DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS

Lucía Morán González

UNIVERSITAT ROVIRA i VIRGILI

Decoding Chemical Processes: The Power of Data-Driven Descriptors

Lucía Morán González

DOCTORAL THESIS
2023

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

Lucía Morán González

Decoding Chemical Processes: The Power of Data-Driven Descriptors

Ph.D. Thesis

Supervised by Prof. Feliu Maseras Cuní

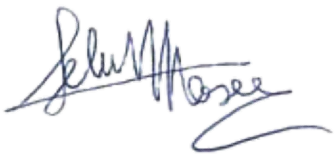
Tarragona, 2023

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

UNIVERSITAT
ROVIRA i VIRGILI

Prof. Feliu Maseras Cuní, Group Leader at the Institute of Chemical Research of Catalonia,

I STATE that the present study, entitled "Decoding Chemical Processes: The Power of Data-Driven Descriptors" presented by Lucía Morán González to receive the degree of Doctor, has been carried out under my supervision at the Institute of Chemical Research of Catalonia (ICIQ) and that it fulfills all the requirements to be eligible for the International Doctor Distinction.



Doctoral Thesis Supervisor
Tarragona, October 30th, 2023

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

Acknowledgements

I would like to thank my supervisor, Prof. Feliu Maseras, for giving me the chance of pursing my PhD in his research group,but above all for being a really good supervisor. I warmly remember when I came for the first time to ICIQ, surprised and unaware of the enriching experience I was about to begin. Thanks for all the support, and the patience with my insecurities and nerves. It has been a very pleasant experience.

Then, of course, I also have to thank all the labmates who have been around through these years. They have become family, a 'particular' compu-family that has made my life full of happiness. I am grateful for meeting such amazing people.

Furthermore, I would also thank my parents and my sister. They have always been my role models in life, supporting me without judgment. They have instilled in me the values of humility, kindness, and integrity.

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

Funding Agencies

This work has been funded by the Insitute of Chemical Research of Catalonia (ICIQ) and by a FI grant from the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya and the European Social Fund.

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

Publications

Morán-González, L.; Rodríguez-Guerra Pedregal, J.; Besora, M.; Maseras, F. Understanding the Binding Properties of N-heterocyclic Carbenes through BDE Matrix App. *Eur. J. Inorg. Chem.*, **2022**, e202100932, 1-6. doi.org/10.1002/ejic.20210093

Morán-González, L.; Besora, M.; Maseras, F. Seeking the Optimal Descriptor for S_N2 Reactions through Statistical Analysis of Density Functional Theory Results. *J. Org. Chem.*, **2022**, 1, 363-372. doi.org/10.1021/acs.joc.1c02387

Morán-González, L.; Maseras, F. A computational search of the ideal metal fragment for monohapto coordination of dihydrogen. *Aus. J. Chem.*, **2023**, doi.org/10.1071/CH23121

Morán-González, L.; Betten, J. E.; Kneiding, H.; Balcells, D. AABBA: Atom-Atom Bond-Bond Bond-Atom Graph Kernel for Machine Learning on Molecules and Materials. *ChemRxiv*, **2023**, doi.org/10.26434/chemrxiv-2023-5wbkr

Morán-González, L.; Maseras, F. Hidden Descriptors from DFT Calculations: Decoding the Underlying Forces of Kinetics and Thermodynamic Processes. *Manuscript in preparation*

Abstract

Chemical descriptors play an integral role in decoding chemistry by revealing valuable insights. With the greater accessibility of computational power, the applicability of these descriptors in data-driven strategies has seen a tremendous rise. Consequently, this progress has shaped the design of catalytic and synthetic processes more efficiently compared to the traditional trial-and-error approaches. These implementations have transformed the chemical featurization into a significant area of research in computational chemistry. However, commonly used descriptors can sometimes overstate the connection between specific parameters and targeted properties, yielding suboptimal statistical models. Besides, the development of optimal descriptors that can address specific chemical problems is underdeveloped. In this Thesis, building on prior work from our group, we have furthered the concept of hidden descriptors as an alternative chemical representation. These entities best characterize the underlying forces of the chemical problem under study. The pioneering work using this methodology covered the study of the electronic interaction between metal and ligands, setting a precedent for subsequent research encapsulated in this work. In the first part (Chapter 3), we developed and employed the *BDE Matrix App* to derive the hidden descriptors of N-heterocyclic carbene ligands, providing insights into their electronic properties. This analysis offered a comprehensive view of the selected chemical species, revealing trends in electronic properties and their modulation by structural

customization. Furthermore, we delved into the potential formation of $L_nM(\eta^1\text{-}H_2)$ complexes using the hidden descriptors. Next, we shifted towards the application of this methodology in the field of organic chemistry (Chapter 4). We focused on the study of bimolecular nucleophilic substitution reactions at carbon center ($S_N2@C$). In this context, we designed an energy barrier ($\Delta G^{\ddagger}$) matrix encompassing more than 600 DFT-based $\Delta G^{\ddagger}$, including diverse nucleophiles. The application of the singular value decomposition (SVD) algorithm over the matrix enabled the elucidation of the hidden descriptors that accurately describe the entering and leaving group abilities. Additionally, we devised a tool to extend the acquisition of the hidden descriptors for nucleophiles out of the initial set. Finally, we changed the paradigm to tackle large data size regimes. Herein, we unwrapped AABBA, a novel tool to provide fixed-length molecular representation from molecular graphs. This method incorporates autocorrelation functions that transform graph data into molecular vectors compatible with machine learning (ML) techniques. Overall, we believe that our contributions can significantly streamline and enhance the understanding of chemical processes.

Contents

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

List of Abbreviations

AI	Artificial Intelligence
BDE	Bond Dissociation Energy
CoMFA	Comparative Molecular Field Analysis
CP	Cyclopronylidenes
DCM	Dichloromethane
DFT	Density Functional Theory
ESP	Electrostatic Potential
EDG	Electron-donating Group
EWG	Electron-withdrawing Group
FMO	Frontier Molecular Orbital
FOM	Figure of Merit
GNN	Graph Neural Networks
GP	Gaussian Processes
HD	Hidden Descriptor
HOMO	Highest Occupied Molecular Orbital
HTE	High-throughput Experiments
KNN	Kohonen Neural Network
LEP	Lone Energy Pair
LKB	Ligand Knowledge Base
LR	Linear Regression
LUMO	Lowest Occupied Molecular Orbital
MAE	Mean Absolute Error

ML	Machine Learning
MLR	Multilinear Regression
MSE	Mean-squared Error
NatQG	Natural Quantum Graph
NBO	Natural Bond Orbital
NLE	Nonlinear Effect
NLP	Natural Language Processing
NHC	N-heterocyclic Carbene
PCA	Principal Component Analysis
PES	Potential Energy Surface
QM	Quantum Mechanics
QSAR	Quantitative Structure-Activity Relationship
QShAR	Quantitative Shape-Activity Relationship
QSPR	Quantitative Structure-Property Relationship
QSRR	Quantitative Structure-Reactivity Relationship
RF	Random Forest
SMILES	Simplified Molecular Input Entry System
SVD	Singular value decomposition
TM	Transition Metal
TS	Transition State
TSS	Transition State Structure
TST	Transition State Theory
ZPVE	Zero Point Vibrational Energy

Chapter 1

Introduction

> The internet wasn't created for mockery, it was supposed to help researchers at different universities share data sets. It was!
>
> — Homer Simpson – The Simpson

1.1 Databases

Human knowledge is constantly evolving. The growth of any field of study is closely tied to the comprehension and categorization of its fundamental concepts.[1] The systematic organization of any source of knowledge stems from our ability to perceive, evaluate and understand reality.

This concept fits well with the view of human beings as an *animal symbolicum*, a term coined by the German neo-Kantian philosopher Ernst Cassirer.[2] Humans are symbol-making animals, that create symbols and classes to understand the universe. This classification of data has driven the progress of society. Scientific endeavors towards the pursuit of suitable collection and organization of information should be viewed not merely as the accumulation of facts. Instead, they should be recognized as a means of conducting research.[3] These scientific ideas are at the root of the FAIR

principles, which stand for data that is Findable, Accessible, Interoperable, and Reusable.[4]

In practice, these ideas place a lot of importance on the generation and construction of datasets, which are potentially useful for many research domains.[5] In chemistry, the Periodic Table devised by Mendeleev in the second half of the 19th century serves as an early and very successful example of intelligent collection of information.[6] During the middle of the 20th century, the emergence of the first computers gave rise to the concept of computerized *database*, *i.e.* organized collection of data stored and accessed electronically. The progress in computer science enabled the acquisition of a huge amount of information on a scale that was unthinkable just a few decades before.

The advancement of technology had a profound impact on the field of chemistry.[7,8] An example of these *Big Data* repositories in chemistry is Reaxys[9], a database of chemical reactions. Notably, two key factors have played an important role in the exponential growth of chemical databases. Firstly, as mentioned above, the development of computer devices played a pivotal role.[10] Secondly, the emergence of high-throughput experiments (HTE) in the late 1990s further catalyzed this expansion.[11,12] Evidence of this growth is displayed in Figure 1.1. A bar plot illustrates the frequency of the *chemical database* topic in scientific articles indexed in the Web of Science repository.[13] From Figure 1.1, it is clear that the number of scientific publications related to the matter has significantly increased. From the past decade, these advancements have established a new branch of research. The combination of data-led approaches and statistical methods have become ubiquitous terms in chemistry.

Despite automation and HTE advancements, empirical screening in critical areas like catalysis still grapples with limitations in data size and normalization of available experimental data.[14] As Wiest and co-workers have recently pointed out "*the use of HTE datasets in ML has some significant drawbacks in that these datasets represent a very narrow part*

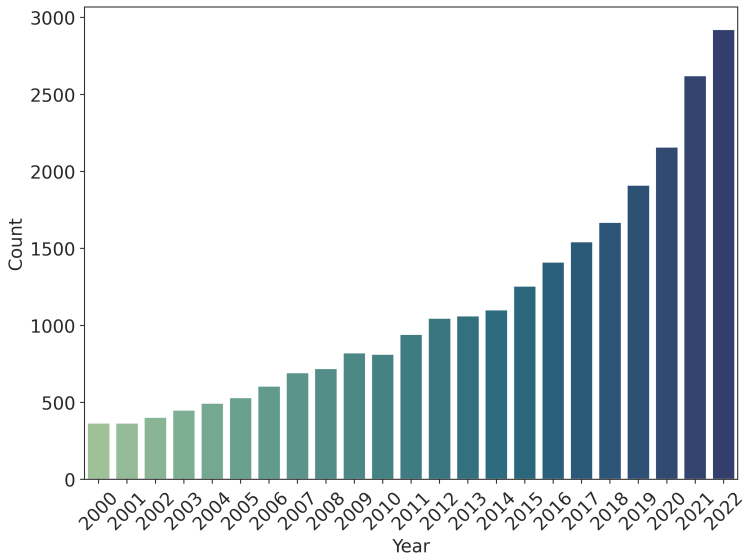


Figure 1.1: Bar plot illustrating the frequency of mentions of the chemical database topic in scientific publications between 2000 and 2022.

of the reaction space, are very time- and resource intensive and present challenges with overfitting of the model".[15] Fortunately, an efficient solution to avoid skewed or small datasets lies in the application of computational chemistry. Its application not only unveils the investigation of mechanistic studies, but also allows the evolution of the data-driven area.[16,17]

Several Quantum Mechanics (QM)-based databases have been employed or hold great promise. Notable examples are QM9 database in organic chemistry[18]; ioChem-BD,[19] and PubChem[20] platforms for storing QM calculations; and Materials Project[21] for computed materials. Regarding transition metal-complexes, datasets such as tmQM[22] and tmQMg[23] have been developed by Balcells and co-workers. Furthermore, Fey and her colleagues have contributed to the field with the Ligand Knowledge Base (LKB) containing monodentate phosphorus ligands (LKB-P)[24,25] and carbenes (LKB-C)[26], alongside the recently published library called *kraken*[27]. With these computational tools, researchers can leverage the

power of simulations to augment, predict, and complement experimental results.

This brief timeline above highlights how data collection has influenced in the understanding and discovery in the field of chemistry. Throughout the course of this Thesis, we will see new approaches to expand data utilization.

1.2 Origin of the molecular descriptors and QSAR

In the previous section 1.1, we highlighted the necessity to classify data in order to extract knowledge from it. In chemistry, this data is usually represented as chemical scaffolds or as experimental and computational parameters. Within the chemical discipline that harnesses the chemical information to design data-related approaches, we find the *chemical descriptors*. Generally, a chemical descriptor is defined as a mathematical object that encodes chemical information of a compound or process. Another term commonly used for this purpose is *molecular descriptor*. While these two terms are usually interchangeably, in fact, they refer to different concepts. The latter is associated with molecular or complex information, while the former can also be related to broader concepts such as melting point, solubility, and others. Unless otherwise stated, we will use in this Thesis both terms indistinctly.

The origin of molecular descriptors can be traced back to the 1860s and 1880s when Crum-Brown presented his Doctoral Thesis *On the Theory of Chemical Combination*.[28] Additionally, August Kekulé proposed for the first time the benzene structure, which established the foundation for the Theory of the Chemical Structure.[29] Later on, the 3D conception of chemical structures was also born. The molecular skeleton was defined according to their topological descriptors.

Indeed, the evolution of the chemical descriptors cannot be understood without considering the Quantitative Structure-Activity Relationship (QSAR). QSAR methods are based on the assumption that the structure

of a molecule, such as its electronic, steric, and geometrical features (p_i), can be related to its physical, chemical or biological properties (P) as the equation 1.2.1 shows.

$$P = f(p_i) \tag{1.2.1}$$

According to the Equation 1.2.1, when the model that relates P and p_i is found, the value of P can be determined by knowing p_i. The basic idea is that p_i is easier to measure than P. Despite the popularity of the acronym QSAR, it is noteworthy that there are other types of sought relationships. For instance, Quantitative Structure-Reactivity Relationship (QSRR) or Quantitative Shape-Activity Relationship (QShAR). In addition, Quantitative Structure-Property Relationship (QSPR) has a broader use since it encompasses all the properties.

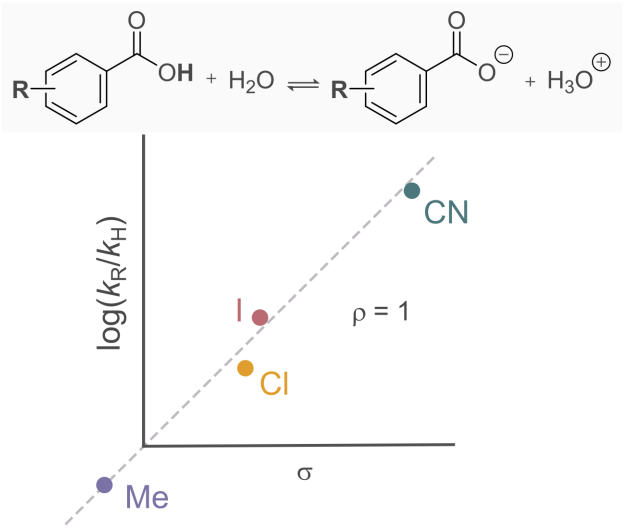


Figure 1.2: Scheme of the ionization reaction of benzoic acid (top) and linear free energy relationship between the Hammet σ and the relative rate constant (bottom).

An early QSAR-related development in chemistry that has been prominent and successful is the Hammet equation.[30] This pioneering work

relates the relative reaction rate of *meta-* or *para-* substituted benzoic acid derivatives (ρ) to tabulated Hammet sigma constants (σ_x) as shown in Figure 1.2. Here, k_R and k_H are the respective rate constant terms for substituted and unsubstituted benzoic acid derivates, δ is the reaction constant and σ_x is the Hammet electric constant of the substituent (x). This linear regression enables to measure the electronic influence of the substituent in the substituted benzoate ion. Thus, it is possible to account for the induction and the resonance effect of the substituents on the reactivity. This milestone approach is still being employed to date,[31] which reinforces the idea that the continual task of developing chemical descriptors, p, is essential to design successful models.

1.3 Types of molecular descriptors

The development of molecular descriptors is based on two concepts: the adequate representation of chemical data and their implementation in models as machine-readable objects.

Regarding the first aspect, the molecular representations depend on their information content. Therefore, there are simple molecular descriptors, such as the molecular mass (m) of the molecule, whereas there are others which are cumbersome to obtain, such as the electrostatic potential (ESP) of the molecule. The information can be data-mined from either experimental sources, including Electronic Lab Notebooks (ELN) or HTE grids,[15] generated from scratch using computational tools,[32] or a mix of both.[33]

Figure 1.3 depicts a schematic representation of the types of molecular structures according to their dimensionality. The description of the chlorobenzene molecule can be approached from various perspectives, offering multiple avenues for investigation. The simplest molecular representation is its formula, C_6H_5Cl. Knowing its chemical formula, we can derive the number of atoms and bonds present, the different types of

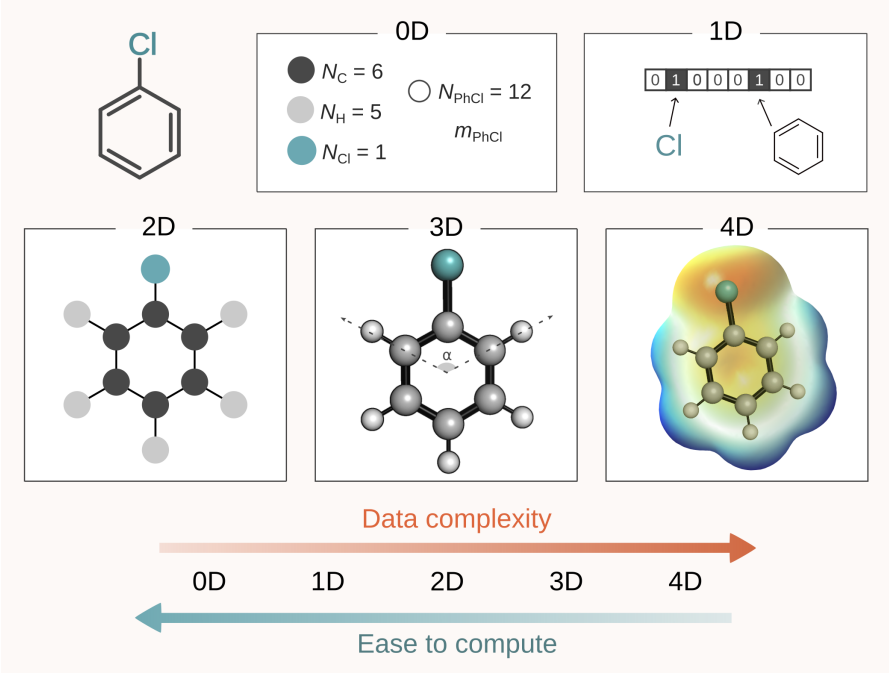


Figure 1.3: Types of molecular descriptors for the chlorobenzene molecule. OD descriptor entails counting atom types; 1D descriptor provides a fingerprint depiction; 2D descriptor shows the molecular graph drawing; 3D descriptor represents *(x,y,z)* coordinates; and 4D descriptor captures the molecule's electrostatic potential (ESP).

atoms, its molecular mass, and others. Moreover, the atomic features such as the covalent radius (S), the electronegativity (χ), and the atomic number (Z) are tabulated values, that can provide insights about the molecule. These are self-explanatory and the easiest to obtain. All these constitutional descriptors and any function of the atomic properties constitute the **0D** descriptors of the molecule (Figure 1.3 0D).

Moving one step forward towards **1D** descriptors, we can split a molecule into functional motives, in similar manner to the retrosynthetic strategy. This type of representation is commonly named *molecular fingerprints* (Figure 1.3 1D) and it is widely employed in machine learning (ML).[34] In 1D representation, the vector featuring representation is based on a binary or count fingerprint that encodes the presence or absence of substructure. Besides, these molecular fingerprints can include physicochemical information.

The following stage of descriptors regards the bond connectivity information. These **2D** descriptors, also known as *topological descriptors*, explicitly include the chemical bond information. Currently, a widely used representation that has grown exponentially in the field of computational chemistry is the *molecular graph* (Figure 1.3 2D).[35,36] Furthermore, Simplified Molecular Input Entry system (SMILES)[37] is probably the most popular string representation of atoms and bonds.[38] The advantage of these representations is its geometry-agnostic approach, which facilitates its computation.

The fourth class of descriptors is calculated from the geometrical or tridimensional representation (x,y,z) of a molecular structure: distances, angles, dihedral, volume. The **3D** descriptors usually provide more information than 0D, 1D, and 2D descriptors. Yet, they require structure optimizations, thus, their collecting process is time-consuming. Most of the physicochemical properties of a molecule are affected by its entire structure. For example, descriptors include electronic features (*e.g.*, NBO features,[23] vibrational frequencies[39], highest occupied molecular orbital

(HOMO) energies). Furthermore, molecular graphs that have geometrical attributes associated with $\mathcal{G}$, $\mathcal{V}$ or $\mathcal{E}$ also belong to that 3D group.

The last group corresponds to the **4D** descriptors. The basic principle is that this type of descriptor considers the interaction between a molecule and a probe. Quantifying such interaction implies mapping a grid space around the molecule. The 4D part of Figure 1.3 shows the molecular electrostatic potential map of chlorobenzene. A famous example of this type of descriptor is the Comparative Molecular Field Analysis (CoMFA) developed by Cramer *et al.*[40] The evaluation of this parameter provides a response to the favorable and unfavorable receptor-ligand interactions.

The second aspect to consider, in the molecular descriptor development is that all the framed information must be converted into machine-readable objects. These chemical parameters must maintain a particular format in terms of dimensionality, numerical expression, and scale, depending on the type of algorithm they will be subjected to (*vida infra*). Therefore, to leverage the mathematical chemical expressions, it is required to follow certain rules. Moreover, the format should ensure the elementary concepts of uniqueness and similarity. Uniqueness for distinguishing subtle differences present in similar molecules, and similarity for describing comparable property values.

In general, using more complex chemical descriptors enhances the prediction task, even for challenging target properties, *e.g.* enantiomeric ratios.[41] Nonetheless, there are instances when this does not hold true. As explained in Equation 1.2.1, the desired relationship is the one that correlates target and feature parameters. Thus, employing complex information does not necessarily guarantee better comprehensibility; indeed, in some cases, it may even saturate the model. This is a consequence of the common pitfall of incorporating irrelevant or biased descriptors in the models. Hence, careful consideration and validation of selected descriptors are imperative to ensure the model's reliability and effectiveness in capturing significant data patterns. Additionally, incorporating more

advanced data into chemical descriptors implies longer extraction times for chemical information. This becomes a significant drawback, especially when handling extensive datasets, where automated techniques are not plausible.

In summary, the pursuit of a balance should be mandatory. When databases with thousands of molecules are represented, fast-to-calculate descriptors are desired *i.e.* 0D, 1D, 2D. On the contrary, if the aim is to search for a complex P target property for a small set of molecules, 3D or 4D might be a potentially better option.

1.4 Development and selection of chemical descriptors

The elaboration and selection of chemical descriptors is a hot topic in chemistry, often referred to as *featurization*. The selection of molecular properties is a critical phase undertaken before any statistical model application.

There are mainly two types of chemical representation derivation: *nonlearned molecular* featurization and *featuring learning* (depicted in Figure 1.4). The first approach is depicted in the upper part of Figure 1.4.[42] This method is also denoted as *feature engineering*, and it is based on manual design and selection of the chemical features. It requires the domain of experts in the field to craft suitable molecular descriptors. These descriptors are built with knowledge of the potential causal factors influencing the target property. Models designed with these types of features offer a greater interpretability of the outcomes. As a result, they are adequate to untangle challenging and specific tasks. For instance, buried volume[43] ($\%V_{bur}$) (Figure 1.4) and Sterimol[44] have their applicability in asymmetric catalysis.[45] Moreover, in a recent publication on the prediction of the solubility of organic solutes,[46] the computation of 3D or 4D descriptors through manual selection was used.

