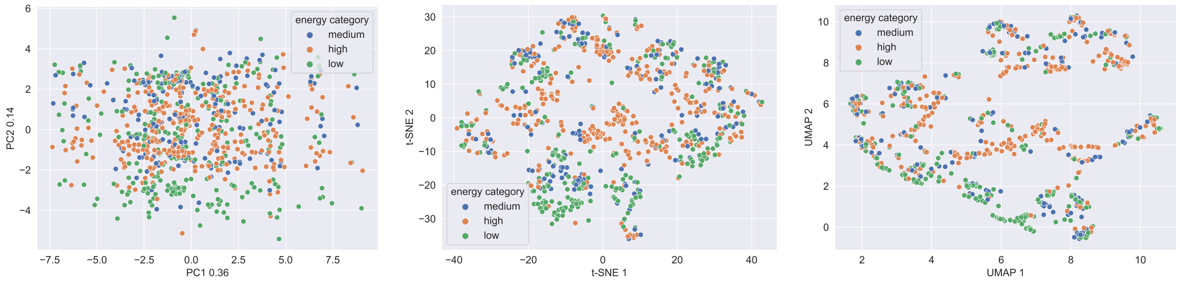
(a)

(b)

Figure 4.3: a) Ligand 138, from family bibop. b) Ligand 167, from family duphos.

4.3. Energy categories

To see if the relative stability of the TM-complexes can be related to the descriptors, all the complexes have been sorted into three different energy categories. The relative energy difference has been split into quartiles, of which the lowest is labelled ’low’, the second lowest is labelled ’medium’ and the other two quartiles are together labelled ’high’. Thus, ’low’ contains the smallest energy difference and ’high’ the largest difference. The score plots for PCA, t-SNE and UMAP are shown in Figure 4.4. For the PCA (Figure 4.4a) we see that the different energy categories are all mixed over the whole graph. The PCs cannot be quite linked to the energy categories. For the t-SNE (Figure 4.4b) we see that the high energy clusters are mainly around the middle of the graph. The low energy shows a small cluster at the bottom of the graph. The medium energy is not really clustering, but appears to be more around the sides and not in the middle where the high energy is. For UMAP (Figure 4.4c) we see something similar, namely most of the ’high’ is in the middle of the graph. ’Low’ and ’medium’ are around the sides, with a small cluster of low energy.



(a)

(b)

(c)

Figure 4.4: Score plots of the first two PCs/dimensions for a) PCA, b) t-SNE and c) UMAP. Colours indicate energy category.

Plotting the metal centres apart from each other gives more insight in this, this is best visible in the UMAP score plot which is shown in Figure 4.5. The same plots for PCA and t-SNE are given in the appendix (Figure A.5 and Figure A.6 respectively). For the iridium complexes the difference between the energy categories is quite large for all three clustering methods. Furthermore, some overlapping can be seen between ruthenium and iridium. The C-N and H-N configurations are roughly in the same area on the graphs. But for manganese these configurations are in a different area.

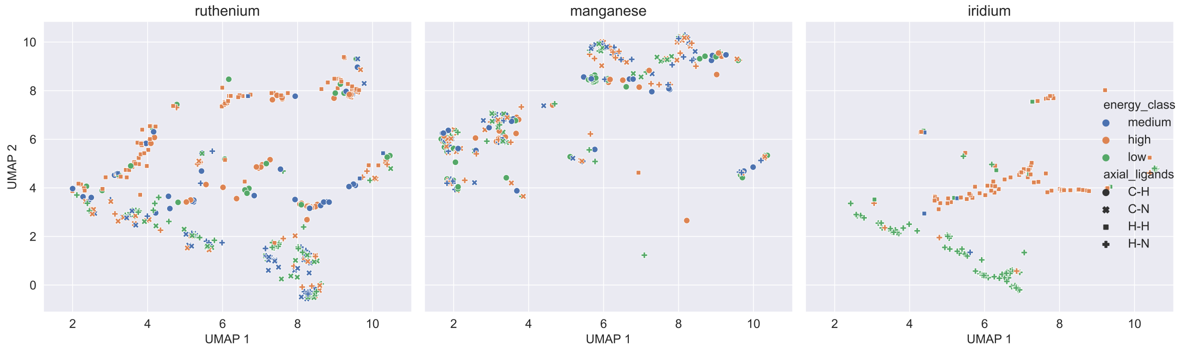
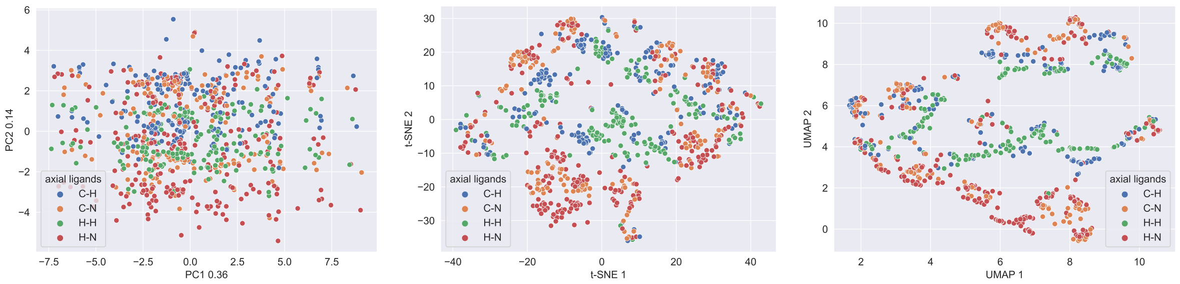


