Universitat Rovira i Virgili

Facultat de Química

Bachelor's Thesis, BSc Chemistry

Dimensionality reduction techniques for mapping diphosphine chemical space using the Ligand Knowledge Base (LKB)

Computational tools for organometallic catalysts optimisation

Mario Villares Cañón

Supervised by Dr. Natalie Fey

Tutorised by Prof. Elena Fernández

Bristol, United Kingdom

Acknowledgments

I would like to express my sincere gratitude to all the individuals who have contributed to the completion of this bachelor's thesis. Their support, guidance, and encouragement have been invaluable throughout this journey.

First and foremost, I would like to extend my deepest appreciation to my supervisor, Dr. Natalie Fey, for her unwavering support, expertise, invaluable guidance throughout the research process and for being such an inspiration. I would like to thank Dr. Carla Saunders who have provided me with support, encouragement, and valuable discussions throughout this research endeavour. And, to the whole Fey Group, whose friendship have been a constant source of inspiration and motivation.

I would like to thank Lukas for being my family in Bristol, for all the support, happy moments and for spelling out fun facts about the most random topic you could find. I'm sure you are going to be a great scientist and good luck with your bachelor's thesis.

Vull agrair al Dr. Jordi Carbó per haver-me introduït en el món de la química computacional, per la teva confiança i generositat, per el teu temps i per donar-me un lloc on poder iniciar-me en la recerca. És tot un plaer aprendre amb tu. Moltíssimes gràcies Jordi, de tot cor. També vull agrair a la Dra. Maria Besora per la seva dedicació i per totes les hores invertides en el meu aprenentatge. També vull agrair a la Dr. Elena Fernández pel tutoratge d'aquest treball.

A l'Albert Masip, l'Albert Solé i a en Toni Salom. Us admiro i estimo a parts iguals. Sou una font d'inspiració, no només pel vostre nivell acadèmic, sinó per lo grans que sou com a persones. Infinites gràcies pel suport rebut i per ser-hi sempre.

Agradecer también a Gonzalo, por toda tu paciencia y tiempo. Sin tu ayuda nada de esto habría sido posible. A tota la resta del Quantum Chemistry Group de la URV per acollir-me i fer-me sentir part del grup des de el primer moment.

A tots els professors i professores que, durant la meva vida com estudiant, m'han inspirat i han fet que avui en dia em vulgui dedicar a la recerca, en especial a la Maria Solanellas, la Teresa Fortuny i la Joana Angós. Gràcies a la vostra motivació a l'hora d'ensenyar vaig decidir estudiar Química. Us estaré eternament agraït.

A la Laura, Nerea, Marta, Abril i Marcos, per ser els millors companys de carrera que podria haver tingut. Amb vosaltres he crescut com a persona i com a químic.

A tota la meva família de la Jove. A l'Enric, la Irina, l'Adri, la Irene, el Pau, la Sardà, el Ferru, el Chals, la Mery, el Willy, la Núria, el Marc, la Paula, el Coco, el Muji, el Torija, l'Uri i el Peiret. Gràcies per ser-hi en els moments bons i en els difícils, fer-nos grans junts és la cosa més divertida del món.

I per acabar, agrair als meus pares per estimar-me, escoltar-me i recolzar-me en totes les decisions que he pres. Gràcies per creure en mi, ensenyar-me a ser persona i a viure en llibertat.

Table of content

Abstract

(ENG)

The chemical space is the multidimensional region where all known and unknown molecules are. Describing the diphosphine chemical space by the usage of the Ligand Knowledge Base (LKB) methodology, DFT-calculated property descriptors and the latter application of dimensionality reduction techniques lead to maps of chemical space that are useful for organometallic catalysts optimisation. Different dimensionality reduction techniques are tested (PCA, UMAP and t-SNE) and the information that such maps contain is determined by clustering algorithms. Structural characteristics have been used to see if the maps show trends. If structurally similar diphosphines are clustered together, this can likely be translated to similar catalytic performance since they will have similar properties. Test for comparing the cluster ability, robustness and extent of retained information are performed. At the end an experimental-based test case is carried out to demonstrate the potential that those generated maps and prediction models have on the task of optimising organometallic catalysts. The whole project was developed at the University of Bristol in Dr. Natalie Fey Group.

(CAT)

L'espai químic és la regió multidimensional on es troben totes molècules (conegudes o no conegudes). Descriure l'espai químic de les difosfines mitjançant l'ús de la metodologia Ligand Knowledge Base (LKB), calcular descriptors de propietat a través del DFT i la posterior aplicació de tècniques de reducció de la dimensionalitat ens proveeixen de mapes de l'espai químic útils per l'optimització de catalitzadors organometàl·lics. En el treball es proven diferents tècniques de reducció de la dimensionalitat (PCA, UMPA i t-SNE) i la informació retinguda en els mapes generats es determinada per algoritmes de clusterització. Les característiques estructurals han estat utilitzades per veure si hi ha tendències als mapes. Si difosfines amb característiques estructurals comunes són clusteritzades juntes, és altament probable que tinguin un comportament catalític similar ja que és l'estructura la que en determina les propietats. Es porten a terme diferents tests per comparar les habilitats de clustering, robustesa i extensió d'informació retinguda. El projecte es completa amb un exemple (test case) inspirat en un treball experimental per tal de demostrar el potencial que els mapes i models predictius generats tenen per tal d'optimitzar catalitzadors organometàl·lics. Tot el projecte ha estat elaborat a la Universitat de Bristol dintre del grup de la Dra. Natalie Fey.

Objectives

The proposed objectives for this bachelor thesis are:

- Perform a dimensionality reduction of the diphosphine chemical space with three different techniques: Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbour Embedding (t-SNE).
- Test the influence of the number of neighbours in UMAP, which is parameter responsible of capturing more or less global structure.
- Understand and compare the chemical information retained by each of the DR techniques for the construction of maps of chemical space that could be used as potential tools for catalysis optimisation and ligand selection purposes.
- Design a test case as an example for the potential uses of these maps as tools for optimising homogeneous catalysis.

Abbreviations

TEP – *Tolman electronic parameter.*

DR – *Dimensionality reduction.*

LKB – *Ligand Knowledge Base.*

MM – *Molecular Mechanics.*

MD – *Molecular Dynamics.*

MC – *Monte Carlo.*

QM – *Quantum Mechanics.*

FF – *Force Field.*

HF – *Hartree Fock.*

DFT – *Density Functional Theory.*

PCA – *Principal Component Analysis.*

t-SNE – *t-distributed Stochastic Neighbour Embedding.*

UMAP – *Uniform Manifold Approximation and projection*

KNN – *K-Nearest Neighbours.*

1

Introduction and goals. The chemical space: precedents and perspective for optimizing homogeneous catalysis.

1.1 Precedents and perspective on the use of computational approaches for optimizing homogeneous catalysis.

There is a great tradition in organometallic catalyst to characterize ligand properties with simple parameters (the words parameter and descriptor are commonly used interchangeably especially in organometallic chemistry). In the late seventies, Tolman[1] proposed two parameters, the cone angle (θ) for capturing steric information and the Tolman electronic parameter (TEP) capturing electronic information measured by infrared spectroscopy corresponding to the vibrational frequency of the highest carbonyl stretching mode of $[Ni(CO)_3L]$ complexes.

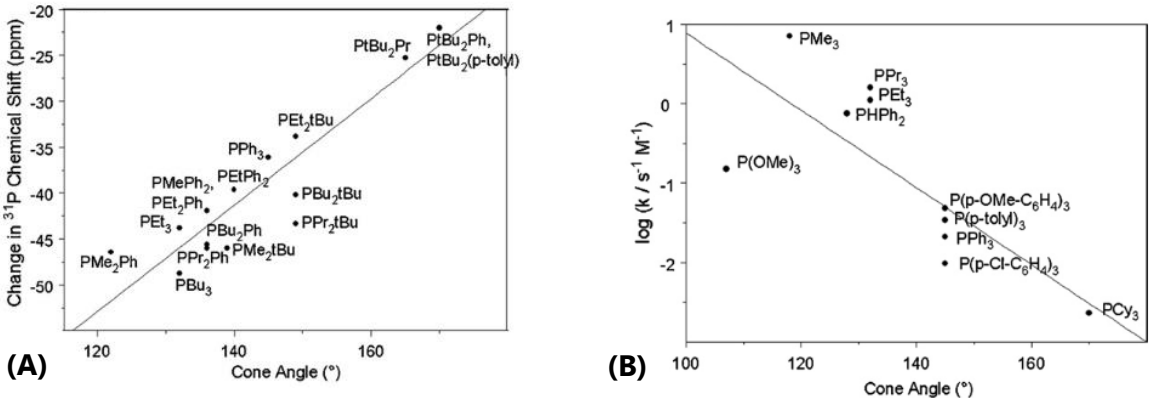


Figure 1. Correlations with cone angles (θ) [2] (adapted from ref. 1): (A) change in phosphorous chemical shifts form 31P NMR spectra of trans – $[RhCl(CO)L_2]$ on coordination ($\delta_{coordinated} - \delta_{free}$). (B) rate constant for the reaction[3] $2Co(DH)_2L + PhCH_2Br \rightarrow PhCH_2Co(DH)_2L + CoBr(DH)_2L$ in benzene, where DH = dimethylglyoxime or 1,2-cyclohexanedione dioxime. Copyright 2009 ELSEVIER.

In that work, Tolman showed that some experimental measurables are quite well described by these parameters, e.g. correlating cone angles (θ) with changes in ^{31}P chemical shifts and with rate constants of a given reaction as shown in Figure 1.

The idea behind this is that with these parameters, Tolman began to derive structure – property relationships. Continuing with this discussion, Tolman plotted the cone angle against the Tolman electronic parameter (TEP). In this work the author showed that ligands showing similar behaviours can be found relatively close to each other in chemical space and contrarily, the ones that were far apart, had different behaviours, as shown in Figure 2.

It is important to note that chemical space is vast, and there is often a wide range of structural and property diversity within it. While molecules with similar characteristics tend to be found close together, there can also be regions of chemical space with diverse structures and properties.

This is the key idea of this study. If a set of representative ligands is selected and their chemical space mapped, when new ligands are introduced, the properties and catalytic behaviour could be derived from the position where they sit in the created maps. Also, the models can be used to generate predictions on relevant catalytic features.[4] Moreover, this opens the door for other

ways to visualize these ligands with more sophisticated statistical approaches for mapping and clustering data.

The chemical space is defined as the theoretical ensemble that encompasses all possible chemical structures, including all known and unknown ligands for all possible catalytic reactions used to build relevant and valuable compounds. Mapping the chemical space involves identifying and categorizing molecules based on their structural and chemical properties. The total number of compounds limited to organic chemical space is estimated to be 10^{60}. However, the vast majority are unstable or synthetically inaccessible.[5]

It is quite simple to understand the global concept, but one may ask about the importance of mapping chemical space and which type of information chemists can extract from it.

As Fey explained in the review *Lost in chemical space? Maps to support organometallic catalysis*[4] there are three connected purposes that push us to map chemistry:

- **Fine-tuning and optimisation.** Ligands are an efficient way of enhancing the performance of a catalyst once a viable reaction has been optimized to be catalytic or a product can tune its properties by changing its substituents. These improvements can be directed by computation, allowing researchers to predict the properties of the studied specie before all synthetic work is carried out.
- **Selection (virtual screening).** Identify the most promising targets for synthesis but can be applied also to the evaluation of new designs before experimental realisation.
- **Exploration.** By automated combinatorial building and evaluation.

To be able to explore, understand, extract relevant chemical information, and even characterize unknown, one must rely on calculated property descriptors. In the context of computational chemistry and molecular modelling, a calculated property descriptor refers to a numerical value or characteristic that is derived from the molecular structure or other related data. In other words, descriptors transform the symbolic representation of a particular molecule into a number that describes it. In our case, descriptors are obtained through computational methods.

Calculated property descriptors are used to quantitatively represent various molecular properties, such as physical or chemical properties. The set of calculated property descriptors for each of the ligands selected constitute a dataset, which will be used to construct the maps and models.

Regarding property descriptors, several approaches to this topic had been reported before.[6] However, most of the parameters are specific to a certain type of ligand which limits the transferability. The Ligand Knowledge Base (LKB) approach developed by several research groups in Bristol[7,8] (A. G. Orpen, J. N. Harvey, E. Harris and N. Fey) includes a set of descriptors which are transferable to different coordination environments, different ligand donor atoms, and easy to determine computationally. This last characteristic is key to allow the consideration of novel designs before synthesis.

Descriptors are based on potential energies, avoiding the expensive computational cost of frequency calculations. Other features such as chelate effect and hemilability, which are challenging to be captured computationally, are not considered since solvent, dispersion or thermodynamical entropic correlations have not yet been introduced to the databases. These latter characteristics constitute one of the limitations that this field is facing nowadays, since a

methodology that provide computational affordable but at the same time accurate enough results needs to be developed in order to introduce these new descriptors.

Chemical space has no boundaries and can have as many dimensions as you want since you are defining it with descriptors. The chemical information is there, only an efficient way of extracting it is needed. Techniques and algorithms provide an image of the chemical space, but an alternative methodology may provide a different map, and so different chemical information.

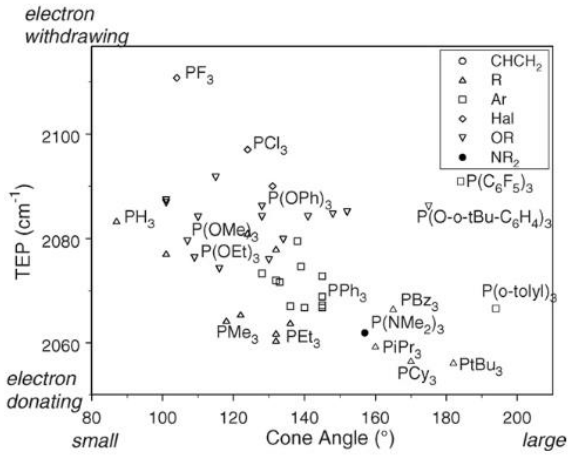


Figure 2. Map of phosphine chemical space (adapted from ref. 1): scatter plot of the cone angle (θ) against the Tolman electronic parameter (TEP). Copyright 2009 ELSEVIER.

Chemical space maps could be built with two descriptors as Tolman did *(see Figure 2)*. However, for descriptor databases with more than two descriptors, different statistical approaches are suitable for converting the data into a chemical space map.

These statistical approaches are dimensionality reduction techniques (DR), that will "move" our data into a low dimensional space making it suitable for analysis.

For the last decade Principal Component Analysis (PCA) has been used for this purpose, but recently, other authors have introduced new DR techniques due to the latest improvements on data analytics and cheminformatics. T. Gensch, M. Sigman, A. Aspuru-Guzik and collaborators proposed a new platform on organophosphorus ligands for catalysis called "The Kraken".[9] The platform combines computational modelling techniques with experimental validation, enabling researchers to efficiently explore a wide range of ligand structures and predict their catalytic properties. The authors highlight the significance of this platform in overcoming the limitations of traditional trial-and-error approaches and reducing the time and resources required for ligand discovery.

During the exploratory analysis of data and inspired by the work of Fey, organophosphorus chemical space maps are created. The authors used PCA and Uniform Manifold Approximation and Projection (UMAP), which is a relatively new DR technique.[10] The authors noticed two main groups in the UMAP chemical space map: electropositive and electronegative phosphorous substituents. Moreover, the authors colour-code the maps according to substituents and clusters appear, allowing to distinguish different ligand classes. No results are reported about the type and extent of chemical information retained on such maps.

Comparing, validating, and testing the model potential based on the abilities of each DR technique of capturing relevant chemical information could be useful for new approaches, not

only on visualizing such maps, but also to help catalysis optimisation purposes, and maybe answer what type of representations of chemical space are most useful and intuitive for routine uses.

1.2 Diphosphine dataset. The LKB – PP and the LKB – PP $_{screen}$.

The LKB contains thousands of ligands, either monodentate or bidentate with different donor atoms. Different subsets of the LKB are published, with slight variation of descriptors depending on the set. The LKB subsets[11–15] are: LKB-P (for monodentate phosphorus donor), the LKB-C (for monodentate carbene donor), the LKB-bid (for bidentate and various donor atoms) and the LKB-PP and LKB-PP$_{screen}$ (for bidentate phosphorus donor atoms), whose differences will be explained at the end of this section. This present work will be focussed on the diphosphine subset *(see Figure 3).* Diphosphine ligands are versatile tools for controlling and enhancing the reactivity of metal catalysts.[16] They provide extra stability to the metallic centre due to the chelate effect, since it is entropically favoured due to the coordination of one molecule with two donor atoms instead of two molecules with one donor atom each,[17] easy modulation of steric and electronic properties by changing their substituents, facilitate the activation of substrates and enhance selectivity towards directed mechanism pathways.[18]

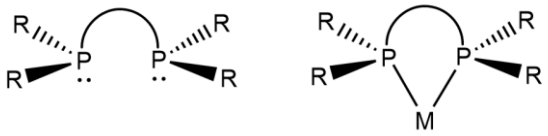


Figure 3. Schematic general representation of a diphosphine. Left: free diphosphine (in coordination conformation) with general R substituents. Right: diphosphine coordinated to a general metallic atom (M) showing the chelate formation.

The set of descriptors that constitutes the LKB-PP are calculated property descriptors. The main reason why descriptors are only based on Density Functional Theory (DFT) calculations is because the main objective of the LKB is to provide useful dataset for synthetic target identification and experimental design, meaning that, for instance, by calculating the set of descriptors for a novel ligand you could predict its properties and catalytic behaviour before carrying out the synthesis.

The published work on the LKB-PP[6,14] highlights the importance of computational descriptors in providing insights into the structural and electronic features of chelating ligands. By accurately capturing the relevant molecular properties, these descriptors enable one to predict ligand behaviour, optimize ligand structures, and guide ligand selection for specific applications. Fey and collaborators outline the development of the computational descriptors, which involve a combination of DFT calculations, molecular modelling, and statistical analyses using PCA to reduce the dimensionality and creating a map of diphosphine chemical space.[13] More information about the descriptors and the methodology in chapter 3.

The LKB-PP dataset has 474 published ligands.[19] After the work of some collaborators and after my internship in Fey group where I was trained in the ligand introduction methodology, the data set contains 508 diphosphine ligands (The ligands I introduced can be found in annex Figure 1). However, the LKB-PP has large groups with similar characteristics such as backbone length and substituents (most aryl). This can cause a bias and may produce negative effects on the analysis of the group of ligands that do not share these specific characteristics. As a result, the chemical space is non-uniformly covered, as shown in Figure 4.