The second strategy relies on computer-aided selection[47], where

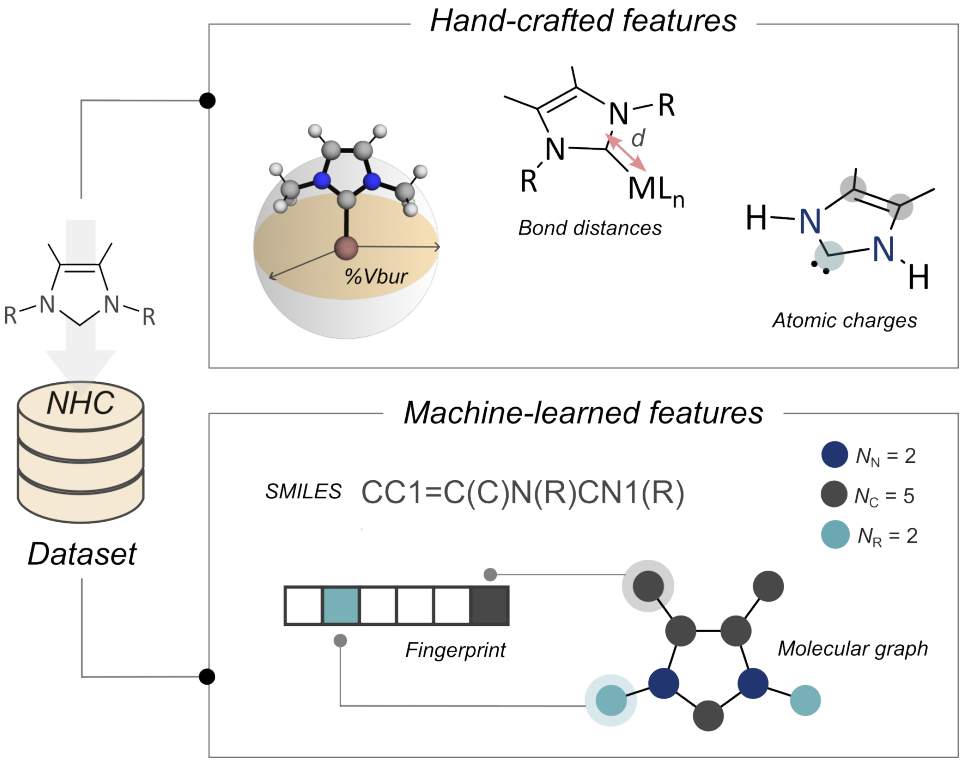


Figure 1.4: Schematic representation of the types of chemical featurization. Hand-craft featurization (top), and machine-learned features (bottom).

the machine takes advantage of inherent logic to produce chemical representations.[48] The lower section of Figure 1.4 illustrates certain machine-learned features. Herein, the chemical structure of the imidazole is taken as the input to yield learned representations. The computer directly takes the structures as the input and returns learned descriptors as SMILES or the fingerprint of the functional groups. These string-based representations do not contain spatial or physical information. Nevertheless, natural language processing (NLP) techniques have achieved accurate models with such descriptors.[49] Additionally, molecular graphs can be automatically generated with machine translation frameworks. The algorithm captures the relevant information from the samples to generate a language-processing representation.

In general, the choice between either approach depends on various factors. It is important to account for the data-size regime as small datasets enable handcrafted feature derivation, whereas thousands of samples require a learning feature process. In some specific scenarios, the bottleneck of the data-driven pipeline is the availability of the chemical descriptors. For example, the commonly used representations for organic molecules, *e.g.* SMILES, are underdeveloped for transition metal complexes. Therefore, more advanced notations should be employed to capture the complexity of these species. Another aspect to consider is the level of expertise required for the task. As mentioned earlier, there is a balanced relationship between the quality of descriptors and their computational cost. High-quality descriptors, such as a structural or mechanistic hypothesis, and physicochemical descriptors, increase the probability of prediction success in models compared to more generalizable representations.[50] Moreover, factors such as compatibility with programs and algorithms, and space restrictions should be taken into consideration as well.[51]

1.5 First stage of data-driven approaches

As mentioned earlier, QSAR approaches opened the way for a branch of chemistry toward statistical treatments.

The concept of the data-led strategies in chemistry is straightforward: recognize a chemical pattern by applying statistical methods, in order to improve model performance. To do so, a target property P must be connected with one or more chemical properties p (Equation 1.2.1). This P is also denoted as a sought-after performance parameter or figure of merit (FOM), while the p refers to the chemical descriptors. Upon discovering the mathematical relationship between the FOM and the descriptor, this connection is then employed to forecast the value of the said property, using the p_i, which, in theory, is simpler to measure.

As shown with the Hammet equation, the univariate linear scaling relationship has been the algorithm for most of the analysis.[52,53] In this case, a single parameter was related to the FOM. Thirteen years before the breakthrough of the Hammet equation in 1937, Brønsted and co-workers already conducted the first quantitative analysis, the so-called Brønsted catalysis law.[54] This LFER relates the ionization constant of acids (K_a's) to the rate of reactions, catalyzed by general acids through a sensitivity factor α ($\log(k_{cat}) = \alpha \log(K_a) + C$).

The outlined studies were built with experimental data. Later on, with the emergence of computers, molecular parameters were also derived from such machines. This progress led to the development of the *chemioinformatics*. This field of chemistry offers computer approaches to extract information from chemical data, to solve chemical scenarios.[55]

In the recent applications of LFERs, volcano plots applied to homogeneous catalysis have gained attention.[56] Corminboeuf *et al.* introduced this frequently applied tool of heterogeneous and electrocatalysis in homogeneous catalysis, with the name of *molecular volcano plots*. This robust tool is based on Sabatier's principle, which claims that the catalyst's

effectiveness lies in its ability to bind a substrate with neither excessive strength nor weakness. Utilizing linear scaling relationships between a descriptor (*e.g.*, relative energy of a catalytic intermediate) and the most challenging step in the catalytic cycle, allows the prediction of the catalytic performance. More recently, Hartwig *et al.* used this tool to rationalize the effect of ligands in the fluoroalkylations of aryl halides.[57]

Nevertheless, the plainness of LFERs limits the ability to connect intricate target scenarios with a single descriptor. For instance, nonlinear effects (NLE) in asymmetric catalysis break the linearity of the LFERs.[58] In fact, most chemical problems are modelled by an ensemble of factors and not only by a linear isolate molecular property.

Thus, multiple parameters were introduced into correlation models to lead to multilinear regression (MLR) analysis.[59] Traced back to 1952, Taft and co-workers envisioned that the base-catalyzed ester hydrolysis is ruled by both steric and electronic factors (Figure 1.5).[60] This resulted in a MLR that derived Taft electronic (σ^*) and steric (E_s) parameters, ($\log(k_s/k_{CH_3}) = \rho^*\sigma^* + \delta E_s$). Recently, Sigman and Doyle screened the oxidative addition of aryl iodides by Ni(I) catalysts with MLR analysis.[33] Combining experimental rate constants with DFT-derived computational descriptors, they derived regression models that gave insights into the local and global features affecting the rate of the reaction step. Furthermore, a MLR model estimates the nucleophilicity of the C-H bond of alkenes using six topological descriptors. The model was developed with target experimental values and computational chemical descriptors.[61]

LR and MLR analysis are widely applied not only in chemistry, but also in many other areas due to their strong benefits. Firstly, the mathematical simplicity of the models enables a rapid interpretation of the chemical processes. Secondly, the rich-designed models, that provide good performance, are robust to predict FOMs. Conversely, the smooth implementation of these algorithms may fail in the characterisation of puzzling chemical scenarios. As previously discussed, molecular descriptors

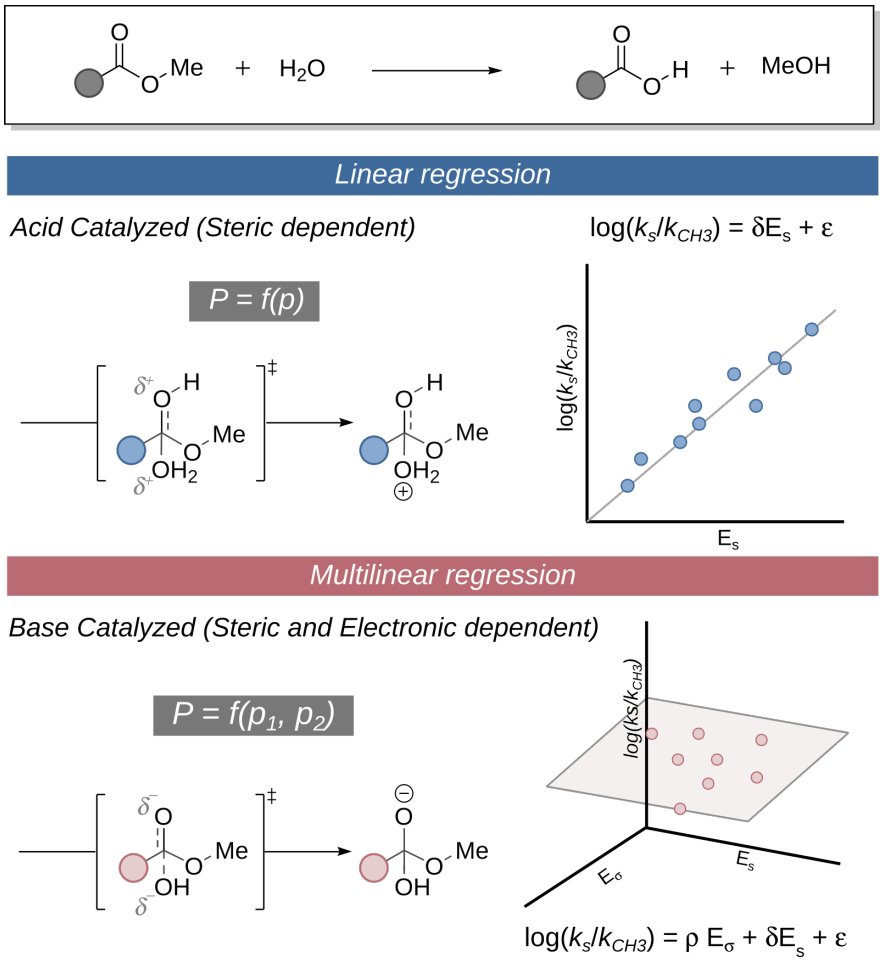


Figure 1.5: Schematic representation of the mechanisms for the ester hydrolysis under acid (top) or base catalysis (bottom). The acid-catalysed reaction is fit with linear regression and the base-catalysed reaction is explained with multilinear regression.

for these two mathematical models are often sophisticated. The challenge of elaborating these intricate chemical descriptors is amplified by the increase in the number of samples within a chemical space. Thus, it was necessary to develop new mathematical operations.

1.6 Modern computational chemistry: data-driven strategies

The amount of raw available data is closely connected to the exponential growth of the computational power. Apart from the enormous benefits of generating thousands of data, it is essential to underline that the aim of scientists is not only to generate data *per se*, but to analyze and deduce chemical insights. Therefore, it was necessary to integrate more powerful mathematical tools in standard research tasks, thus, addressing previously untractable datasets.

In the previous Section 1.5, the efficiency of the LR and MLR analysis to unravel data patterns was demonstrated. Unfortunately, these treatments may not be appropriate for large datasets. In that context, the application of modern computing within chemoinformatics or *chemometrics* has a pivotal role in dealing with the growing demands. Chemoinformatics techniques emerged during the 1960s and 1970s having a great impact in QSAR approaches.[62] To clarify, chemometrics refers to the search for chemical trends without using chemical structure, but both chemoinformatics and chemometrics are usually employed indistinctly. Later on, the term *machine learning* has overshadowed the utilization of chemoinformatics to this day.

The machine learning notation started to appear in chemical literature around 1988.[63] This term, albeit sometimes unclear, commonly refers to the area of artificial intelligence (AI) that uses computer power to automatically improve performance with the experience. Another general definition accounts for the extraction of knowledge from data by harnessing statistical computing capabilities.

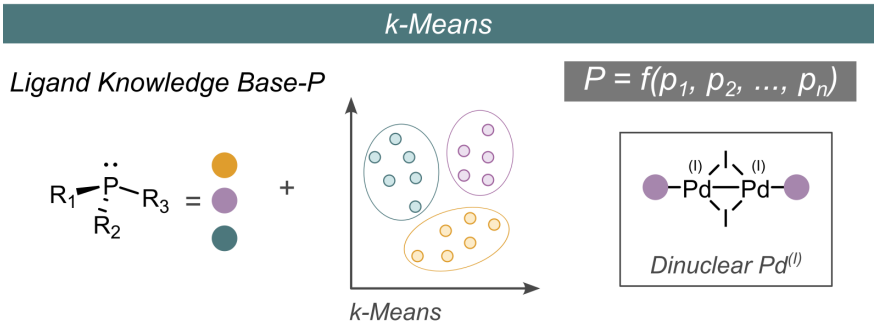


Figure 1.6: Scheme of the variables employed during the prediction of suitable monodentate phosphine ligands for building dinuclear Pd(I) complexes.[69]

The coexistence of this heuristic-based approach with chemoinformatics blurs their distinctions. Recently, it was stated that chemoinformatics encompasses LRs, whereas nonlinear techniques, together with the processing of vast datasets fall within ML.[52] However, no generally accepted conclusive definitions are available yet.

Nowadays, ML has become a rather *mainstream* term in science. Particularly in the field of chemistry, it has accelerated the areas of organic synthesis,[64] catalysis,[65] material science[66] and drug discovery,[67] among others. The myriad of sophisticated ML algorithms enables the selection of suitable architectures, based on the specific chemical problem to be addressed.

The classification of the solubility behaviour of inorganic alkali-polyoxometalates pairs in water was accomplished through a Kohonen Neural Network (KNN).[68] Furthermore, Schoenebeck and co-workers employed the LKB-P[24,25] and k-Means techniques to explore the speciation of palladium catalysts (Figure 1.6). This exploration led to the synthesis of unreported dinuclear Pd(I) species.[69]

When dealing with regression tasks, neural networks (NNs) provide a quantitative prediction of target properties. For instance, in the study of first-row transition metal multiplicities, the spin-state ordering and specific

bond lengths were predicted using NNs.[70] Herein, NN with a suitable selection of empirical inputs predicted the spin-state splittings of TM complexes, with around 3 kcal·mol^{-1} of error.

Furthermore, Principal Component Analysis (PCA) is a common approach to reduce the dimensionality of chemical space and leverage new chemical descriptors[26,71] More complex architectures like Gaussian Processes (GP)[32] and graph neural networks (GNN)[72] have lately been incorporated into the field of chemistry solving with accurate results chemical problems. Despite these advancements, further progress in ML methods remains essential to attain higher levels of precision, adaptability, and explainability.

Whilst it is crucial to select a suitable ML algorithm in any data-driven application, the choice of a critical chemical representation may be even more significant.[73] In this regard, a reported model predicting the yield of the C−N cross-coupling, using random forest (RF), fell into a pitfall.[74] It was proven that such a model does not depend on the selected chemical descriptors for that work, but on random-valued features.[75] As a consequence, an adequate representation of molecules is essential. Furthermore, in these ML strategies, there is a recurring practice based on first predicting outcomes, and then, interpreting the results, if needed. Nonetheless, this *modus operandi* is prone to provide a *good* explanation for the *wrong* reasons, or viceversa.

All in all, the success of a model relies on whether the algorithm resonates well with the working data. Thus, establishing a symbiosis between the chosen method, the properties of the data, and the hypothesis under investigation is the cornerstone of any task.

Last but not least, it is important to acknowledge that the ongoing unprecedented technological advancements −that we are currently experiencing− are the result of *Open Knowledge* philosophy. Notably, most of the algorithms enclosed in this introductory Chapter are publicly available to the scientific community. Widely recognized Python libraries such as

PyTorch,[76] scikit-learn,[77] and Numpy[78], among others, have contributed to the explosion of novel discoveries, in our specific case, in chemistry.

1.7 Aims and objectives

The main objective of this Thesis is to extract and rationalize knowledge from a vast amount of chemical data. Specifically, we will focus on discovering the forces governing the thermodynamics and kinetics of chemical phenomena using descriptors. To achieve this, DFT calculations together with data-driven strategies are employed. From this overarching goal, more specific objectives can be proposed.

- Investigation of widely applied reactions from a novel perspective.
- Design of a pipeline to generate new chemical descriptors.
- Discovering the chemical concepts governing the new chemical descriptors.
- Development of prediction models to estimate properties of chemical species in a straightforward manner.
- Application of the new chemical descriptors to elucidate diverse chemical scenarios.

This Thesis is divided into two main parts. The first part (Chapter 2) provides an explanation of the tools employed throughout this work. The fundamental concepts of reactivity in the field of computational chemistry are outlined. Next, we present a brief overview of the Quantum Mechanics (QM) methods employed to date. Furthermore, the mathematical objects and operations applied in this Thesis are described, together with their origin and applications. An comprehensive explanation about the hidden descriptor (HD) method is also included.

In the second part, three main applications using DFT and data-driven strategies are detailed. Chapter 3 presents the use of the hidden descriptors to address two distinct chemical scenarios. We apply this strategy to characterize a wide range of N-heterolytic carbene (NHC)

ligands. Additionally, we strive to identify metal fragments capable of forming stable complexes with an unusual disposition of the dihydrogen ligand (η^1-H_2). Chapter 4 delves into the study of the underlying forces of bimolecular nucleophilic substitution reactions at carbon centres. Hidden descriptors are employed to quantify and understand the behaviour of chemical fragments engaged in the reaction, and to predict energy barriers of the concerted mechanism. Finally, Chapter 5 presents the graph kernel atom–atom bond–bond bond–atom (AABBA). This tool is applied to a reported Vaska's dataset to generate molecular representations. These are subsequently used to predict H_2-oxidative addition target properties.

The findings of these Chapters are expected to significantly enhance the understanding and design of the studied chemical processes, offering a novel and alternative computational approach.

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

Chapter 2

Theoretical background and method development

In this house we obey the laws of thermodynamics!

— Homer Simpson, The Simpsons

Chemistry research entails multiple fields of specialization: organic synthesis, inorganic, analytical, biochemical and computational. This Thesis embraces and applies the latter field, which has become a prominent branch in the past decades.

The state-of-the-art in computational chemistry encompasses diverse tasks, including the prediction of molecular properties,[32] simulation of molecular trajectories,[79] or creation of databases[16] (Chapter **1**). The synergy between mathematical methods with fundamental quantum mechanical laws has unlocked a profound understanding of matter itself.[80] Nevertheless, Dirac pointed out in 1929, the remarkable challenges that quantum chemistry faced, "*The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known (...) It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which*

can lead to an explanation of the main features of complex atomic systems without too much computation."[81] Difficulties in achieving this goal can arise from either hardware limitations for executing precise tasks, or from the intricate nature of reality, that cannot be fully described by physical laws. The acknowledgement of these limitations has led to the development of various strategies in order to achieve the ultimate goal: understanding and predicting the behaviour of matter both micro- and macroscopically. This chapter briefly explains the theory and mathematical strategies employed in the Thesis.

2.1 Concepts on thermodynamics and kinetics

In the current work, we performed *electronic structure calculations* to unravel chemical problems. This type of computation uses two pillar concepts in quantum mechanics (QM): the solution of the Schrödinger equation $\widehat{H}\Psi = E\Psi$ and the Born-Oppenheimer approximation.

In this context, the wavefunction Ψ represents the positions of electrons and nuclei in a chemical system. However, in a real system, the coupling of electrons and nuclei renders the exact solution of the Schrödinger equation practically unattainable. Considering that the nuclei are heavier, they are expected to move much slower than the electrons. Therefore, nuclei can be regarded as stationary at any point along the *potential energy surface* (PES). This assumption constitutes the Born–Oppenheimer approximation. The PES provides the energy for a given configuration of atoms. This approach decouples electronic motion from nuclear motion, and makes the Schrodinger equation solvable. This is key in the understanding of chemical reactivity.

Thus, the resulting wavefunction depends on $3N - 6$ coordinates, representing three positional coordinates per each of the N atoms within the system, excluding three degrees of freedom accounting for the translation of the molecule across the space, and an additional three for rotational aspects.

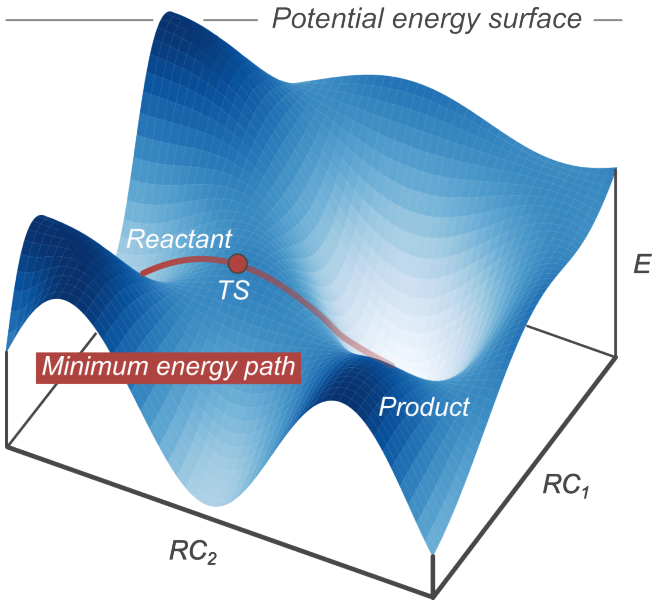


Figure 2.1: 3D potential energy surface with dark blue color corresponding to higher energies and light blue color to lowest energy.

In the case of diatomic or linear molecules, the rotation only requires two coordinates to be described. Hence, here the equation depends on $3N - 5$ coordinates to render the movement of the nuclei. These movements are the *vibrational normal coordinates*. However, even for a reduced chemical system, mapping the whole PES is impracticable due to the expansive dimensionality of the PES. A simply scenario would imply a diatomic molecule, N becomes equal to 2, and thus, the equation is dependent on $3 \cdot 3 - 5 = 1$ reaction coordinate. On the other hand, if considering a polyatomic molecule with 6 atoms such as the ethanol, $N = 6$, the equation is dependent on $3 \cdot 6 - 6 = 12$ coordinates. This highlights the exponential increase in the complexity of the equation.

Given the time-consuming task mentioned above, computational chemistry focuses on locating interesting features on the PES. Figure 2.1 illustrates a PES described by two reaction coordinates (RC), as well as the energy of each point on the PES. The wells in the PES (in light blue

color) indicate electronically stable configurations of a molecule, *minima*, in mathematical terms. The lowest potential energy point hosts the reactant and product states of the reaction. Characterizing the electronic energy of the reactants and products provides insight into the chemical transformation. Their energy difference reveals the most stable state and the expected direction of the reaction.

Chemical transformations often deal with bond formation and cleavage. In this Thesis, we show that the energetic analysis of the minima of each fragment involved in the breaking process can shed light onto the chemical phenomenon. In this regard, bond dissociation energy (BDE) is one of the key parameters for studying such reactions. BDE measures the strength of a chemical bond $A-B$, which determines, for instance, the thermodynamically favored pathway and selectivity. Its standard definition involves the enthalpy change in gas-phase when $A-B$ is broken into fragments A and B, through homolysis, typically resulting in radical species: $A-B \rightarrow A^{\cdot} + B^{\cdot}$.

In addition, minima points can be connected by pathways shaped along the PES. These indicate chemical transformations and are defined by large vibrational motion. The easiest path of this process corresponds to the lowest energy reaction path (red path in Figure 2.1). Throughout this path, the second derivative of energy relative to any dimension can take a positive, zero, or negative value, while all other second derivatives must remain positive. The first-order saddle point on this trajectory refers to the transition state (TS) and it is characterized by having a single negative second derivative, and thus, a single imaginary vibrational frequency. This spot marks the highest energy point along the reaction coordinate, as well as the minima along all other dimensions. Thus, the TS governs the chemical transformation between the two minima.