Figure 4.5: UMAP score plot of the first two dimensions. Colours indicate energy category and marker style ligand configuration.

This overlap between ligand configuration stirred some curiosity, so Figure 4.6 was made to see if this is a coincidence or not. The figure is the same as Figure 4.4, but the colours now indicate the ligand configuration. For the PCA (Figure 4.6a) we see that there is not any clustering happening. For the t-SNE and UMAP (Figure 4.6b and Figure 4.6c respectively) this is different. There is a clear distinction between C-H & H-H and C-N & H-N. The C-H and H-H configurations are mainly in the middle of the graph, and the C-N and H-N configurations are around the sides. This could indicate that t-SNE and UMAP are better at showing relations regarding the ligand configuration.



(a) (b) (c)

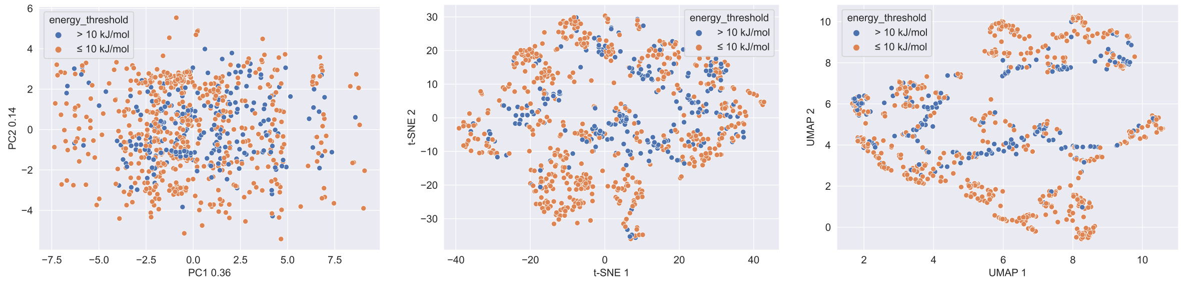
Figure 4.6: Score plots of the first two PCs/dimensions for a) PCA, b) t-SNE and c) UMAP. Colours indicate ligand configuration.

4.4. Energy threshold

Alongside the graphs with the energy categories, the same graphs have been made but a threshold of 10 kJ/mole was used instead of the three categories. This threshold is based on the previous research where it was used to examine if multiple ligand configurations could be considered stable during reactions and thus contribute to the catalytic behaviour.

Overall these graphs give the same results as before. Figure 4.7 shows the score plots for PCA, t-SNE and UMAP with the colours indicating whether a point falls within the threshold or not. Similar to Figure 4.4a, the PCA does not show much clustering. For the t-SNE and UMAP the points outside the threshold are in the middle of the graphs while the points within the threshold are around the sides. This corresponds to Figure 4.4b and Figure 4.4c.

Figures A.7, A.8 & A.9 (see appendix) show for each analysis method the score plots, with energy threshold, where the metal centres are plotted apart from each other. Again, the distinction between within the threshold and outside is best visible for UMAP. Additionally, the overlapping between the C-N and H-N configurations for ruthenium and iridium is still visible, whereas manganese shows a contrast.



(a) (b) (c)

Figure 4.7: Score plots of the first two PCs/dimensions for a) PCA, b) t-SNE and c) UMAP. Colours indicate energy threshold.

5
Conclusion and outlook

5.1. Conclusion

The aim of this thesis was to use dimensionality reduction methods on generated descriptors values for TM-complexes with different possible ligand configurations. To do so, the unsupervised clustering methods PCA, t-SNE and UMAP have been used. The research was divided into four parts. The first part focused on using PCA to determine the variance in descriptors. The second part looked at whether different ligand families can be distinguished bases on the descriptors. The third part consisted of linking the descriptors to different classes of stability, and the fourth part focused on linking the descriptors to whether multiple ligand configurations coexist by applying an energy threshold.