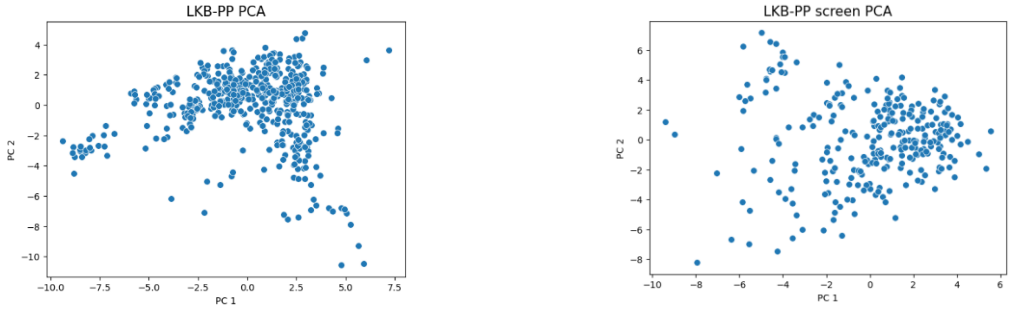
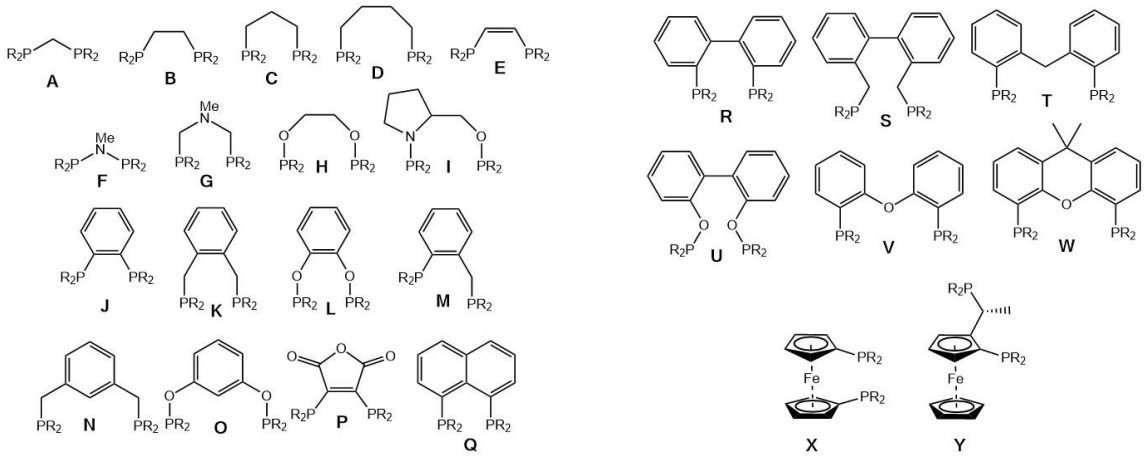


Figure 4. Left: 2D - PCA representation of LKB – PP dataset. Right: 2D – PCA representation of the LKB – PP_{screen} . It can be seen that, for the case of the LKB – PPscreen the diphosphine chemical space is better uniformly covered.

That is why the dataset used for generation of chemical space maps, applying different DR techniques and the further analysis in is not the LKB – PP. Instead, the LKB – PP $_{screen}$ is used, which was developed by Fey and Jover in 2013.[20] The idea of the LKB – PP $_{screen}$ is to have a more equilibrated dataset in the sense of ligand substituent and backbone. The paper[13] emphasizes the importance of this screening approach in guiding ligand design and optimizing ligand selection for specific catalytic applications. By evaluating a broad range of substituents and backbones, insights are gained into their effects on metal-ligand bonding, ligand flexibility, and steric hindrance. This knowledge can then be used to tailor ligands with improved catalytic activity, selectivity, and stability. The same descriptors are used as in the LKB-PP, detailed information about descriptors and methodology in section 3.

Fey and Jover combined 25 backbones and 11 substituents, leading to a total of 275 ligands, as shown in Figure 5. The backbones and substituents were decided according to representative references in organometallic catalysis based on an extensive bibliographical search. The number of descriptors calculated for each ligand is 28 and they are the same as in the LKB-PP. Detailed information about descriptors and methodology in section 3.



Number	*Substituents (R)*
1	Me, tBu
2	CF_3, C_6F_5
3	Ph, o-Tol, p-Tol, p-CF_3-Ph
4	OMe, OPh
5	NMe_2

Figure 5. *(Top)* Backbones in the LKB – PP_{screen} identified with letters from A-Y for colour coding. *(Bottom)* 11 substituents used in the LKB – PP_{screen} and the corresponding numbering used: 1-corresponding to C_{sp3} substituents, 2 – halogen containing compounds, 3 – Aromatics, 4 – Oxy groups, 5 – Amine group. The backbone lengths go form 3-8 counting the P atoms as part of the backbone.

2

Theoretical basis.

2.1 Computational basis. Molecular Mechanics and Force Fields (MM), conformational searches and Density Functional Theory (DFT).

Molecular Mechanics Force Fields (MM) and conformational searches.

Molecular mechanics force fields, also known as empirical force fields, are computational models used in molecular simulations to describe the interactions and behaviour of molecules. They are based on classical mechanics principles and approximate the potential energy of a system as a sum of terms that represent different types of interatomic interactions.

The fundamental concept behind molecular mechanics force fields is that the total energy of a molecular system can be expressed as the sum of bonded and non-bonded terms. These terms capture the contributions from bonds, angles, dihedral angles, and non-bonded interactions such as Van der Waals forces and electrostatic interactions, shown in equation 1:

$$E_{MM} = E_{str} + E_{bend} + E_{tor} + E_{VdW} + E_{elec} + E_{cross} \quad (1)$$

Where E_{str} corresponds to stretching energy, E_{ben} to bending energy, E_{tor} to torsion energy E_{VdW} to Van der Waals contribution energy and E_{elect} to electrostatic interactions energy. E_{cross} corresponds to the coupling contribution of the bonding terms.

- **Bonding interactions.** Bonded terms describe the stretching and bending of chemical bonds within a molecule. They include bond stretching terms, which model the stretching of bonds as springs, and angle bending terms, which account for the bending of angles between bonded atoms. Dihedral terms capture the rotation around a bond, representing the potential energy associated with the torsional angles in a molecule. These terms are used to describe the conformational flexibility of molecules and the barriers to rotation around specific bonds.
- **Non-bonding interactions.** Non-bonded terms account for interactions between atoms that are not directly connected by chemical bonds. This includes Van der Waals interactions, which represent attractive and repulsive forces between atoms including dipole – dipole and London Dispersion Forces, and electrostatic interactions, which describe the interactions between charged atoms or groups.

The parameters in molecular mechanics force fields are typically derived from experimental data, quantum mechanical calculations, and empirical fitting to reproduce experimental observations. These parameters define the force constants, equilibrium values, and other properties of the interactions.

It is important to note that molecular mechanics force fields are empirical models and have limitations. They are based on simplified assumptions and do not capture all aspects of molecular behaviour, such as electronic effects.

Molecular mechanics are used in computational chemistry to carry out conformational searches. The goal of a conformational search is to identify low-energy conformations, which correspond to stable or energetically favourable structures. This is typically achieved by evaluating the energy of each conformation using in our case, a force field (although QM calculations can also be used).

There are different types of conformational searches such as: systematic search, random search, molecular dynamic (MD) simulations and Monte Carlo (MC) simulations.

MC simulations use probabilistic sampling to explore the conformational space. By randomly changing dihedral angles or atomic positions and accepting or rejecting these changes based on energy considerations, MC simulations can sample different conformations. Conformational searches are useful for exploring the different spatial arrangements and potential energy surfaces of molecules, providing insights into their stability, reactivity, and interactions.

Density Functional Theory (DFT).

Thomas-Fermi's models tried for the first time in 1927 to use electronic density (ρ) instead of using the wave function (ψ) in order to solve the problem of the electronic molecular structure. However, density functional theory (DFT) developed by Hohenberg, Kohn and Sham demonstrated that any system can be described starting from its electronic density, allowing a great improvement in quantum mechanical methods. DFT is able to take in account the electronic exchange and correlation without increasing significantly the computational cost, making affordable to obtain geometries and energies for medium-sized systems.

DFT theory describes electronic states as a function of the electronic density, which depends on three spatial coordinates (x, y, z). This was a great advantage with respect to Hartree-Fock (HF) and *ab initio* methods since they used the wave function, which depends on 3N coordinates (being N the number of electrons) to describe electronic states.

The functional E[ρ] will allow us to produce an energy value (E) starting from a function (ρ), and so by knowing the functional, the exact energy will be obtained. Equation 2 shows the functional expression:

$$E[\rho] = T[\rho] + V_{Ne}[\rho] + V_{ee}[\rho] \quad (2)$$

Where T represent the kinetic energy, V_{Ne} the electron – nuclei attraction, and V_{ee} the electron – electron repulsion, which can be decomposed in coulombic contribution (J[ρ]) and exchange contribution (K[ρ]). However, kinetic energy for an electronic density function cannot be described neither the exchange contribution for the electron – electron repulsion. As usual in computational chemistry, approximations are used to know the energy, therefore, the energy will not be exact. [21]

The first problem is the kinetic energy. Kohn and Sham in 1965 proposed that the kinetic term could be defined as s sum of two terms: $T_s[\rho]$ and $T_c[\rho]$.[21] The first consider system of independent electrons (non-interacting) and the second one is accounting for the interactions that were missing. The problem is that DFT is no longer dependent on 3 spatial coordinates, now it depends on 3N coordinates because Ts term reintroduces Molecular Orbital Theory (MOT) and so the concept of wave function.

For this reason, in equation 3, the kinetic correlation and the exchange contribution are introduced in term $E_{xc.}$

$$E[\rho] = T_s[\rho] + V_{Ne}[\rho] + J[\rho] + E_{xc}[\rho] \quad (3)$$

Since then, the development of functionals has been the key point for the achievement for more accurate calculations. Functionals are usually classified in four main types:

- **Local Spin Density Approximation (LDA).** The expression only depends on the electronic density (ρ)

- **Generalized Gradient Approximation (GGA).** The expression depends on the electronic density (ρ) and its gradient (the first derivative).
- **Meta-functionals (Meta-GGA).** It is the same as the GGA but also introducing the second derivative.
- **Hybrid functionals.** Since the exchange contribution is well calculated by HF methods, hybrid functionals take in account a percentage of it, combined with the exchange correlation energy from other sources (*ab initio* or empirical)

To make obtain good results, experimental parameters are often introduced in some functionals. It would be impossible to have a universal one since the set of experimental measures that the functional would require will be infinite. For this reason, different functionals are used depending on the system.

2.2 Statistical basis. Dimensionality reduction techniques (DR) and clustering algorithms.

The LKB-PP_{screen} contains 28 descriptors, and therefore 28 dimensions. These descriptors allow one to map the chemical space. However, it is challenging to visualize more than 3 dimensions. To do such task DR techniques are used to take that high dimensional space and convert it into a plot with 2 or 3 dimensions as maximum. In this thesis, 3 DR techniques have been tested: Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbour Embedding (t-SNE). Clustering techniques were then used to see what information is retained, since reducing dimensionality is only possible if some information is lost.

As a general idea PCA is able to capture global relationships, t-SNE is able to capture local relationships and UMAP can capture both, making this latter DR technique the most complex of the list.

Local relationships capture the idea that data points that are similar or share local relationships should be grouped together in the visualization, while global relationships refer to the overall organization and relationships between different clusters or groups of data points in the original high-dimensional space. Preserving global structure ensures that the larger-scale patterns, such as broad clusters, separations, or trends, are maintained in the lower-dimensional projection.

Principal component analysis (PCA).

PCA transforms the original parameters, capturing the distance between objects when dimensionality reduction is performed. It identifies the principal components, which are orthogonal directions that capture the most significant variation in the data, making it widely used for data compression, visualization, and feature selection. It was proposed by Karl Pearson in 1901.[22]

Capturing maximum variation means that the projection of the data over an axis can cover the maximum space possible. In this way the information lost is minimized as shown in Figure 6. The technique chooses the axis that provides the maximum variation. In Figure 6 A the projection is concentrated in the middle of the axis, but in Figure 6 B it is spread along the whole axis. Representation on Figure 6 B will be chosen as first component.

Second component will be orthogonal to the first one, the third will be orthogonal to first and second components and so on. That is why the first component retains maximum information (maximum explained variance).

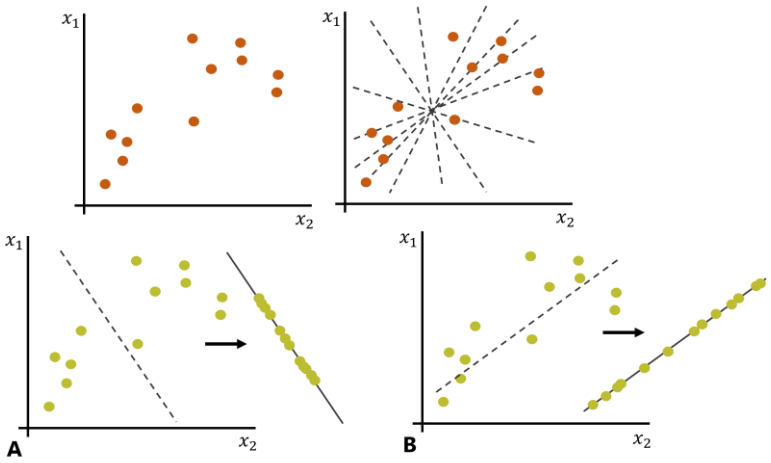


Figure 6. Schematic representation of reduction of dimensionality with PCA. Top plots represent the raw data and all possible components that could be chosen as first PC. (A) Projection of data over the highlighted axis giving lower explained variance. (B) Projection of data over the highlighted axis giving maximum explained variance.

The function that is used to reduce dimensionality is a linear combination of loadings and scores, as explained in equation 4:

$$PC_m = X_1 v_{1,m} + X_2 v_{2,m} + \ldots + a_n v_{n,m} \quad (4)$$

Where X is the original data and $v_{n,m}$ the loadings, being m the number of the component, and n the variable number.

Scores are the values of the linear combination of the individual original data on the new axes system and loadings are the weight of each variable in each component. By linear combination each PC Is obtained. For this reason, PCA is a linear dimensionality reduction technique.

t-distributed Stochastic Neighbour Embedding (t-SNE).

t-SNE (t-Distributed Stochastic Neighbour Embedding) is a dimensionality reduction technique commonly used for visualizing high-dimensional data in a lower-dimensional space. It was introduced by Laurens van der Maaten and Geoffrey Hinton in 2008.[23]

The main goal of t-SNE is to reveal the underlying structure or patterns in complex datasets by representing them in a more easily interpretable form. Unlike traditional linear methods such as PCA (Principal Component Analysis), t-SNE is particularly effective at capturing non-linear relationships and preserving local structure.

The t-SNE algorithm works by constructing a probability distribution over pairs of high-dimensional data points and a similar probability distribution over pairs of their corresponding points in the lower-dimensional space. It then minimizes the divergence between these two distributions using gradient descent, aiming to find a mapping that best preserves the pairwise similarities. t-SNE computes probabilities p_{ij} (equation 5 and 6) that are proportional to the similarities of objects x_i and x_j as follows:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \quad (5)$$

$$p_{j|i} = \frac{exp\left(-||x_i - x_j||^2/2\sigma_i^2\right)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)} \quad (6)$$

Where N is the dimension of the space and σ the effective neighbours that each point has (perplexity), and k parameter refers to the number of nearest neighbours considered for a neighbourhood graph.

As shown in Figure 7, the unscaled distance is measured (Figure 7 A) and then by means of a normal t-distributed probability function for a certain random state which in this case has started in the red coloured point, centred in the distribution, (Figure 7 B) leads to 1 dimensional representation (Figure 7 C) where data has reduced its dimensionality clustering the groups that had similarities.

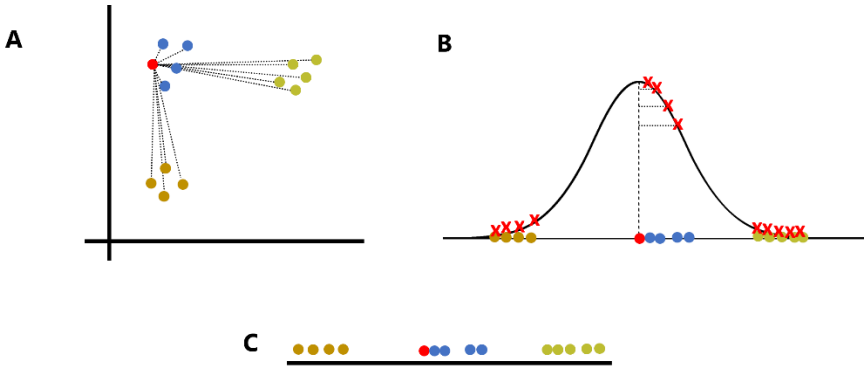


Figure 7. Schematic representation of the reduction of dimensionality for t-SNE. (A) Representation of raw data, in this case with only two variables for simplification purposes. Dotted lines correspond the distances measured. (B) The distances are then used in the normal t-distributed probability function that projects each point with respect the red one, which is set by a random state. (C) Result of the projection, and one-dimensional representation of the data.

One important aspect of t-SNE is that it is a stochastic algorithm, meaning that different runs of the same dataset may produce slightly different results. It is essential to keep this in mind when interpreting the visualizations generated by t-SNE since all data points are randomly projected in one dimension and then the probabilistic function clusters them together based on their local relationships.

t-SNE has a few key parameters that can be adjusted to influence the behaviour and performance of the DR technique. Here are the main parameters used in t-SNE:

- **Perplexity**. Can be thought of as a measure of the number of "effective neighbours" that each data point has. A higher perplexity value suggests that each point should consider a larger number of neighbours during the dimensionality reduction process. Perplexity values oscillate between (5 - 50) suggested by van der Maaten & Hinton.
- **Learning Rate.** The learning rate controls the step size at each iteration of the optimization process. It determines how quickly t-SNE updates the mapping in each iteration.
- **Number of Iterations.** This parameter determines the maximum number of iterations the algorithm will perform during the optimization process. Increasing the number of iterations can improve the convergence of t-SNE, but it also increases the computation time.

- **Metric.** The choice of distance metric used to compute pairwise similarities between data points can have an impact on the results. The most commonly used metric is Euclidean distance, but other metrics like cosine distance or correlation distance can be used depending on the nature of the data.
- **Method.** It refers to the calculation of the gradient algorithm. By default it uses the Barnes-Hut approximation, but it only allows t-SNE to build models with a maximum of 3 components. If more components are desired, the exact Barnes-Hut algorithm needs to be used instead.
- **Random State.** t-SNE is a stochastic algorithm, and the random seed parameter determines the starting point of the random number generator. Fixing the random seed ensures reproducibility, meaning that running t-SNE with the same seed will produce the same result.

It is important to note that the parameter values should be chosen carefully based on the characteristics of the data and the desired outcome. Adjusting these parameters can significantly affect the resulting visualization, so experimentation and understanding the impact of each parameter are crucial when applying t-SNE to a specific dataset.

Uniform Manifold Approximation and Projection (UMAP).

UMAP (Uniform Manifold Approximation and Projection) is a dimensionality reduction technique that is widely used for visualizing and analysing high-dimensional data. It was introduced by Leland McInnes, John Healy, and James Melville in 2020.[10]

UMAP is designed to preserve both local and global structure of the data, making it effective at capturing complex relationships and patterns. The UMAP algorithm works by constructing a low-dimensional representation of the data that is topologically equivalent to the original high-dimensional space. It achieves this by optimizing a cost function that balances the preservation of pairwise distances and the preservation of local neighbourhood structures.

The idea and the overall shape of algorithms is quite like t-SNE. However, the inner mathematics are much more complex, and the functions used to reduce dimensionality are based on statistical similarities. As shown in Figure 8, in order to calculate the statistical similarities, the distances between data points are needed (Figure 8 A). The raw distance and its projection over 1 dimensional axis provide the global relationships of the data (Figure 8 B), and the function assures that data points with similarities are packed closely (Figure 8 C) introducing the effect of local relationships.