According to the Transition State Theory (TST) formulated by Erying,[82] the rate constant for most reactions can be calculated by assessing the free energy difference ($\Delta G^{\ddagger}$) between a transition state and the

preceding ground state with the lowest energy,

$$k = \kappa \frac{k_B T}{h} e^{-\Delta G^{\ddagger}/RT} \tag{2.1.1}$$

In the Equation 2.1.1 k is the rate constant, κ is the transmission coefficient, T is the temperature, k_B, h and R are the Boltzmann, Planck and gas constants, respectively.

In the ideal TST, the transmission coefficient is considered as a unit, where no recrossing occurs at the barrier. While the process of calculating the ground state electronic structure might be attainable, the quest for finding a TS supposes a great challenge. Furthermore, there are transition states with such high energies, that they cannot be accessible critical points for achieving the product of the reaction. In such case, the TS would not influence the reaction. In addition, QM calculations are modeled for gas phase situations. Progress in incorporating continuum dielectric fields to simulate solvent effects (*i.e.* implicit solvation) has increased the accuracy of PESs. However, these calculations are not straightforward when complex interactions occur between ionic intermediates or/and transition states, and solvent molecules.

Considering the significance of defining these crucial points along the PES, this Thesis focuses on identifying minima and transition states. This characterization is paramount for understanding the chemical behavior of molecules.

2.2 Overview of the electronic calculation methods

The crucial aspects of chemical reactivity stem from the systematic characterization of the PES. Methods must be set up to carry out electronic calculations and thus, shape the PES. We will briefly summarize the principal theories and methods in computational chemistry. Further

derivation of the equations of the methods can be found in more detailed elsewhere.[80,83]

As previously stated, the Schrödinger equation considers some approximations to be solved for real chemical systems. On one side, *ab initio* methods based on Ψ, establish well-defined simplifications that allow solving the Schrödinger equation. Ultimately, the total wave function, for single reference system, is regarded as the product of orbitals. As a result, the solution provides their respective energies without any empirical parameters.

The simplest wavefunction-based method is the Hartree-Fock (HF). (bottom in Figure 2.2). In HF electrons are not explicitly correlated, rather each electron experiences the repulsion of the remaining electrons as an average electronic field. Although HF is often considered insufficient for accurate predictions in chemical systems, this theory established the basis for other more sophisticated methods. The systematic improvements of HF led to the post-HF methods. These methods achieve greater accuracy than their predecessors by incorporating electron *correlation*, effectively accounting for the electron-electron interactions. Within post-HF, the most prominent approach is full configuration interaction (FCI). FCI considers the exact numerical solution of the Schrödinger equation by taking into account all feasible excited state configurations. However, this latter cornerstone method is impractical for most chemical systems due to the exponential scaling of accessible configurations. A solution to that limitation is to reduce the available states by selecting the predominant electronic configurations. Configuration interaction (CI) method uses these subsets to perform the calculations. Alternative post-HF approaches include Møller–Plesset (MP) or Coupled Cluster (CC) methods. It is necessary to highlight the Coupled Cluster technique, CCSD(T), in computational chemistry. This method is regarded as the gold-standard approach and is employed as a benchmark reference.[84] While the aforementioned post-HF methods are more cost-effective compared to FCI, they remain

computationally demanding. Hence, they are not the primary choice for practical computational homogeneous catalysis studies. Figure 2.2 left summarizes the wave-function methods, as well as their accuracy relationship.

The second strategy to solve the Schrödinger equation involves the Density Functional Theory (DFT). This strategy does not put the focus on Ψ, instead, it relies on functionals of electron density: $E = F[\rho(r)]$. Electron density is defined as the probability of finding one electron in an infinitesimal volume. The theory originated in 1964 with the Hohenberg-Kohn theorem,[85] linking the ground state energy of an electronic system to its electron density through an exact yet unknown functional. Later in 1965, Kohn and Sham developed consistent equations where the density is constructed from a set of orbitals.[86] This implementation allowed the application of DFT to complex molecules. Due to these advancements, Kohn received the Nobel Prize in Chemistry in 1998 alongside Pople.[87] An important advantage of the method is that the $\rho(r)$ is represented by three spatial coordinates rather than $3N$ as shown in the previous methods. The integral over all the space yields the number of electrons in the system. The main drawback lies in the unknown exact expression of the functional. The term denoted as the exact correlation coefficient (E_{XC}) within the functionals is elusive. It would account for the differences between the exact FCI energy and the energy of the simplest system with non-interacting electrons in the ground state. Unfortunately, the missing of the exact term inherently implies the inclusion of arbitrary errors in the DFT approach. This theory lacks systematic improvements. Consequently, inexact functionals derive the energy from the electron density.

Since its origin, DFT has spawned a wide variety of functionals. Different strategies and modifications have given rise to the description of Jacobs's ladder.[88] Figure 2.2 right illustrates this intuitive scheme where the functionals are hierarchically arranged according to their improved descriptions of the electronic structure. The ladder starts from the simplest

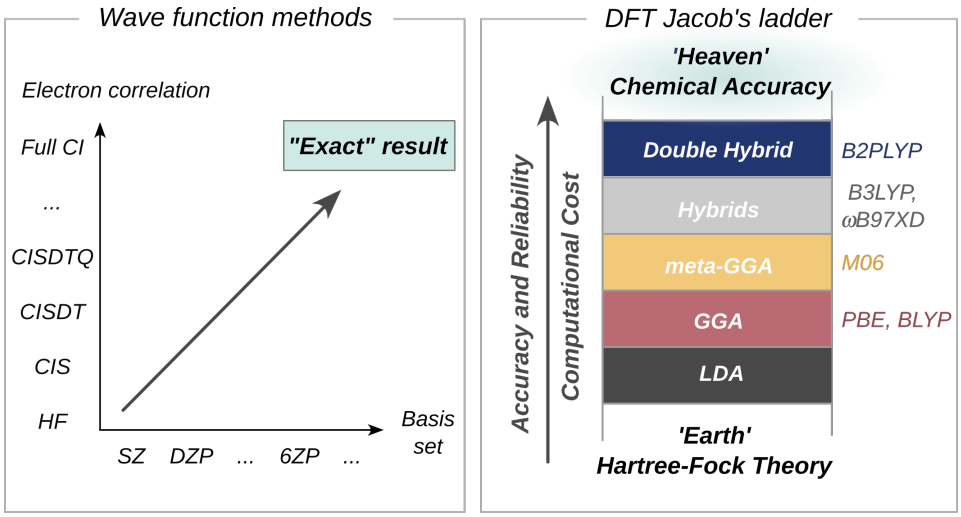


Figure 2.2: Wave-function based methods converging to the exact solution. *ZP* denotes zeta polarized basis set, *S* refers to Singles, *D* to Doubles; *CI-* refers to CI with singles *S*, doubles *D*, triples *T* (left). Jacob's Ladder where the density functional theory approaches are sorted hierarchically in function of their accuracy (right).

approximation in the realm of computational chemistry and goes up, reaching the most accurate method. The lowest level, and thus, the least accurate is local density approximation (LDA), which is based on the uniform electron gas model. Next, the Generalized Gradient Approximation functionals (GGA) consider the non-homogeneity of the electron density. This strategy can introduce empirical data (B) or quantum mechanical principles (PBE). Moreover, there are the Hybrid functionals and the Meta-GGA functionals. Other functionals, such as the hyper-GGA are *a priori* more accurate from the theoretical perspective. The list is headed by double hybrid functionals to reach the chemical accuracy.

Whilst, it holds true that Jacob's ladder offers a general scale to choose the DFT functional according to the necessities of the system, indeed, determining the exactness of DFT is a matter still under discussion. However, it is noteworthy to mention that DFT entailed a breakthrough

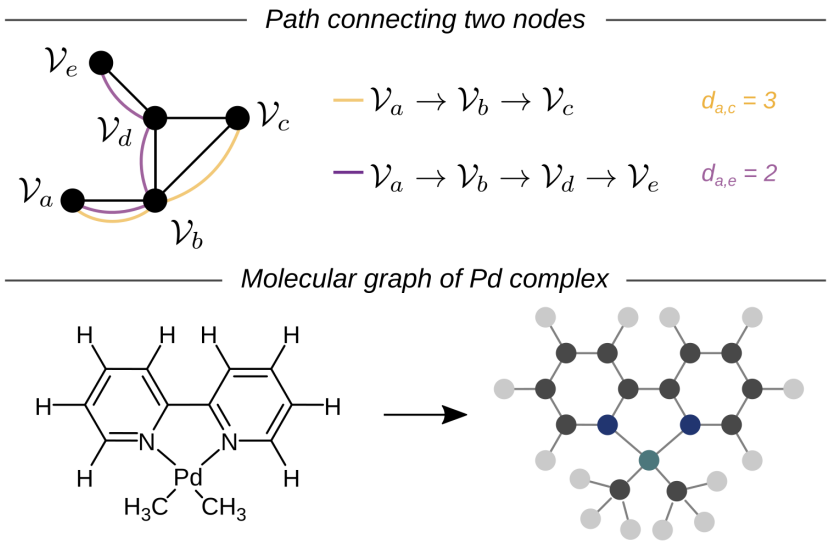


Figure 2.3: Example of vertex sequences on a given graph including two walks (top). Molecular graph representation of a palladium complex. Color-coded of the nodes: • nitrogen, • hydrogen, • carbon, and • palladium (bottom).

in chemistry. The two considered approaches, $\Psi-$ and $\rho(r)-$based theories are generally employed across computational chemistry depending on the demands of the problem. Throughout this Thesis, we have entirely focused on DFT calculations based on their optimal ratio between accuracy and computational cost.[89] Given the extensive volume of calculations involved, opting for a less time-consuming method was found to be more convenient.

2.3 Molecular graphs

Graphs, expressed as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, are mathematical objects defined as a set of vertices or nodes, $\mathcal{V}$, and a set of edges, $\mathcal{E}$, which connect two nodes of the graph.[90] These entities are commonly represented by drawings where the nodes are dots, and the lines linking the points are the edges (Figure 2.3).

Moreover, this mesh of points and lines can be *traversed* by different

paths. It is possible to read a graph jumping sequentially from one node to another. Figure 2.3 shows a *walk* going from node a to node c passing through b. Within this path, it is observed that nodes a and b are separated by one edge, thus, the distance between nodes a and b is $d_{a,b} = 1$. However, in the path connecting nodes a and e, this definition is not obvious, since there are two accessible routes either $\mathcal{V}_a \rightarrow \mathcal{V}_b \rightarrow \mathcal{V}_c \rightarrow \mathcal{V}_d \rightarrow \mathcal{V}_e$ or $\mathcal{V}_a \rightarrow \mathcal{V}_b \rightarrow \mathcal{V}_d \rightarrow \mathcal{V}_e$. The rule of thumb for determining the distance between two nodes is to search for the shortest possible walk between them. Therefore, in the previous example the $d_{a,e} = 3$.

These concepts are successfully transferred to the chemical domain establishing the *molecular graphs*. In this context, atoms are considered nodes, and chemical bonds are converted to edges (Figure 2.3). Indeed, the graph drawing resembles a molecular structure. The advantage of this representation is its geometry-agnostic approach, since it lacks spatial coordinates, facilitating its computation.

The most crucial part in the definition of molecular graphs are the edges, that rule the connectivity. These topological descriptors need to convert chemical bonds to edges in a satisfactory manner. In organic molecules, this is straightforward since the valence or degree of connectivity of atoms is well-defined, but this task is more challenging in transition metal complexes where the valences are fuzzier. To address this aspect suitable criteria have been adopted. For instance, the use of Bader's Quantum Theory of Atoms in Molecules (QTAIM) has settled the connectivity in polyoxometalate-based graphs.[91] Furthermore, electronic patterns relied on NBO analysis provided a proper connectivity definition of TMCs in the diverse tmQMg database.[23]

In addition, graphs can encode more information beyond to the topological structure. These supplementary features of the molecular graphs are called *attributes*. Each element of the graph, and the entire graph itself, can have associated attributes. By assigning attributes, such as atomic number or atom size, the graph assumes a chemical-looking entity.

2.4 Machine learning classification

In the first Chapter, the timeline of data-analysis techniques displays the potential of ML in chemistry. ML models are classified according to the type of problem they solve. Examples include *supervised*, *unsupervised* and *reinforcement learning*.

Supervised learning models map $\widehat{f} = \mathcal{X} \overset{ML}{\rightarrow} \mathcal{Y}$ to connect a set of input ($\mathcal{X}$) and outputs ($\mathcal{Y}$), data. These models can be used for either regression tasks, where the output is predicted in a continuous range; or classification, where the goal is to predict a class. The model is trained using pairs of input/output data during the training set. Then, the ultimate objective is to predict never-before-seen data by the model.

The second set of ML algorithms involves unsupervised learning. Here, the input is known, but not the output. Thus, the aim is to extract knowledge from $\mathcal{X}$ without previous instructions. Consequently, it is uncertain whether the output generated is correct or not. Among the models, we find regression and clustering algorithms. A widespread application of this type of ML model is dimensionality reduction for visualization purposes and chemical descriptor generation.[26,92] This reduction retains the meaningful properties of the dataset with fewer degrees of freedom. Additionally, the detection of outliers is another commonly exploited use of this model family.

The last group of machine learning models regards reinforcement learning (RL) tasks. This class of methods combines aspects of both unsupervised and supervised learning. During the process, the input/output pairs are not labelled. Therefore, the RL agent perceives and interprets the data environment and it learns by receiving feedback in the form of punishments and rewards.

The selection of ML models depends firstly on the purpose of the chemical problem, and secondly on the available data. Likewise, choosing the appropriate method can be challenging. In such instances,

programs like *ROBERT* can be valuable.[93] The code provides a framework for automatically performing certain ML models and analyzing their performances systematically.

2.5 Singular value decomposition

Processing, analyzing, and effectively utilizing the vast amount of chemical data generated is a challenge. As discussed in the Introduction, data-driven approaches are cutting-edge strategies that rely on mathematical techniques and databases. Throughout the current Thesis, one of our focus has been placed on the application of singular value decomposition (SVD).

SVD is an algorithm classified as an unsupervised ML model. As explained earlier, this type of methods only relies on the input data $\mathcal{X}$. Widely utilized for reducing data dimensionality, SVD hierarchically organizes the data to maintain significant information, while eliminating noise. Hence, when the objective is to extract knowledge from a dataset, the SVD approach is appropriate. This mathematical tool finds application and demonstrates its impact across various fields (*vida supra*).

Before delving into the mathematical details of the operation, a representative example of this method is presented. When observing a figure, we often overlook fine-grained shapes, but retain the global concept. Images are constructed using pixels, each representing numerical data. Thus, each figure constitutes a dataset. By subjecting a figure to SVD, we effectively reduce the file size while preserving the general shape of the picture. Figure 2.4 provides an example of this application using an image of León's cathedral facade. *Prior* applying SVD, we can specify the amount of information to withhold. Figure 2.4 shows images that vary the retained weight. The image containing 17 % of the original picture is unfeasible to analyze. Increasing the information weight to 32 %, evaluation is cumbersome but its content can still be inferred. At 54 % of the original information, the photo, while somewhat blurred, closely

resembles the original. This proof-of-concept highlights the property of the SVD in preserving the most important information of the data. We did not need a high-quality photo to recognize León's cathedral.

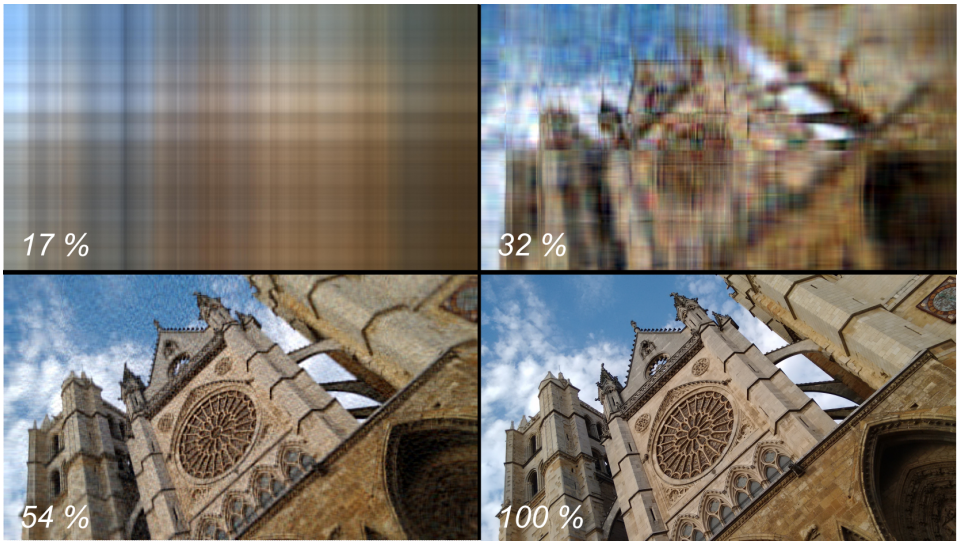


Figure 2.4: Representation of the singular value decomposition for dimensionality reduction in an image of León's cathedral facade, at various levels of truncation.

Upon providing a general idea of the method, we proceed with its mathematical foundation and its chemical application. In linear algebra, singular value decomposition (SVD) is a powerful technique that breaks down a real or complex matrix into three matrices.[94] Given a real $m \times n$ input matrix $\boldsymbol{A}$, upon application of SVD, the following decomposition is provided,

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}} \tag{2.5.1}$$

where $\boldsymbol{U}$ $m \times m$ and $\boldsymbol{V}^{\boldsymbol{T}}$ $n \times n$ are unitary matrix, thus, $\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^{\boldsymbol{T}} = \boldsymbol{I}$ and $\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^{\boldsymbol{T}} = \boldsymbol{I}$, and $\boldsymbol{\Sigma}$ $k \times k$ is a diagonal matrix with nonnegative elements, sorted from largest to smallest (see Figure 2.5). $\boldsymbol{V}^{\boldsymbol{T}}$ and $\boldsymbol{U}^{\boldsymbol{T}}$ are the transpose matrix of $\boldsymbol{V}$ and $\boldsymbol{U}$, respectively.

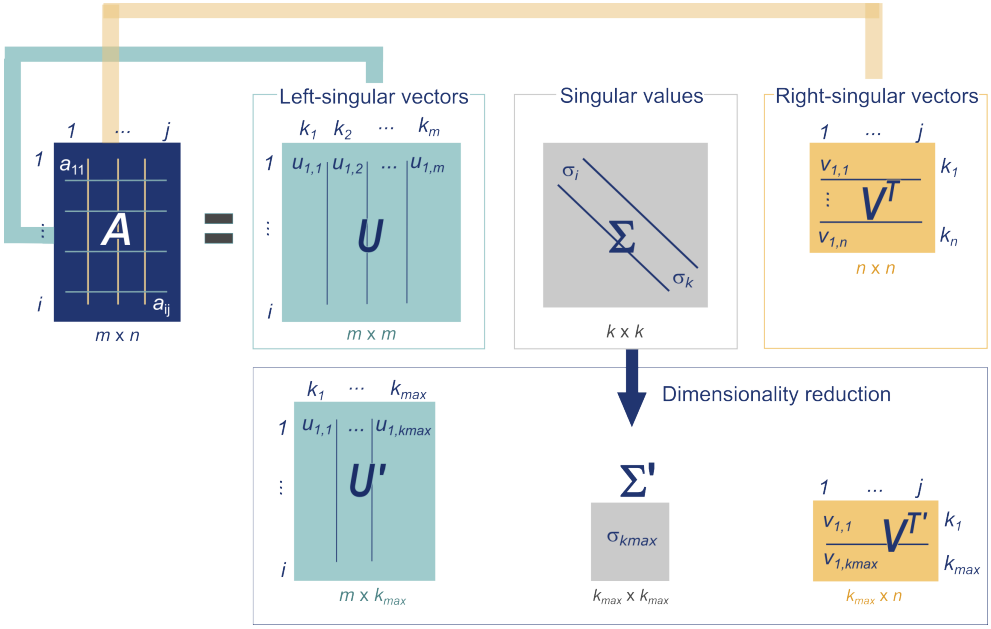


Figure 2.5: Schematic representation of singular value decomposition over a matrix $\boldsymbol{A}$ to obtain three matrices $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V^T}$.

Columns of the $\boldsymbol{U}$ are the left singular vectors ($\boldsymbol{u}$) and columns of $\boldsymbol{V}$ are the right singular vectors ($\boldsymbol{v}$), while the diagonal elements of $\boldsymbol{\Sigma}$ (σ_i) are the singular values of the matrix $\boldsymbol{A}$. (Figure 2.5). Therefore, we can expressed the above equation 2.5.1 as bellow,

$$\boldsymbol{A} = \boldsymbol{u_1}\boldsymbol{\sigma_1}\boldsymbol{v_1^T} + \boldsymbol{u_2}\boldsymbol{\sigma_2}\boldsymbol{v_2^T} + ... + \boldsymbol{u_k}\boldsymbol{\sigma_k}\boldsymbol{v_k^T} \tag{2.5.2}$$

The number of nonzero singular values ($N_{\sigma_i \neq 0}$) is the rank of the $\boldsymbol{A}$, and the σ_i values are sorted in a hierarchical order ($\sigma_1 > \sigma_2 > ... > \sigma_k$). This confers the capability to highlight the most significant σ_1 to the least important σ_k. Moreover, the rows of $\boldsymbol{V^T}$ are the eigenvectors of $\boldsymbol{A^T A}$ and, the columns of $\boldsymbol{U}$ are the eigenvectors of $\boldsymbol{AA^T}$. In both matrices either $\boldsymbol{A^T A}$ or $\boldsymbol{AA^T}$ their corresponding eigenvalues are given by $\boldsymbol{\Sigma^2}$. This property provides relationships with other mathematical factorization (*vide infra*).

The design of this decomposition operation is to apply a linear transformation of the data, as both $\boldsymbol{U}$ and $\boldsymbol{V^T}$ can be chosen to be rotations/reflections of the space. This transformation enables the original points to be projected onto a new basis, where the primary vector directions align with the directions of maximum variance. Hence, directions with low variance correspond to smaller axes. This characteristic creates a hierarchical structure within the procedure. Through this process, the SVD allows a data compression, extracting valuable insights into the properties of the original matrix $\boldsymbol{A}$, increasing the understanding of the data and preserving the maximum amount of information. This notable feature, known as *truncation* or *dimensionality reduction* endows the method with a powerful tool for obtaining crucial information, *i.e.* the primary vectors, while eliminating less informative variables. This treatment, can be expressed as the following equation:

$$\hat{\boldsymbol{A}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V^T}} \tag{2.5.3}$$

In the context of this Thesis, this procedure effectively captures the fundamental aspects of chemical problem at hand. Figure 2.5 illustrates the dimensionality reduction procedure after the decomposition of matrix $\boldsymbol{A}$. The determination of the truncation, represented by the number of representative variables $\boldsymbol{k_{max}}$, does not have a fixed ruled. Authors usually employ a threshold to collect information that explains more than 95% of the data variance, but this decision is subjective.[71]

Furthermore, SVD provides the left and right singular vectors of a matrix, denoted as $\boldsymbol{u_i}$, $\boldsymbol{v_j}$, respectively. This property becomes particularly interesting since our objective is to apply SVD to a matrix $\boldsymbol{A}$ where both rows (represented as i in Figure 2.5) and columns (indicated as j in Figure 2.5) encompass data variables of equal significance. Figure 2.5 displays this characteristic for further elucidating the concept.

Nowadays, the application of SVD is carried out using different computational tools. Throughout this thesis, Numpy[95] Python library was the tool employed with the below snippet code:

```
import numpy as np
U, S, VT = np.linalg.svd(A)
```

However, it is crucial to outline the steps involved in the operation to further understand the method. Given a dataset represented by an $m \times n$ matrix $\boldsymbol{A}$, the SVD includes the following operations:

1. Compute a square matrix by multiplying the transpose of $\boldsymbol{A}$ by itself, denoted as $\boldsymbol{A^T A}$

2. Find the eigenvalues, represented as $\boldsymbol{\lambda}$, of $\boldsymbol{A^T A}$ by solving the equation $det(\boldsymbol{A^T A - \lambda I}) = 0$, where $\boldsymbol{I}$ is the identity matrix.