The variance in descriptors showed with PCA that the ligand configurations could have an influence on mostly electronic descriptors and the ligand family might influence mostly steric descriptors. The different ligand families can be distinguished to a certain extent using PCA. t-SNE and UMAP did not show this ability to distinguish. However, t-SNE and UMAP showed better clustering in terms of stability. The complexes with the relatively largest energy differences seemed to cluster in the middle of the score plots, whereas the complexes with the relatively smallest energy differences were somewhat clustered around the sides. Additionally, there is some similarity between the iridium and ruthenium complexes, whereas the manganese complexes do not share this similarity. For iridium and ruthenium, the C-N and H-N configurations are overlapping, as well as the C-H and H-H configurations. This overlap is best visible with UMAP, but also a bit with t-SNE. This was confirmed with other score plots where there was coloured based on ligand configuration. At last, the application of the energy threshold showed similar results to the stability.

All in all, the results indicate that PCA is able to distinguish between ligand families, but that the stability is not well classified with this method. For the t-SNE and UMAP, the results show that these two methods are able to differentiate between ligand configurations and between relative stability.

5.2. Outlook

The outcome of this research invites for further research on the ligand configurations. Within t-SNE and UMAP there are a few parameters that can be adjusted and optimised, which probably will give different results. Additionally, it might be interesting to do a statistical analysis on descriptors and their ranges within configurations. So, to determine which ligands are more prone to showing flexibility in the configurations, and for which ones this is not readily obvious. At last, running a machine learning task on the stability to use the descriptors to determine relative stability or even the flexibility of a certain ligand type might give new insights.

Acknowledgements

The past ten weeks have been an incredible learning curve for me. Due to having trouble finding a thesis project, I am very grateful that this project was available at the Inorganic Systems & Engineering group.

First, I want to thank Adarsh Kalikadien, my daily supervisor, for all the guidance and help. He had never doubt that this thesis would come to a good end. Adarsh helped me with seeing the bigger picture and setting the goal straight. But also helped me with the small things, which would have cost me much time had I have to figure it out on my own.

Next, I would like to thank Professor Evgeny Pidko, who was always enthusiastic about my results and made sure I was too. He provided new insights and discussions which helped me getting closer to my goal.

At last, I am thankful for the other bachelor students who I shared an office with. Even though we were not working on the same project, sometimes we could help each other or give advise on something. Thank you David, Mas and Nina.

- Joyce Sweere
Delft, June 2024

Bibliography

[1] Stephen Mccord, Cynthia Labrake, and David Vanden Bout. Catalyst. URL `https://ch302.cm.utexas.edu/kinetics/index.php#catalysts/intro-catalysis.html`.

[2] Zhen Ma and Francisco Zaera. Heterogeneous catalysis by metals. In *Encyclopedia of Inorganic and Bioinorganic Chemistry*, pages 1–16. John Wiley & Sons, Ltd. ISBN 978-1-119-95143-8. doi: 10.1002/9781119951438.eibc0079.pub2. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119951438.eibc0079.pub2`.

[3] U. Hanefeld and L. Lefferts. *Catalysis: An Integrated Textbook for Students*. Wiley. ISBN 978-3-527-34159-7. URL `https://books.google.nl/books?id=reQ5DwAAQBAJ`.

[4] DJ Cole-Hamilton and RP Tooze. Homogeneous catalysis—advantages and problems. *Catalyst Separation, Recovery and Recycling: Chemistry and Process Design*, pages 1–8, 2006.

[5] G. Duca. *Homogeneous Catalysis with Metal Complexes*. URL `https://link.springer.com/book/10.1007/978-3-642-24629-6`.

[6] Qing-An Chen, Zhi-Shi Ye, Ying Duan, and Yong-Gui Zhou. Homogeneous palladium -catalyzed asymmetric hydrogenation. 42(2):497–511. doi: 10.1039/C2CS35333D. URL `https://pubs.rsc.org/en/content/articlelanding/2013/cs/c2cs35333d`.

[7] Yan-Yun Li, Shen-Luan Yu, Wei-Yi Shen, and Jing-Xing Gao. Iron-, cobalt-, and nickel-catalyzed asymmetric transfer hydrogenation and asymmetric hydrogenation of ketones. 48(9):2587–2598. ISSN 0001-4842. doi: 10.1021/acs.accounts.5b00043. URL `https://doi.org/10.1021/acs.accounts.5b00043`.

[8] Andrew L Clevenger, Ryan M Stolley, Justis Aderibigbe, and Janis Louie. Trends in the usage of bidentate phosphines as ligands in nickel catalysis. *Chemical Reviews*, 120(13):6124–6196, 2020.

[9] Tina Overton, Jonathan Rourke, and Fraser A. Armstrong. *Inorganic Chemistry*. Oxford University Press. ISBN 978-0-19-876812-8.

[10] Weicheng Zhang, Yongxiang Chi, and Xumu Zhang. Developing chiral ligands for asymmetric hydrogenation. 40(12):1278–1290. ISSN 0001-4842. doi: 10.1021/ar7000028. URL `https://doi.org/10.1021/ar7000028`.