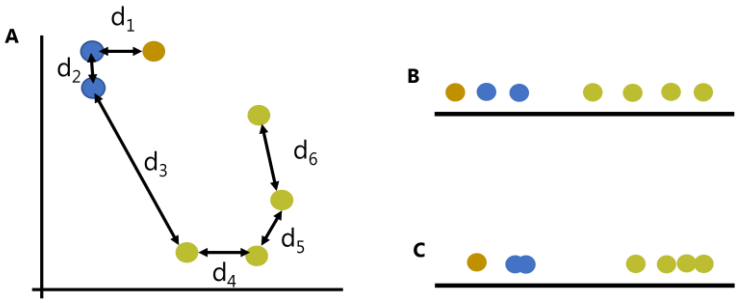


Figure 8. Schematic representation of the reduction of dimensionality with UMAP. (A) Raw data and measure of distances between points. (B) Projection of data to one dimensional axis based on the distances. (C) Application of the similarity function making that data points with common characteristics are packed together.

One of the key advantages of UMAP is its scalability, allowing it to handle large datasets efficiently. It can also handle both numerical and categorical data, making it versatile for various types of data analysis tasks. Since the functions that reduce dimensionality are not linear (equation 7), UMAP is a non-linear DR technique as well.

$$C = \sum_{ij} log\left(\frac{1}{w_{ij}}\right) + log\left(\frac{1}{1 - w_{ij}}\right) \quad (7)$$

Where C is the cost function and w_{ij} represents each of the data points.

Then the derivative of the function respect each data point is done, obtaining the stochastic descendent gradient, which guides de direction where to move the data points for lowering dimensionality.

As happened with t-SNE UMAP provides several parameters that can be tuned to influence the behaviour of the algorithm. These include the number of neighbours, the minimum distance, and the metric (same concept as in t-SNE) used to measure distances between data points. Adjusting these parameters can impact the resulting visualization and should be chosen carefully based on the characteristics of the data and the desired outcome.

- **Number of neighbours.** It effectively controls how UMAP balances local versus global structure. Low values will push UMAP to focus more on local structure by constraining the number of neighbouring points considered when analysing the data in high dimensions, while high values will push UMAP towards representing the big-picture structure while losing fine detail.
- **Minimum distance.** The parameter controls how tightly UMAP is allowed to pack points together. It, quite literally, provides the minimum distance apart that points are allowed to be in the low dimensional representation. Low values of minimum distance will result into more packed map representations. It usually goes from 0.0 to 1.0.

As seen UMAP and t-SNE rely on the same concept but with different execution. UMAP is known for its ability to preserve global structures, scalability, and efficient computation. t-SNE excels at visualizing local structures and for high dimensional data it provides great clustered maps. The choice between UMAP and t-SNE depends on the specific goals of the analysis and the characteristics of the dataset, and that is why both will be tested.

Comparing PCA, UMAP and t-SNE is a difficult task. The concept of explained variance does not exist for non-liner DR techniques since it relies on loadings for its calculation. Moreover, and for the same reason, the weight of variables in each component cannot be obtained.

k-means and hierarchical clustering algorithms.

In combination with DR techniques, clustering algorithms are used as a test to see the retained information. Both are unsupervised algorithms, meaning that the data has not been labelled.

k-means clustering is an iterative algorithm that partitions data points into k clusters, where k is a user-defined parameter representing the desired number of clusters. First it randomly initialises k cluster centroids in the feature space, then it assigns each data point to the nearest centroid based on a distance metric, which is usually Euclidean distances. After that, the centroid is recalculated by calculating the average of the previously assigned points. These

steps are repeated up to convergence of centroids, or until the maximum number of iterations is reached.

The result of k-means clustering is a set of k clusters, each represented by its centroid. K-means aims to minimize the within-cluster sum of squared distances, making each point in a cluster close to its cluster centroid as shown in equation 8. K-means is computationally efficient but sensitive to the initial centroid placement and may converge to local optima.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \quad (8)$$

Where k is the number of clusters, n the number of data points, c the centroid for cluster j and J the objective function.

Hierarchical clustering builds a tree-like structure (dendrogram) of data points based on their similarity. There are two main types of hierarchical clustering:

- **Agglomerative Hierarchical Clustering (bottom-up).** Initially, each data point is treated as a separate cluster. The algorithm successively merges the most similar clusters until all points are contained in a single cluster.
- **Divisive Hierarchical Clustering (top-down).** This approach starts with a single cluster containing all data points and recursively splits the cluster into smaller subclusters until each data point forms its own cluster.

Different algorithms and formulae are used depending on the linkage. The one that is commonly used and the one used in this bachelor thesis is the ward linkage, as described in equation 9.

$$d_{ij} = d(\{X_i\}, \{X_j\} = ||X_i - X_j|| \quad (9)$$

Where X_i and X_j are two points of data.

Hierarchical clustering doesn't require the number of clusters to be predefined. It results in a dendrogram that represents a hierarchy of nested clusters. The dendrogram can be cut at different levels to obtain different numbers of clusters.

3

Methodology and design. DFT – calculated property descriptors and dimensionality reduction techniques.

3.1 Methodology followed in the development of DFT – calculated property descriptors in the Ligand Knowledge Database (LKB-PP).

Property descriptors need to be transferable and be able to capture and incorporate information for different coordination environments. To do so, in the LKB approach used for bidentate ligands, DFT calculations based on 5 different representations per ligands are done. These representations are free ligand, two metallic fragments and two helium wedges as show in Figure 9.

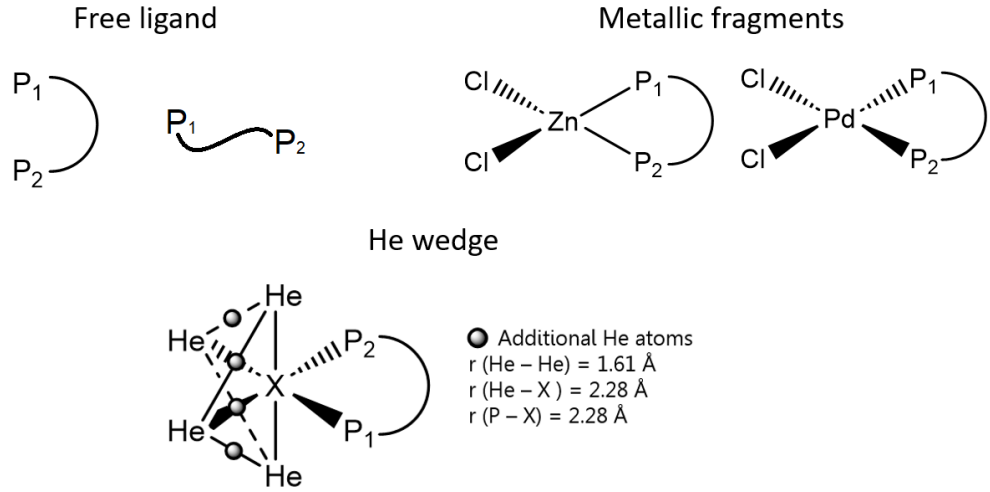


Figure 9. Free ligand (showing both possibilities, chelating and non-chelating conformation), metallic fragment and He wedge models in the LKB-PP that will allow to obtain a set of DFT – calculated property descriptors. Phosphine substituents are removed to have a clearer image.

Before any starting geometry for the ligand and representative complexes is constructed, Molecular Mechanics (MM) stochastic type conformational searches were performed, to screen conformational space for free ligands and their tetrahedral zinc complexes [$ZnCl_2$(PP)]. These searches were aimed at eliminating strained, high-energy conformers as input geometries, rather than reliably locating the global minimum for both MM and DFT. The conformational noise impact was already estimated in previous work for the LKB-PP.[6] The conformer of lowest MM energy was then used as input for DFT optimizations, modifying the metal fragment when required for constructing the rest of the representations, with the help of already prepared scripts.

A set of descriptors combining steric and electronic properties comes from the metallic fragments. The zinc dichloride fragment [$ZnCl_2$(PP)] acts as a Lewis acid since it has no contribution from Crystal Field Stabilization Energy (CFSE). The zinc fragment forces the ligand to be in a binding confirmation, adopting something close to the "natural" bite angle.[24] The palladium dichloride fragment [$PdCl_2$(PP)], has in contrast a strong preference for a square planar coordination geometry (*see Figure 9).* In previous work, the Fey group tried additional calculations with higher coordination numbers, but actually turned out that the Pd complex gives a reasonable model for octahedral complex.[15]

Ligands themselves are optimised as free ligands, and in a binding conformation through the zinc dichloride complex [$ZnCl_2$(PP)] by simply subtracting the total energy of the complex and the energy of the zinc dichloride itself as show in equation 10.

$$BE = E_{TOT}(fragment) + E_{TOT}(L) - E_{TOT}(complex) \quad (10)$$

Two steric descriptors are calculated by using a wedge of helium atoms as a model. These helium atoms are located where *cis* ligands would be found in an octahedral complex, mimicking way the steric hindrance of an idealised coordination environment.

As shown in Figure 10, both start with the optimised ligand structure obtained in the zinc dichloride complex. The zinc dichloride is replaced by a dummy metal atom, adopting a trigonal planar geometry. The resulting structure with the application of some constraints lead to the two steric descriptors.

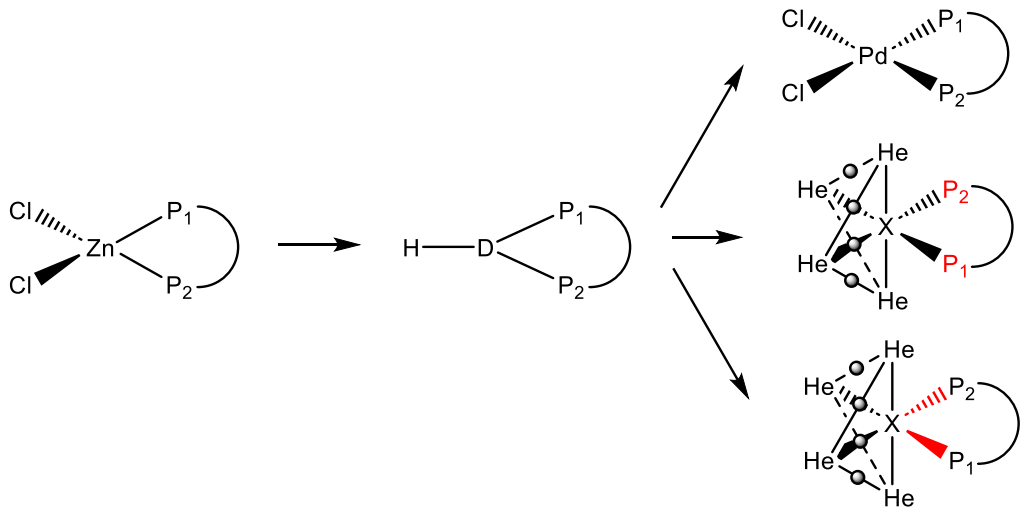


Figure 10. Schematic representation for the development of Palladium complex and the two helium wedges. The replacement of $ZnCl_2$ by a dummy atom (D-H) is done in Maestro software (See computational details). Then by using the corresponding fragment script, the dummy atom is replaced, which gives the three missing models. For the helium wedge descriptors, the constrained atoms and distances are highlighted in red.

As shown in Figure 10 there exist two He complexes. The first one is the He_8_wedge, which freezes the donor atom position optimised in the positions obtained in the zinc complex calculation giving rise to a reasonable approximation of the natural bite angle which cannot adjust in response to steric strain. The second one is the nHe_8 which constrains the X-P distance to 2.28 Å *(see Figure 9)* but allow donor atoms to move capturing if ligands can respond to steric hindrance by changing the bite angle. In other words, to see if they can adjust to the steric pressure. This distance is the one provided by previous studies carried out by Tolman.[1] *(See Data and code availability section for an example of these input geometries)*.

Electronic descriptors are obtained from the Highest Occupied Molecular Orbital (HOMO), Lowest Occupied Molecular Orbital (LUMO) and proton affinity (PA) via a truncation methodology, which has been developed in the past years in the Fey group.

This truncation methodology introduces a simplification of the real chemical structure of the ligands. Previous work[8,25] has been done to prove that this simplification is not generating significant differences on the calculated electronic property descriptors. It simply truncates the diphosphine and treat each phosphorus separately *(see Figure 11)* which saves a considerable

amount of time and work. Moreover, if the truncation methodology was not used, the electronic descriptors for each individual donor atom could not be computed.

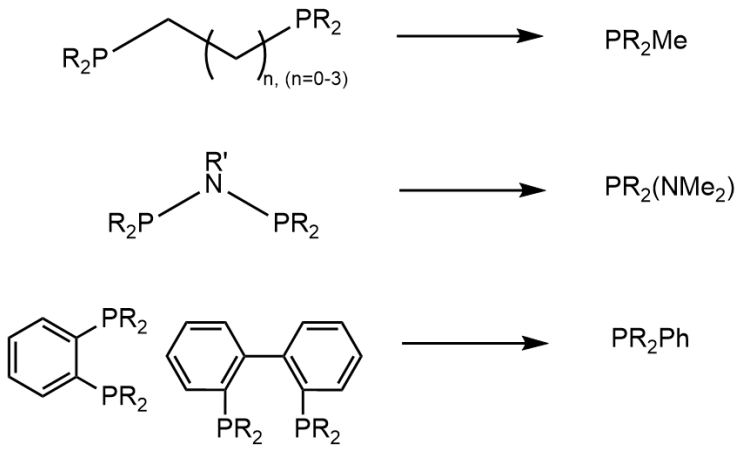


Figure 11. Schematic representation of the truncation methodology for ligands in the LKB-PP and LKB-PP_{screen}.

Once the electronic descriptors are obtained, the rest of descriptors are calculated by a series of scripts acting on the before mentioned representations, resulting in a total of 28 calculated property descriptors summarized and described in Table 1.

Table 1. Descriptor data harvested from calculations on complexes (Scheme x). Donor atoms in fixed positions, fixed "D–X" distance = 2.28 Å (P). Fixed "D–X" distances (as for He8_wedge), D atom position free to move. (A = general nomenclature for P substituents)

descriptor	derivation (units)
	Split ligands (P_1, P_2)
E_{HOMO_P1}, E_{HOMO_P2}	energy of highest occupied molecular orbital (hartrees)
E_{LUMO_P1}, E_{LUMO_P2}	energy of lowest unoccupied molecular orbital (hartrees)
PA_P1, PA_P2	proton affinity, PA) $E(LP_n)$ - $E([HLP_n]$ +) (kcal mol^{-1})
	Free ligands
He_8_wedge	Interaction energy between ligand in chelating conformation and wedge of 8 He atoms, maintaining donor atom position similar to [$ZnCl_2$] complex,a EHe_8_wedge = $E(He_8(P1{\sim}P2))$ − E(He8) − $E((P1{\sim}P2))$ (kcal mol^{-1})
nHe_8	Interaction energy between ligand in chelating conformation and wedge of 8 He atoms, donor atoms free to move at fixed X–P distances,b $EnHe_8$ = $E(He_8(P1{\sim}P2))$ − $E(He_8)$ − $E((P1{\sim}P2))$ (kcal mol^{-1})
	Zinc Complexes ([$ZnCl_2$(PP)])
BE(Zn)	bond energy for dissociation of {PP} from fragment (kcal mol^{-1})
Zn-Cl	r(Zn-Cl) (Å)
∠P1-Zn-P2	ligand bite angle in complex (deg)
ΔP1-A(Zn), ΔP2-A(Zn)	change in av r(P-A) cf. (PP) (Å)
ΔA-P1-A(Zn), ΔA-P2-A(Zn)	change in av ∠(A-P-A) cf. (PP) (deg)
ΔZn-P1, ΔZn-P2	Change in Zn–P distances cf. reference ligand (Å)
Q(Zn)	NBO charge on $ZnCl_2$ fragment
	Palladium complexes ([$PdCl_2$(PP)])
BE(Pd)	Bond energy for dissociation of P1∼P2 ligand from metal fragment (kcal mol−1)
Pd-Cl	Average Pd–Cl distance (Å)
∠P1-Pd-P2	Ligand bite angle in complex (deg)
ΔP1-A(Pd), ΔP2-A(Pd)	change in av r(P-A) cf. free (PP) (Å)
ΔA-P1-A(Pd), ΔA-P2-A(Pd)	change in av ∠(A-P-A) cf. free (PP) (deg)
ΔPd-P1, ΔPd-P2	Change in Pd–P distances cf. reference ligand (Å)
Q(Pd)	NBO charge on [$PdCl_2$] fragment

The data provided by the 28 calculated descriptors is then introduced to the LKB-PP dataset, contributing to the expansion of the LKB-PP.

3.2 Methodology for the development and comparison of chemical space maps via DR techniques and design of test case.

3.2.1 Development of maps of the chemical space via PCA, UMAP and t-SNE.

A powerful concept in chemistry is that a compound's structure determines its properties. Structural characteristics will be used to see if the maps show trends. If structurally similar diphosphines are clustered together, this can likely be translated to similar catalytic performance since they will have similar properties.[26] The structural features that will be used are substituents and backbone. It is a good point to start because previous work[7] on PCA has shown that trends are established when colour-coding those features.

The LKB-PP_{screen} dataset consists of 28 descriptors. Each of the descriptors contributes one dimension. Since it is impossible to visualize 28 dimensions at the same time, technique is needed to reduce these 28 dimensions into 2 or 3 as maximum. The techniques that have been studied in this Bachelor Thesis are PCA, UMAP and t-SNE.

A Jupyter notebook has been built by the author using Python language *(see Data and code availability, Script 1).* This notebook provides the DR data and the representation of the first two components for a 2D map, or the first three for a 3D map. Colour-coding can be changed to any suitable feature: descriptors themselves, backbone type, backbone length or phosphine substituent. The coding corresponds to the labels for backbones and substituents in Figure 5, the identifiers for all the ligands are also in DataDRtechniques.xlsx *(see Data and code availability section).*

To test if the different DR techniques are capturing different chemical information, two clustering algorithms have been used. These are k-means and hierarchical clustering. For the hierarchical clustering, agglomerative hierarchical clustering (bottom-up) with ward linkage was used. The clustering numbers obtained have been added to the colour-coding possibilities and will help on the mentioned analysis. The clustering algorithms are implemented in this same notebook *(see Data and code availability, Script 1).*

For UMAP, 3 maps will be generated in order to test the effect of capturing more or less global structure, by varying the number of neighbours, and how this translates onto the maps and in the models established with that DR technique.

3.2.2 Comparison of PCA, UMAP and t-SNE.

The aim of the comparison is to check the robustness, accuracy and extent of retained information of the models when applying each DR technique. This comparison will be complemented with a test case, in which the potential applications and performance of the models will also be tested.

Comparison of these DR techniques cannot be done directly since PCA is a linear DR technique and UMAP and t-SNE are non-linear *(see chapter 2, Theoretical basis)*. To be able to compare such models three comparison tests have been proposed:

- **Silhouette score coefficient.** The coefficient assesses how well separated the clusters are. If ligands in the same cluster have similar characteristics, it might be useful to know where a new synthesized ligand sits in an established model. If clusters are well separated, the analysis and visualization is easier. Silhouette score coefficient is calculated following equation 11:

$$s = \frac{b - a}{max(a, b)} \quad (11)$$

Where a is the mean distance between a sample and all other points in the same cluster and b Is the distance between a sample and all the other points of the nearest cluster.