3. The singular values of $\boldsymbol{A}$, expressed as $\boldsymbol{\sigma_i}$, are the square roots of the eigenvalues of $\boldsymbol{A^T A}$.

$$\boldsymbol{\sigma_i} = \sqrt{\boldsymbol{\lambda_i}} \tag{2.5.4}$$

4. Normalize the eigenvectors of $\boldsymbol{A^T A}$ which are the right singular vectors of $\boldsymbol{A}$, $\boldsymbol{v_j}$.

5. Compute the left singular vectors, $\boldsymbol{u}$, using equation 2.5.5, and normalize them,

$$\boldsymbol{AV\Sigma^T = U} \tag{2.5.5}$$

   where $\boldsymbol{\Sigma^T}$ is equivalent to $\boldsymbol{\Sigma}$ since $\boldsymbol{\Sigma}$ is a diagonal matrix.

6. Build the matrices by combining their respective singular vectors and singular values to represent $\boldsymbol{A}$ as $\boldsymbol{U\Sigma V^T}$.

Nowadays, efficient algorithms exist to calculate SVD of $\boldsymbol{A}$ without having to form the matrix $\boldsymbol{A^T A}$.

2.5.1 SVD *vs* PCA

The application of dimensionality reduction practices in chemistry such as t-Stochastic Neighbor Embedding (t-SNE), and Principal Component Analysis (PCA) is a general practice in data-led strategies.[26,69,96]

In the frame of the linearly transformations, there is a common misunderstanding that assumes that SVD and PCA are referred to the same procedure. Figure 2.6 depicts and schematic representation of both methods. In order to shed light on the subject, an explanation of this PCA practice will help to understand the similarities and differences between these methods.

Starting from the broad overview of these two concepts, while SVD is a linear algebraic operation, PCA is a technique that employs algebraic operations such as eigenvalue decomposition or SVD, to interpret the data. Secondly, it is true that PCA can be performed by carrying out SVD over the covariance matrix $\boldsymbol{C}$ of a specific matrix, i.e. $\boldsymbol{C = A^T A}$, however, the outcomes of the practices are not the same. Aiming to explain from their mathematical basis their differences, we briefly comment the steps to perform PCA:

1. First, to center the data of the matrix, $\boldsymbol{A}$, by calculating its mean by rows or columns u_i and subtracting it.

$$\mu_j = \frac{1}{n_i} \sum_{1}^{i} a_{ij} \tag{2.5.6}$$

2. Obtain the covariance matrix of $\boldsymbol{A}$, denoted as $\boldsymbol{C}$, by performing $\boldsymbol{C} = \boldsymbol{A^T A}$.

3. Retrieve the eigenvalues, represented as $\boldsymbol{\lambda}$, and eigenvectors, which are referred as *loading vectors* in PCA, from matrix $\boldsymbol{C}$. This operation can be achieved by diagonalizing matrix $\boldsymbol{C}$ or, alternatively, by conducting SVD of matrix $\boldsymbol{C}$. The vectors are also normalized.

   - By diagonalization of the matrix $\boldsymbol{C}$, the result is,

$$\boldsymbol{D} = \boldsymbol{V C V^T} \tag{2.5.7}$$

     where $\boldsymbol{D}$ is the diagonal matrix that includes the eigenvalues, and $\boldsymbol{V}$ is the matrix that contains the eigenvectors.

   - By SVD:

$$\begin{aligned} \boldsymbol{A^T A} &= \boldsymbol{(U \Sigma V^T)^T U \Sigma V^T} \\ &= \boldsymbol{V \Sigma^T U^T U \Sigma V^T} \\ &= \boldsymbol{V \Sigma^2 V^T} \end{aligned} \tag{2.5.8}$$

   Therefore in both operations, we get the eigenvalues included in either $\boldsymbol{D}$ or $\boldsymbol{\Sigma^2}$ matrix, and the eigenvectors which are the columns, the vectors of the matrix $\boldsymbol{V}$.

4. Project the data of $\boldsymbol{A}$ onto the new basis vectors of $\boldsymbol{V}$, to get a new matrix $\boldsymbol{T}$ which is the *score matrix* (purple matrix in Figure 2.6).

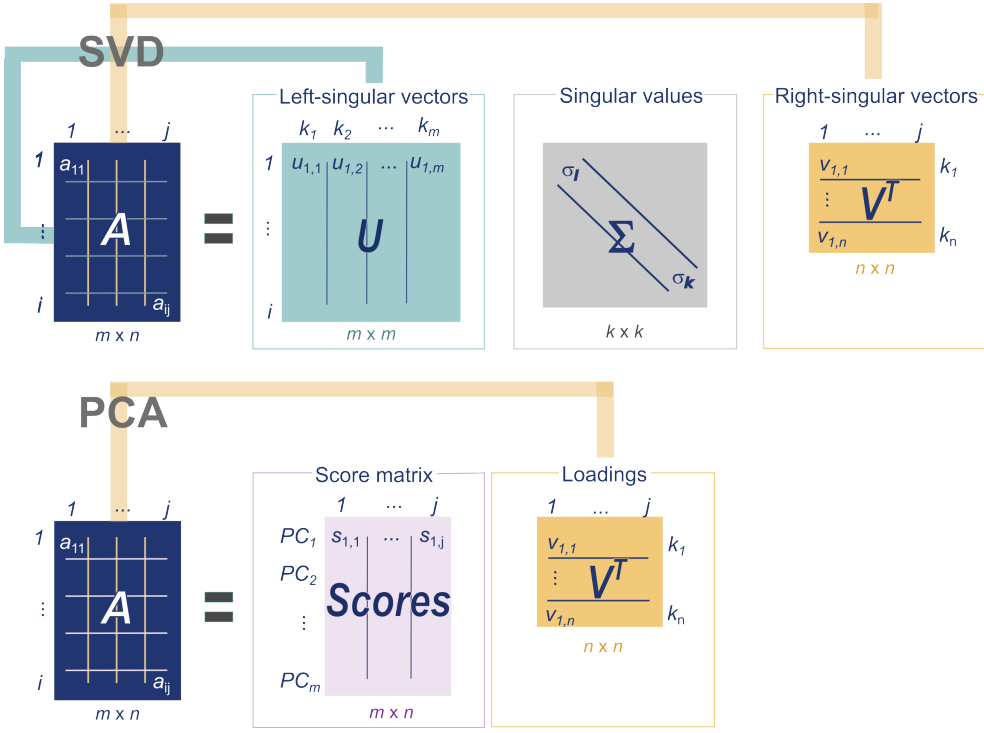


Figure 2.6: Schematic representation of singular value decomposition over a matrix $\boldsymbol{A}$ to obtain three matrices $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V^T}$ (top) and representation of PCA application over a matrix $\boldsymbol{A}$ to derive two matrices $\boldsymbol{T}$ and $\boldsymbol{V^T}$ (bottom).

$$\boldsymbol{T} = \boldsymbol{AV} \tag{2.5.9}$$

The outlined derivation demonstrates that the loading vectors of PCA and the left singular vectors of SVD can be identical, provided that the data is centered *prior* the application of SVD. The key distinction for applying SVD, instead of PCA, lies in the origin of the information they generate. Figure 2.6 reflects their differences. SVD yields three matrices, $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V^T}$, which primarily capture information regarding the rows, weights, and columns of matrix $\boldsymbol{A}$, respectively. In contrast, PCA results in two matrices, $\boldsymbol{T}$ and $\boldsymbol{V^T}$, mainly deriving from the columns of matrix $\boldsymbol{A}$.

Understanding this property in the context of this Thesis is crucial. The objective throughout this compilation is to investigate chemical phenomena that involve two distinct variables. For instance, the study of the metal–ligand bond (M–L) regards two moieties that hold equal significance. If this bond is featured through its energy, (e_{ij}), where i is the metal fragment and j is the ligand, then, it is possible to create a matrix $\boldsymbol{E}$ introducing other values (e_{ij}) per each partner of ith metal fragment (rows) and jth ligand (columns). By expanding this matrix and applying SVD, it results in three matrices: $\boldsymbol{U}$ which contains information about the metal fragment (rows), $\boldsymbol{V^T}$ which refers to the ligands (columns), and $\boldsymbol{\Sigma}$ which defines the weight of the singular values. This approach allows for simultaneous analysis of both the metal fragments and ligands. In contrast, if PCA is the method selected, it will provide information exclusively about the ligands (columns) either with the $\boldsymbol{T}$ or $\boldsymbol{V^T}$ matrices.

2.5.2 Applications of SVD

SVD is a powerful technique widely used today in a lot of models. Applications of singular value decomposition are manifold, spanning across various disciplines. This section provides an overview exploration of its diverse implementations.

As introduced with Figure 2.4, SVD can perform image compression. Since images are an ensemble of numbers, we can treat them as 3D matrices. Thus, applying SVD over them is almost trivial. The reduction in the number of singular vectors helps in limited image storage, while preserving the overall image quality.

The following application is required in our chemical study. Dimensionality reduction and data comprehension are key to extract knowledge from matrices containing wide-ranging accumulated data. The selection of the necessary k singular values allows to retain crucial information for the data, while reducing redundant or noisy information. The effectiveness of this approach has been demonstrated

in various contexts, such as studying interactions in protein complexes[97], decomposition of the matrices simulating wave function of entangled systems,[98], or analysis of orbital interactions in different chemical systems.[99,100]

SVD is widely applied in signal-processing tasks. In this realm, noisy signals are inherent problems in spectra. Here, SVD aids in the reduction of such noise, facilitating the analysis and comprehension of spectra.[101,102]

Furthermore, latent semantic analysis (LSA) is a technique of natural language processing (NLP), that identifies the main topics of a text document and measures the similarity of documents containing comparable concepts. This text analytics implies to map the word frequency of the documents to build a concept-frequency matrix. Then, SVD is applied over the matrix to give insights into the documents through the singular value and singular vectors derived.

A recent notable application of SVD comes from its integration within social networks. Recommendation algorithms relying on user-generated content, are widely used in internet apps, and some of them are grounded on SVD. SVD effectively captures pivotal features of the provided data, enabling the classification of users based on attributes such as their preferences, age, gender, or any other valuable characteristics that marketing strategies seek to spot.[103]

The broad range of applications provides a solid basis for justifying the utilization of SVD across various domains. In our study, we apply it to address computational chemical challenges.

2.6 Hidden Descriptor method

The careful selection of chemical descriptors is still a subject under study. The fundamental premise is that by thoroughly exploring a diverse range of descriptor sets, one or a combination of them will accurately reproduce the target property. However, the inherent complexity of this topic stems

from two main reasons. First, the challenge of finding relationships between descriptors and a target property. Subsequently, if the correlation is achieved, distinguishing whether the relationship is due to causation or mere chance implies another layer of complexity. The second reason lies in the inclusion of irrelevant or skewed variables that may affect the rationalization of the chemical transformation, as well as the predictive and generalization ability of the model (top Figure 2.7).

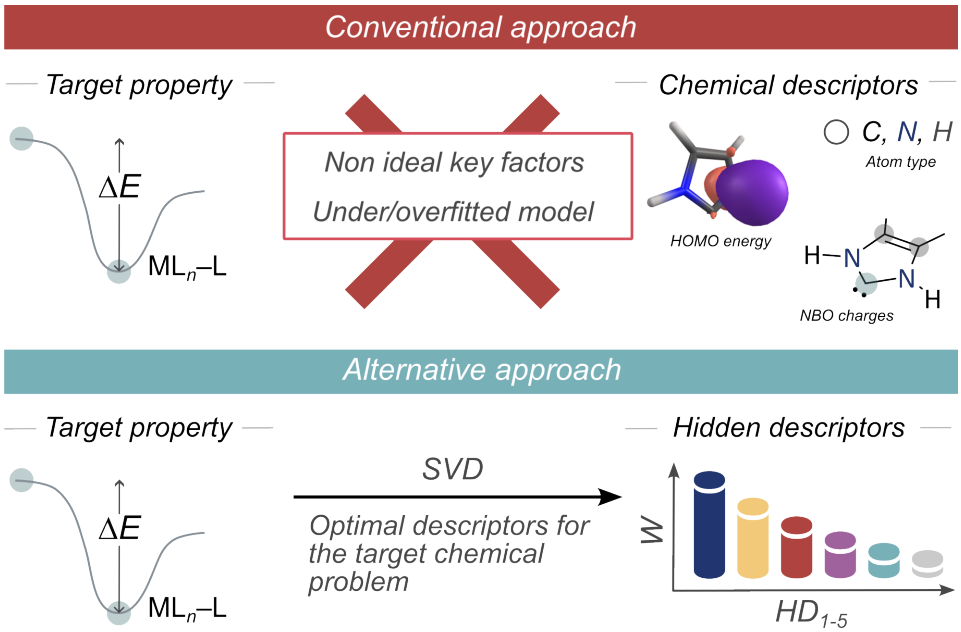


Figure 2.7: Conventional and alternative approaches in the use of chemical descriptors.

Concerning these potential pitfalls in the use of predefined descriptors, our group proposed a shift of paradigm in both the development and use of chemical descriptors. The method entails the application of statistical techniques of a relatively large number of DFT results to identify the mathematical optimal descriptors. The so-called *hidden descriptor* (HD) method was developed as a tool to find the optimal descriptors for a specific chemical scenario (lower part of Figure 2.7). The chemical problem must

be rationalized by a designated target property. This property will be computed to build the dataset, ensuring coverage across a diverse chemical space.

This method was first published in 2018.[104] In its first application, it was studied the interactions between metal fragment and ligand moieties through the metal–ligand (M–L) bond. Optimal descriptors were found that accurately described the electronic interactions between these two moieties utilizing the HD method. To do so, the authors selected an array of metal fragments and ligands and computed all the possible combinations of bond dissociation energies (BDE). The selection of BDE as a target property was based on its capacity to evaluate the stability and reactivity of the moieties involved in the bond. The resulting BDE values were introduced in a matrix, subjected to SVD. Upon an analytical study, it was assigned five hidden descriptors for describing metal and ligand electronic properties, HD_M and HD_L, respectively. *A priori*, Lakuntza *et al.* did not know if HD_M and HD_L were associated with well-known properties or if they were just simply numerical scales. This is the explanation of the *hidden* name, since firstly HD were unknown. Further examination led to the conclusion that each of the five HDs for metal fragment and ligand were mostly related to fundamental chemical concepts. Moreover, an extension of HD for metals and ligands out of the scope of the initial chemical space is possible with a prediction tool. This strategy allows obtaining new HDs without performing SVD anymore.

So far, we have detailed the initial study on the subject. This led us to think that the same methodology can be extrapolated to different chemical scenarios. We can capture the important features of chemical process, *i.e.* HD, and use the derived chemical descriptors to distinguish the behaviour of the compounds considered in the dataset. Moreover, we can predict HD and use them to predict the target property in a reverse mode. In that vein, we outlined the protocol for obtaining the hidden descriptors for any chemical phenomenon. Figure 2.8 illustrates the following steps:

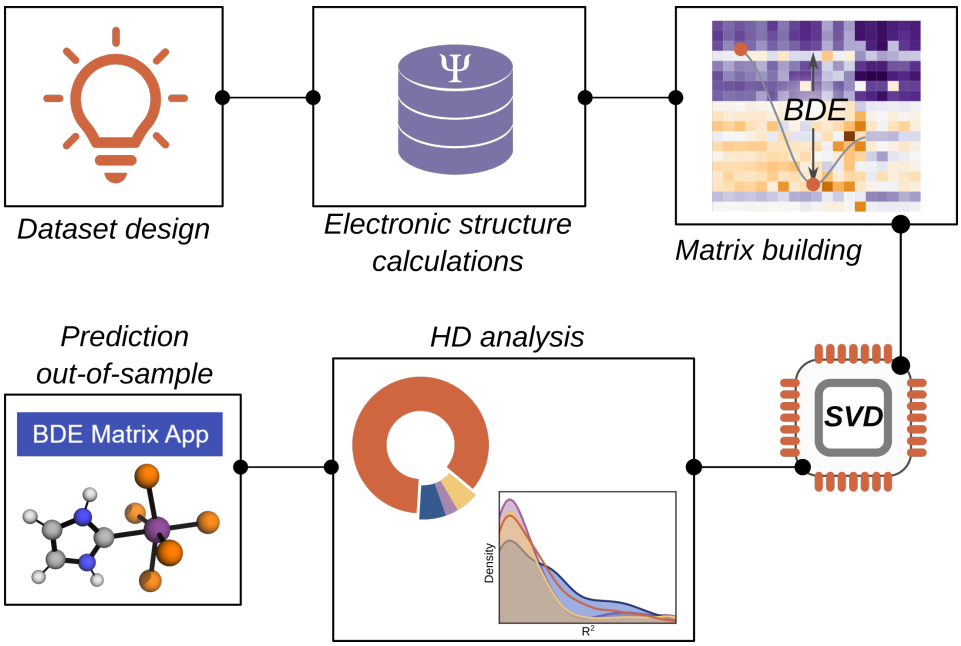


Figure 2.8: Workflow of the procedure for obtaining hidden descriptors and using them to predict the target properties.

(i) defining the desired property and the chemical space where computing such property; (ii) conducting electronic structure calculations within the dataset; (iii) construction of the matrix that contains the target feature for each pair of chemical variables *i.e.* in the previous example, M and L are variables that constitute the target property, the bond dissociation energy; (iv) upon data curation, execution of SVD Python code; (v) afterwards, an in-depth qualitative and quantitative evaluation of the obtained hidden descriptors is accomplished; (vi) prediction of HD for unsampled chemical species and target properties.

In the previous Section 2.5, a detailed explanation of the singular value decomposition operation was present. We concluded that a matrix $\boldsymbol{A}$ is decomposed in three matrixes, $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V}$. In the context of this Thesis, these matrices have chemical meaning. $\boldsymbol{A}$ is a chemical matrix built with the target property a. Each target value is defined according to two chemical variables i and j relevant to the event, a_{ij}. Thus i property is the row-variable and j is the column-factor. Matrix $\boldsymbol{U}$ is linked to the row-variable i of the matrix $\boldsymbol{A}$, and matrix $\boldsymbol{V}$ is connected to the column-variable j. The $\boldsymbol{\Sigma}$ matrix contains the chemical weights $\boldsymbol{W}$ of the vectors of the matrices, thus, the weights of the hidden descriptors. Placing this within the context of the M–L bond study, the variable i refers to the metal fragments. Each entry ith in the rows corresponds to a distinct metal fragment. On the other hand, the column-variable is designated to the ligands, and here, each entry jth indicates a different ligand. All gaps of the matrix are filled with each possible combination of M and L fragments.

To illustrate this idea, Figure 2.9 shows the three matrices, with chemical significance. $\boldsymbol{U}$ equals to $\boldsymbol{M}$, $\boldsymbol{V}$ is $\boldsymbol{L}$, as well as, Σ matrix that is represented by the $\boldsymbol{W}$ which quantifies the relative chemical importance of the vectors in these matrices. Upon decomposition, *truncation* of the matrices is conducted (lower part of Figure 2.9). The reduced matrices contain, thus, less number of vectors k_{max}. In the Figure 2.9, matrix $\boldsymbol{M}$ is pruned to $\boldsymbol{HD}_M$ with a dimension of $i \times k_{max}$, and matrix $\boldsymbol{L}^T$ to $\boldsymbol{HD}_L$ with

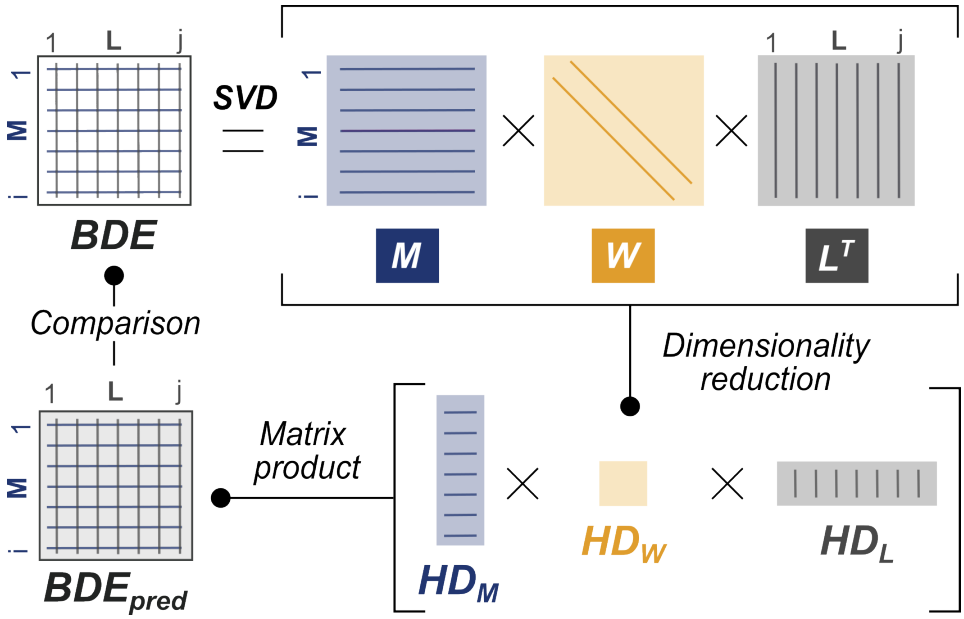


Figure 2.9: Schematic representation of HD procedure for the metal–ligand bond.

dimensions $k_{max} \times j$; the diagonal matrix $\boldsymbol{W}$ is truncated to $\boldsymbol{HD}_W$ with the size $k_{max} \times k_{max}$. These condensed matrices hold the hidden descriptors vectors, which will provide the relevant interaction of the M–L bond.

2.7 Neural networks

Within the vast amount of ML architecture, one of the most prominent algorithms is the feed-forward neural networks. This supervised ML model is also regarded as a class of *deep learning*. Particularly, in this Thesis, we focused on the application of a relatively simple method, namely *multilayer perceptron* (MLP). This strategy undergoes classification and regression tasks, and herein, we are going to apply the latter application. We will briefly introduced the basic concepts of MLP.

As earlier mentioned, the supervised models map a set of inputs $\mathcal{X}$ and outputs $\mathcal{Y}$ to provide predictions of subsequent input, $\mathcal{X}_{test}$, based on the

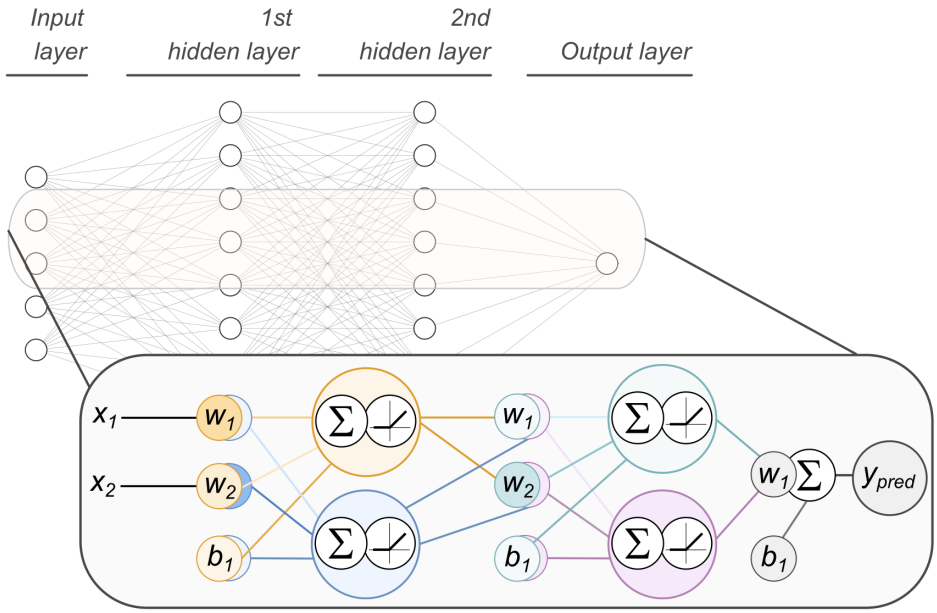


Figure 2.10: Schematic representation of the MLP architecture in the forward pass with two hidden layers, where x denotes the input elements, w the weights, b the biases, y_{pred} the predicted value, $\sum$ means the linear summation and the linear draw shows the activation function.

learned patterns. The learning process takes place due to the transmission of information along the *hidden layers* that connects the input and output layers (Figure 2.10).