[11] Cheng Hou, Yinwu Li, and Zhuofeng Ke. Transition metal center effect on the mechanism of homogenous hydrogenation and dehydrogenation. 511:119808. ISSN 0020-1693. doi: 10.1016/j.ica.2020.119808. URL `https://www.sciencedirect.com/science/article/pii/S0020169320310070`.

[12] Christopher Masters. *Homogeneous Transition-metal Catalysis: A Gentle Art*. Springer Science & Business Media. ISBN 978-94-009-5880-7.

[13] Ian J. S. Fairlamb and Jason M. Lynam. *Organometallic Chemistry: Volume 35*. Royal Society of Chemistry. ISBN 978-1-84755-103-0.

[14] Frédéric Guégan, Vincent Tognetti, Laurent Joubert, Henry Chermette, Dominique Luneau, and Christophe Morell. Towards the first theoretical scale of the trans effect in octahedral complexes. 18(2):982–990. doi: 10.1039/C5CP04982B. URL `https://pubs.rsc.org/en/content/articlelanding/2016/cp/c5cp04982b`.

[15] Benjamin J Coe and Susan J Glenwright. Trans-effects in octahedral transition metal complexes. 203(1):5–80. ISSN 0010-8545. doi: 10.1016/S0010-8545(99)00184-8. URL `https://www.sciencedirect.com/science/article/pii/S0010854599001848`.

[16] Sarah Callaghan. Toward machine learning-enhanced high-throughput experimentation for chemistry. 2(3):100221. ISSN 2666-3899. doi: 10.1016/j.patter.2021.100221. URL https://www.sciencedirect.com/science/article/pii/S2666389921000350.

[17] Marco Foscato and Vidar R. Jensen. Automated in silico design of homogeneous catalysts. 10(3):2354–2377. doi: 10.1021/acscatal.9b04952. URL https://doi.org/10.1021/acscatal.9b04952.

[18] Anoop Kumar Kushwaha. A brief review of density functional theory and solvation model. URL https://chemrxiv.org/engage/chemrxiv/article-details/6231c8042c5010f92a7d0bd6.

[19] David S. Sholl and Janice A. Steckel. *Density Functional Theory: A Practical Introduction*. John Wiley & Sons. ISBN 978-1-118-21104-5.

[20] Errol Lewars. *Computational Chemistry*. Springer. ISBN 978-3-319-30914-9. URL https://link.springer.com/book/10.1007/978-3-319-30916-3.

[21] Thomas Engel and Philip Reid. *Physical Chemistry*. Pearson. ISBN 978-1-292-02224-6.

[22] Maylis Orio, Dimitrios A. Pantazis, and Frank Neese. Density functional theory. 102(2):443–453. ISSN 1573-5079. doi: 10.1007/s11120-009-9404-8. URL https://doi.org/10.1007/s11120-009-9404-8.

[23] John A. Pople, Peter M. W. Gill, and Benny G. Johnson. Kohn—sham density-functional theory within a finite basis set. 199(6):557–560. ISSN 0009-2614. doi: 10.1016/0009-2614(92)85009-Y. URL https://www.sciencedirect.com/science/article/pii/000926149285009Y.

[24] Jonathan Schmidt, Carlos L. Benavides-Riveros, and Miguel A. L. Marques. Machine learning the physical nonlocal exchange–correlation functional of density-functional theory. 10(20):6425–6431. doi: 10.1021/acs.jpclett.9b02422. URL https://doi.org/10.1021/acs.jpclett.9b02422.

[25] Konstantinos D. Vogiatzis, Mikhail V. Polynski, Justin K. Kirkland, Jacob Townsend, Ali Hashemi, Chong Liu, and Evgeny A. Pidko. Computational approach to molecular catalysis by 3d transition metals: Challenges and opportunities. 119(4):2453–2523. ISSN 0009-2665. doi: 10.1021/acs.chemrev.8b00361. URL https://doi.org/10.1021/acs.chemrev.8b00361.

[26] Jing Yang, Liang Z. Tan, and Andrew M. Rappe. Hybrid functional pseudopotentials. 97(8):085130. ISSN 2469-9950, 2469-9969. doi: 10.1103/PhysRevB.97.085130. URL https://link.aps.org/doi/10.1103/PhysRevB.97.085130.

[27] Christopher J. Cramer. *Essentials of Computational Chemistry: Theories and Models*. Wiley, 2 edition. ISBN 978-0-470-09182-1.

[28] K. I. Ramachandran, Gopakumar Deepa, and Krishnan Namboori. *Computational Chemistry and Molecular Modeling: Principles and Applications*. Springer Science & Business Media. ISBN 978-3-540-77302-3.