- **Split and clustering.** Effectively what it is done is creating a prediction model**.** It takes as X matrix the components of each DR technique and performs a classification using K Nearest Neighbours (KNN) algorithm, feeding as Y matrix each of the cluster classification obtained by k-means and hierarchical algorithms, in order to test if the established model is able to accurately classify when new ligands are given as input. A Y matrix is needed because KNN is a supervised clustering algorithm and so a classification of the ligand must be given. This gives us confidence to trust where the ligand sits in the map. The data set is split into training (80%) and test (20%) in a random fashion. With the training X and Y matrices a multivariate linear regression model is constructed. Then the test X matrix is introduced in the model, and Y matrix is predicted as an output. Y predicted and the Y test matrices are compared, and the accuracy score is obtained. The test is run over 10 random states. Finally, the mean accuracy score is provided, calculated following equation 12:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (12)$$

To also test the robustness of the model, cross – validation is performed run by selecting 10 random states and averaging the accuracy scores.

- **Reconstruction of the original data.** It works in a similar way to the split and clustering test but now the Y matrix is each of the descriptors that were originally used respectively. The idea is to see if the DR data is able to provide more or less accurately the original data, to be able to know how well information has been captured by each DR technique in the model, since the concept of explained variance of PCA do not exist in UMAP and t-SNE due to the non – linear nature of the functions used to reduce dimensionality.

3.2.3 Design of test cases.

As stated before, test cases will help to complete the comparison of the different models as well as showing the potential uses that these types of maps have for catalysis optimisation purposes. The main problem when trying to test models with experimental results is to find a considerable amount of useful data. Two main problems can be identified when a data search is carried out:

- **The type of data that are published.** Generally, only excellent results are reported, which is not good since the whole range of possibilities needs to be covered in order to fit or test a model.
- **Conditions of experimental data.** Rarely the conditions of different reactions are the same, which introduces variability that is not related to diphosphine properties itself. Since these descriptors are not introduced, only data generated with the same reaction conditions can be used, which constrains this further.

The data presented in Table 2 is extracted from the work of Kamer, Van der Leeuwen and Reek about Rh hydroformylation of octane with Xantphos ligands.[27] The conditions are the same for

all reactions, but the number of ligands studied limits the construction of any possible tool because a global prediction model cannot be constructed with only nine ligands. Moreover, since all are Xantphos derivatives, the bite angles and percentage of linear aldehyde ranges, which are the most relevant catalytic features, are really narrow. This may lead to not trustworthy analysis when plotting these ligands and searching for trends in the chemical space maps. For these reasons, data have been computationally generated for a larger set of ligands.

Table 2. 1-Octene hydroformylation using Xantphos ligands (1-10)[a]. Copyright 2010 ELSEVIER.
[a]Conditions: CO/H2 = 1, P(CO/H2) = 20 bar, ligand/Rh = 5, substrate/Rh = 637
[Rh] = 1.00 mM, number of experiments = 3. In none of the experiments hydrogenation was observed.
[b]Natural bite angle.
[c]Linear to branched ratio and turnover frequency were determined at 20% alkene conversion.
[d]Turnover frequency = (moles of aldehyde)(moles of Rh)$^{-1}$h^{-1}
The ligands correspond to: 1 – Homoxantphos, 2 – Phosxantphos, 3 – Sixantphos, 4 – Thixantphos, 5 – Xantphos, 6 – Isopropxantphos, 7 – Benzylnixantphos, 8 – Nixantphos, 9 – Benzoxantphos

Ligand	β_n (°)[b]	l/b ratio[c]	% linear aldehyde[c]	% isomer[c]	tof[c,d]	ee:ea ratio
1	102.0	8.5	88.2	1.4	37.0	3:7
2	107.9	14.6	89.7	4.2	74.0	7:3
3	108.5	34.6	94.3	3.0	81.0	6:4
4	109.6	50.0	93.2	4.9	110.0	7:3
5	111.4	52.2	94.5	3.6	187.0	7:3
6	113.2	49.8	94.3	3.8	162.0	8:2
7	114.1	50.6	94.3	3.9	154.0	7:3
8	114.2	69.4	94.9	3.7	160.0	8:2
9	120.6	50.2	96.5	1.6	343.0	6:4

For the test case, the activation of the precatalyst has been chosen as the reaction to be optimised, as shown in the reaction equation in Figure 12. The active form of the catalyst needs to keep a balance between being stable enough to undergo the desired catalytic rection and active enough to have high efficiency. By using the LKB models that had been built it is possible to predict the viability of the catalyst activation before any experimental work is done, being a helpful tool for optimising the reaction.

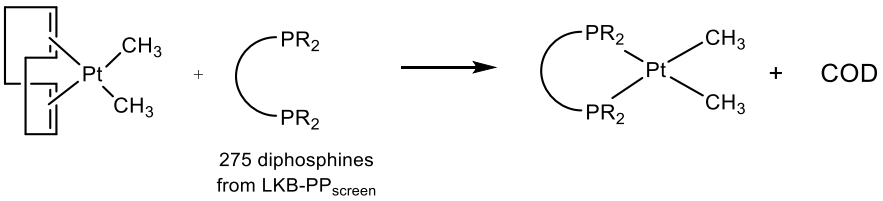


Figure 12. Substitution reactions equation of [$Me_2Pt(COD)$] with diphosphines. COD=cyclooctadiene.

The test case chosen is inspired by an experimental study on the thermochemistry of ligand substitution reactions of aminobis(phosphines), $Ph_2P(R)NPPh_2$, with [$Me_2Pt(COD)$][28]. A key point in catalytic cycles is the catalyst activation. The precatalyst, in this case [$Me_2Pt(COD)$] as shown in Figure 12, is converted into the active catalytic specie [$Me_2Pt(PP)$].

In that work, the enthalpy of substitution of the reaction shown in Figure 12 has been determined experimentally, for six different ligands using the same conditions.

Of the six ligands that were tested experimentally, one of them ($Ph_2P(Me)NPPh_2$) is already in the LKB-PP_{screen} and so already employed to build the models. The other five, which are not in the LKB-PP_{screen}, will be used as new input ligands. These five (see Figure 13) will simulate the new ligands that the experimentalist propose for the optimisation of the reaction.

The ligands that were not captured in the LKB-PP (2, 3 and 4) were introduced according to the methodology in section 3.1. Molecular Mechanics (MM) stochastic conformational searches were performed, to screen conformational space for free ligands and their tetrahedral zinc complexes [$ZnCl_2$(PP)]. Then the rest of the models described in section 3.1 were constructed using Maestro Software and scripts and finally the DFT calculations were performed with Jaguar package (s*ee computational details section)*.

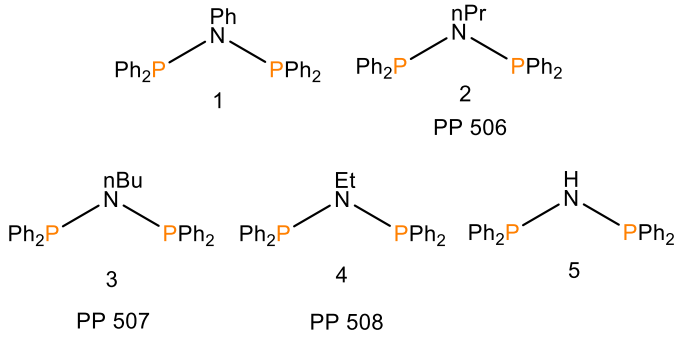


Figure 13. Test ligands coming from the experimental work on the thermochemical study of ligand substitution reactions of aminobis(phosphines), $Ph_2P(R)NPPh_2$, with [$Me_2Pt(COD)$].[28] Ligands are numbered from 1 to 5, which is the nomenclature that will be used in the analysis. Ligands 1 and 5 were in the LKB-PP so the descriptor values were just copied. Ligands 2, 3 and 4 were not in the LKB-PP. They were calculated according to the methodology in section 3.1. Truncated ligand used for the calculation of electronic descriptors can be found in Annex Table 1. The PP identification is consistent with the last ligands introduced to the LKB – PP (which are in Annex Figure 1).

Data have been generated by DFT calculations, in order to make analysis on the maps generated and build a prediction model. The enthalpy will not be calculated, instead the potential energy will be obtained (E_{SCF}), since only optimization calculations are run. This is because the free ligand calculation has already been done since it is needed to construct the $LKB\text{-}PP_{screen}$ dataset itself, and it is an optimisation, not a frequency calculation. From Statistical Thermodynamics the expression of the enthalpy (equation 13) requires the zero-point energy calculation, which is dependent on the vibration frequency of each bond.

$$H = N_A(\varepsilon_0 + zpe) + N_A k\left[\frac{3}{2}T + nT + \sum_{s=1}^{3N-6(5)} \frac{\theta^s}{e^{\frac{\theta_s}{T}} - 1}\right] + N_A kT \tag{13}$$

$$\mathrm{ZPE} = \frac{1}{2}h\nu_j \tag{14}$$

Where N_A is the Avogadro number, ε_0 is the electronic energy at level 0, zpe is the zero-point energy, k is the Boltzmann constant, T is the temperature and θ^s is a vibrational constant. The n in equation 13 will have different values depending on the geometry of the molecule, if linear n = 1, if non-linear n = 3/2, and in a similar way will happen with the sum operator, 3N – 5 for linear and 3N – 6 for non – linear being N the number of atoms of the molecule.

The ZPE, as seen in equation 14, where h is Plank's constant and ν_j the frequency for a vibrational state j, will not be 0 for the ground vibrational state, making the whole enthalpy dependent on frequency.

The potential energy of substitution for this reaction will be calculated as shown in equation 15, for the 275 ligands of the $LKB - PP_{screen}$.

$$E_{substitution} = E_{SCF\ [Me2Pt(Ph2P(R)NPPh2)]} + E_{SCF\ (COD)} - E_{SCF\ Ph2P(R)NPPh2} - E_{SCF\ [Me2Pt(COD)]} \tag{15}$$

For time purposes it is more convenient to run only 277 calculations (275 Pt complexes, the [Me_2Pt(COD)] and COD) rather than 553.

However, besides the fact that the potential energy is calculated, same trend would be expected with enthalpies. Molecules with similar structures are likely to have similar vibrational modes and so, frequency calculations add considerable computational cost and do not provide much information respect the optimisation calculation.

Once all these calculations are obtained the analysis that will be performed is:

- Check the correlations between substitution potential energies and the descriptors of the LKB-PP_{screen}.
- Establish a regression model by splitting our data set (using the same methodology as in section 3.2.2) and compare the different DR techniques, to see which will be best for predicting the substitution potential energies of new ligands.
- Plot new ligands in the established models, without changing it. This means that the same model that has been used for all the analysis will be used to transform the new data and plotted in the same maps. From here, the position where they sit and the possible clusters that they form will be analysed.
- Above all, seeing if results are coherent with the previous comparison tests will also be helpful to complete the analysis.

To sum up, in Figure 14 a schematic workflow is presented, combining the methodology of the LKB and the proposal for the this work, which focuses on points 3, 4 and 5, giving a complete view of the project.

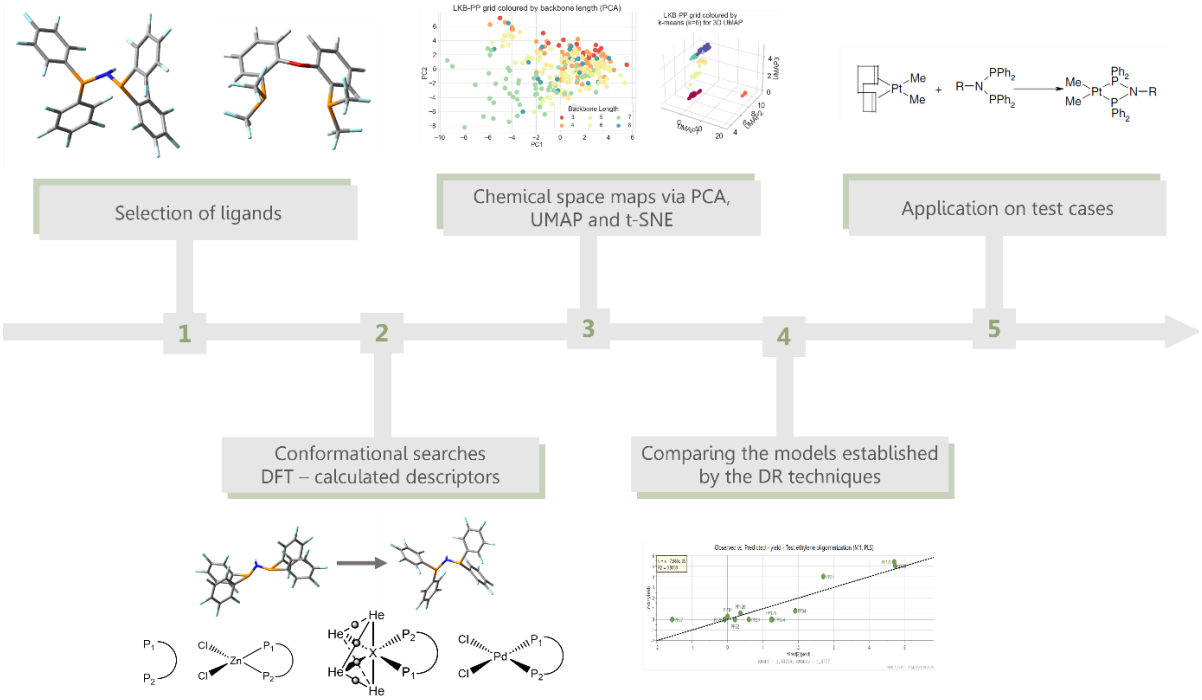


Figure 14. General workflow followed on the LKB methodology and in the bachelor thesis.

4

Results and discussion: Chemical space maps, comparison, and test case results.

4.1 The search for trends based on substituents and backbone length of diphosphines.[i]

Before generating any map, it is necessary to decide the number of components that will be used to build the models of each DR technique. For PCA this is rather easy, components are kept until around 90% of explained variance is obtained (see Annex Figure 2). In this present case the PCA model contains 7 components. However, for UMAP and t-SNE the concept of explained variance does not exist. When deciding the number of components, a quick check by changing that number of components is performed. Around 10 component the maps had a rather stable shape, but then by adding some more, the plots appeared "compressed" indicating an overfitted model. That is why 10 components have been used for UMAP and t-SNE.

As explained in the theoretical basis, some parameters tune the behaviour of UMAP and t-SNE. For t-SNE, perplexity is set at 12 to preserve more local structures, the learning rate is "auto" by default (which is the standard way of proceeding), the number of interactions is set in 5000 since van der Maaten & Hinton pointed that is the number around stability is achieved,[23] the method has been forced to solve the exact Barnes-Hut since the model has been build with 10 components. Regarding the metric t-SNE uses Euclidean distances and is defined by default.

In the case of UMAP three models have been generated, with neighbours 5, 10 and 15, and the minimum distance is set as 0.5. For both, t-SNE and UMAP, the random state is 42, which is the standard procedure as well.

The first analysis of the obtained chemical space maps consists of colour-coding them according to two structural features: the substituents and the backbone length, since some trends were observed for PCA in previous work.[13]

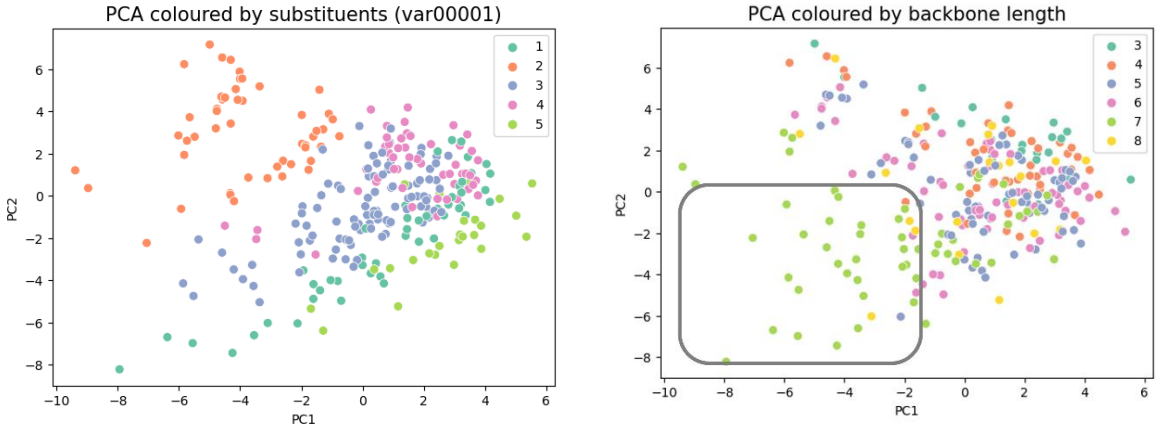


Figure 15. PCA model coloured by substituents (left) and backbone length (right). The substituents are numbered according to table of Figure 8. Detailed classification of substituents and backbone lengths can be found for all the ligands in the data sheet specified in *Data and code availability* section. The numeric classification of substituents is presented in the column var00001 in the corresponding file.

[i] The plots and results presented have been generated by Script 1: *DR techniques and clustering algorithms* where the different identifiers for each ligand are used to colour-code the maps, see Data and code availability section for more details.

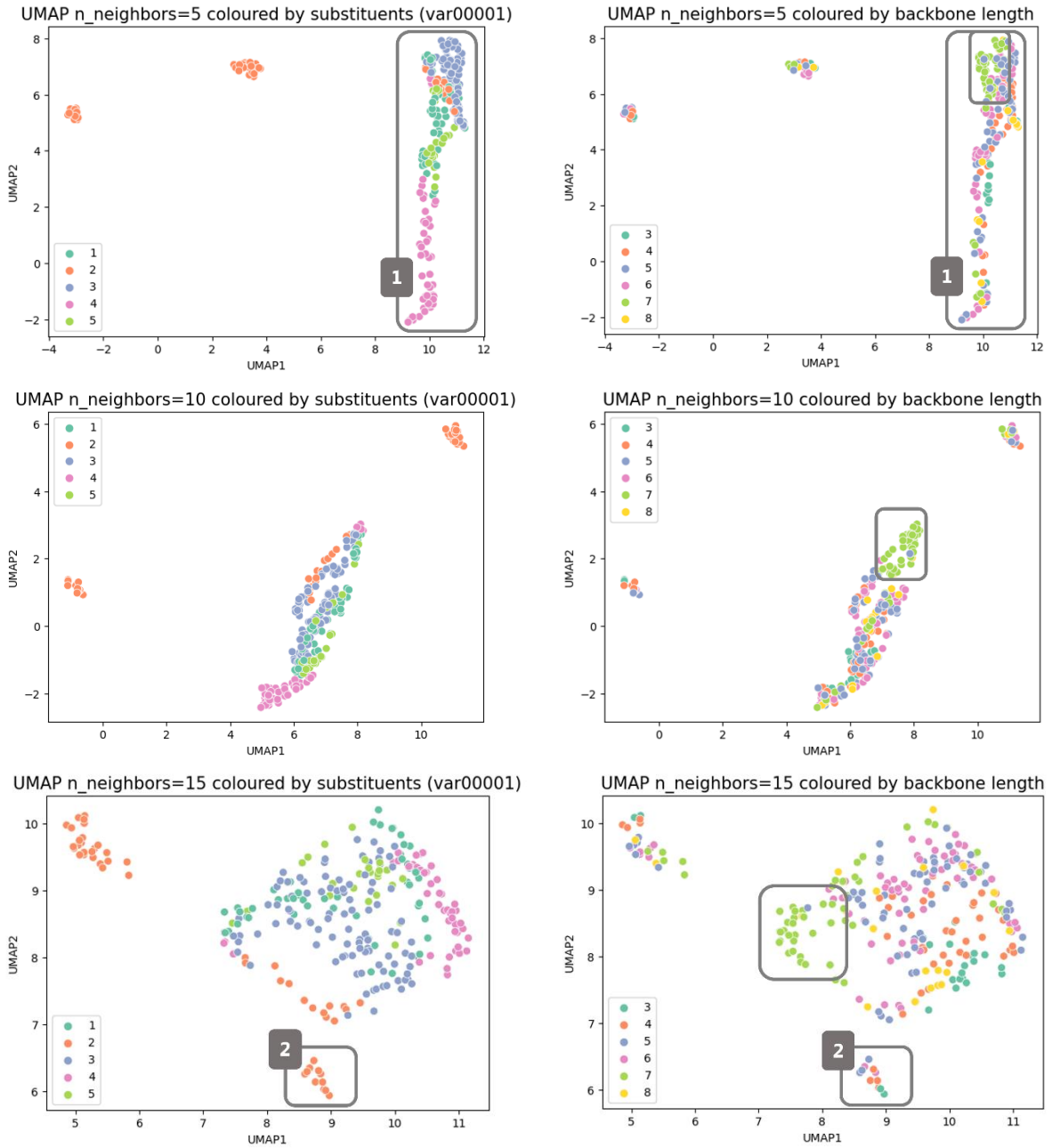


Figure 16. UMAP model coloured by substituents (left) and backbone length (right). Each pair of maps correspond to number of neighbours = 5, 10, 15 respectively. The substituents are numbered according to table of Figure 5. Detailed classification of substituents and backbone lengths can be found for all the ligands in the data sheet specified in *Data and code availability* section. The numeric classification of substituents is presented in the column var00001 in the corresponding file.