These intermediates layers are constituted by neurons, and each of them transforms the values from the previous layer (x_i) with a weighted linear summation (Equation 2.7.1), followed by a non-linear activation function.

$$z = \sum_{n}^{i=1}(x_i \times w_i) + b \tag{2.7.1}$$

The non-linear functions can adopt different shapes, being the most popular the rectified linear unit (ReLU) and the *tangens hyperbolicus* (tanh). The weights are set randomly before the training starts. Thus, the weight

initialization can lead to slightly different models' performance. To retain the same w values, it is possible to fix the randomness of the seed. The output layer receives the values from the last hidden layer and transforms them into output values. Once the output is generated, the selected *loss function*, often the mean-squared error (MSE), computes the differences between the outcome and the truth value. This unidirectional process is the *forward* task. Upon the estimation of the output data, the results are evaluated and the weights are updated via the *backpropagation* strategy, namely the *backward* pass.

This reverse path is based on the minimization of the loss function. To do so, the neural network model's parameters, *i.e.* weights and bias, are fine-tuned repeatedly until achieve a plateau of accuracy. The level of adjustement is determined by the gradients of the loss function with respect to those parameters and the learning rate (lr), with the following equation:

$$w = w - lr\frac{\partial L}{\partial w} \tag{2.7.2}$$

This cycle is repeated in a loop, where each iteration is named as *epoch*. Therefore, the number of repetitions is determined by the number of epochs, and all the data is read in each epoch.

Furthermore, there are more important parameters that need to be adjusted before the training. The whole collection denotes the *hyperparameters*. Within it, there are the number of nodes in the hidden layer (n), the number of layers (l) in the architecture, and the size of the batches if needed, among others. The batches separate the training dataset in groups that facilitate the gradient computation, and the subsequent update of the models' parameters. Apart from that, it is important to consider the number of parameters in the model, *i.e.* weights and biases, to avoid some anomalies such as *underfitting* – scarce number of parameteres, or *overfitting* – exceed the number of variables. The equation to compute

it:

$$n_1 n_2 + n_2 n_3 + ... + n_{l-1} n_l + (n_1 + n_2 + ... + n_l) \tag{2.7.3}$$

For instance, an architecture of 2 hidden layers, each with $n = 128$, and with a input size of 33 will contain $33 \cdot 128 + 128 \cdot 128 + 128 \cdot 1 + (128 + 128) = 20,992$ parameters to be adjusted. Moreover, the homogenization of the data is critical in any statistical model, thus standarization techniques are set as fundamental pretraining step. The algorithm and tutorials are freely available to be used. It is worth mentioning the PyTorch,[76] or Keras[105] libraries.

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

Chapter 3

Metal-ligand interaction

Second comes right after first.

— Buzz Aldrin – The Simpson

3.1 Introduction

Bond dissociation energy is a measure that holds significant importance as it offers insights into the strength of chemical bonds. By quantifying the energy interaction of the two moieties involved in a bond, BDE provides information about their relative stability and reactivity. This parameter is particularly useful in the frame of transition metal (TM) chemistry. The reactivity of TM complexes undergoes ligand exchange processes, with thermodynamics ruled by differences in BDE.

In the Methodology Chapter, the origin of the HD strategy was introduced. Former members of the group designed the HD methodology using the BDE quantity. In that study, BDE was considered in its heterolytic form, *i.e.* where no radical species are formed, between a metal fragment (M or ML_n) and a ligand (L).[104] This transformation is depicted as $(M–L)^x \rightarrow M^{(x-y)} + L^y$. Therefore, BDE target property was regarded as

dependent on two variables: metal fragment i and ligand j. This definition allowed the construction of a $\boldsymbol{BDE}$ matrix.

According the procedure outlined in Methodology Chapter, the HD analysis elucidated the fundamental binding forces. These results yielded vectors of five hidden descriptors which describe each metal fragment, HD_{Mk}, and each ligand, HD_{Lk}, of a particular bond in water, where k denotes the number of the hidden descriptors. Furthermore, the prediction of the BDE property was attained using the estimated HD values, within an average absolute error of 1.3 kcal·mol^{-1} and a maximum error of 6.7 kcal·mol^{-1}. The first four hidden descriptors happen to be related to conventional chemical concepts in the field of metal–ligand bond. The first hidden descriptor, HD_1, is associated with the σ donation, the second HD, HD_2, is related to the π interactions, the third, HD_3, is identified with the *cis*-influence; and the fourth, HD_4, correlates with a covalency term. For the fifth HD, HD_5, no traditional chemical concept examined was found to correlate with it. The sign of the HD determines the direction of the property. For instance, a positive sign of HD_1 denotes a donor capacity, but a negative sign refers to an accepting σ ability. The negative sign of HD_2 values implies π acceptor tendency to the moiety, while the positive reflects π donor capacity. It is crucial to emphasize that the SVD method follows a hierarchical approach, signifying that the σ donation mainly governs the M–L bond, followed by the π interactions, and so forth.

Further extension of that work enabled to derive HD values for metal fragments and ligands beyond the original dataset. This strategy aids in investigating the chemistry of the species from dual perspectives. First, the characterization of HDs of ligands and metal fragments leads to envision their electronic properties, thus, providing valuable insights into the behaviour of these chemical species. On the other hand, the HDs can be employed to predict BDE, which as previously said, is a quantity that evaluate the stability of the metal complex. Two distinct cases have been tested to examine the two mentioned approaches.

In the first part of the Chapter, N-heterocyclic carbene (NHC) ligands were submitted to the HD analysis. NHC grasped our attention because of their outspread use in transition metal chemistry. From the synthesis of the first air-stable ylidic carbene,[106], they have played a role in important processes such as cross-coupling,[107,108] olefin metathesis,[109,110] or asymmetric catalysis[111,112]. Their modularity facilitates their design and their constant involvement in chemical transformations.[113] In certain practices, NHCs have assumed the function of phosphine ligands due to their similarities.[114,115] Experimental[116] and computational[117–120] studies have explored the binding properties of NHCs in TM complexes. Herein, energy and charge partitions analysis has served as a tool in computational investigations.[121] These findings leveraged a better understanding of the function of these ligands. NHCs are an intriguing class of neutral ligands containing a divalent carbon atom. A significant feature of this type of species is their strong σ donation capacity. Thanks to their versatile topology this property can be tuned. The main bonding characteristics of the M–NHC bond are depicted in the left part of Figure 3.1. The σ donation from the ligand's lone pair to an empty d_{z^2} orbital of the metal is the dominant force. This interaction occurs together with a negligible π back donation from the metallic d_{xz} (or d_{yz}) orbital to the π orbital of the ligand. In addition, a combination of occupied and empty π orbitals on the NHC may donate density to electron-deficient metal fragments through the $\pi \rightarrow d_{z^2}$ (or d_{yz}).[122]

In the second part, the HD strategy aims to untangle the coordination mode of H_2 with TM complexes. The understanding of the fashion coordination of this small H_2 molecule is key to optimize hydrogenation reactions.[123–125] Nowadays, it is well-known the multistep nature of H_2 activation in homogeneous catalysis where the dihydrogen and hydride-bonded intermediates are found. Yet, this was not always clear.[126–128] The dihydrogen-bonded complex remained elusive until 1984 when Kubas *et al.* characterized a sigma complex with η^2 coordination (η denotes hapticity).[129]

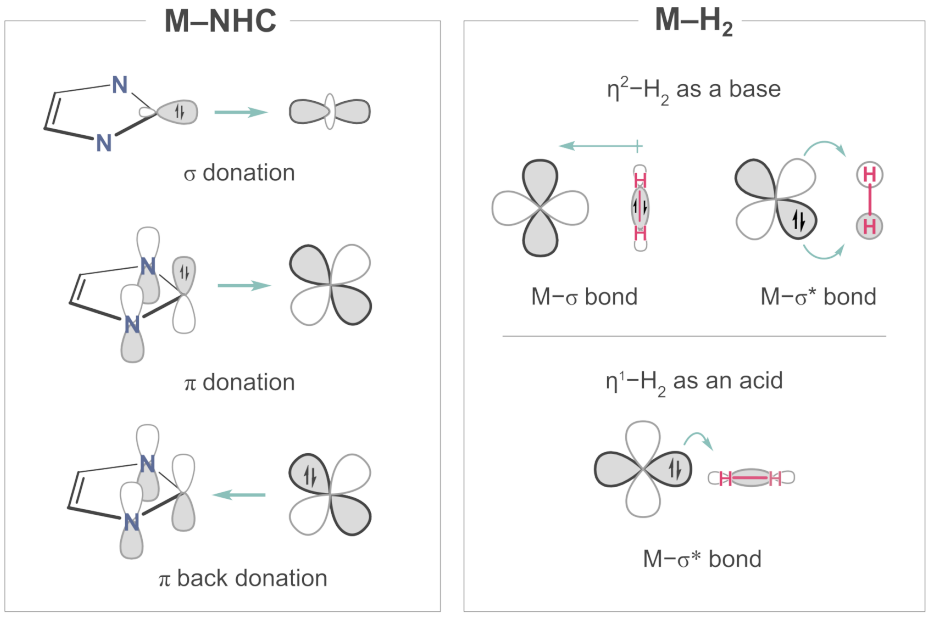


Figure 3.1: Molecular orbital scheme for the interaction between a NHC ligand and a TM center through a $\sigma \rightarrow d_{z^2}$, d_{xz} (or d_{yz}) $\rightarrow \pi^*$ back donation, and $\pi \rightarrow d_{xz}$ (or d_{yz}) (left); and the interaction between a metal complex with a H_2 molecule in the η^2–H_2 and η^1–H_2 configurations (right).

This breakthrough in inorganic chemistry prompted questions about the degree of coordination of H_2 with a metal center.[130–132]

On the top-right side of Figure 3.1, the dihapto coordination of the H_2 with a TM complex is illustrated. The H_2 molecule is bound to the metal via a 3-centered bond with C_{2v} symmetry. The ligand and the metal fragment interact through a donation from the occupied σ orbital of the H_2 to the empty orbital of the metal; and via back donation from the metal to the empty σ^* orbital of the dihydrogen. Here, the hydrogen acts as a base, however, if the back donation is too strong, cleavage of the dihydrogen can occur and the oxidative addition takes place. In addition, the stability of the molecule within a TM complex stems from the nature of the metal center, as well as the solvent and the presence of counterions.[133,134] Alternatively, the coordination of H_2 in a monohapto fashion (η^1–H_2) remains unexplored. In this scenario (bottom-right part of Figure 3.1) the H_2 molecule acts as an acid, accepting electrons from the metal via a σ^* bond. Limited research has been conducted to date on the subject, resulting in scarce data. Investigations that examined the activation of H_2 molecule by a Pd atom in inert matrices at low temperature, identified that H_2 can coordinate with metal species in this alternative η^1–H_2 mode, with a local $C_{\infty v}$ symmetry.[135] Nevertheless, the presence of the η^1–H_2 was finally attributed to the kinetic stability of the complex in the inert matrices.[136] All these findings raised the question of whether the mono– and dihapto–configurations are isomers. Can it be expected in some metal complexes?, or by contrast, is η^1–H_2 highly unstable and challenging to characterize?

Throughout this Chapter, we will derive the HDs from the BDE parameter for the mentioned chemical systems. In the first part, HDs will assist in the comparison of the electronic properties of the NHCs with other ligands. In the second part, HDs will reveal the preference for the conventional or the unconventional coordination of the H_2 ligand within the $L_nM(H_2)$ complex.

3.2 Computational details - BDE Matrix App

The five hidden descriptors described in the first study,[104] can be easily predicted for ligands and metal fragments that have not been previously characterized. To achieve this, it is not necessary to apply the whole SVD treatment as in its implementation. That initial strategy is time-consuming and is used to set a new chemical space. Therefore, researchers within the group designed a mathematical equation for obtaining efficiently HD values.[137] This equation utilizes a limited set of DFT calculations to produce the HD values. The tool is implemented in an open-publicly app named *BDE Matrix App* available at https://maserasgroup-repo.github.io/bdeapp/.[138]

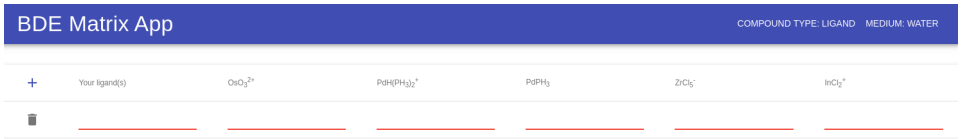


Figure 3.2: BDE matrix App.

The *BDE Matrix App* takes as input each of the BDEs of a given target ligand (metal fragment) paired with five metal fragments (ligands) of reference (*ref*). The outcome includes the five hidden descriptors for this ligand, denoted as $HD_{L(1-5)}$ (metal fragment, $HD_{M(1-5)}$). These descriptors are modelled in water media, thus, certain calculations must be performed in continuum water solvent as specified below. In this Chapter, we only focused on deriving HDs for ligands in water. Therefore, we need to identify the five metal fragments of reference to calculate their BDE with the ligands. In former studies, it was concluded that the set of reference is formed by: OsO_3^{2+}, $PdPH_3$, $PdH(PH_3)^{2+}$, $ZrCl_5^-$, and $InCl_2^+$ metal species. The formula used for the *BDE Matrix App* is a multiple linear regression,

$$HD_{Lk} = \sum_{ref=1}^{5} \alpha_{k,ref} \cdot BDE_{ref,L} + \beta_k \tag{3.2.1}$$

The parameters $\alpha_{k,ref}$ and β_k are integrated in the application. In water solvent, the average and maximum error associated with the prediction of the BDE using predicted hidden descriptors is 1.3 kcal·mol^{-1} and 6.7 kcal·mol^{-1} in water. It is important to remark that SVD was not employed in any part of this Chapter making this strategy very efficient.

The computational method selected in this Chapter follows the specifications established to compute the BDEs of the original dataset.[104] All electronic structure calculations were perfomed using Gaussian 09 package.[139] The geometry optimizations and frequency calculations were conducted using B3LYP-D3[140,141] functional, where empirical dispersion correction was introduced by means of D3 version of Grimme's dispersion.[142] The basis set 6-31+G(d)[143,144] was applied to all elements between H and Cl, and Stuttgart/Dresden effective core potential (ECP), together with the SDD basis set for the heavier atoms.[145,146] It is worth mentioning that the optimization of the metal complexes with monohapto coordination was challenging, and some constraints were needed in order to get a η^1-like structure. The optimization calculations performed for four of the referenced metal complexes: $[PdPH_3(\eta^1 - H_2)]$, $[PdH(PH_3)_2(\eta^1 - H_2)]^+$, $[ZrCl_5(\eta^1 - H_2)]^-$, and $[InCl_2(\eta^1 - H_2)]^+$ required relax the convergence criteria threshold. In the case of $[OsO_3(\eta^1 - H_2)]^{2+}$ no specific demand was included. Solvent effects were simulated implicitly via the PCM[147,148] model for water. The pyssian library was employed as a managing tool for the processing of the input and output files.[149] The vibrational frequency calculations were performed on optimized geometries with the default temperature (298.15K) and 1 atm of pressure to characterize stationary points. A minima state was found when no imaginary frequencies were obtained, while a transition state was determined if one imaginary frequency was present. All energies reported correspond to the potential energies in the aqueous phase plus zero-point energy corrections (ZPE) from the vacuum calculations. A data set collection of computational results is available in the ioChem-BD repository,[19] and is accessible via

https://dx.doi.org/10.19061/iochem-bd-1-220.

3.3 Binding properties of N-heterocyclic ligands

In the analysis of the NHC ligands, we selected a set of twenty-two ligands that covers the common NHC families. Our research focus lies in the study of the electronic properties of these ligands. Hence, we did not consider bulky NHCs to avoid any possible interference.

The top part of Figure 3.3 depicts the considered NHC. For the sake of clarity, we adopted the NHC nomenclature employed by Gusev.[118] This is based on the following scheme abbreviation: [parentheterocycle](substituents)$_n$N(substituents)$_m$. Among the chosen parentheterocycles we found imidazole (Im), imidazoline (sIm), pyrazole (Pyraz), abnormal-imidazole (aIm), triazole (Triaz), pyridine (PyC4), saturated pyrimidine (sPm), benzimidazole (BIm), and dipyridoimidazole (DPyIm). The substituents introduced were alkyl (Me–, Et–, ^{i}Pr–), aryl (Ph–), and functionalized F–, CN–, NO_2– and NMe_2– groups.

The ligands can be classified into different family types as well. The imidazole, imidazolidine, and triazole cores belong to the normal NHC ligands (nNHC). Within the NHC species with reduced heteroatom stabilization, we encounter the Pyraz partners. We also incorporated the remote NHCs (rNHC) and cyclic(alkyl)(amino)carbenes (CAAC) that encompass pyridine-derived structures. The abnormal NHC (aNHC) hosts the NHC in which the nitrogen atoms are not in the expected positions. As previously stated, NHCs are highly modular leading to a broad range of ligand scaffolds. Throughout the Chapter, the Gusev's nomenclature along with the NHC family classification was employed.

3.3.1 Calculation of the HDs for NHCs

The derivation of the HDs involves computing certain BDEs for each target chemical ligand. All BDE values are collected in the Appendix A. The

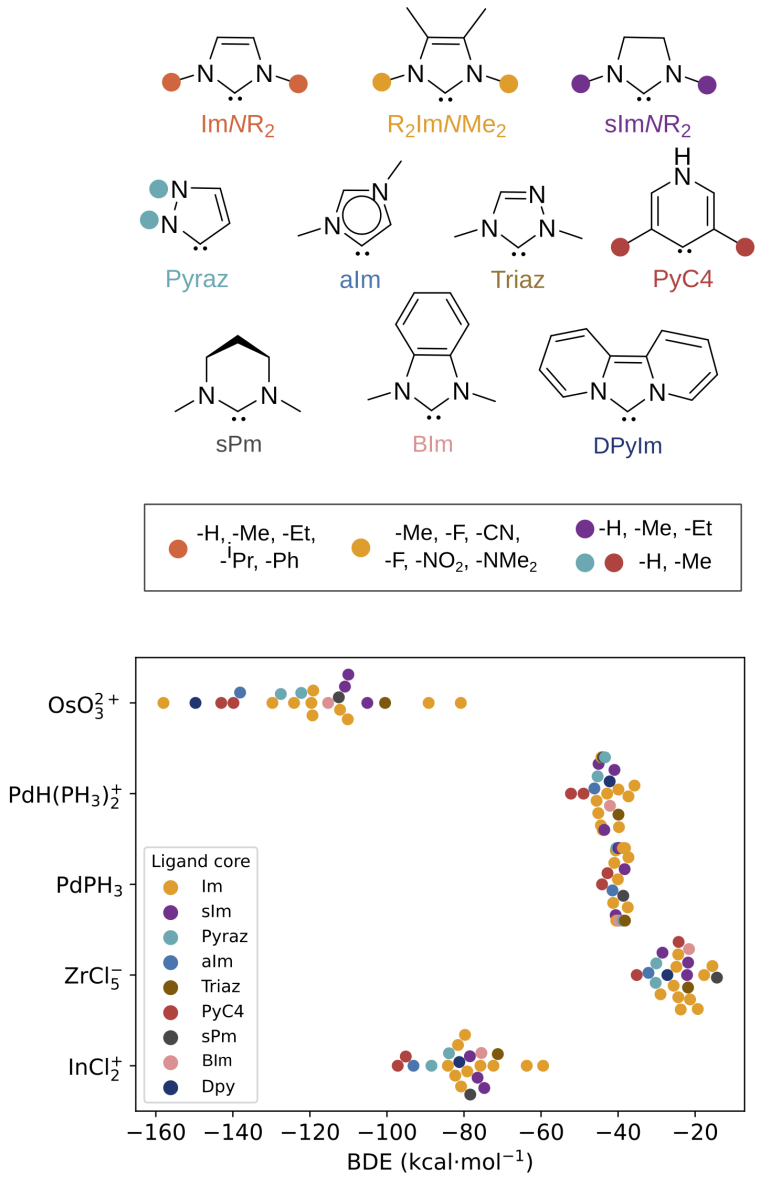


Figure 3.3: NHC structures (top). Scatter plot of the bond dissociation energies (in kcal·mol^{-1}) for each NHC ligand of the set (bottom). Color-coded representation: • • imidazole (Im), • imidazoline (sIm), • pyrazole (Pyraz); • abnormal-imidazole (aIm); • triazole (Triaz); • pyridine (PyC4); • saturated pyrimidine (sPm); • benzimidazole (BIm) and • dipyridoimidazole (DPyIm)

scatter plot in Figure 3.3 shows the DFT-BDE values between the considered NHCs and each metal fragment.

A first inspection of the Figure 3.3 plot reveals that the BDE values are primarily influenced by the metal component rather than the ligand. The osmium complex yields the lowest and broadest BDE values among the metal fragments. In this context, the lowest BDE refers to the strongest bonds and thus, the less labile ligands. Furthermore, the $InCl_2^+$ fragment provides energies ranging from -100 and -60 kcal·mol^{-1}. We also observed that the transition metals of the group 10 and the zirconium complex are the least efficient partners in forming stable bonds with NHCs.

Regarding the NHC classes, we realized that the pyrazole-derivate ligands display the most robust bonds within every metal fragment. Unsaturated imidazoles are spread across the spectrum of BDEs associated with each metal fragment. The remaining values are more difficult to analyse. This plot aids in identifying potential outliers or sources of errors in the calculations.

3.3.2 Analysis of the HDs

Upon calculating the bond dissociation energies, we input them into the *BDE Matrix App*. The resulting HD for each ligand is showcased in Table 3.1.

Table 3.1: Hidden descriptor values of the NHC ligands.

NHC	**HD**$_{L1}$	**HD**$_{L2}$	**HD**$_{L3}$	**HD**$_{L4}$	**HD**$_{L5}$
ImNH$_2$	0.193	-0.160	-0.123	0.038	-0.092
ImNMe$_2$	0.200	-0.220	-0.038	-0.025	-0.089
ImNEt$_2$	0.204	-0.243	-0.080	-0.032	-0.047
ImN^iPr$_2$	0.207	-0.243	-0.105	-0.030	-0.034
ImNPh$_2$	0.204	-0.295	0.011	-0.145	-0.079
Im(NO$_2$)$_2N$Me$_2$	0.151	-0.270	-0.117	0.068	-0.101
ImCN$_2N$Me$_2$	0.162	-0.259	-0.106	0.040	-0.089

ImF_2*N*Me_2	0.185	-0.227	-0.039	-0.020	-0.131
ImMe_2*N*Me_2	0.210	-0.220	0.005	-0.082	-0.123
ImNMe_2*N*Me_2	0.227	-0.188	0.197	-0.269	-0.378
sIm*N*H_2	0.190	-0.192	-0.185	0.046	-0.027
sIm*N*Me_2	0.194	-0.269	-0.091	-0.043	-0.029
sIm*N*Et_2	0.197	-0.285	-0.123	-0.028	0.019
PyrazC3*N*H_2	0.210	-0.174	-0.114	0.004	0.004
PyrazC3*N*Me_2	0.219	-0.185	-0.101	0.009	0.036
sPm*N*Me_2	0.192	-0.320	0.025	-0.082	0.043
Pyc4*N*H	0.238	-0.175	-0.136	-0.021	0.081
Pyc4-3,5-Me_2*N*H	0.239	-0.292	-0.011	-0.117	0.081
BIm*N*Me_2	0.193	-0.245	-0.037	-0.050	-0.115
DPyIm	0.219	-0.155	0.135	-0.197	-0.372
aIm*N*Me_2	0.230	-0.159	-0.064	0.001	0.009
1,2,4-Triaz*N*Me_2	0.180	-0.237	-0.114	0.028	-0.044

An initial evaluation of Table 3.1 indicates that the first and the second HD_L provided the expected values. HD_{L1} parameter has a positive sign indicating a sigma donating capacity, while the negative sign of the HD_{L2} denotes a general π acceptor behaviour. This outcome aligns with the anticipated properties of such ligands.