[29] Stig Rune Jensen, Tor Flå, Dan Jonsson, Rune Sørland Monstad, Kenneth Ruud, and Luca Frediani. Magnetic properties with multiwavelets and DFT: the complete basis set limit achieved. 18(31):21145–21161. ISSN 1463-9084. doi: 10.1039/C6CP01294A. URL https://pubs.rsc.org/en/content/articlelanding/2016/cp/c6cp01294a.

[30] Jingjing Zheng, Xuefei Xu, and Donald G. Truhlar. Minimally augmented karlsruhe basis sets. 128 (3):295–305. ISSN 1432-2234. doi: 10.1007/s00214-010-0846-z. URL https://doi.org/10.1007/s00214-010-0846-z.

[31] Alberto Otero de la Roza and Gino DiLabio. A comprehensive overview of the DFT-d3 london-dispersion correction. In *Non-Covalent Interactions in Quantum Chemistry and Physics*, pages 195–219. Elsevier. ISBN 978-0-12-809835-6. doi: 10.1016/B978-0-12-809835-6.00007-4. URL https://www.sciencedirect.com/science/article/pii/B9780128098356000074.

[32] H. Bernhard Schlegel. Geometry optimization. 1(5):790–809. ISSN 1759-0884. doi: 10.1002/wcms.34. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.34.

[33] Francesco Fracchia, Gianluca Del Frate, Giordano Mancini, Walter Rocchia, and Vincenzo Barone. Force field parametrization of metal ions from statistical learning techniques. 14(1):255–273. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00779. URL https://doi.org/10.1021/acs.jctc.7b00779.

[34] A. K. Rappe, K. S. Colwell, and C. J. Casewit. Application of a universal force field to metal complexes. 32(16):3438–3450. ISSN 0020-1669. doi: 10.1021/ic00068a012. URL https://doi.org/10.1021/ic00068a012.

[35] Jesús Jover and Jordi Cirera. Computational assessment on the tolman cone angles for p-ligands. 48(40):15036–15048. doi: 10.1039/C9DT02876E. URL https://pubs.rsc.org/en/content/articlelanding/2019/dt/c9dt02876e.

[36] Cone angle. URL https://digital-chemistry-laboratory.github.io/morfeus/cone_angle.html.

[37] Alexandre V. Brethomé, Stephen P. Fletcher, and Robert S. Paton. Conformational effects on physical-organic descriptors: The case of sterimol steric parameters. 9(3):2313–2323. doi: 10.1021/acscatal.8b04043. URL https://doi.org/10.1021/acscatal.8b04043.

[38] Natalie Fey. The contribution of computational studies to organometallic catalysis: descriptors, mechanisms and models. 39(2):296–310. doi: 10.1039/B913356A. URL https://pubs.rsc.org/en/content/articlelanding/2010/dt/b913356a.

[39] Adrián Gómez-Suárez, David J. Nelson, and Steven P. Nolan. Quantifying and understanding the steric properties of n-heterocyclic carbenes. 53(18):2650–2660. ISSN 1364-548X. doi: 10.1039/C7CC00255F. URL https://pubs.rsc.org/en/content/articlelanding/2017/cc/c7cc00255f.

[40] Hervé Clavier and Steven P. Nolan. Percent buried volume for phosphine and N-heterocyclic carbene ligands: steric properties in organometallic chemistry. *Chemical Communications*, 46(6):841–861. ISSN 1364-548X. doi: 10.1039/B922984A. URL https://pubs.rsc.org/en/content/articlelanding/2010/cc/b922984a.

[41] K.L. Kapoor. *Quantum Chemistry and Molecular Spectroscopy*. Macmillan, 4 edition. ISBN 978-0-230-32332-2. URL https://elearning.raghunathpurcollege.ac.in/files/0329716315869707630.pdf.

[42] Joseph J. Feher. *Quantitative Human Physiology: An Introduction*. Academic Press. ISBN 978-0-12-801154-6.

[43] Ferath Kherif and Adeliya Latypova. Chapter 12 - principal component analysis. In Andrea Mechelli and Sandra Vieira, editors, *Machine Learning*, pages 209–225. Academic Press. ISBN 978-0-12-815739-8. doi: 10.1016/B978-0-12-815739-8.00012-2. URL https://www.sciencedirect.com/science/article/pii/B9780128157398000122.

[44] Vrinda Kalia, Douglas I. Walker, Katherine M. Krasnodemski, Dean P. Jones, Gary W. Miller, and Marianthi-Anna Kioumourtzoglou. Unsupervised dimensionality reduction for exposome research. 15: 32–38. ISSN 2468-5844. doi: 10.1016/j.coesh.2020.05.001. URL https://www.sciencedirect.com/science/article/pii/S2468584420300349.

[45] Rasmus Bro and Age K. Smilde. Principal component analysis. 6(9):2812–2831. doi: 10.1039/C3AY41907J. URL https://pubs.rsc.org/en/content/articlelanding/2014/ay/c3ay41907j.