It is quite clear that all maps, regardless of the DR technique used, are capturing information related to the substituents *(see Figure 15, 16 and 18)*. In general, when colour-coding according to that structural feature differentiated clusters have been seen without excessive overlap.

In the case of backbone length, the different clusters appeared highly overlapped, which may indicate that there is no high correlation between the length of the backbone and the descriptors since the maps have not been able to capture this type of information.

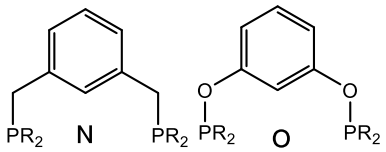


Figure 17. Backbones N and O, from the construction of the LKB-PP_{screen}. These two backbones appeared clustered together as indicate the square boxes in all DR maps.

There is however a relevant cluster that has appeared in all the maps created regardless the DR technique used. A group of ligands with backbone length equal to 7 is always appearing together, corresponding to the backbones N and O *(see Figure 17).* The region where these backbones were sitting are squared in the backbone length plots for all DR technique maps. These backbones are rigid structures that induce a large bite angle to the phosphine when coordinated, causing steric problems to the complex. As a result, this differentiated cluster was obtained.

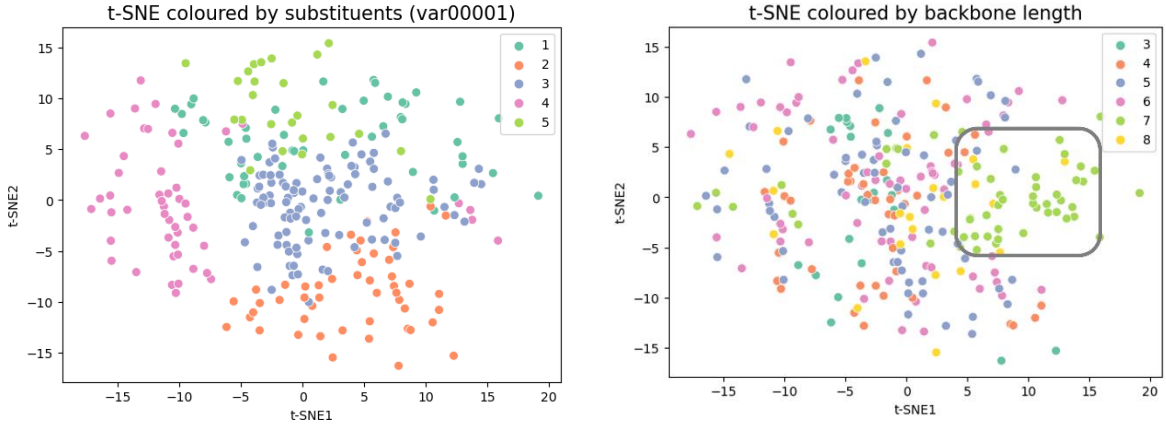


Figure 18. t-SNE model coloured by substituents (left) and backbone length (right). The substituents are numbered according to table of Figure 5. Detailed classification of substituents and backbone lengths can be found for all the ligands in the data sheet specified in *Data and code availability* section. The numeric classification of substituents is presented in the column var00001 in the corresponding file.

In PCA *(see Figure 15)* the substituent clusters presented a diagonal trend, which may be a side effect of the linearity of the functions that are used to reduce the dimensionality of the chemical space. This is not observed neither in UMAP or t-SNE because of the nature of their functions.

Regarding the increase of numbers of neighbours in UMAP *(see Figure 16)*, by going from 5 to 15 neighbours, a decompression of the principal cluster (square 1 for UMAP, number of neighbours = 5) is observed. The two small clusters two are maintained. When the number of neighbours is increased up to 15, one of these clusters, ends being close to the larger one indicating that UMAP has now found more similarities (square 2 for UMAP, number of neighbours = 15), which indicates the incorporation of more global information.

For t-SNE *(see Figure 18)* it is true that a better clustering was expected, since it captures local structure. The maps obtained present a lot of overlap between clusters, and no denser zones are seen. However, this is not a surprise at all. Other authors[29] that have done work with t-SNE have already explained that the abilities of t-SNE are enhanced when working with a high dimensional matrix of data, which is not the case for the LKB.

In the following section, and with the help of clustering algorithms, more detailed analysis of the maps will be performed.

4.2 Analysing the chemical information retained with k-means and hierarchical clustering algorithms. [ii]

The results of the last section guided the procedure for analysing the maps provided by the clustering algorithms. The focus has been on finding relationships according to the substituents of the diphosphines.

For choosing the appropriate number of k (for k-means) and n (for hierarchical), the elbow method was used. Three (k, n=3) was the optimal number of clusters, which was the same for all DR techniques and algorithms. But being optimal does not mean being useful. The number of substituents is 11 and their classification gives a total of 5 groups *(See table in Figure 5, chapter 1),* so three is probably not the most sensible selection. Maps with k and n = 3, 5 and 6 where generated and a quick analysis of them was performed. A k, n = 4 was not considered since the same reasoning as for 3 could be applied. In the end k, n = 6 was chosen, since it gave maps that were clear, and more variation was observed between clustering algorithms which may give some more space for the analysis. The silhouette score coefficient (see Table 3, section 4.3) indicated that there was no real impact on the clustering performance when deviating from the optimal cluster number. Moreover, k, n = 6 had generally a slightly better coefficient than 5. Detailed explanation will be provided in the next section.

In the end, when working with unsupervised clustering algorithms, the answer of what the right number of clusters is may not be straight forward, and somewhat arbitrary. Other possible explorations for choosing the number of k, n may be developing the same analysis that it is done in this Bachelor Thesis with the optimal number of k, n or increasing to an even high number of clusters. However, this falls out of the scope of this project but was noted as a possibility to give better description of the models and as something to take into account for future work on similar projects.

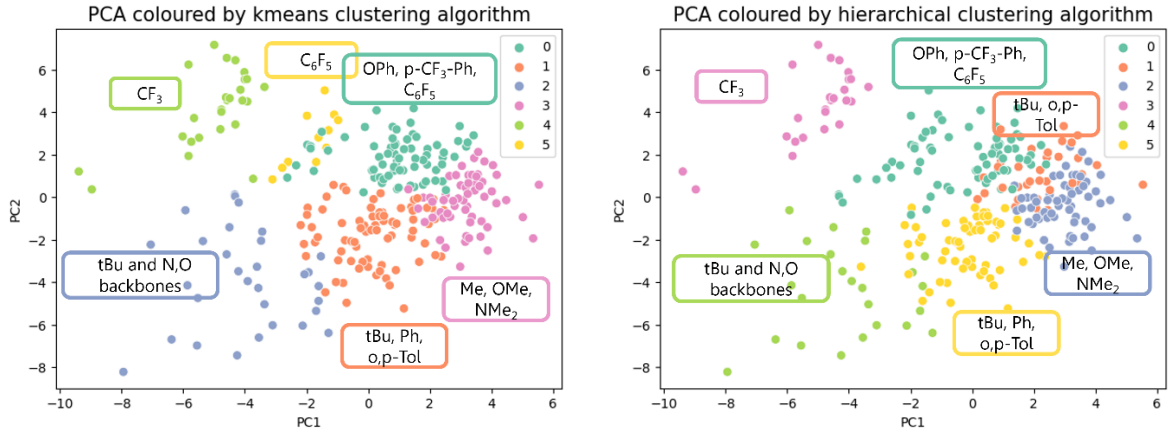


Figure 19. PCA model coloured by k-means clustering algorithm (left) and hierarchical clustering algorithm (right). The major substituent or substituents of each cluster are written in the same colour box. Full detailed information about the composition of the clusters regarding the substituents can be found in *Data and code availability,* cluster analysis sheet. The legend corresponds to the clusters obtained when the algorithm is applied.

In Figure 19, 20 and 21 the maps of chemical space obtained from the different techniques are shown. At a first glance, no extreme differences have been noticed between the different clustering algorithms in each DR technique map.

[ii] The plots and results presented have been generated by the Script 1: *DR techniques and clustering algorithms* (for running the algorithms and colour-codding according to the obtained clusters), see Data and code availability section for more details.

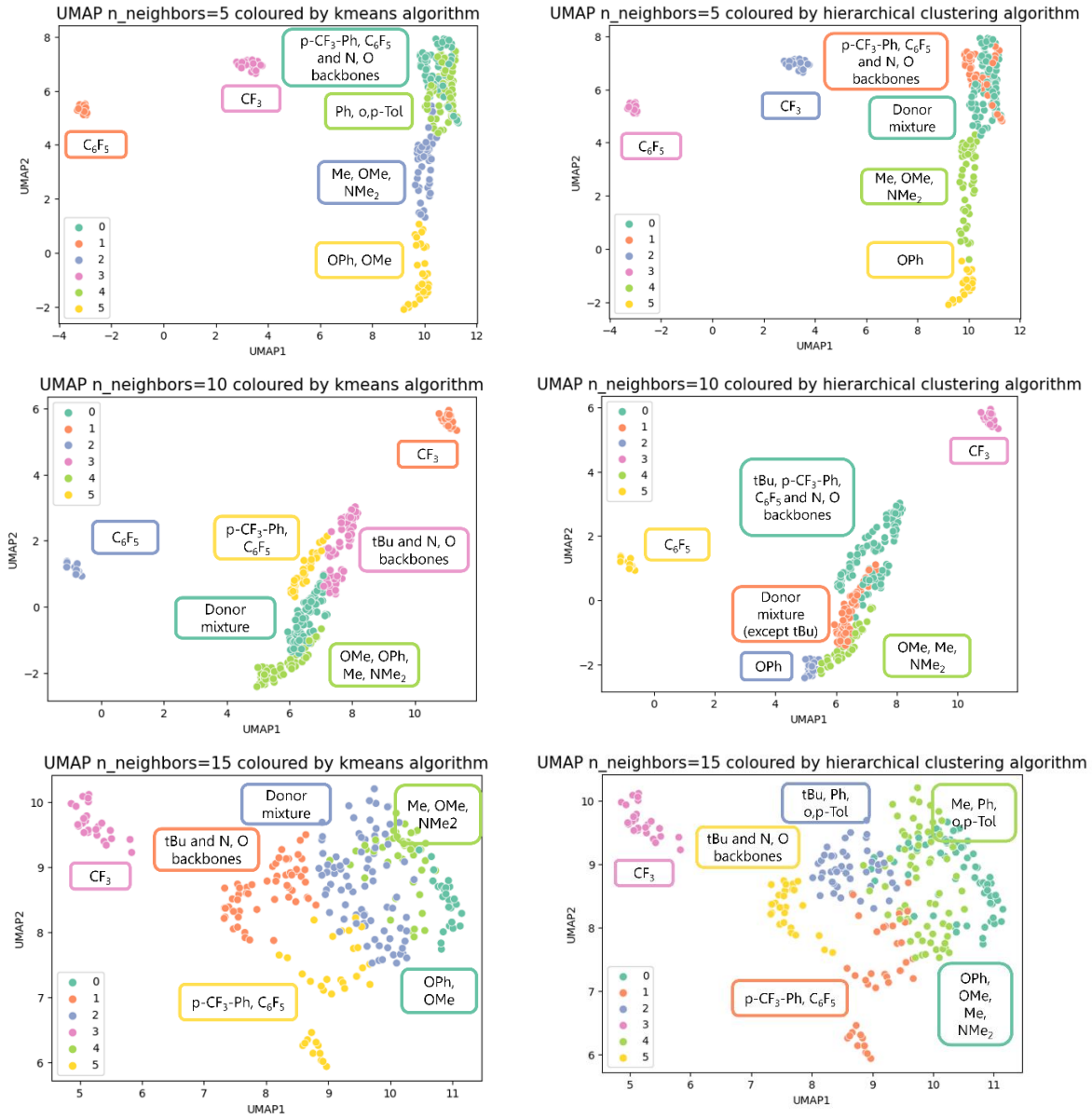


Figure 20. UMAP model coloured by k-means clustering algorithm (left) and hierarchical clustering algorithm (right). Each pair of maps correspond to number of neighbours = 5, 10, 15 respectively. The major substituent or substituents of each cluster are written in the same colour box. Full detailed information about the composition of the clusters regarding the substituents can be found in in *Data and code availability,* cluster analysis sheet. The legend corresponds to the clusters obtained when the algorithm is applied.

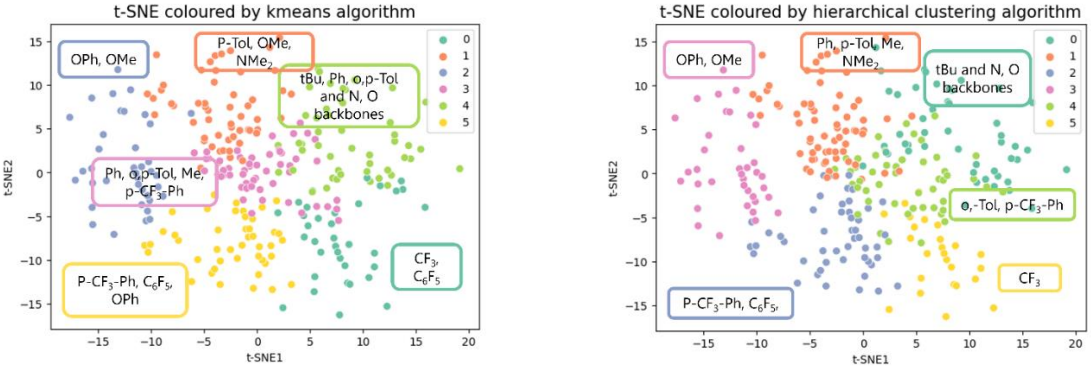


Figure 21. t-SNE model coloured by k-means clustering algorithm (left) and hierarchical clustering algorithm (right). The major substituent or substituents of each cluster are written in the same colour box. Full detailed information about the composition of the clusters regarding the substituents can be found in in *Data and code availability,* cluster analysis sheet. The legend corresponds to the clusters obtained when the algorithm is applied.

Some information is shared and presented in the same way for all the maps. All CF_3 substituents appeared always clustered together, and this cluster contains no other substituents apart from CF_3 except for t-SNE (k-means) where some C_6F_5 are also allocated to the group *(see Data and code availability, cluster analysis sheet)*, which indicates that this cluster is highly directed by electronic effects. Since CF_3 is the most electron withdrawing substituent of the set, the electronic descriptors present substantial differences compared to the rest of the dataset as seen in Figure 22 A, taking the HOMO descriptor as an example, causing the permanent formation of the cluster.

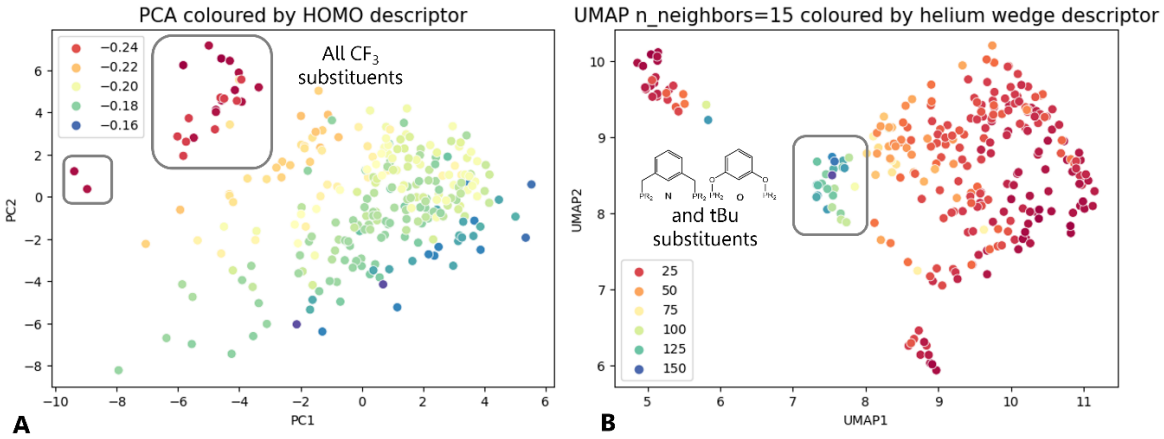


Figure 21. Maps of chemical space coloured by descriptors. (A) PCA model colour-coded by HOMO descriptor. (B) UMAP n_neighbors = 15 colour-coded by helium wedge descriptor. The legend corresponds to the range of values of the descriptor. All maps presented show the same results, these are just examples.

Also, as noticed in the last section analysis, the N, O backbones are always shown together. This cluster can enlarge or reduce the number species but when the number of ligands is increased the major substituent that is present corresponds to tBu. This observation means that this particular cluster is sterically controlled, As shown in Figure 22 B, it is clearly seen that the steric descriptor helium wedge has the highest values where the N, O backbone ligands are found.

For PCA *(see Figure 19, and the corresponding legends)* the main difference between the two clustering algorithms is the transformation of cluster 0 and 5 of k-means into cluster 0 and 1 for hierarchical approaches. In the case of k-means cluster 5 it contains only C_6F_5 substituents while 0 contains OPh, p-CF_3-Ph, and C_6F_5 as well. The reason why some C_6F_5 are forming their own cluster is because of electronics effects, where higher electron-withdrawing character is found in cluster 5, as show in Figure 22 A. However, some overlap between these two clusters is observed. In that overlap region only C_6F_5 substituents coming from both clusters (0 and 5 PCA, k-means) are found, indicating that the model is able to distinguish between electronic effects for the same diphosphine substituent.

In the case of hierarchical clustering, cluster 0 contains electron withdrawing groups, and all the C_6F_5 substituted ligands are grouped together while cluster 1 contains electron donor groups. The rest of clusters remains essentially the same.

When the distribution of clusters in the space is considered, especially for the hierarchical algorithm, they seem to be ordered according to steric and electronic effects. Profiting from the fact that PCA enables one to make a biplot, the direction of descriptors is analysed (see Figure 23 A). Strongly electronic descriptors are grouped in the same position (square 1 for

Figure 23 A) while strongly steric descriptors are also together and located almost orthogonal with respect to the electronic ones (square 2 for Figure 23 A).