In the initial dataset reported by our group, HDs were derived by applying the SVD method. Consequently, in that particular study, the weight of each hidden descriptor (HD_{Wk}) represented the average influence of diverse 42 ligands and 43 metal fragments. In contrast, in the current Chapter, our focus narrows down to a specific type of ligands. As a consequence, we quantified the % contribution of each HD to the BDE for the particular M–NHC bond set. Comparison of these newly derived contributions with those documented in the initial chemical space reveals the properties of these NHC ligands. Equation 3.3.2 renders the amount of BDE for each k value,

$$bde_{k,ij} = HD_{Lk,i} \cdot HD_{Wk} \cdot HD_{Mk,j} \tag{3.3.1}$$

Within Equation 3.3.2, HD_{Wk} and $HD_{Mk,j}$ are SVD-derived HDs, while $HD_{Lk,j}$ are the new values obtained in this Chapter. To quantify the % influence of each $bde_{k,ij}$ on the computed DFT-BDE value, we conducted Equation 3.3.2 simplifying the notation of HD_W to W. In other words, we aimed to know the percentage of BDE that each HD_k explains (% $W_{k,ij}$).

$$\%W_{k,ij} = \frac{bde_{k,ij} \cdot 100}{BDE_{DFT,ij}} \tag{3.3.2}$$

We then collected % $W_{k,ij}$ for all BDE values between each pair of metal fragment of reference i and each NHC ligand j (where N_{BDE} denotes the total number of computed BDEs). Next, we averaged that value to obtain a percentage of the weights for all the NHC set.

$$\%\overline{W_k} = \frac{\sum_{ij=1}^{N_{BDE}} |\%W_{k,ij}|}{N_{BDE}} \tag{3.3.3}$$

$$\%\overline{W} = \sum_{k=1}^{5} \%\overline{W_k} \tag{3.3.4}$$

Finally, we normalized,

$$\%W_k(NHC) = \frac{\%\overline{W_k} \cdot 100}{\%\overline{W}} \tag{3.3.5}$$

The weight of the hidden descriptor within the NHC set (% $W_{k,NHC}$) (displayed as blue bars in Figure 3.4) was compared to the weight of the ligands from the preceding study (% $W_{k,DFT}$) (displayed as purple bars in Figure 3.4).

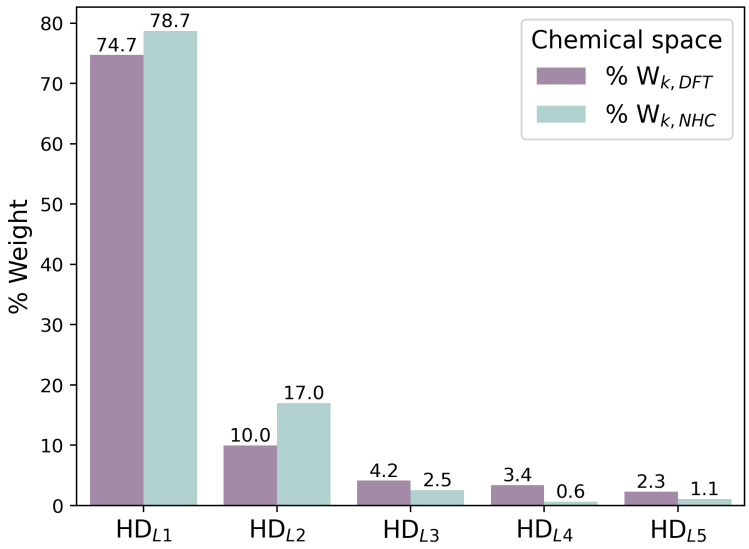


Figure 3.4: Bar plot of the percentage of the hidden descriptor weights in the reported chemical space[104] (purple bars) and in the NHC collection (blue bars).

Bars corresponding to the HD_{L1} and HD_{L2} exhibit greater heights in the NHC set than in the reported chemical space. In contrast, the remaining bars follow the opposite trend. In the original set, the first two chemical descriptors accounted for 84.7 % of the chemical contribution. In the NHC ensemble, this contribution increases up to 95.7 %, reflecting the higher importance of these two parameters within the NHC collection. Conversely, the HD_{NHCk} for $k = 3$, 4, and 5 have a lower contribution to the BDE value. Thus, the chemical properties associated with them are less significant in the M–NHC interaction.

Due to the increased importance of the hidden descriptors 1 and 2, the following HD analysis concentrates on the study of these two values. Moving back to the Table 3.1, we located at the top of the HD_{L1} scale, the 6-membered PyC4-3,5-Me_2NH fragment (0.239) and the five-membered aImNMe_2 (0.230). These are the strongest σ donors in our group. In contrast, the less donating NHC is Im$(NO_2)_2NMe_2$ with HD_{L1} of 0.151.

This result agrees with the experimental values, as the ligands on the top are CAAC and abnormal NHC, which were expected to be the most donating. In the case of $Im(NO_2)_2NMe_2$, we assumed its poor capacity to donate electrons because of the presence of the NO_2– group. This electron-withdrawing group (EWG) retains the electron density preventing the charge transfer to the metal center.

In the investigation of the HD_{L2} column, the DPyIm holds the highest value (-0.155). Its ligand structure is a fused tricycle NHC with high aromaticity and electron density delocalization, enabling lower π acceptance capacity. On the other hand, the greatest π acceptor is $sPmNMe_2$ with HD_{L2} of -0.320. The structure of the latter is opposite to the one identified on the top, being a saturated six-membered structure.

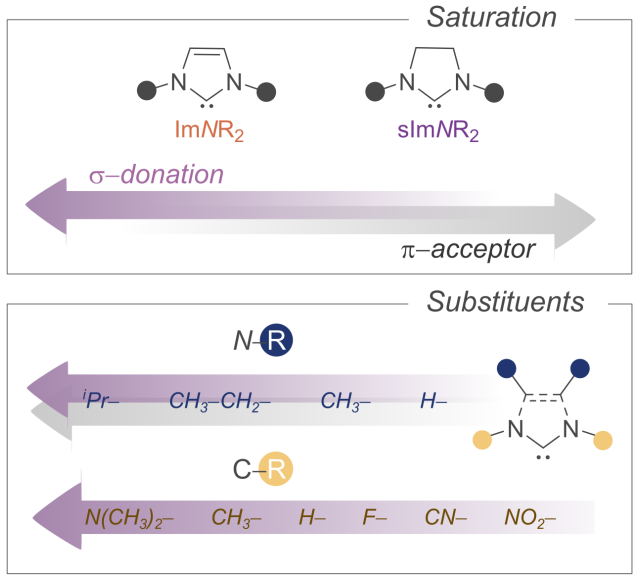


Figure 3.5: General trends for σ donation and π acceptance for the NHC set.

Regarding the entire collection of values, it is worth noting some patterns (Figure 3.5). For instance, unsaturated imidazole carbenes ($ImNR_2$ being R = H–, Me–, Et–) are stronger σ donors and weaker π acceptors than their saturated imidazoline analogues ($sImNR_2$ being R = H–, Me–, Et–). The

alkyl substituents on the nitrogen atoms customize the global properties of the ligands. The longer the chain length (H– < Me– < Et– < ^{i}Pr–), the greater the sigma donating and π accepting features of Im*N*R$_2$ and sIm*N*R$_2$, and of pyridine- and pyrazole-derived fragments. In addition, the substituents in the C4 and C5 positions tunes to higher extent the σ donation capacity of the five-membered rings. The EWGs decrease the σ donation ability and the electron-donating groups increase it (NO$_2$– < CN– < F– < H– < Me– < NMe$_2$–) along the ImR$_2$*N*Me$_2$ group. Lastly, CAACs and abnormal ligands exhibit a stronger σ donation contribution compared to the nNHCs. After evaluation of the resulting HDs for NHCs, we confirmed the parallelism between this outcome and the findings reported in different studies.[150–152]

3.3.3 Descriptors for σ donor ability and HDs

The analysis carried out in Section 3.3.2 validates the effectiveness of the HD method with this type of chemical species. We have already mentioned that HD method offers a simplified approach since it only requires the computation of a few DFT calculations. However, it is remarkable that certain widely employed conventional descriptors, *e.g.* HOMO energy, are even more straightforward to acquire. To assess the performance of our descriptor against more popularly used, we carried out statistical analyses.

BDE is inherently connected to the binding affinity of the ligands. This value is the sheer parameter to describe the bonding tendency of the ligands. Moreover, it is established that the binding affinity of the NHC is strongly correlated with the σ donation feature. This evidence is further reinforced by the calculation of % W$_{NHC1}$ which accounts for 78.7 % of the overall metal-ligand interaction. Therefore, we collected all the computed BDEs in this Chapter, between each NHC and each metal fragment, and calculated their average values with respect to each NHC ligand, $\overline{BDE_j}$. Then, we utilized this value to evaluate the descriptors of the sigma donating property. We sought conventional chemical descriptors related to that property. Our

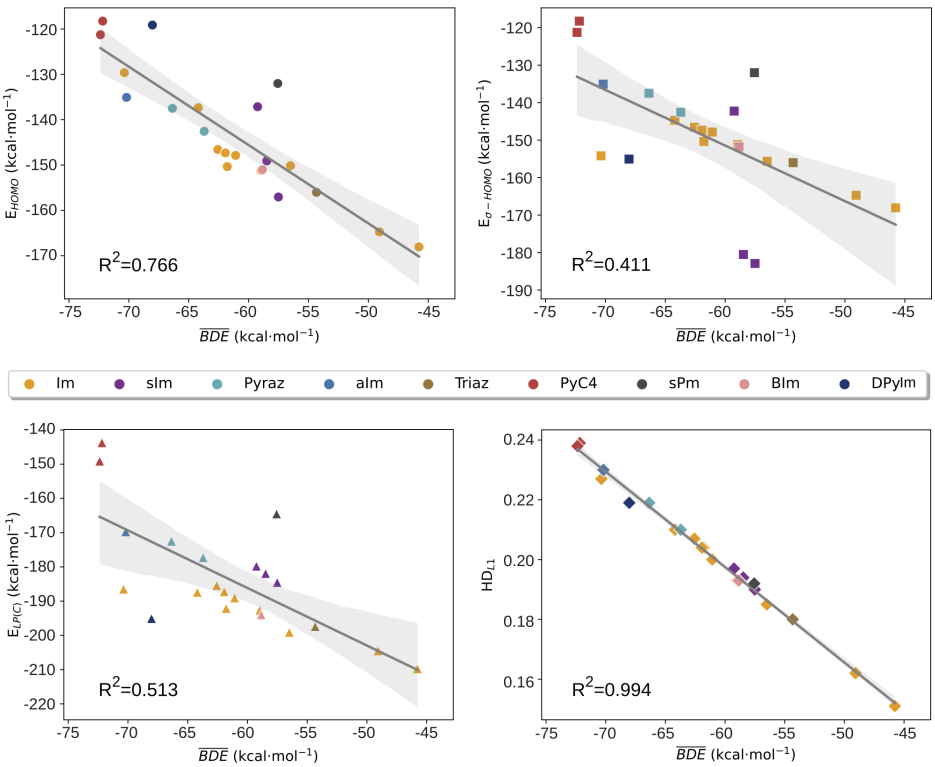


Figure 3.6: Scatter plots of the average BDE per each ligand ($\overline{BDE_j}$) *vs* the values of the HOMO energy (left-upper plot in circle dots); the values of the σ-HOMO energy (right-upper plot in square dots); the values of the NBO energy of the LP(C) (left-bottom plot in triangle); values of the HD$_{L1}$ (right-bottom plot in diamonds).

expectations were to find good correlations between these easy-handled descriptors and the $\overline{BDE_j}$.

Among the standard descriptors, we encountered features associated with the frontier molecular orbitals (FMO). The energy of the highest occupied molecular orbital (E$_{HOMO}$) is a commonly used parameter to rationalize chemical species reactivity. We first evaluated the linear correlation of such descriptor, yielding a R^2 of 0.765 (top-left) plot in Figure 3.6). We noticed that the raw energy values of HOMO are not suitable for the fitting. In this attempt, the correlation was fully conducted to correlate

the descriptors with the σ donation characteristic, however, the orbitals may happen to bear different symmetry. Hence, we examined the symmetry orbital to gather the highest occupied molecular orbital with the correct symmetry. This new descriptor was denoted as the energy of the HOMO of the σ orbitals ($E_{\sigma-HOMO}$). Top-right plot of Figure 3.6 illustrates the correlation between $\overline{BDE_j}$ and $E_{\sigma-HOMO}$ with a R^2 of 0.411. This redefined function was also unsuccessful.

Next, we gathered the orbital energies derived from the Natural Bond Orbital (NBO) analysis. NBO results provide a valence bond-type description of the system giving the most accurate possible "natural Lewis structure" associated with the total electron density. Within this theoretical framework, NBO analysis localizes atomic (i.e. lone pairs and vacancies) and bond (i.e. 2- and 3-center non-bonding, bonding, and anti-bonding) orbitals interacting with each other. Therefore, we identified the lone pairs (LP) located over the carbene carbon of the NHC. Bottom-left plot of Figure 3.6 shows the correlation between the sp^2-hybridized lone energy pair (LEP) of the ylide carbon and the $\overline{BDE_j}$. The R^2 denotes a poor relationship between these two concepts. In this case, the molecular orbital linked to the LP of the carbene carbon contributes to a minor extent due to the mixing of atomic orbitals distant from the carbon center. The variations in the results can be attributed to the unique characteristics inherent to each NHC. Indeed, after an in-depth inspection of the plots, we realized that there are subsets within the same type of ligands with good correlation. In light of the findings, we acknowledge the challenging task of associating orbital-based descriptors with parameters of a different nature.

Finally, HD_{L1} is associated with the $\overline{BDE_j}$ parameter yielding a R^2 of 0.994 (bottom-right plot of Figure 3.6). It's important to note that our study does not intend to discredit the utility of E_{HOMO}; instead, it did not provide the expected results in this specific investigation. This conclusion is also a valuable proof that there are instances where E_{HOMO} is not adequate enough, and the application of the HD method helps in the understanding

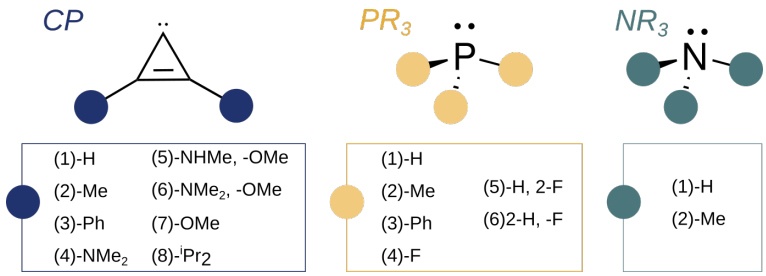


Figure 3.7: Cyclopropenylidenes (CP), phosphines (PR_3), and amines (NR_3). The boxes contain the substituents considered in this study.

of the σ donation property.

3.3.4 Comparison between ligands employed in TMC

Throughout the Chapter, the hidden descriptor methodology was utilised to evaluate the electronic role of various NHC ligands. Their characteristics were identified and compared based on their skeletal structure. Additionally, the HD values have the ability to distinguish ligands, regardless of their nature.

Aiming to evaluate that property, we have selected different classes of ligands employed in TM chemistry: cyclopropenylidines (CP), phosphines (PR_3), and amines (NR_3) with diverse substituents, –R (Figure 3.7). Cyclopropenylidines are a type of carbocyclic carbenes that have gained attention due to their involvement in catalysis[153,154] and their relevance in the field of intellestar chemistry.[155] We curated a set of phosphines for comparison, considering that NHCs have been considered as "phosphine mimic" ligands.[113] Furthermore, amine ligands were included to observe the effect of the nitrogen atom directly acting in the donation. The expanded set encompassed 38 ligands and all of them are classified as σ donors and π acceptors.

We then applied the protocol by first computing the BDEs between the incorporating ligands and our metal species and second, by stemming the HDs. Table 3.2 collects all the HD values of the new set.

Table 3.2: Hidden descriptor values of the cyclopropenylidine, phosphine and amine ligands.

Ligand	**HD**$_{L1}$	**HD**$_{L2}$	**HD**$_{L3}$	**HD**$_{L4}$	**HD**$_{L5}$
CP_1	0.157	-0.214	-0.140	0.103	-1.236
CP_2	0.181	-0.194	-0.110	0.063	-0.089
CP_3	0.195	-0.175	-0.084	0.012	-0.064
CP_4	0.221	-0.129	0.030	-0.043	-0.143
CP_5	0.193	-0.156	-0.032	0.041	-0.140
CP_6	0.200	-0.151	-0.037	0.019	-0.075
CP_7	0.178	-0.180	-0.088	0.075	-0.098
CP_8	0.230	-0.220	0.043	-0.108	-0.063
PH_3	0.092	-0.192	0.175	0.052	-0.277
PMe_3	0.156	-0.184	0.125	-0.063	-0.267
PPh_3	0.156	-0.225	0.208	-0.164	-0.406
PF_3	0.030	-0.262	-0.052	0.245	-0.509
PHF_2	0.063	-0.254	-0.006	0.130	-0.486
PH_2F	0.082	-0.230	0.063	0.072	-0.394
NH_3	0.111	-0.091	-0.018	0.177	0.093
NMe_3	0.114	-0.103	0.094	0.165	0.117

The range of substituents in CP endows 3-membered cycles with flexible properties. Minor modifications in their structure greatly affect their features because of the proximity of the –R to the carbene carbon, and the constrained cycle core. The CP_1 holds the smallest value of HD$_{L1}$ (0.157). This ligand possesses hydrogen substituents which are the weakest donors considered in the set. CP_8 contains ^{i}Pr, the longest alkyl chain within the CP group, and it has the highest value of HD$_{L1}$. Trends for the HD$_{L2}$ are also consistent with the expected results, being CP_8 the preferred π acceptor.

A further examination of Table 3.2 reveals that the fluorine bound to the phosphorous atom decreases the capacity of the phosphines to donate

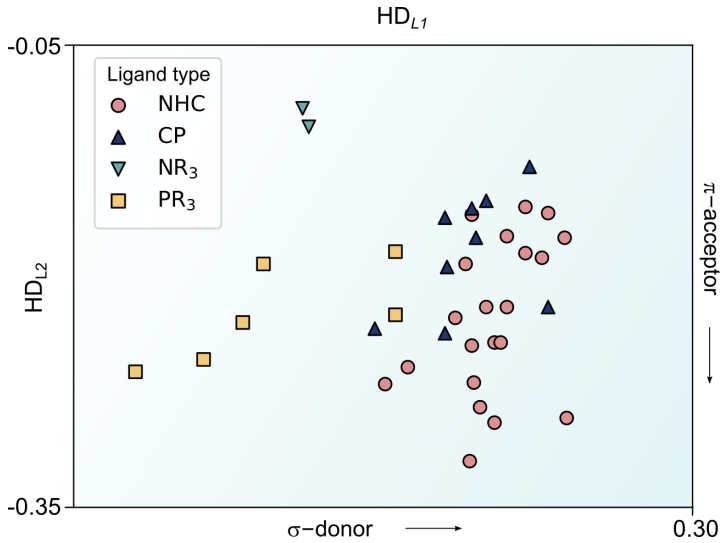


Figure 3.8: Scatter plot of the HD$_{L1}$ and HD$_{L2}$ for the 38 ligands considered in the study.

σ electron density. The increment in the number of fluor atoms enhances this behaviour (HD$_{L1}$: $PH_2F > PHF_2 > PF_3$). The π acceptor property diminishes in the opposite direction. The selected two amine fragments anticipated the outcomes since the presence of alkyl group augments the donation ability of the ligand. Furthermore, the capacity to accept π electrons is enriched as well.

Figure 3.8 shows a scatter plot with the 38 ligands submitted to the study. The NHCs exhibit the highest σ donor property together with CPs. However, they differ in the HD$_{L2}$ because CPs offer less π acceptor capacity. The adjacent nitrogen atoms of NHC may aid in the accommodation of the π electron density. Additionally, the phosphine set is clearly the less donating group. This is indeed affected by the F– substituents introduced in such ligands. The two amines are far from being designed as π accepting groups, while they fall within the σ donating range of the phosphines.

3.4 Searching for metal fragment candidates for monohapto-dihydrogen ligand

As introduced at the beginning of the Chapter, the monohapto coordination of H_2 remains elusive. We attempt to find a metal fragment that stabilizes a complex with a η^1-H_2 ligand ($L_nM(\eta^1\text{-}H_2)$).[156] Here, the HD method is used to identify such ideal metal fragment.

3.4.1 Calculation of the HDs

First, it is necessary to acquire the hidden descriptors of the unexplored H_2 coordination mode. To accomplish this, the same protocol as explained in Section 3.2 is implemented. The computed BDEs between the η^1-H_2 and the metal fragments of reference furnish the HDs. The resulting values are collected in Table 3.3.

Table 3.3: Hidden descriptor values for both η^1-H_2 and η^2-H_2 ligands, along with the respective differences between them.

$\mathbf{HD}_{Lk}$	$\mathbf{HD}_L(\eta^1\text{-}\mathbf{H_2})$	$\mathbf{HD}_L(\eta^2\text{-}\mathbf{H_2})^a$	$\Delta\mathbf{HD}_L((\eta^1\text{-}\mathbf{H_2})$ - $(\eta^2\text{-}\mathbf{H_2}))$
1	-0.015	-0.013	-0.002
2	-0.058	-0.143	0.085
3	-0.022	0.013	-0.035
4	0.312	0.118	0.194
5	0.032	-0.021	0.053

[a]Extracted data from previous publication.[104]

From Table 3.3, it is evident that the differences between the HDs for both ligands are minimal. The first hidden descriptor only differs by 0.002 units. In both ligands, the σ donation interaction is almost negligible. The HD_{L2} varies in 0.085 units. In this case, the conventional H_2 reflects a higher tendency to accept π electrons from the metal fragment. Therefore, metal fragments entitled to donate π electrons will slightly favour the dihapto coordination. This phenomenon is understood for the orbital disposition in

the ligands (Figure 3.1) where the σ^* orbital that locates the electrons is closer in the C_{2v} symmetry than in the $C_{v\infty}$ symmetry, thereby, promoting the interaction. The HD_{L3} value is smaller for monohapto than for dihapto H_2. In this case, the unconventional ligand exhibits smaller *cis* repulsive interaction with the metal fragments. This decreased behaviour arises from the linear arrangement of the hydrogen atoms where the σ^* orbital does not contact the other ligands of the metal fragment. The HD_{L4} quantifies the degree of covalency, where the η^1 ligand holds higher values for this characteristic than the η^2-H_2.

Within Table 3.3, we observed that the ligand configuration confers different electronic properties that can be explained by the hidden descriptor method. Consequently, these derived property values are expected to impact the predicted BDE. Our analysis will be based on BDE parameter because it distinguishes between the two types of ligand configuration and reflects the stability of the complexes.