[46] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. 9(86):2579–2605. ISSN 1533-7928. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[47] Jyoti Pareek and Joel Jacob. Data compression and visualization using PCA and t-SNE. In Vishal Goar, Manoj Kuri, Rajesh Kumar, and Tomonobu Senjyu, editors, *Advances in Information Communication Technology and Computing*, pages 327–337. Springer. ISBN 9789811554216. doi: 10.1007/978-981-15-5421-6_34.

[48] Andy Coenen and Adam Pearce. Understanding UMAP. URL https://pair-code.github.io/understanding-umap/.

[49] Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. In Abderrahim El Moataz, Driss Mammass, Alamin Mansouri, and Fathallah Nouboud, editors, *Image and Signal Processing*, pages 317–325. Springer International Publishing. ISBN 978-3-030-51935-3. doi: 10.1007/978-3-030-51935-3_34.

[50] Krishan Pal and Mayank Sharma. Performance evaluation of non-linear techniques UMAP and t-SNE for data in higher dimensional topological space. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 1106–1110. doi: 10.1109/I-SMAC49090.2020.9243502. URL `https://ieeexplore.ieee.org/abstract/document/9243502`.

A

Additional plots

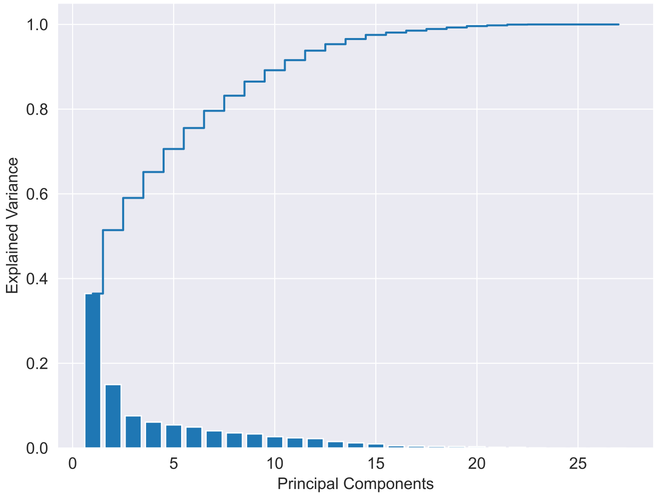
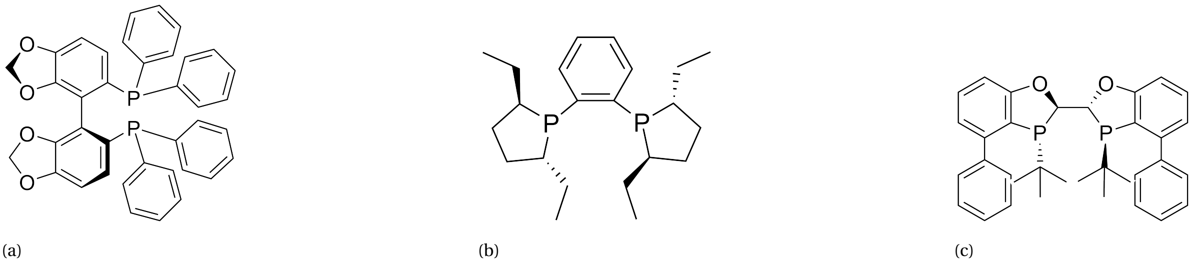


Figure A.1: Explained variance for each PC, together with cumulative explained variance

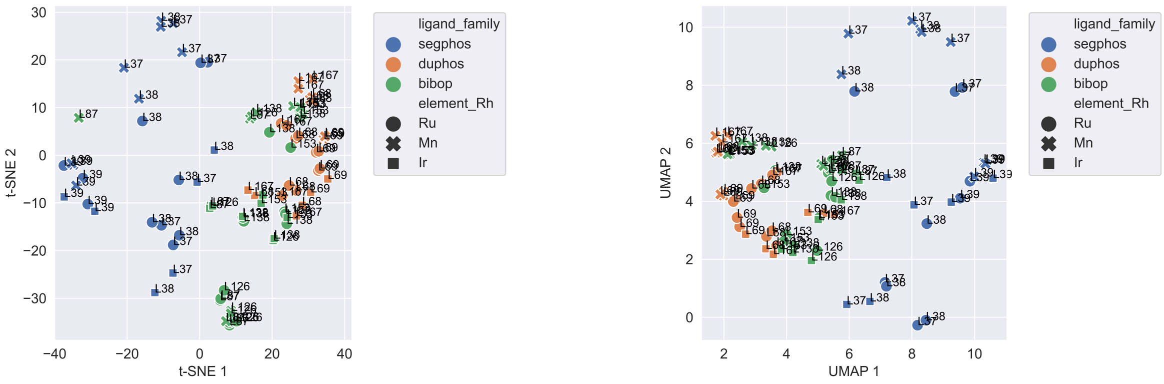


(a)