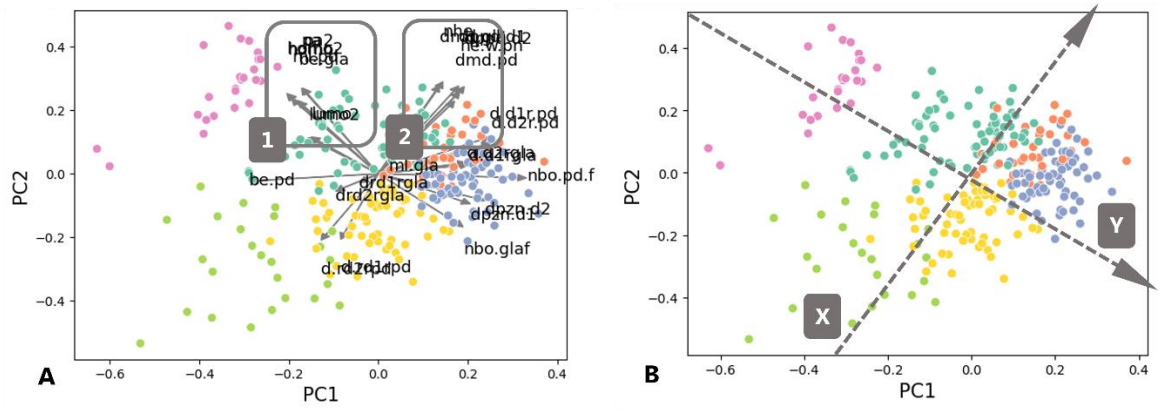


Figure 23. Biplot and trends for PCA coloured by hierarchical clustering algorithm. (A) Superposition of arrows (loadings) and the scores creating a biplot. Square 1 corresponds to purely electronic descriptors: homo, lumo, pa, homo2, lumo2 and pa2. In Square 2 purely steric descriptors are found among others: nhe, helium wedge. (B) Trend X correspond to steric effects and trend Y correspond to electronic effects.

Electronic and steric axes are established in Figure 23 B. For the steric one (X), it goes from congestion to steric freedom. In the case of the electronic dimension (Y), it goes from electron withdrawing to donating, both trends following the direction of the arrows. The map obtained in Figure 23 B has potential uses to be transformed into a powerful tool when combined with catalytic features for a given reaction. This will be deeply explored in the test case.

For UMAP *(see Figure 20),* the same general cluster classification as in PCA can be found. For neighbours 5 and 10, the main notable difference has been that OPh substituents constitutes a different cluster in the case of the hierarchical classification. For 5 and 10 neighbour maps it can be seen that this group is either clustered with OMe, Me and NMe_2 or closer to the late two groups.

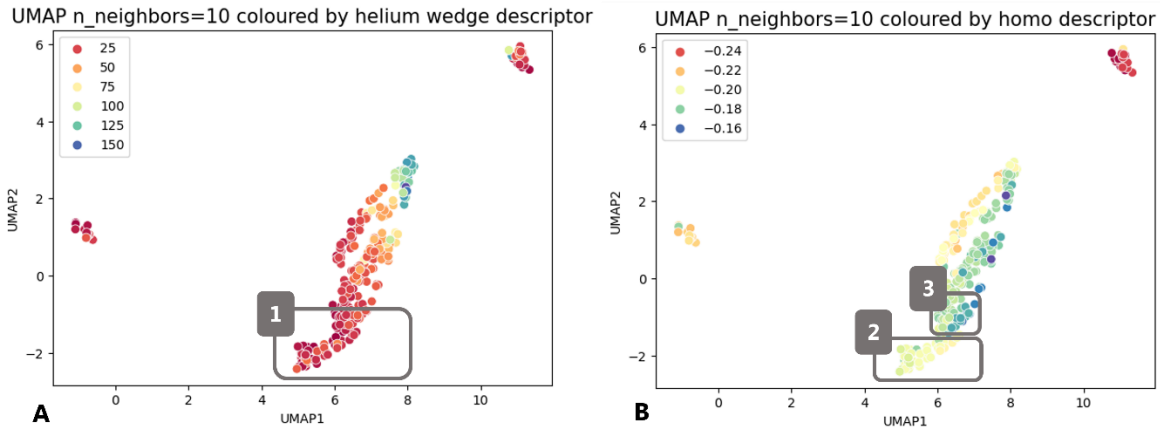


Figure 24. (A) UMAP n_neighbours = 10 coloured by helium wedge descriptor. Square 1 correspond to the region where OPh, OMe, Me and NMe_2 are located. No steric differences are observed between these substituents. (B) UMAP n_neighbours = 10 coloured by homo descriptor. Square 2 corresponds to OPh and OMe that differs form square 3, which corresponds to Me and NMe_2, because of their electronic nature. OPh and OMe have a withdrawing character while Me and NMe_2 have a donor nature.

OPh is located there due to the steric similarities with all these mentioned groups as Figure 24 A shows. When clustered together with OMe as in k-means, number of neighbours = 5 and 15, the algorithm shows that was able to capture their similarities due to their electron withdrawing character. To illustrate that, the map for UMAP and number of neighbours = 10 is

used as example, since the exact same trends are obtained in the different neighbours' plots. By taking an electronic descriptor, HOMO in this case, and colour coding the map, as shown in Figure 24 B a clear difference on electronic behaviour is seen between OPh and OMe (square 2 for Figure 24 B) and NMe_2 and Me (square 3 for Figure 24 B), which explains the separation of this groups in some of the maps of diphosphine chemical space generated by UMAP.

However, for UMAP hierarchical clustering algorithm and number of neighbours = 5 and 10 *(see Figure 20),* OPh substituent was constituting an independent cluster. This may be attributed to the aromatic character of the group. The fact the clustering algorithms can differentiate groups according to other characteristics such aromaticity, in the case of OPh, indicates that underlying chemical information is inside the maps. Such information does not come out as easy as other that may be more influent.

Increasing the number of neighbours caused other reactions that recall the cluster disposition used with hierarchical PCA. In Figure 20, for UMAP number of neighbours = 15, some C_6F_5 substituents no longer formed an independent cluster as happened for number of neighbours = 5, 10. For number of neighbours = 15 the independent cluster formed by C_6F_5 substituents joined the cluster that contains mainly p-CF_3-Ph since their electronics are similar. This makes sense because when more global structure is introduced, the differences between C_6F_5 and p-CF_3-Ph became less relevant, forming a unique cluster.

Although being identified as the same, as seen in Figure 20 for number of neighbours = 15, there was some distance inside the same cluster separating the C_6F_5 substituents that had slightly different electronic descriptor values from the rest. This indicated the capability of UMAP of incorporating local and global structures when the reduction of the dimensionality is performed.

In the case of t-SNE, as shown in Figure 21, no significant differences with respect to the other DR techniques is observed. Since t-SNE captures local information, the similarities between OPh and OMe became more important than the relation with others and so, these appeared clustered together independently of the algorithm used because of their similar electronic effects.

To sum up, the two tested clustering algorithms have shown that the different DR techniques capture the same type of chemical information, which is related to diphosphine substituents. However, some differences were noticed, mainly related to the global/local information retention when reducing the dimensionality, which was used to understand the behaviour of the data and even establish trends for PCA. This does not mean that this is the only type of chemical information retained, but it is what have been seen in this study. Other clustering algorithms can be used to explore the information retained.

During the "Automation in Chemistry" meeting celebrated on May 10th at the University of Bristol, I had the opportunity to talk with Prof. Sophia Yaliraki from Imperial College about this specific problem. She participated in the elaboration of an unsupervised methodology to choose which is the best clustering technique according to the data used.[30] As said, testing clustering algorithms could be a topic itself. Unfortunately, this is also out of the scope of this present thesis, but it is an idea to keep in mind for future work.

4.3 Comparison of the DR techniques.[iii]

As explained in section 3.2, an indirect methodology was needed to compare the different approaches because of the nature of the functions that each technique uses to reduce dimensionality.

The silhouette score coefficient is presented in Table 3. For k, n = 3, 5 and 6, the silhouette score is shown since, as explained in the beginning of the previous section, they were used for deciding the number of clusters.

For PCA and t-SNE, there is no real difference when deviating from optimal numbers of clusters. In the case of UMAP, the scores may suggest that there is a loss of clustering separation when not choosing 3, which is true, but even with this loss of cluster separation, UMAP still has the mist separated of the three dimensionality reduction methods, suggesting that our choice of n = 6 is sufficient.

Table 3. Silhouette score coefficient for all techniques and both clustering algorithms. For UMAP the table also presents number of neighbours = 5, 10 and 15.

PCA			**t-SNE**		
k, n	*k-means*	*hierarchical*	*k, n*	*k-means*	*hierarchical*
3	0.29	0.24	3	0.19	0.17
5	0.23	0.22	5	0.16	0.14
6	0.25	0.22	6	0.16	0.15

UMAP (n_neighbors = 5)			**UMAP (n_neighbors = 10)**			**UMAP (n_neighbors = 15)**		
k, n	*k-means*	*hierarchical*	*k, n*	*k-means*	*hierarchical*	*k, n*	*k-means*	*hierarchical*
3	0.60	0.60	3	0.71	0.71	3	0.41	0.38
5	0.44	0.44	5	0.42	0.38	5	0.42	0.38
6	0.45	0.43	6	0.44	0.38	6	0.39	0.37

UMAP shows the best clustering separation performance as seen in Figure 20 with number of neighbours = 5, 10, where little overlap is observed. If cluster separation is desired, UMAP may be the preferred choice. The good separation is better seen when a 3D plot is build as shown in Annex Figure 3.

To test if the models obtained are able to correctly classify new ligands according to the obtained clusters in the last section, a test is performed. KNN is used as classification algorithm as explained in methodology section 3.2. A selection of 55 ligands was used as the test set, and the accuracy coefficients are summarized in Table 4. All techniques show excellent performance when clustering unknown ligands, so all models can be used for clustering purposes. *(See Data and code availability, Script 2)*

Table 4. Accuracy coefficient obtained from the split and clustering with KNN algorithm test. The results are the mean over 10 random states. Red values correspond to cross-validation (cv=6).

PCA		**t-SNE**	
kmeans	*hierarchical*	*kmeans*	*hierarchical*
0.92	0.95	0.91	0.93
0.91	0.96	0.91	0.91

UMAP (n_neighbours = 5)		**UMAP(n_neighbours = 10)**		**UMAP(n_neighbours = 15)**	
kmeans	*hierarchical*	*kmeans*	*hierarchical*	*kmeans*	*hierarchical*
0.97	0.99	0.98	0.99	0.93	0.98
0.98	0.98	0.98	0.99	0.93	0.99

[iii] The plots and results presented have been generated by the Script 1: *DR techniques and clustering algorithms* (for the silhouette score coefficient), and by Script 2: *Test for clustering and reconstruction of data* (for clustering and reconstruction of data), see Data and code availability section for more details.

The cross-validation process splits the whole set into 6 groups and take one of them as the test set. The same procedure is repeated, taking the next group as test, until all have played this role. The average of the six results is provided, which are the values in red (Table 4). Cross validation values are also excellent which may indicate that the models are robust.

Finally, to see the extent of information retained, a regression over each descriptor was done. Each point in the plot corresponds to the mean of the correlation coefficient over 10 random states for the descriptor.

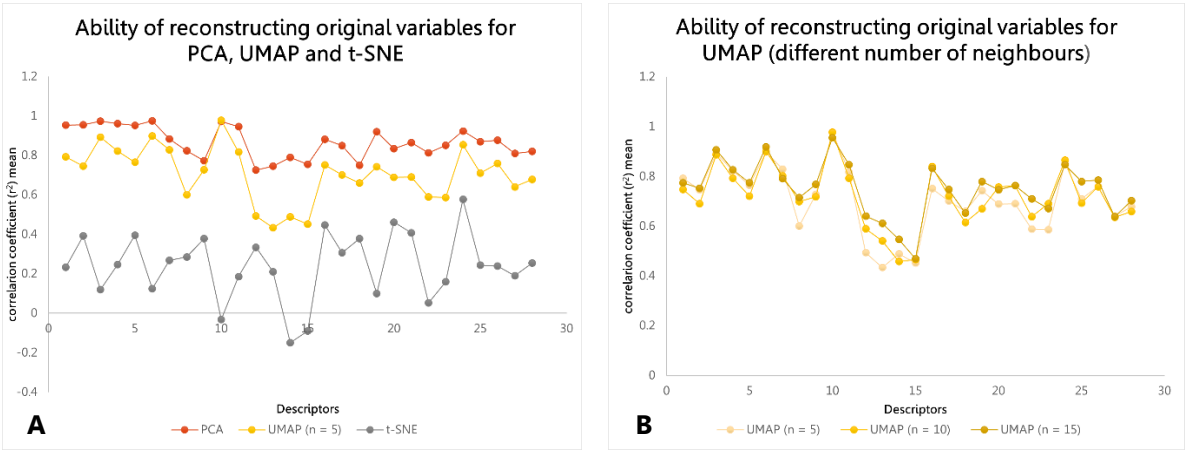


Figure 25. Plots for the ability of reconstructing the original variables. (A) Comparison of the regression coefficients for PCA, UMAP and t-SNE. The models are constructed with 7 components for PCA (as before) but for UMAP and t-SNE all the components (28 components) are used since the "explained variance" is unknown. (B) Comparison of the same test for UMAP with number of neighbours = 5, 10 and 15 respectively. All components are selected as well.

As seen in Figure 25 A, PCA is the best for reconstructing the original data, giving a mean regression coefficient of 0.90 or above for most of the descriptors. Contrarily, t-SNE is clearly failing since most of its regression coefficients do not go above 0.50. UMAP however, has decent results and some of them come quite close to PCA. By looking at Figure 25 B, it is noticeable that introducing more global structure, the model is able to perform slightly better, which may indicate that more information is retained when increasing the number of neighbours.

PCA and UMAP in Figure 25 A, had similar shapes, which may indicate that the same type of information is captured in the two models.

However, PCA is retaining more information than UMAP. It is logical to think that PCA may perform better for prediction purposes, but it may be possible that, since UMAP has captured the same type of information but to a lesser extent, it will have more flexibility when trying to accommodate a model. This will be tested in the next section.

4.4 Test case: Substitution energy of aminobis(phosphines), Ph2P(R)NPPh2, with [Me2Pt(COD)]. [iv]

Once all the calculations were performed, and the results checked (DataDRtechniques file, LKBPP screen data sheet, substitution energy column) the substitution energy was introduced to the dataset as a characteristic feature of each ligand, similar to what was done with substituents or backbone lengths in previous sections.

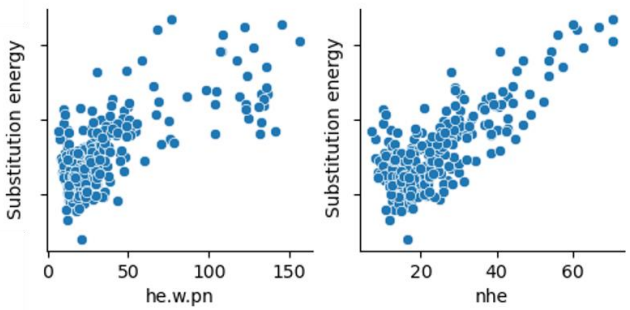


Figure 26. he.w.pn and nhe correlation with substitution energy plots, extracted from the pair plot. The rest of the pair plot correlations can be check in the referenced (iv) jupyter notebook.

A pair plot test was then done, which checks the correlations between the substitution energy and each descriptor. The pair plot correlates each descriptor with the feature selected, in this case the substitution energy, As seen in Figure 26, there was clear correlation between steric descriptors and the substitution energy. The bulkier or more rigid the diphosphine, the more steric hindrance it has when trying to coordinate the Pt metallic centre. This is obviously less favourable, which implies an increase of the energy.

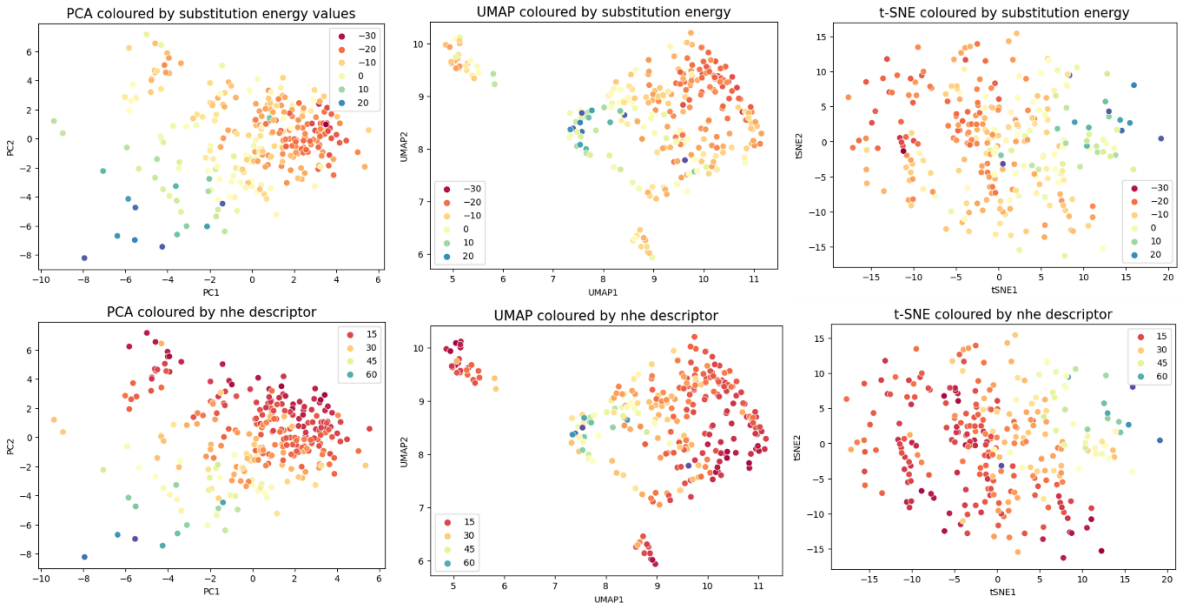


Figure 27. Top row shows the PCA, UMAP with number of neighbours = 15 and t-SNE maps colour-coded according to substitution energy values (kcal/mol). The bottom row shows same techniques colour-codded according to nhe descriptor. The legends correspond to the range of substitution energies and descriptor values respectively.

[iv] The plots and results presented are generated by the Script 1: *DR techniques and clustering algorithms* (for colour-codding the plot according to substitution energy), Script 3: *Prediction model substitution energy* (for the pair plot and prediction model) and Script 4: *Plotting new ligands without changing the model* (for plotting the new ligands in the established models), see Data and code availability section for more details.

Figure 27 demonstrates graphically how well correlated the steric parameter nhe and substitution energy were since very similar patterns are found for both plots.