3.4.2 Comparison between end-on- and side-on-bonded dihydrogen in metal complexes

BDE is a metric that measures the stability of a given bond, therefore, we can employ this value to discriminate the unstable isomer of H_2 ligand. Besides the insights about the electronic role of the ligands, hidden descriptors are inputs to predict BDE of the M−L bond. Equation 3.4.1 outcomes the predicted BDE, BDE_{pred}, between a metal fragment and the H_2 ligand, where k denotes the number of hidden descriptors considered.

$$BDE_{pred} = \sum_{k=1}^{5}(HD_{Mk} \cdot HD_{Wk} \cdot HD_{H_2k}) \tag{3.4.1}$$

The hidden descriptors for the metal fragments, HD_{Mk}, were extracted from the former study of the group.[104] In total, we considered twenty-three ML_n, resulting in the same number of predicted BDEs for metal complexes

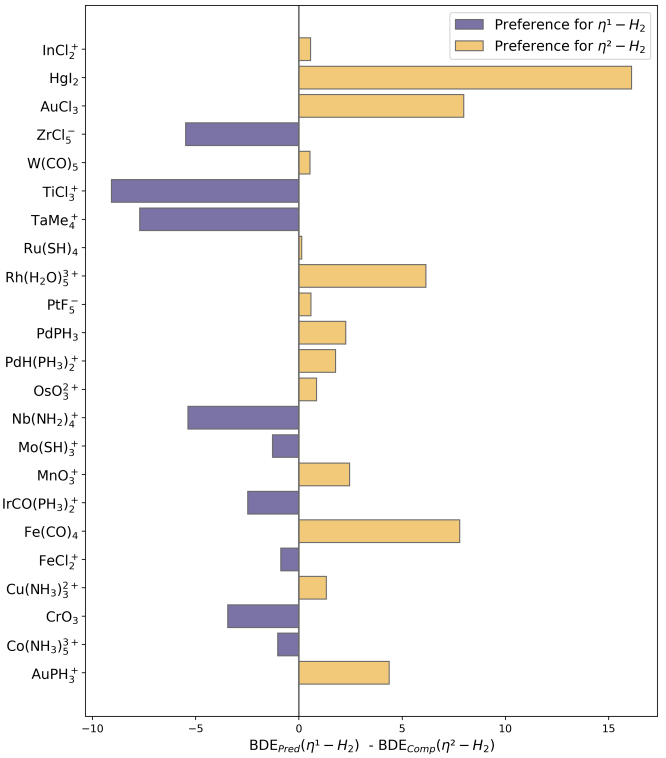


Figure 3.9: Bar plot of the bond dissociation energy difference between the bonds L_nM-η^1-H_2 and L_nM-η^1-H_2

with η^1-H_2. The BDEs for the metal complex with η^2-H_2 are gleaned from the reported article.[104] All the BDE_{pred} values are reported in the Appendix A.

Our attempt is to identify the metal fragments that favour end-on-bonded configuration rather than side-on-bonded. For this purpose, we performed a simple substraction of BDE($L_nM(\eta^2$-$H_2)$) to the BDE($L_nM(\eta^1$-$H_2)$). Figure 3.9 collects the outcome of the operation. Herein, the negative values (purple bars in Figure 3.9) correspond to instances where the unconventional rearrangement is favoured, while the positive values (yellow bars in Figure 3.9) state for the opposite case. Among the twenty-three tested complexes, we found that fourteen metal complexes have a

preference for the conventional configuration, whereas nine of them favour the monohapto disposition.

A careful evaluation of Figure 3.9 reveals a greater understanding of the metal . The following nine metal complexes *a priori* favours the atypical ligand configuration: $IrCO(PH_3)^{2+}$ with d^8 electron configuration, $Co(NH_3)_5^+$ comprising d^6, $FeCl_2^+$ containing d^5, $Mo(SH)_3^+$ with d^3 and four d^0 metals: $ZrCl_5^-$, $TiCl_3^+$, $TaMe_4^+$, $Nb(NH_2)_4^+$, CrO_3. Notably, they do not exhibit an uniform characteristic, apart from the higher number of d^0 metal fragments present. Identifying that 39 % of the entire metal set promotes the underdeveloped arrangement is promising. We are aware of the simple approach of the HD method and its inherent limitations, however, we do not want to fall into the misconception that a more complex method would necessarily yield more accurate results. Hence, we proceeded to investigate the nature of these nine metal fragments using our current strategy.

So far, the HD method filtered the whole set of metal fragments. The BDE-based approach pointed out nine metal fragments that favoured the η^1 hapticity. However, as the BDE values are predictions, they might contain errors. In addition, η^1 and η^2 arrangements are isomers, whereby the differences between them are smaller, and even minimal errors are significant. Bearing this in mind, we decided to compute the nine identified potential complexes through DFT optimizations. Geometry optimizations were challenging and they required constraints. It was necessary to freeze the angle along the M–H–H axis at 178.0° for $[FeCl_2(\eta^1 - H_2)]^+$, $[Mo(SH)_3(\eta^1 - H_2)]^+$, $[TiCl_3(\eta^1 - H_2)]^+$, $[TaMe_4(\eta^1 - H_2)]^+$ and $[Nb(NH_2)_4(\eta^1 - H_2)]^+$. For the remaining complexes, $[IrCO(PH_3)(\eta^1 - H_2)]^{2+}$, $[Co(NH_3)_5(\eta^1 - H_2)]^+$, $CrO_3(\eta^1 - H_2)$, the angle was fixed at 179.0°. Table 3.5 shows the computed BDE values for the $L_nM(\eta^1\text{-}H_2)$ and the reported DFT energies for $L_nM(\eta^2\text{-}H_2)$ with the nine metal fragments.

Only four out of the nine candidates fulfilled the expected outcome, as indicated by the bold values in Table 3.5. The four metal complexes that followed the anticipated trend were: $FeCl_2^+$, $ZrCl_5^-$, $TaMe_4^+$, and $Nb(NH_2)_4^+$.

Table 3.5: Bond dissociation energies in kcal·mol^{-1} computed with DFT calculations and predicted with HD method between the nine selected metal complexes and the H_2 ligand in η^1 and η^2 hapticities.

Metal	**BDE$_{DFT}$**		**BDE$_{HD}$**	
fragment	**L$_n$M(η^1-H$_2$)**	**L$_n$M(η^2-H$_2$)^a**	**L$_n$M(η^1-H$_2$)**	**L$_n$M(η^2-H$_2$)^a**
$IrCO(PH_3)^{2+}$	12.8	3.3	0.8	10.6
$Co(NH_3)_5^+$	8.3	4.4	3.4	14.1
$FeCl_2^+$	**10.4**	**11.5**	**10.6**	**13.9**
$Mo(SH)_3^+$	15.9	5.0	3.7	14.8
$ZrCl_5^-$	**1.7**	**9.3**	**3.8**	**7.9**
$TiCl_3^+$	16.0	13.8	4.7	17.2
$TaMe_4^+$	**15.0**	**16.2**	**8.5**	**11.1**
$Nb(NH_2)_4^+$	**5.2**	**12.1**	**6.7**	**7.9**
CrO_3	10.2	3.1	-0.4	15.8

[a]Extracted data from previous publication.[104]. Lower values for η^1-H_2 ligand in bold.

A 3D representation of these metal fragments with the associated linear ligand is depicted in Figure 3.10.

Computing the BDE parameters supposed a second filter that eliminated five metal fragments from the set. At first glance, the removal of five candidates may suggest that the HD method failed to predict energies. Yet, it is important to consider several factors in order to fully analyze this outcome. Firstly, as already mentioned in the Section 3.2, the prediction of BDE using HD method has associated an average and maximum error of 1.3 kcal·mol^{-1} and 6.7 kcal·mol^{-1} within the water phase in the initial set.

In the current work, the average absolute error in predicting the BDE is 6.8 kcal·mol^{-1} for L$_n$M(η^1-H$_2$) and 6.2 kcal·mol^{-1} for L$_n$M(η^2-H$_2$). Apparently, these values appear relatively high, so it is crucial to place them within the frame of BDE parameter. The values of the initial BDE set span from -280 kcal·mol^{-1} to 20 kcal·mol^{-1}. Here, the average absolute error corresponds to 2.3 % of the total range, which is deemed acceptable. Furthermore, due to the constraints imposed during the DFT optimizations for η^1-H$_2$, small energy errors were expected. Attempting to analyse the electronic features of the selected candidates, Table 3.6 displays their hidden descriptors.

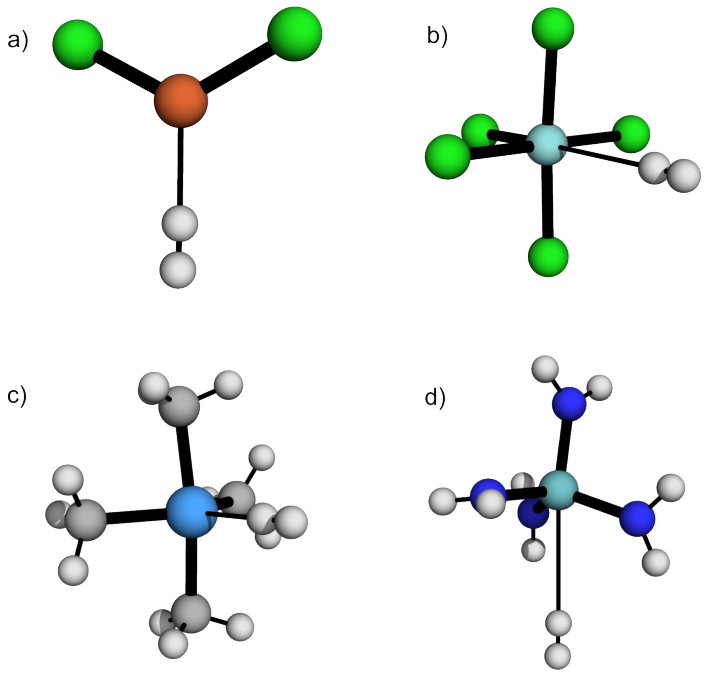


Figure 3.10: 3D Representation of metal complexes with η^1-H_2 ligand. Following the order top, left-to-right, the metal centers of the complexes are (a) • Fe ,(b) • Zr,(c) • Ta, and (d) • Nb. The color-coded for the non-metal atoms is: • C, • H, • N, and • Cl.

Table 3.6: Hidden descriptor values of the candidate metal fragments.

$\mathbf{HD}_{Mk}$	$\mathbf{FeCl_2^+}$	$\mathbf{ZrCl_5^-}$	$\mathbf{TaMe_4^+}$	$\mathbf{Nb(NH_2)_4^+}$
1	-0.198	-0.112	-0.158	-0.112
2	-0.230	-0.256	-0.371	-0.245
3	-0.004	0.350	0.229	0.388
4	0.161	-0.058	0.055	0.087
5	-0.073	0.081	-0.187	0.075

Concerning the ligands of the four homoleptic metal fragments, we observed that Ta and Nb complexes hold donating groups. In contrast, the Zr and Fe complexes contain chloride, an EWG ligand. The σ donation capacity of H_2 ligand is minimal (as seen in Table 3.3). Thus, it is expected to find metal fragments that do not rely on the electron density from this ligand for their stabilization. Regarding HD_{M1} values, we observed that behaviour since this variable ranges from -0.112 to -0.198 in this subset. In this context, we considered the inverted ligand field (ILF) situation. In this case, the sigma bonding orbital is predominantly centered on the metal. Thus, metal fragments that promote this situation may provide an alternative approach to forming the unusual complex.

On the other hand, HD_{M2} yields negative values, indicating the stronger π acceptor character of these metal centers. This result seems counterintuitive as the C_{2v} ligand interacts strongly with π acceptor fragments, due to its arrangement. This observation holds true, and it is the source of the problem. The increased π interaction between the metal and the orbitals of the hydrogen atoms would lead to the cleavage of the H−H bond, switching to a more stable dihydride complex. Therefore, in such cases, the monohapto configuration is preferred to avoid bond breaking.

Three of the four metal complexes, with the exception of the $FeCl_2^+$, exhibit stronger positive values for the third descriptor. This positive behaviour agrees with the linear disposition of the ligand, which reduces the potential repulsion between the H_2 orbitals and the ligands bound to the metal center. This repulsion would be more pronounced for the η^2-H_2 ligand, where the 3-centered disposition facilitates the interaction with the rest of the ligands. The HD_{M4} parameter is relatively small in this set, covering from -0.058 to 0.161, in contrast to the wider range of values in the initial set (-0.233 to 0.368).

So far, we have figured out the preferred isomer of H_2 with twenty-three metal fragments of different nature. Nevertheless, even if we identified four metal fragments that favours the sought η^1 isomer, we needed to investigate

the stability of such metal complexes. The computed BDE values (Table 3.5) were between 1.7 and 15.0 kcal·mol^{-1}. These results do not satisfy the criteria anticipated at the begining of the section. To consider a relatively stable bond, and thus, a potential existing metal complex, we need to identify negative BDE, approximately lower than -10 kcal·mol^{-1}. This is required to further account for the entropic contributions opposing coordination. Therefore, with our outcome, the H_2 ligand will separate from the four metal centers.

Upon evaluating the interaction between this atypical mode of the ligand with twenty-three diverse metal fragments, we ended up assuming that none of them facilitate the formation of the $L_nM(\eta^1\text{-}H_2)$. The challenging task of finding a suitable metal fragment is out of the scope of our metal species group. Aiming to discover the ideal electronic features that would accommodate that ligand in a complex, we explored alternative statistical investigation.

3.4.3 Search for the ideal metal fragment for monohapto dihydrogen metal complex

We aim to explore the features of the ideal fragment out of the chemical space delimited by the twenty-three metal fragments. The electronic attributes are defined as quantitative hidden descriptor variables. In the previous section, the HD_{Mk} values corresponded to the twenty-three metallic species defined earlier. Henceforth, these variables HD_{Mk} have to be defined. Assigning values to an empty HD array will generate combinations of HD_{Mk} that refer to hypothetical metal fragments. Figure 3.11 shows the HD_{Mk} arrays for theoretical metals, M_1, M_2, ..., M_N, where N refers to the total number of metal vectors described. We selected a numerical range between -0.650 to 0.650 for HD_{Mk} parameters based on the chemical space outlined in Lakuntza's study.[104]

For the HD_{M1}, these values are pruned to negative values (-0.650 to

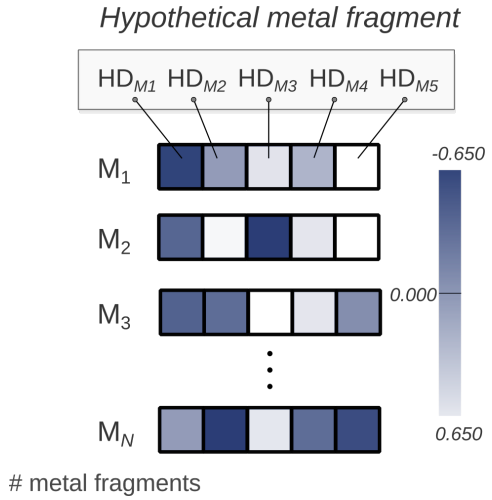


Figure 3.11: Scheme of of the development of N combinations of HD_{Mk} values for describing hypothetical metal fragments.

0.000) since positive values are not anticipated for this parameter. A negative sign denotes the metal fragment's capability to accept σ electron density, while a positive sign would imply a metal fragment endowed with a σ donation capacity, which is not expected *a priori*.

The remaining HD_k parameters, HD_2, HD_3, HD_4, and HD_5 span the entire range of values with incremental intervals of 0.050. The complete set of combinations results in a total of (N) of 6908733 vectors, each comprising five elements.

Upon establishing the arrays of HD_{Mk}, it is necessary to subject these combinations to a twofold evaluation: stability *vs* dissociation, and stability *vs* hapticity. The first criterion is related to the likelihood of finding a newly described metal complex to be stable. This aspect is crucial in assessing the feasibility of the proposed metal complex. The second is based on comparing the preferred isomer for the hypothetical metal fragment. Thus, we gain insights into the promising candidates.

We started by examining the stability *vs* dissociation. This evaluation distinguishes between potentially present bonds and unstable ones that

break apart. Regarding Equation 3.4.1, BDE values can be predicted by introducing the three variables: HD_{Mk}, HD_{Wk} and HD_{H_2k}. As we were interested in analysing the stability of η^1-H_2 ligand, HD_{Lk} is $HD_{(\eta^1-H_2)k}$, leading to the customised formula,

$$BDE_{L_nM(\eta^1-H_2)_{pred}} = -22.935HD_{M1} - 11.832HD_{M2} - 1.870HD_{M3} + 21.528HD_{M4} + 1.504HD_{M5} \quad (3.4.2)$$

Equation 3.4.2 has to be filled with the new combinations of HD_{Mk} in order to predict BDEs. The resulting BDEs encompass numbers between -22.7 to 38.7 kcal·mol^{-1}. As explained during the Chapter, our focus is on BDEs with negative signs. Consequently, we refined the BDE dataset to encompass values below -10 kcal·mol^{-1}. The filtering imposed reduced the HD_{Mk} arrays to only 4.2 % of the initial combinations. We analyzed 287338 HD_k arrays, each associated with its respective BDEs. Notably, 287k combinations met the stability criterion.

We visualized the 287k metal vectors by using a plot that displays their five elements, HD_{Mk}, where k ranges from 1 to 5. However, we realized that such a graph would be overly complex and challenging to interpret. Therefore, we opted for a more manageable approach by utilizing two 2D plots, as illustrated in Figure 3.12, which represent the outcomes in terms of the descriptor values. Furthermore, the previously defined twenty-three metal fragments were included in the plots to provide context. Figure 3.12 neglects the HD_{M5} variable, which has a minimal effect on predicting BDE, and its chemical meaning could not be identified.

The Figure is split into two plots. The left plot displays the values of the HD_{M1} *vs* HD_{M2}, whilst the plot on the right illustrates the third and fourth hidden descriptors of the metal fragments. Different types of points appear in the scatter plots. The light green points refer to the 287k derived combinations (as indicated by the L_nM_{ideal} dots in the legend). The

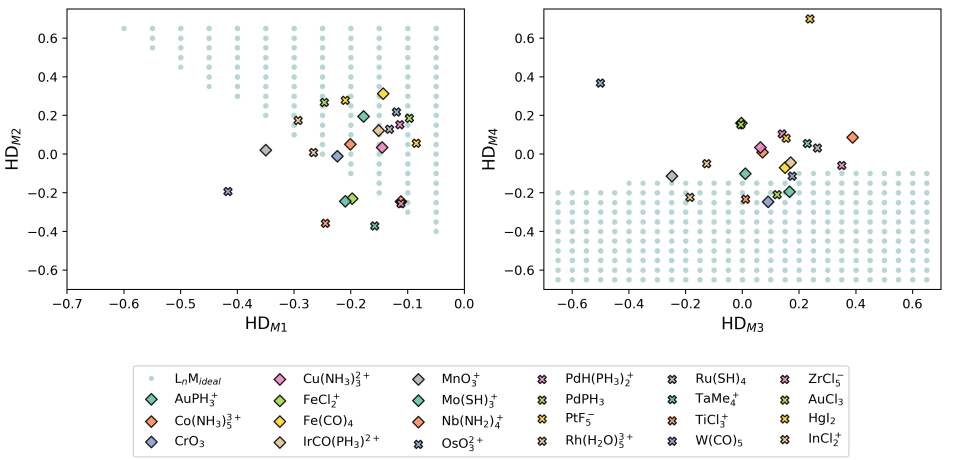


Figure 3.12: Scatter plots containing the values of the HD$_{M1}$ and HD$_{M2}$ (left), and displaying the values of the HD$_{M3}$ and HD$_{M4}$ (right). In the legend, there are the 23 metal fragments considered explicitly in this study, and the remaining green dots correspond to the predicted values for the potentially stable L$_n$M(η^1-H$_2$) complexes.

L$_n$M$_{ideal}$ dots on the plot represent a HD$_{Mk}$ array that yields the necessary BDE value below -10 kcal·mol^{-1} to form a metal complex with a monohapto H$_2$ ligand. The remaining dots, distinguished by their different shapes and colors, indicate the hidden descriptors of the metal chemical space from the preceding study. It is remarkable that each dot in the plot represents a HD$_{Mk}$ array, of which three elements are not visible in the graph.

In the first inspection, certain actual metal fragments were observed within the regions marked by the green points, hinting at promising outcomes. Nonetheless, it is important to note that even if the green dots share the same HD$_{M1}$ and HD$_{M2}$ values with a specific metal fragment, such as Co(NH$_3$)$_5^{2+}$ (an orange diamond on the left plot of Figure 3.12), they may not necessarily be the same metal. This outcome does not automatically imply that Co(NH$_3$)$_5^{2+}$ would be a suitable metal fragment for η^1-H$_2$ ligand. The results of Equation 3.4.2 are influenced by the HD$_{M3}$, HD$_{M4}$ and HD$_{M5}$ parameters as well. In the case of Co(NH$_3$)$_5^{2+}$, its complete HD$_{Mk}$ vector

does not furnish the desired BDE with η^1-H_2 ligand (BDE $(Co(NH_3)_5^{2+} - \eta^1$-$H_2$ = 3.4 kcal·mol^{-1}). Although limited by the constraints of a 2D plot, our analysis aimed to examine the desired features independently. When focusing on the HD_{M1} axis, green dots are concentrated in the vicinity of zero, indicating minimal desired σ interaction. On the other hand, regarding HD_{M2} parameter, the values are clustered in the regions with larger positive values, indicating to a preference for π donor metals.

Regarding the right plot of Figure 3.12, the L_nM_{ideal} dots present a different distribution compared to the left plot. The x-axis, which corresponds to the HD_{M3} variable, is populated without restrictions along the whole range of values. However, the ordinates confine the sought values to the lower part of the plot. In other words, the *cis* influence does not significantly affect the BDE result. Instead, the covalency term selects the metal fragments towards those favouring covalent interactions with the H_2 ligand. Understanding these aspects is crucial for identifying and characterizing an ideal fragment to form the $L_nM(\eta^1$-$H_2)$ complex. Additionally, we could also understand the outlined trends regarding Equation 3.12. To achieve a negative BDE value, the first term, HD_{M1} should be minimal or positive, HD_{M2} has to render a positive value, the HD_{M3} is not relevant to the result due to its low coefficient, and the fourth variable requires a negative value.

Upon discussing general trends about the stability criteria, it became necessary to examine the findings regarding the preference for either the η^1- or η^2-H_2 isomer. We decided to perform a similar analysis to that of the η^1- configuration but for the η^2-mode. The following Equation 3.4.3 is analogous to Equation 3.4.2, but here it applies to the dihapto ligand.

$$BDE_{L_nM(\eta^2-H_2)_{pred}} = -19.877HD_{M1} - 29.172HD_{M2} + 1.105HD_{M3} + 8.142HD_{M4} - 0.987HD_{M5} \quad (3.4.3)$$

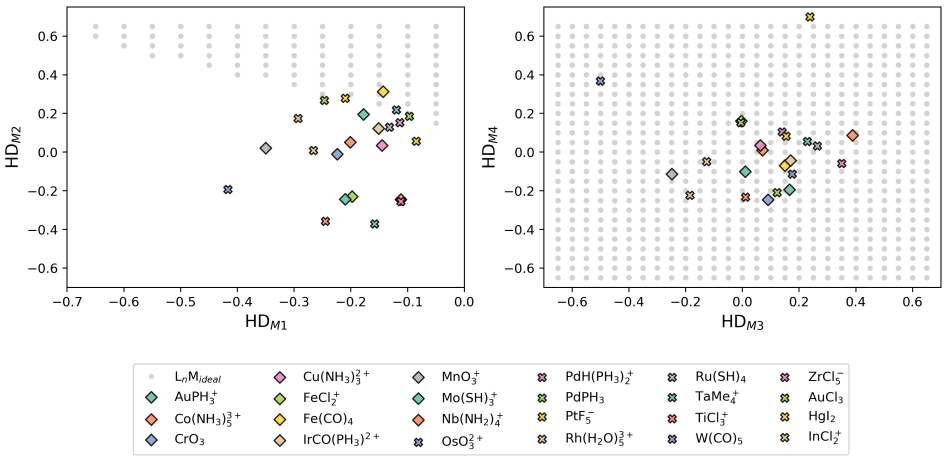


Figure 3.13: Scatter plots containing the values of the HD$_{M1}$ and HD$_{M2}$ (left) and displaying the values of the HD$_{M3}$ and HD$_{M4}$ (right). In the legend, there are the 23 metal fragments considered explicitly in this study, and the remaining grey dots correspond to the predicted values for the potentially stable L$_n$M(η^2−H$_2$) complexes.