(b)

(c)

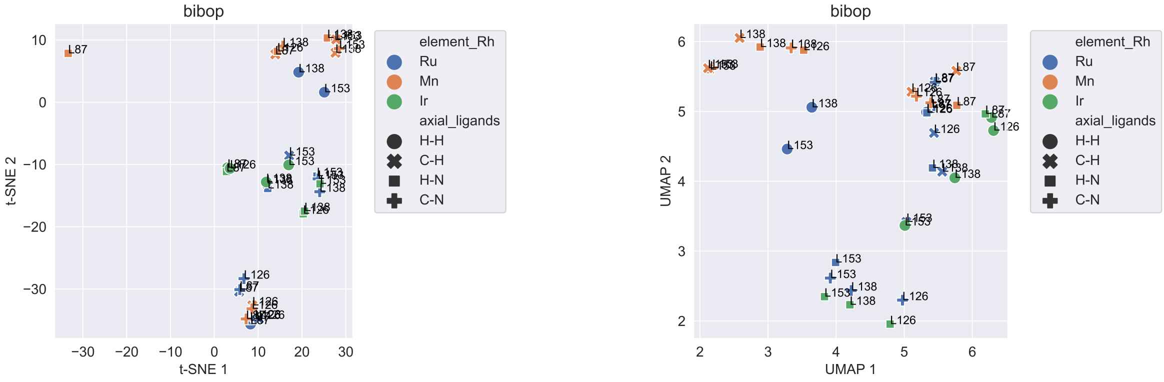
Figure A.2: Example of what ligands look like in the families a) segphos, b) duphos and c) bibop



(a)

(b)

Figure A.3: Score plots of the first two dimensions for a) t-SNE and b) UMAP for ligand families segphos, duphos and bibop. Colours indicate ligand family and marker style metal centre.



(a)

(b)

Figure A.4: Score plots of the first two dimensions for a) t-SNE and b) UMAP for ligand family bibop. Colours indicate metal centre and marker style ligand configuration.

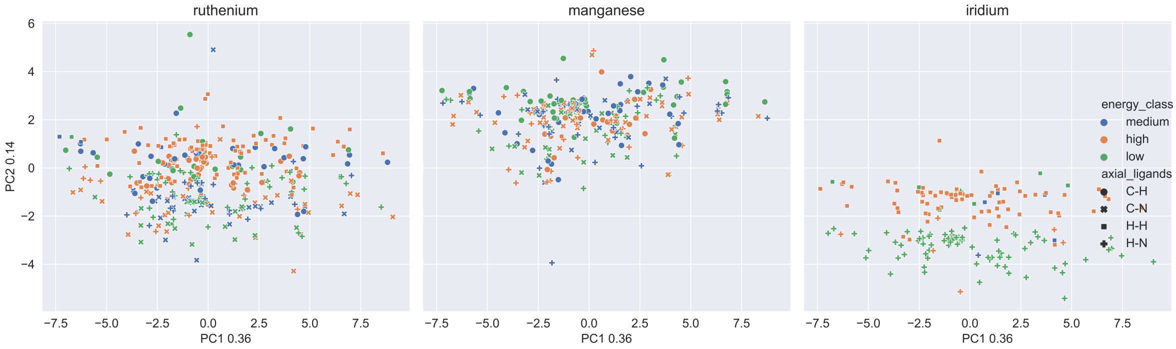


Figure A.5: PCA score plot of the first two PCs. Colours indicate energy category and marker style ligand configuration.

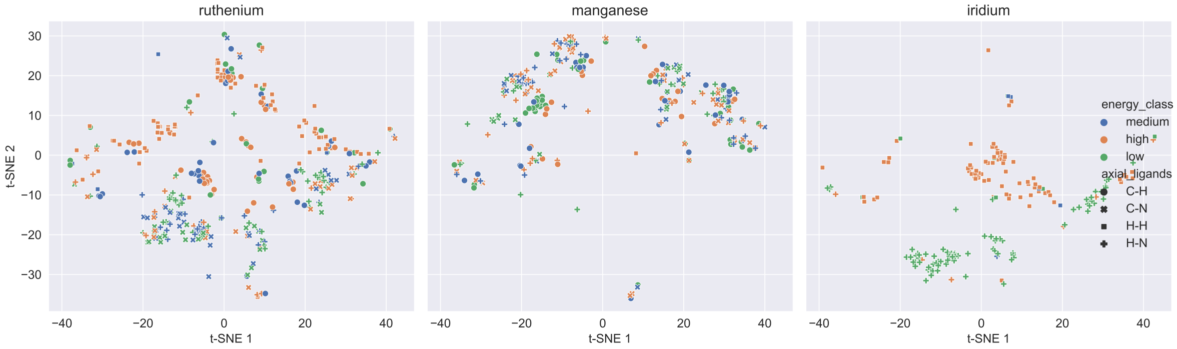


Figure A.6: t-SNE score plot of the first two dimensions. Colours indicate energy category and marker style ligand configuration.