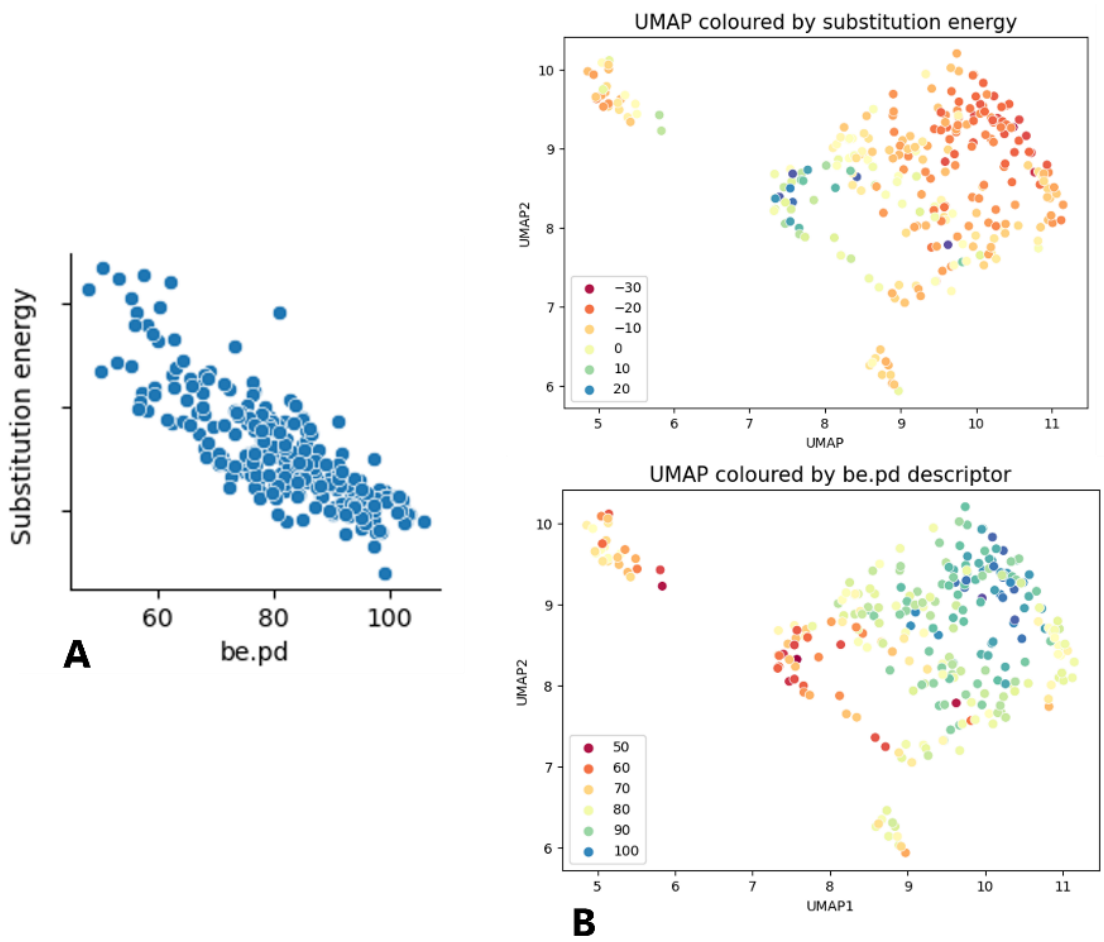


Figure 28. (A) correlation plot between be.pd descriptor and substitution energy. (B) Top: UMAP number of neighbours = 15 colour-coded according to substitution energy (kcal/mol). Bottom: Same technique colour-coded by be.pd descriptor. The legends correspond to the range of substitution energies and descriptor values respectively. There is no preference for choosing UMAP to make this figure, it is just because of visualisation purposes.

An inverse correlation was also found as shown in Figure 28, in this case with the bonding energy descriptor for the Pd model. All the Pt structures are square planar. That is why the correlation is found in the Pd descriptor and not in the Zn one, since Pd has a strong preference for square planar complexes. The bonding energy gets stronger when the diphosphine can better accommodate to the Pt metallic centre. That is why high substitution energies presented low binding energies.

A multivariate regression model has been created to test the performance of each DR technique on predicting substitution energy values. The dimensionality reduction technique models have been used as X matrix (PCA with 7 components, UMAP and t-SNE with 10 components) and the substitution energy value for each ligand has been used as Y matrix. The dataset is split as explained in methodology section 3.2.

The predicted values are then compared with the real substitution energy values and plotted on a graph *(see Figure 29)*. The results of the regression coefficients are shown in Table 5.

Table 5. Mean over 10 random states of the regression coefficients for each DR reduction technique.

r^2 coefficient for each DR technique				
PCA	*t-SNE*	*UMAP (n = 5)*	*UMAP (n = 10)*	*UMAP (n = 15)*
0.84	0.49	0.51	0.51	0.70

PCA is the model that best performed in the prediction test. t-SNE clearly failed and indicated that should not be used for predictive purposes. For UMAP the results were not good either, but it is coherent with previous obtained results, the models enhance their performance if more global structure is introduced, which leave some space for continuing exploring the capability of UMAP as prediction technique.

Regression plot for PCR model and residuals can be seen in Figure 29. All maps can be checked in the corresponding Jupyter notebook. A residual plot shows the difference between the observed response and the fitted response values. The ideal residual plot, called the null residual plot, shows a random scatter of points forming an approximately constant width band around the identity line, which roughly corresponds with the residual plot obtained.

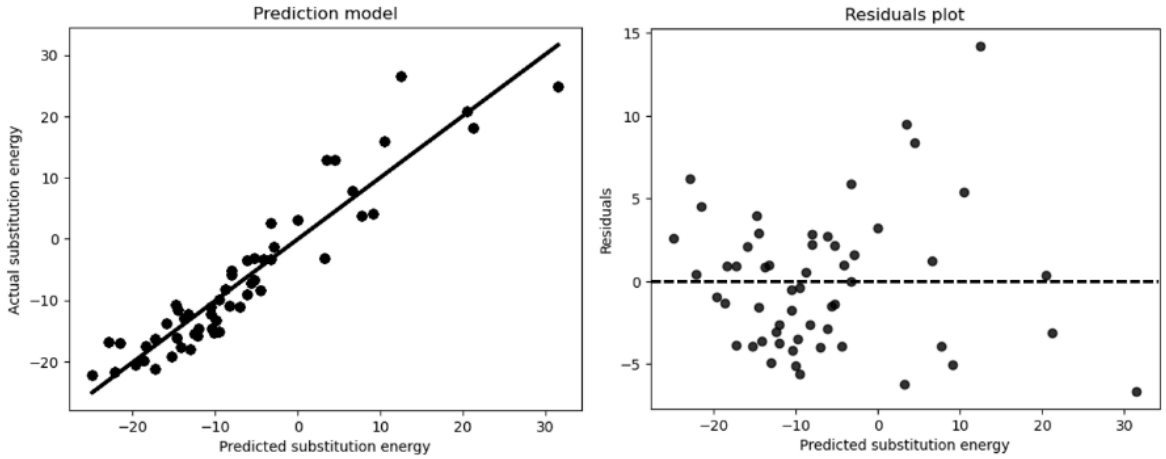


Figure 29. Prediction plot and residuals plot for PCA model in random state = 6. The random state corresponds to the maximum regression coefficient (r^2 = 0.88).

Since the test case is based on an experimental work, the ligands that were tried have been used now to plot them in the established maps and their positions analysed. Combining these new ligands and the information that has been obtained from all the analysis can show the potential uses of the tool for optimisation purposes.

The results of plotting the new ligands on the established models are presented on Figure 30. Before any plot was obtained, it was expected that, since all had the same substituents, they were located in the cluster where Ph substituents were present according to the results obtained in the previous analysis. This was true for t-SNE where all ligands are located in the same region since t-SNE captures local information, the model is highly directed by similarities between structurally close species, and so they are clustered according to the substituents.

However, this was not repeated in case of PCA and UMAP, two different groups were observed. Ligands 2, 3, 4 and 5 (see Figure 13, section 3.2) were clustered in the region where Ph substituents are located but ligand 1 was sitting quite far from those. Ligand 1 was sitting on the region where withdrawing ligands were located. This indicated that ligand 1, was effectively, withdrawing. After checking for possible error, the descriptor values of ligand 1 and withdrawing substituent ligands were compared and similarities were found.

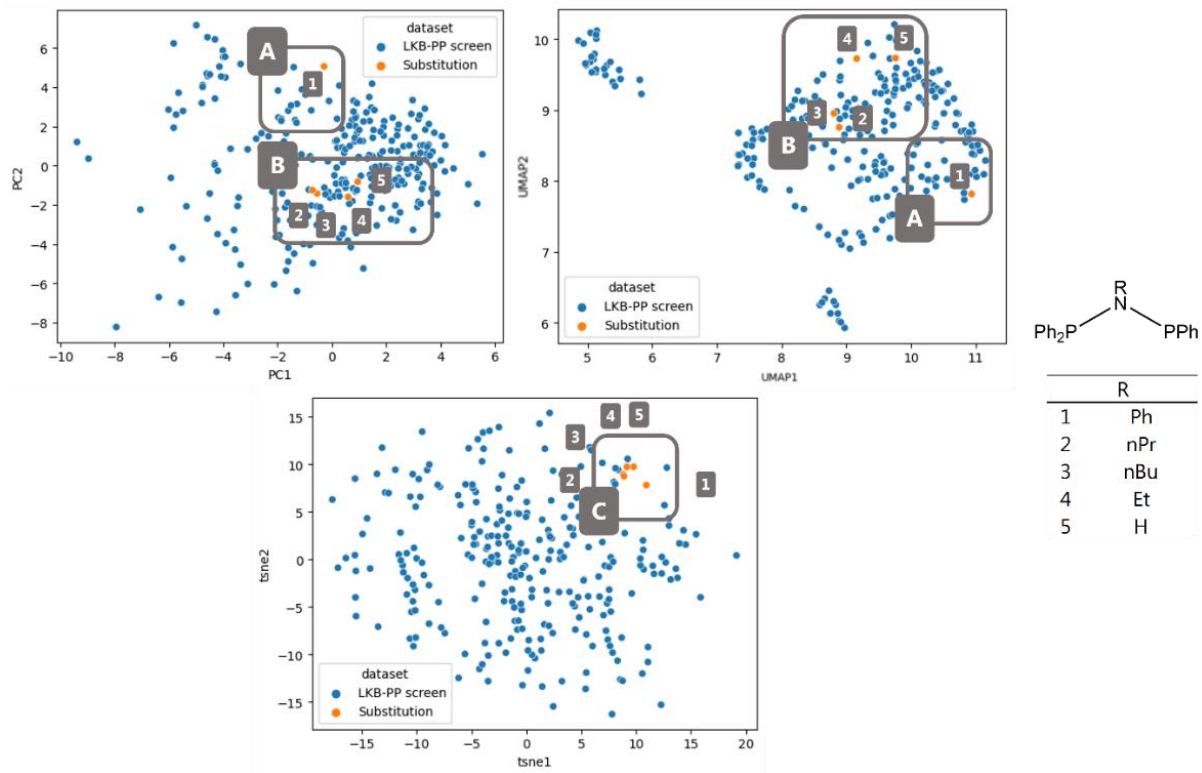


Figure 30. PCA, UMAP with number of neighbours = 15 and t-SNE models (blue) and the test ligands sitting on those models (orange). For UMAP with number of neighbours = 5 and 10, same results are found. This could be check by choosing the desired number of neighbours in the corresponding script.

One of the similarities that were found was the bending energy for the Zn complex model, which was very similar between ligand 1 and ligands containing withdrawing substituents. Compared to the rest of test ligands, the bending energy for ligand 1 was approximately the half. This indicated a withdrawing character. When the Zn structure was checked, the nitrogen in the backbone was fully planar, indicting a sp^2 hybridization, which is a characteristic geometry when the electron pair is conjugated to a π system. The phenyl ring attached to the nitrogen is almost perpendicular to the nitrogen plane but bent enough to allow to establish a conjugated system as shown in Figure 31. Electron density coming from the nitrogen is delocalized along the aromatic ring generating an electron deficient site in that atom. As a result, electron density from each phosphorus atoms try to compensate this effect, generating an electron withdrawing diphosphine.

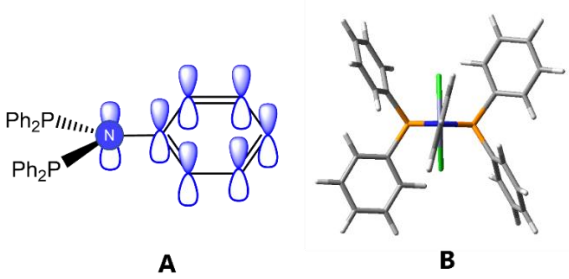


Figure 31. Schematic representation of the withdrawing nature of ($Ph_2P(Ph)NPPh_2$). (A) Molecular orbital representation showing the capability of conjugation. (B) 3D representation of $ZnCl_2(Ph_2P(Ph)NPPh_2)$ model to show the bent phenyl group of the backbone.

These results are interesting and excellent to show the possible applications of the developed tools. The map in Figure 32 is combining the previous information and the trend for substitution energy can be observed, providing a graphical support for quick understanding of the results. What was apparently a set of similar ligands with only small variations on the

backbone, ends with one of their ligands showing clearly different electronic behaviour. Returning to the experimental example, the experimentalist may have selected these ligands because they are similar, however the maps have shown that one of them ($Ph_2P(Ph)NPPh_2$) have different electronic effects.

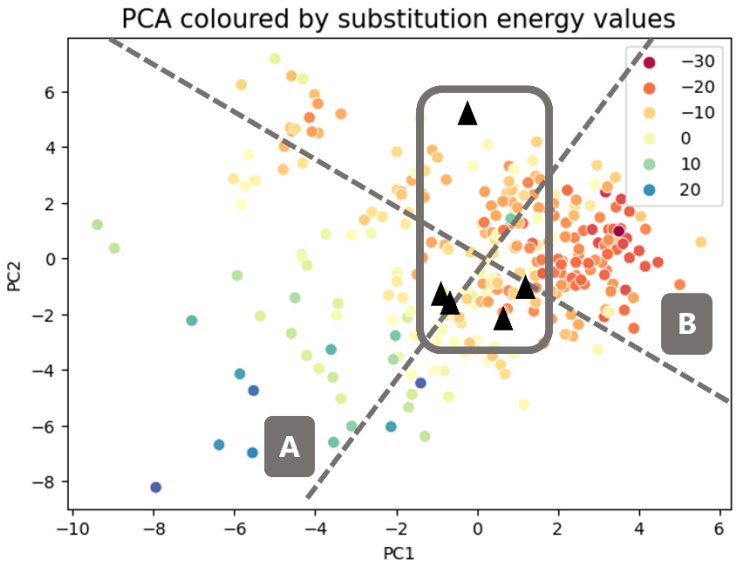


Figure 32. Proposal of map for catalytic optimisation. The map combines the steric and electronic trends found in section 4.1 and the substitution energy trend. The square corresponds to a similar region on substitution energies where the test ligands are located. A and B correspond to steric and electronic axes respectively.

The electron-withdrawing nature of $Ph_2P(Ph)NPPh_2$ does not affect the catalyst activation since it was mainly sterically controlled, and so all the ligands appeared in the same range of substitution energy values as shown in the box of Figure 32, and no real difference is found when checking the experimental results for the enthalpy value.[28] However, if these ligands are used for a certain catalytic cycle, and it has been proven before that only strong donor ligands are providing positive catalytic results, by looking at the map on Figure 32, ligand 1 can be immediately discarded for any experiment.

The results on Figure 32 indicate that, apart from the fact that the chemical information that could be seen applying clustering algorithms was related to substituents, ligands could sit in different regions of the chemical space maps if differences related to their properties are present. This means that there is more chemical information than the one that is shown by algorithms as it was discussed at the end of section 4.2.

Since the new ligands tried have all similar backbones and substituents, similar ranges of substitution energies were obtained in Figure 32. However, if more diverse set of ligands were tried, they will sit along the map according to their steric characteristics. Then, by knowing which is the range of optimal energy for the precatalyst activation, a set of ligands could be directly selected as experimental targets, helping to optimise the process.

Another thing the experimentalist could do is target selection. By an optiomal region on the map or a ligand that has shown good performance, the experimentalist could select a set of ligands that are introduced on the LKB-PP_{screen} based on what ligands are around the good/optimal one.

With these results the experimentalist can change the decision before carrying out any experiment, and in this way experimental design can be directed and optimised.

5

Summary and conclusions

(ENG)

By analysing the data in the LKB-PP_{screen} it has been concluded that all the DR techniques tried present trends regarding the substituents of the different diphosphines, and it is confirmed when clustering techniques are applied. No significant differences have been shown by the two clustering algorithms applied (k-means and hierarchical clustering). There appeared slight differences, related to the steric and electronic properties of ligands.

The maps of diphosphine chemical space and models in the test case are affected depending on the type of information relationships (global or local) that is captured by each DR technique. For the LKB-PP_{screen} dataset it can be concluded that the introduction of more global structures enhances the diphosphine chemical space maps and models used to built machine learning based prediction tools.

It is the case of PCA, that has shown nice visualization due to low overlap and good clustering, excellent retention of chemical information and the best performance for prediction models, with the added value of explained variance and biplots that helped the data analysis and the comprehension of ligand behaviour in the chemical space. Thank to those later characteristics, trends have been easily established making PCA the most complete, versatile, and flexible dimensionality reduction technique, capable of giving good results in different scenarios.

Also, regarding global structures, UMAP has shown a considerable improvement when more global structure was introduced by increasing the number of neighbours. Results have concluded that UMAP presented the least overlap of clusters among the three DR techniques. Despite that, no other advantage makes UMAP a preferred choice when compared with PCA.

t-SNE gives, in general terms, the same information as UMAP and PCA. The differences between the plots are attributed to the ability of capturing local structures. t-SNE cannot be used to reduce dimensionality, as comparison and test case have shown. Focusing on local structures does not provide different chemical information, nor improvement of prediction models.

At the end, and with the help of the experimentally inspired test case the applicability of the LKB to derive machine learning tools capable to optimise catalyst selection has been shown. In this Bachelor thesis a very simple optimisation example has been developed, but one may think of much more sophisticated and complex optimisation purposes such as elucidation of catalytic cycles or target molecule prediction. However, a lot of work is still needed to fully develop such topics.

Meanwhile, exploration of data through maps of chemical space and the use of those for the introduction of novel ligands may help to point out features and characteristics that may not be obvious for researches, and the combination of maps and trends can be an easy, but powerful tool for visualization and guidance when experimental work is designed.

It is important to keep in mind that these are the conclusions for the data in the LKB – PP_{screen}, so they may not be extrapolated to other datasets or databases. However, similar results should be expected for diphosphine, or phosphine databases generated with same or similar descriptors as the LKB – $PP_{screen.}$

(CAT)

A través de l'anàlisi de dades del LKB-PP$_{screen}$ s'ha pogut concloure que totes tècniques de reducció de la dimensionalitat provades presenten tendències respecte els substituents de les difosfines. No s'han trobat diferencies significatives entre els dos mètodes de clúster. Les principals diferencies eren degudes a propietats estèriques i electròniques dels lligands.

Els mapes i models creats al test case es veuen afectats depenent del tipus de relacions a la informació (estructures locals o globals) que es capaç de capturar cada tècnica de reducció de la dimensionalitat.

És el cas de la PCA, que ha demostrat una bona visualització deguda a la poca superposició i a un bon clustering, una excel·lent retenció de la informació i els millors resultats en els models de predicció, combinat amb la possibilitat de obtenir biplots i gràfics de la desviació explicada. Totes aquestes característiques fan que la PCA sigui la més completa, versàtil i flexible de totes les tècniques de reducció de la dimensionalitat, capaç de donar bons resultats en diferents casos.

Respecte la conservació d'estructures globals de la informació, UMAP ha demostrat una considerable millora quan més estructura global és incorporada als models a través d'augmentar el nombre de neighbours. Els resultats presentats conclouen que UMAP és la tècnica que presenta la menor superposició. Tot i això, no hi ha cap altre avantatge que la faci preferible davant la PCA.

La t-SNE proveeix, en termes generals, el mateix tipus d'informació que la PCA o l'UMAP. Les diferencies estan atribuïdes a la capacitat de la t-SNE d'incorporar estructures locals de la informació. La técnica t-SNE no ha de ser utilitzada per reduir la dimensionalitat tal i com mostren els resultats del test case. Podem concloure doncs, que incorporar més estructures locals no comporta cap benefici a l'hora d'utilitzar el LKB-PP$_{screen}$ per crear mapes i models predictius.

Finalment, i amb l'ajuda del test case inspirat en un treball experimental, la aplicabilitat del LKB per tal de obtenir eines basades en el machine learning capaces d'optimitzar catalitzadors organometàl·lics ha estat demostrada. En aquest treball de fi de grau l'exemple d'optimització és molt senzill, però podem pensar altres propostes molt més complexes i sofisticades com per exemple la elucidació de cicles catalítics complets o la predicció de molècules diana. Malgrat això, queda encara molta recerca per tal de desenvolupar aquets camps.

Mentrestant l'exploració de les dades a través de mapes de l'espai químic i l'ús d'aquets per introduir nous lligands pot ser una eina molt útil per tal de descobrir característiques i tendències no-trivials pels investigadors. Combinar els mapes i les tendències és una eina visual molt útil per tal de proveir un mètode d'optimització fàcil, visual i ràpid.

És important tenir en compte que aquestes conclusions són per les dades que es troben al LKB-PP$_{screen}$ cosa que fa que hi ha la possibilitat de que no siguin extrapolables a altres conjunts de dades. De tota manera, s'esperaria veure els mateixos resultats en el LKB-PP ja que conté difosfines i els mateixos descriptors.

Data and code availability.

The code and the data referenced in the bachelor's Thesis, as well as the helium input geometries, can be found in the following GitHub repository:

https://github.com/MarioVillares/BachelorThesis.git

The README.md file contains guide all the information to go through all the data files and scripts.

Computational details.

Multiple conformations are available in such complex structures, and this type of analysis is highly demanding for DFT methods. For this reason, Molecular Mechanics (MM) was used, with the Merk Molecular Force Field (MMFF) force field from Spartan software.[31] Stochastic conformational search was performed with a criterion of stopping after 500 iterations.

The wedge of helium descriptors are calculated by the generation of two different structures that come from the same zinc dichloride complex (described in section 3.1). All this process is done by using Maestro software[32] (Version 12.0.012; Schrödinger, LLC, New York, NY).

All calculations used the Jaguar package[33] and the standard Becke-Perdew (BP86) density functional.[34] The functional used overbinds slightly for molecules, compensating for the lack of dispersion corrections and ensuring that most geometry optimisations are successful.

The Jaguar triple – ξ form of the standard Los Alamos ECP basis set (LACV3P) was used on the transition metal atoms, employing the 6-31G* basis for all other atoms. "Loose" convergence (5 times larger than default criteria) was used for all geometry optimizations.

Dispersion corrections were not implemented in commercial software, for very large ligands.

References

(1) Tolman, C. A. Steric Effects of Phosphorus Ligands in Organometallic Chemistry and Homogeneous Catalysis.

(2) Fey, N.; Orpen, A. G.; Harvey, J. N. Building Ligand Knowledge Bases for Organometallic Chemistry: Computational Description of Phosphorus(III)-Donor Ligands and the Metal-Phosphorus Bond. *Coord Chem Rev* **2009**, *253* (5–6), 704–722. https://doi.org/10.1016/J.CCR.2008.04.017.

(3) Halpern, J.; Phelan, P. F. *Reactions of Bis (Dioximato) Cobalt (II) Complexes with Organic Halides. Influence of Electronic and Steric Factors upon Reactivity*. https://pubs.acs.org/sharingguidelines.

(4) Fey, N. Lost in Chemical Space? Maps to Support Organometallic Catalysis. **2015**. https://doi.org/10.1186/s13065-015-0104-5.

(5) Lipinski, C.; Hopkins, A. *Navigating Chemical Space for Biology and Medicine*; 2004. www.nature.com/nature.

(6) Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Murray, P.; Orpen, A. G.; Osborne, R.; Purdie, M. Computational Descriptors for Chelating P,P- And P,N-Donor Ligands. *Organometallics* **2008**, *27* (7), 1372–1383. https://doi.org/10.1021/om700840h.

(7) Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J. J.; Murray, P.; Hose, D. R. J.; Osborne, R.; Purdie, M. Expansion of the Ligand Knowledge Base for Monodentate P-Donor Ligands (LKB-P) †. *Organometallics*. https://doi.org/10.1021/om100648v.

(8) Fey, N.; Tsipis, A. C.; Harris, S. E.; Harvey, J. N.; Orpen, A. G.; Mansson, R. A. Development of a Ligand Knowledge Base, Part 1: Computational Descriptors for Phosphorus Donor Ligands. *Chemistry - A European Journal* **2005**, *12* (1), 291–302. https://doi.org/10.1002/CHEM.200500891.

(9) Gensch, T.; Dos, G.; Gomes, P.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'addario, M.; Sigman, M. S.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *Cite This: J. Am. Chem. Soc* **2022**, *2022*, 1205–1217. https://doi.org/10.1021/jacs.1c09718.

(10) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**.

(11) Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J. J.; Murray, P.; Hose, D. R. J.; Osborne, R.; Purdie, M. Expansion of the Ligand Knowledge Base for Monodentate P-Donor Ligands (LKB-P). *Organometallics* **2010**, *29* (23), 6245–6258. https://doi.org/10.1021/om100648v.

(12) Fey, N.; Haddow, M. F.; Harvey, J. N.; McMullin, C. L.; Orpen, A. G. A Ligand Knowledge Base for Carbenes (LKB-C): Maps of Ligand Space. *Journal of the Chemical Society. Dalton Transactions* **2009**, No. 39, 8183–8196. https://doi.org/10.1039/b909229c.

(13) Jover, J.; Fey, N. Screening Substituent and Backbone Effects on the Properties of Bidentate P,P-Donor Ligands (LKB-PPscreen). *Dalton Transactions* **2013**, *42* (1), 172–181. https://doi.org/10.1039/c2dt32099a.

(14) Jesús Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J. J.; Murray, P.; Hose, D. R. J.; Osborne, R.; Purdie, M. Expansion of the Ligand Knowledge Base for Chelating P,P-Donor Ligands (LKB-PP) †. *Organometallics* **2012**, *31*, 23. https://doi.org/10.1021/om300312t.

(15) Fey, N.; Koumi, A.; Malkov, A. V.; Moseley, J. D.; Nguyen, B. N.; Tyler, S. N. G.; Willans, C. E. Mapping the Properties of Bidentate Ligands with Calculated Descriptors (LKB-Bid). *Dalton Transactions* **2020**, *49* (24), 8169–8178. https://doi.org/10.1039/d0dt01694b.

(16) Gillespie, J. A.; Dodds, D. L.; Kamer, P. C. J. Rational Design of Diphosphorus Ligands - A Route to Superior Catalysts. *Dalton Transactions*. 2010, pp 2751–2764. https://doi.org/10.1039/b913778e.

(17) Ribas Gispert, J. *Coordination Chemistry*, 1st ed.; Verlag GmbH, KGaA, W., Eds.; WILEY-VCH: Barcelona , 2008.

(18) Christoph Elschenbroich. *Organometallics* , 3rd Edition.; Wiley-VHC: Wiesbaden , 2005.

(19) Newland, R. J.; Smith, A.; Smith, D. M.; Fey, N.; Hanton, M. J.; Mansell, S. M. Accessing Alkyl- and Alkenylcyclopentanes from Cr-Catalyzed Ethylene Oligomerization Using 2-Phosphinophosphinine Ligands. *Organometallics* **2018**, *37* (6), 1062–1073. https://doi.org/10.1021/acs.organomet.8b00063.

(20) Jover, J.; Fey, N. Cite This: Dalton Trans. **2013**, *42*, 172. https://doi.org/10.1039/c2dt32099a.

(21) Kohn, W.; Sham, L. J. *Self-Consistent Equations Including Exchange and Correlation Effects*.

(22) Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine, Series 6, 2(11)*. 1901, pp 559–572.

(23) Van Der Maaten, L.; Hinton, G. *Visualizing Data Using T-SNE*; 2008; Vol. 9.

(24) Birkholz, M. N.; Freixa, Z.; Van Leeuwen, P. Bite Angle Effects of Diphosphines in C–C and C–X Bond Forming Cross Coupling Reactions. *Chem Soc Rev* **2009**, *38* (4), 1099–1118. https://doi.org/10.1039/b806211k.

(25) Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Murray, P.; Orpen, A. G.; Osborne, R.; Purdie, M. Computational Descriptors for Chelating P,P-and P,N-Donor Ligands 1. https://doi.org/10.1021/om700840h.

(26) Fey, N.; Papadouli, S.; Pringle, P. G.; Ficks, A.; Fleming, J. T.; Higham, L. J.; Wallis, J. F.; Carmichael, D.; Mézailles, N.; Müller, C. Setting P-Donor Ligands into Context: An Application of the Ligand Knowledge Base (LKB) Approach. *Phosphorus Sulfur Silicon Relat Elem* **2015**, *190* (5–6), 706–714. https://doi.org/10.1080/10426507.2014.983599.

(27) Kamer, P. C. J.; Van Leeuwen, P. W. N. M.; Reek, J. N. H. Wide Bite Angle Diphosphines: Xantphos Ligands in Transition Metal Complexes and Catalysis. **2001**. https://doi.org/10.1021/ar000060.

(28) Balakrishna, M. S.; Priya, S.; Sommer, W.; Nolan, S. P. Synthesis and Thermochemical Study of Ligand Substitution Reactions of Aminobis(Phosphines), Ph2P(R)NPPh2, with [Me 2Pt(COD)]. *Inorganica Chim Acta* **2005**, *358* (9), 2817–2820. https://doi.org/10.1016/j.ica.2005.03.021.

(29) Karlov, D. S.; Sosnin, S.; Tetko, I. V.; Fedorov, M. V. Chemical Space Exploration Guided by Deep Neural Networks. *RSC Adv* **2019**, *9* (9), 5151–5157. https://doi.org/10.1039/c8ra10182e.

(30) Peach, R. L.; Yaliraki, S. N.; Lefevre, D.; Barahona, M. Data-Driven Unsupervised Clustering of Online Learner Behaviour. *NPJ Sci Learn* **2019**, *4* (1). https://doi.org/10.1038/s41539-019-0054-0.

(31) Young, D. *Computational Chemistry* , SPARTAN.; Wiley-Interscience, 2001; Vol. Appendix A. A.

(32) Schrödinger Release 2023-1: Maestro. Maestro, Schrödinger, LLC : New York, NY 2021.

(33) Bochevarov, A. D.; Harder, E.; Hughes, T. F.; Greenwood, J. R.; Braden, D. A.; Philipp, D. M.; Rinaldo, D.; Halls, M. D.; Zhang, J.; Friesner, R. A. Jaguar: A High-Performance Quantum Chemistry Software Program with Strengths in Life and Materials Sciences. *Int J Quantum Chem* **2013**, *113* (18), 2110–2142. https://doi.org/10.1002/QUA.24481.

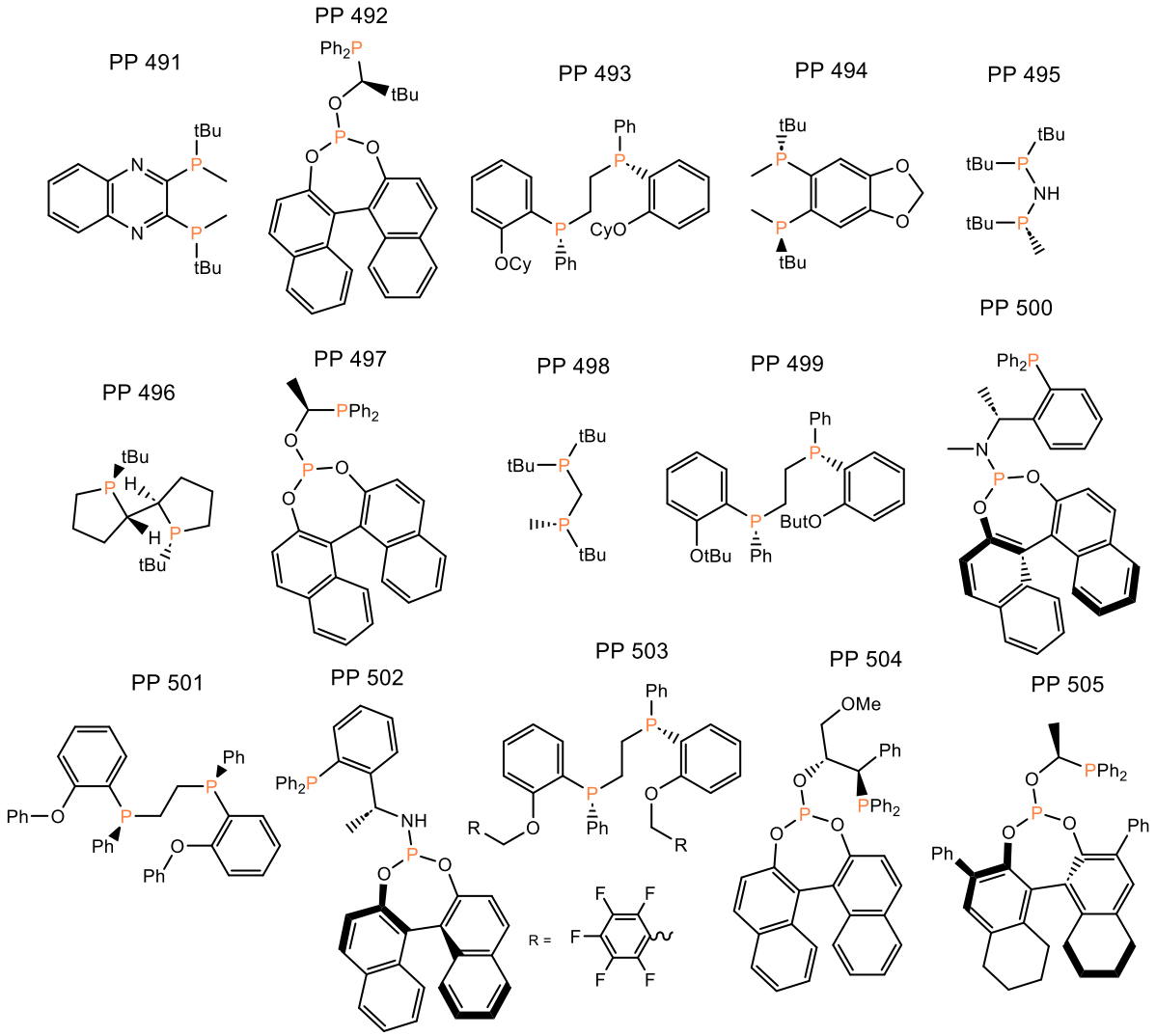
(34) Slater, J. C. Quantum Theory of Molecules and Solids Vol. 4: The Self-Consistent Field for Molecules and Solids. *Physics Today* **1974**, *27* (12).

(35) Xu, L. C.; Zhang, S. Q.; Li, X.; Tang, M. J.; Xie, P. P.; Hong, X. Towards Data-Driven Design of Asymmetric Hydrogenation of Olefins: Database and Hierarchical Learning. *Angewandte Chemie - International Edition* **2021**, *60* (42), 22804–22811. https://doi.org/10.1002/ANIE.202106880.

Annex

Annex Figure 1 presents the ligands introduced following the methodology of section 3.1 during my internship in Fey group before starting the Bachelor Thesis, 15 ligands were selected from the asymmetric hydrogenation database (AHOs)[35] and introduced to the LKB – PP dataset. Molecular Mechanics (MM) stochastic type conformational searches were performed *(See computational details)*, to screen conformational space for free ligands and their tetrahedral zinc complexes [$ZnCl_2$(PP)]. Once the models described in section 3.1 were constructed, DFT calculation were performed using Jaguar (See computational details).

In Annex Table 1, the truncated ligands used for the final 20 introduced ligands (15 in the internship and 5 for the test case) are shown. The numbering corresponds to the identifier shown in Annex Figure 1 and Figure 10 respectively.

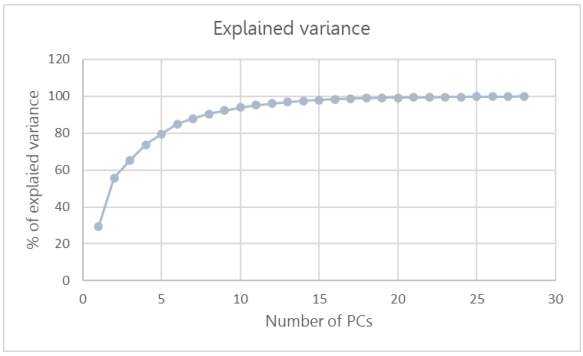


Annex Figure 1. 15 introduced diphosphine ligands to the LKB-PP database, extracted from the asymmetric hydrogenation database (AHOs). The numbers are consistent with the full LKB-PP database.

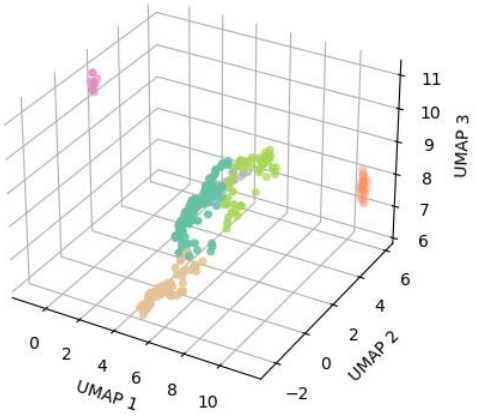
Annex Table 1. Truncated ligands. If the phosphine is symmetric descriptors homo, lumo and proton affinity will be the same that homo2, lumo2, proton affinity 2 respectively.

Truncated ligand	Corresponding diphosphine in which it was used
PPh_2Me	PP492, PP493, PP497, PP499, PP501, PP503, PP505
PMe_2Ph	PP494
PMe_2NH_2	PP495
PPh_3	PP502
PMe_3	PP496, PP498
$P(OPh)_2NMe_2$	PP500
$P(OPh)_2NH_2$	PP502
$P(OPh)_2OMe$	PP492, PP504, PP505
$PMe_2(CH{=}CH_2)$	PP491
PPh_2NH_2	PP506, PP507, PP508

PC	% EV	% EVA	PC	% EV	% EVA
PC1	29.5	29.5	PC15	0.5	98.0
PC2	26.2	55.7	PC16	0.4	98.4
PC3	9.6	65.3	PC17	0.3	98.7
PC4	8.4	73.7	PC18	0.2	98.9
PC5	5.9	79.6	PC19	0.2	99.1
PC6	5.3	84.9	PC20	0.1	99.2
PC7	3.1	88.0	PC21	0.1	99.3
PC8	2.4	90.4	PC22	0.1	99.4
PC9	1.9	92.3	PC23	0.1	99.5
PC10	1.7	94.0	PC24	0.1	99.6
PC11	1.2	95.2	PC25	0.1	99.7
PC12	0.9	96.1	PC26	0.1	99.8
PC13	0.8	96.9	PC27	0.1	99.9
PC14	0.6	97.5	PC28	0.1	100.0



Annex Figure 2. Left: Table with all the values of the explained variance for each PC (%EV) and the accumulated explained variance (%EVA). Right: plot of the explained variance against the number of PCs.



Annex Figure 3. 3D representation of UMAP with number of neighbours = 10, coloured by k-means clustering algorithm.