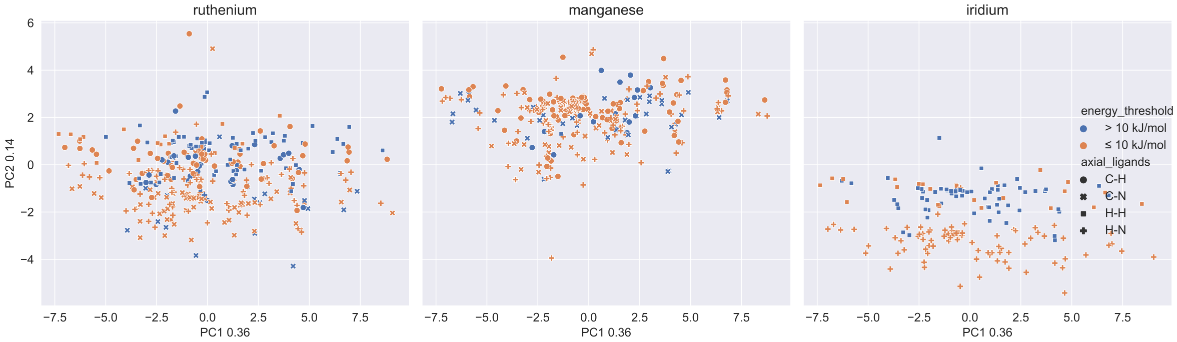


Figure A.7: PCA score plot of the first two PCs. Colours indicate energy threshold.

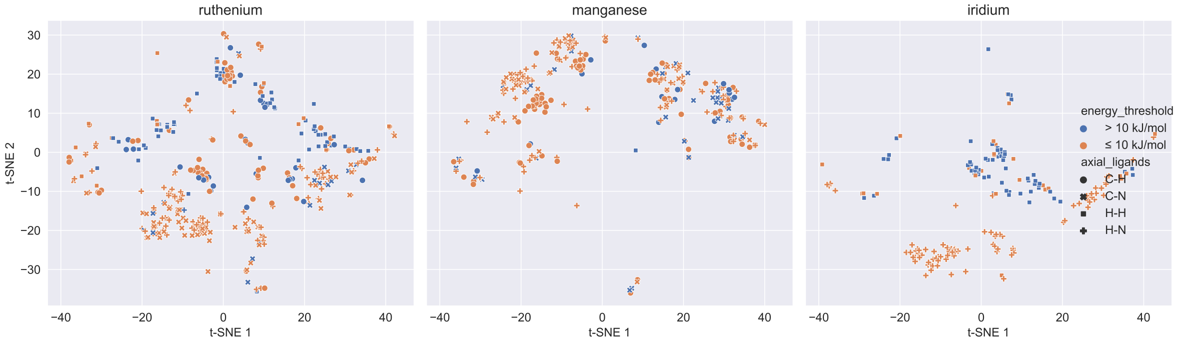


Figure A.8: t-SNE score plot of the first two dimensions. Colours indicate energy threshold.

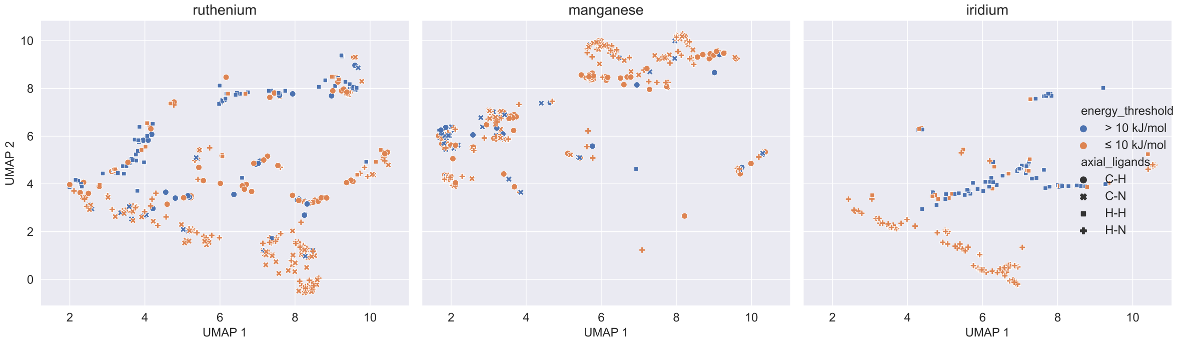


Figure A.9: UMAP score plot of the first two PCs. Colours indicate energy threshold.

B
Use of generative AI

Generative artificial intelligence (AI) was used for this thesis to help with coding. For a part of the problems encountered when making the plots, AI was asked how to solve these problems and get the code working. Generative AI was not used for any other part of this thesis.