The objective is to gain insights into the metal fragments that also will form stable metal complexes with η^2-H$_2$. By filling the gaps of the Equation 3.4.3 with the same array combinations as before ($N = 6908733$), we obtained BDEs ranging from -24.6 to 38.5 kcal·mol^{-1}. This represents a difference between η^1- and η^2-H$_2$ isomers. The range of BDE values before filtering is slightly narrower for the monohapto ligand than for the dihapto. This suggests a smaller capacity of the hypothetical metal fragments to develop stable complexes with the non-traditional ligand. After reducing the combinations to a set that encompasses BDE values below -10 kcal·mol^{-1}, we maintained 641434 vectors, 9.3 % of the initial combinations. Again, the number of potential combinations surpasses the ones of the understudied ligand.

Figure 3.13 shows the HD$_{Mk}$ values of the L$_n$M$_{ideal}$ points that furnish metal complexes with dihydrogen molecules bound in an η^2 fashion. Comparing Figures 3.12 and 3.13, we can draw some observations. Firstly,

a higher quantity of grey dots (Figure 3.13) is observed than green dots (Figure 3.12). Secondly, the left plot in Figure 3.12 shows a greater concentration of L_nM_{ideal} dots than in Figure 3.13. Yet, the distribution trend remains similar. HD_{M2} must be presented with a positive value, indicating that in both cases the metal must be a π donor ligand. On the other hand, the right plots are crucial in determining the chosen isomer. The plot on the right side of Figure 3.12 is distinctly separated according to the HD_{M4} sign. Nevertheless, in the right plot of Figure 3.13, most of the HD_{M3} and HD_{M4} combinations under consideration are the best option to attain the $L_nM(\eta^2\text{-}H_2)$ complex.

After examining the overall characteristics of each hapticity, it was necessary to assess the inclination towards conventional or unconventional ligands using numerical parameters.

In fact, tracing back to Figure 3.9, we subtracted the $BDE(L_nM(\eta^2\text{-}H_2))$ from the $BDE(L_nM(\eta^1\text{-}H_2))$ to determine the most likely isomer to form stable metal complexes. Performing the same mathematical operation on Equations 3.4.2 and 3.4.3 results in the following equation,

$$\Delta BDE_{pred}((\eta^1 - H_2) - (\eta^2 - H_2)) = -3.058HD_{M1} + 17.340HD_{M2} - 2.975HD_{M3} + 13.386HD_{M4} - 2.491HD_{M5} \quad (3.4.4)$$

Equation 3.4.4 predicts the difference in energy for each of the H_2 configurations. If the mathematical operation results in a negative value, this will correspond to metal fragment instances that favour the end-on-bonded configuration. Whereas the positive values stand for the side-on-bonded disposition. The number of negative ΔBDE was 3139714 (45.4 %) between -22.7 to 37.0 kcal·mol^{-1}. Upon filtering the values to selectively obtain the ones lower to -10 kcal·mol^{-1} for $BDE(L_nM(\eta^1\text{-}H_2)$, we obtained a total of 120788 combinations, which only accounts for 1.7 % of the total number of combinations. The results are illustrated in Figure 3.14 using

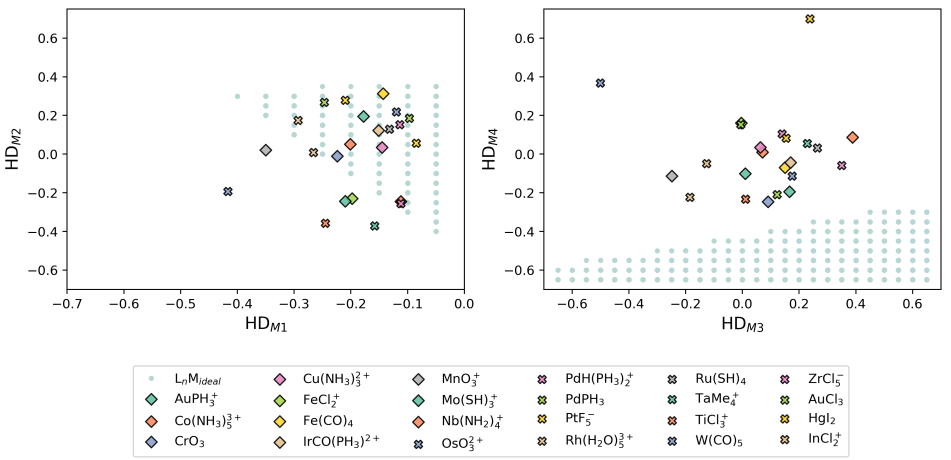


Figure 3.14: Scatter plots containing the values of the HD_{M1} and HD_{M2} (left) and displaying the values of the HD_{M3} and HD_{M4} (right), after the filter of stability and the preference for the isomer. In the legend, there are the 23 metal fragments considered explicitly in this study, and the remaining green dots correspond to the predicted values for the potentially stable $L_nM(\eta^1\text{-}H_2)$ complexes.

the same template as Figure 3.12.

The first concern regarding Figure 3.14 is that the number of green dots has dramatically decreased compared to Figure 3.12. This implies that the preference for one or other isomer is a key factor in the search of the ideal ML_n for η^1-H_2.

In the left plot of Figure 3.14 the 23 metal fragments still fall within the green area. However, in the right plot, none of the defined metal fragments are within the grid region. This implies that HD_{M1}, HD_{M2}, and HD_{M3} for the hypothetical metal fragment are located in ranges of hidden descriptors that fit with the known metal fragments. Instead, HD_{M4} plays a pivotal role. The covalency property happens to concentrate the candidates L_nM_{ideal} in the negative values, thus, in the metal fragments favouring mainly covalent interactions.

To provide a joint understanding of the result, we created a 3D plot in Figure 3.15 using the first four hidden descriptors. The plot indicates clear

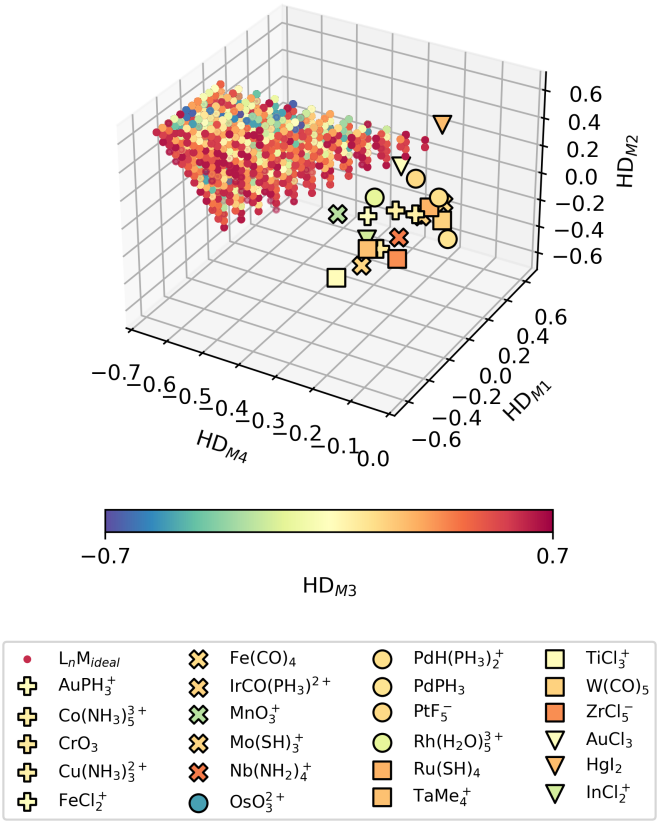


Figure 3.15: 3D plot of the hidden descriptors for metal complexes HD$_{Mk}$. In the legend, there are the 23 metal fragments considered explicitly in this study, and the remaining dots correspond to the predicted values for the potentially stable L$_n$M(η^1-H$_2$) complexes.

separation between the dots and the set of 23 metals. Additionally, the color of the closest L_nM_{ideal} dots does match with any of the defined known metals. Therefore, we concluded that none of the proposed metal species are suited to our problem.

3.5 Conclusions

This chapter focused on the application of *BDE Matrix App* to the study of N-heterocyclic carbenes and monohapto dihydrogen ligand in metal complexes. The first part of this Chapter was dedicated to the analysis of the electronic properties of N-heterocyclic carbene ligands. To achieve that, we calculated the bond dissociation energies between a set of twenty-two NHCs with the set of five metal fragments indicated in the *BDE Matrix App*. Such metal fragments were OsO_3^{2+}, $PdPH_3$, $PdH(PH_3)^{2+}$, $ZrCl_5^-$, and $InCl_2^+$. Upon computing a total of 110 DFT-BDE points, we introduced that values into the App to obtain the HD of each of the NHC ligands. Even if these ligands are endowed of σ donating and π accepting capacity, HD values assist in differentiating within them. We identified as the strongest sigma donor PyC4-3,5-Me$_2$*N*H (HD_{L1} = 0.239), and the least donor Im(NO$_2$)$_2$*N*Me$_2$ (HD_{L1} = 0.151). Among the π accepting property, we located as the greatest π acceptor sPm*N*Me$_2$ (HD_{L2} = -0.320), and the poorest DPyIm (HD_{L2} = -0.155). This provided a global picture of chemical species selected. Notably, we also elucidated the trends of the electronic properties within each family core. For instance, saturated and unsaturated imidazoles can be structurally customized with EWG and EDG in their N-subsituents to vary their σ and π interactions. Moreover, the modification of C4 and C5 substituents affects the sigma properties of the ligands together with the presence of unsaturations on the ring of the ligand. Next, we compared the ability of the first hidden descriptors to account for σ donation with other recognized descriptors such as E_{HOMO} and NBO-properties. In our case, the HD_{L1} unveiled a better understanding of such properties in our

chemical space. Finally, NHCs were subjected to a comparison with other well-known ligands. The results showed that the NHCs are the greatest σ donors together with the CP. However, CP ligands offered a lower flexibility to vary their π interactions compared to the NHCs. The phosphines selected were identified as the least donors.

In the second part, we attempted to identify metal fragments to form potentially stable $L_nM(\eta^1\text{-}H_2)$ complexes using the hidden descriptors. To do so, we employed the *BDE Matrix App* to obtain the hidden descriptors of the H_2 in the monohapto and dihapto arrangements. It was necessary to calculate the BDEs between the H_2 in each respective disposition, with five metal fragment of reference already mentioned. The five generated hidden descriptors provided valuable insights into the electronic variations of H_2 based on its configuration within the complexes. Our exploration focused on discerning metal fragments that stabilize the $L_nM(\eta^1\text{-}H_2)$ complex by evaluating stability against dissociation, and preference for a particular isomer. Upon an assessment of BDE between the $\eta^1\text{-}H_2$ ligand and a set of twenty-three known metal fragments, we concluded that none metal proposed was a suitable candidate for monohaptic binding. Expanding our investigation beyond these twenty-three metal fragments enabled the exploration of alternative electronic motives in metals. We formulated hypothetical metal fragments by incorporating HD_{Mk} values into vacant arrays, resulting in a vast number of arrays, 6908733 in total. These arrays presented the potential to identify theoretical metal fragments characterized by the five hidden descriptors. Evaluating the hypothetical metal fragments, *i.e.* HD_{Mk} arrays, against the criteria of stability and preference for the isomer, diminished the likelihood of finding a metal species that closely matched our requirements. Unfortunately, none of the proposed combinations yielded similar features with the already known metal fragments to achieve the desired $L_nM(\eta^1\text{-}H_2)$ complex. Interestingly, in the discrimination of proposed HD_{Mk} combinations, HD_{M4} emerged as a pivotal element. This property, associated with the covalency percentage

in the M–L bond, needed to exhibit a strongly negative value to achieve the desired metal complex. After an in-depth analysis, we did not pinpoint a suitable metal candidate for η^1-H_2. However, this study helped in the comprehension of the required metal electronic properties for such ligand. Further studies varying the ligands belonging to known metal fragments or exploring overlooked metal fragments that hold σ donating properties can reveal new ways of coordination chemistry domain.

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

Chapter 4

Bimolecular nucleophilic substitution

> We can do anything now that scientists have invented magic.
>
> — Marge Simpson – The Simpson

4.1 Introduction

In the previous chapter, the hidden descriptor method was successfully applied to unravel the thermodynamic aspects of the M−L bond. In this chapter we go one step further by investigating the kinetic aspects of an organic reaction. To accomplish this goal, we selected the bimolecular nucleophilic substitution (S_N2) reaction at sp^3 carbon centers (S_N2@C).

This is a well-known transformation that participates in the synthesis of a broad range of functionalised products. The mechanism of the reaction is well understood and is depicted in Figure 4.1. The orbital interaction causes that a lone pair of the entering group (EG) interacts with the empty σ^* antibonding orbital of the leaving group (LG), which is polarized towards the carbon. This transformation implies a Walden inversion which leads to

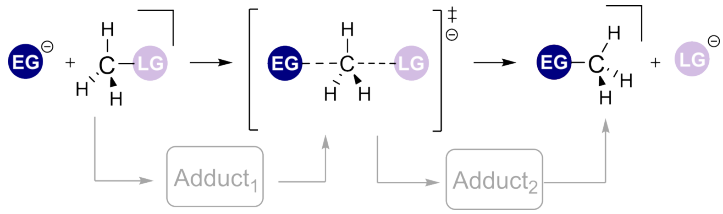


Figure 4.1: Scheme of the S_N2@C reactions. EG is the entering group, LG is the leaving group, and Add_1 and Add_2 are the reactant and product adducts, respectively.

a change of the configuration at the carbon center.

The aforementioned backside attack is widely understood by the community. It is worth mentioning that, it is also possible to carry out a frontside attack, where the configuration is retained.[157] However, in most cases, the latter method is not preferred because of the higher barrier.

The exact shape of the free energy profile depends on the entering group, the leaving group, and the solvent. Figure 4.2 illustrates the representative free energy profile shapes throughout the reaction at carbon centers.[158,159] An unimodal potential free energy profile is present under polar solvents where the reaction proceeds from reactants, transition state to products (grey line in Figure 4.2). In contrast, for apolar solvents, it is common to characterize double-well profiles, where the reactants fall into the reaction complex (Add_1) and move forward the TS until the next product complex (Add_2), resulting in the products (blue line in Figure 4.2).

From the beginning of computational chemistry, many studies have tackled aspects of this reaction starting from HF-based methods[160,161], DFT[162–164] to ML[165]. It is worth highlighting the in-depth work of Bickelhaupt and co-workers addressing this reaction through the application of different techniques, such as the activation strain model (ASM).[157,166]

Along this Chapter, we will be focusing on the derivation of HDs for nucleophiles in the S_N2@C reaction. Upon application of the SVD algebraic operation, we will identify the hidden descriptors that accurately correlate

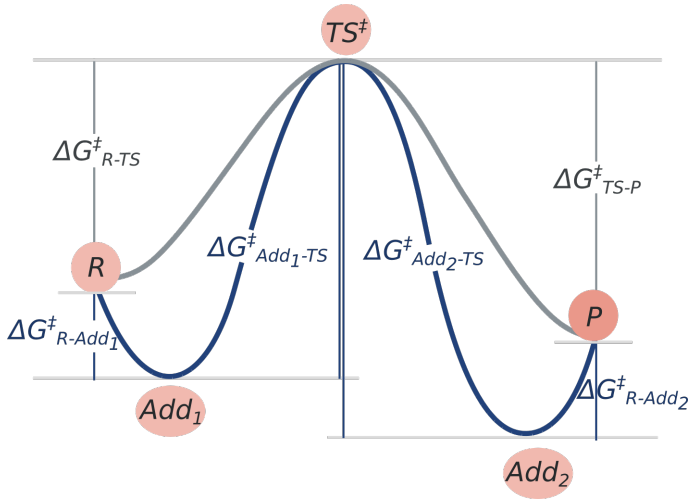


Figure 4.2: Schematic representation of the free energy profile within the S_N2@C reaction in the apolar solvent (blue line) and in the polar solvent (grey line). R denotes the reactants, TS is the transition state, P is the product, Add refers to the adduct complex.

with the kinetics of the reaction. Finally, we will develop a program that straightforwardly predicts HDs for out-of-sample nucleophiles. The general goal of this work is to appraise the main driving forces underlying the bimolecular nucleophilic substitution (S_N2) reaction.

4.2 Computational details

All DFT calculations were performed using Gaussian 16 package.[167] For geometry optimizations and frequency calculations B3LYP-D3[140,141] (D3 states for Grimme-D3 dispersion corrections) with 6-311+G(d)[168–170] was used for all the elements except for Br and I, where an ECP[171] together with the LANL2DZdp[172] basis set was employed. The vibrational frequency calculations were computed on optimized geometries with the default temperature (298.15K) and 1 atm of pressure to ensure the nature of the stationary points. A minima is identified when no imaginary frequencies

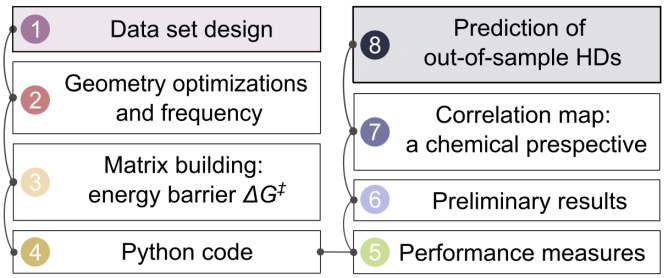


Figure 4.3: Scheme of the workflow followed along this Chapter.

are obtained and a transition state if one imaginary frequency is present. Solvent effects were introduced implicitly *via* the SMD[173] model for water and dichloromethane (DCM). The pyssian library was employed as a managing tool for the processing of the input and output files.[149] All the reported energies correspond to the free energies in solution and in kcal·mol^{-1}. A data set collection of computational results is available in the ioChem-BD repository,[19] accessible via http://dx.doi.org/10.19061/iochem-bd-1-215.

4.3 HD method

Figure 4.3 depicts the sequential steps performed for this particular chemical problem: (1) define the set of reactions to compute, (2) carry out the optimization and frequency calculations, (3) construct the energy barrier matrix, (4) conduct the singular value decomposition with the Numpy library in Python, (5) evaluate the resulting decomposition and decide the number of optimal descriptors, (6) relate the hidden descriptors with chemical concepts, (7) predict new HDs for new nucleophiles and (8) design a publicly open app and make it available to the scientific community.

4.4 Choice of the nucleophiles

The careful selection of chemical species is crucial in any data-driven approach. In this study, we selected a series of twenty-six different chemical fragments acting as both entering and leaving groups. Then, we computed the reactions at the methyl center considering all possible combinations between these chemical species. Our selection was based on previous computational studies,[174–176] and as a result, we built a square matrix with twenty-six rows (one for each entering group) and twenty-six columns (one for each leaving group) for the two solvents considered: water and dichloromethane (DCM).

We acknowledge that the term *nucleophile* can be mistaken for the concept of *entering group*. However, since we are considering the chemical fragment's ability to enter and leave a carbon center, we redefined the reaction elements to entering groups and leaving groups. Therefore, the term nucleophile collects all the chemical species capable of entering and releasing from the carbon center of the reaction.

The nucleophiles considered included halogens (F^-, Cl^-, Br^-, I^-), hydroxide groups (HO^-, MeO^-, EtO^-), α-nucleophiles ($HCOO^-$, CH_3COO^-, HOO^-, $HC(=O)OO^-$), tosylate (TsO^-), triflate (TfO^-), thiolates (HS^-, MeS^-, EtS^-), selanide (HSe^-), azanide (H_2N^-), formamidate ($CHOHN^-$), amines (H_3N, $(CH_3)_3N$), phosphino (H_2P^-), arsinide (H_2As^-), phenyl ($H_5C_6^-$), trifluoromethyl anion (F_3C^-) and cyanide (NC^-). We ended up considering 26 nucleophiles, and their combinations gave rise to a total of 676 reactions.

4.5 Matrix of free energy barriers

4.5.1 Definition of the reference state

As explained in Section 4.1 (*vida supra*), there are two potential reference states for the calculation of the free energy barriers: the separate reactants

(R) or the reactant complex (Add). Aiming to determine the most stable intermediate of the transformation, we performed a preliminary study with seventeen nucleophiles acting as EG and as LG (energies are depicted in the Appendix B). The activation energies were considered as the difference between the TS and the reactants, $\Delta G^{\ddagger}_{TS-R}$, and as the difference between the TS and the Add, $\Delta G^{\ddagger}_{TS-Add}$. To evaluate the adequate energy reference, Figures 4.4 and 4.5 show the energy barrier difference between these two terms.

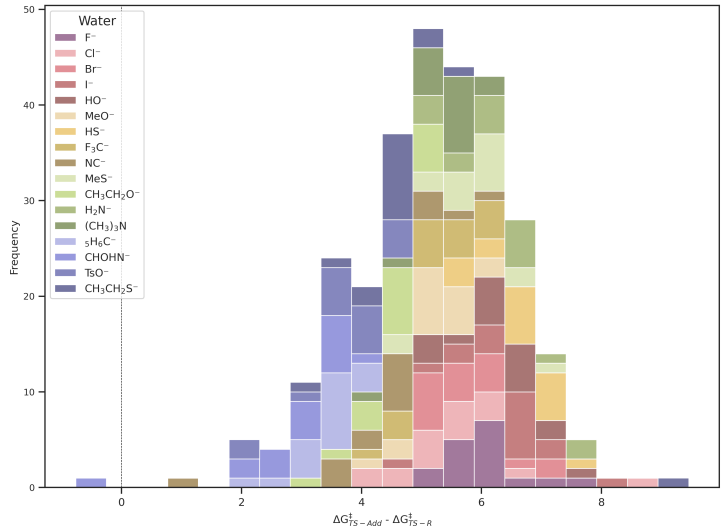


Figure 4.4: Histrogram of the free difference (in kcal·mol^{-1}) between $\Delta G^{\ddagger}_{TS-R}$ and $\Delta G^{\ddagger}_{TS-Add}$ in water. Colors denote the reactions according to their entering groups groups.

In the pair of bar plots, bars located at positive values correspond to instances where $\Delta G^{\ddagger}_{TS-Add}$ is lower than $\Delta G^{\ddagger}_{TS-R}$, indicating that the energy of the adduct is higher than that of the separate reactants, while negative values refer to the opposite trend. In both Figures 4.4 and 4.5, most values are located to the right side of the horizontal black line, thus, the $\Delta G^{\ddagger}_{TS-Add}$ is lower than $\Delta G^{\ddagger}_{TS-R}$. It is noteworthy that the energy separation between Add and R is more pronounced in polar solvents

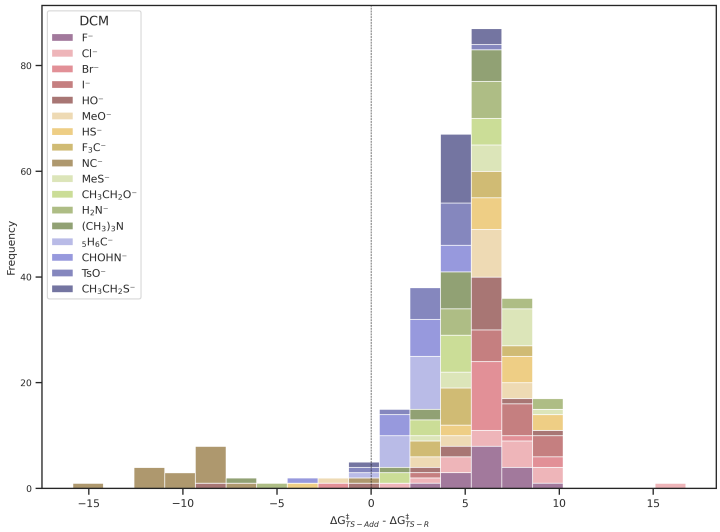


Figure 4.5: Histogram of the free energy difference (in kcal·mol^{-1}) between $\Delta G^{\ddagger}_{TS-R}$ and $\Delta G^{\ddagger}_{TS-Add}$ in dichloromethane. Colors denote the reactions according to their entering groups groups.

compared to apolar media. This observed behavior is consistent with our initial definitions that when transitioning towards less polar solvents, the energy differences between these two minima decrease, and the rate constant starts to be governed by the adduct complex. For instance, in the dichloromethane medium, the cyanide EG (depicted in brown color in Figure 4.5) achieves greater stability by forming a stable adduct with the electrophile rather than remaining separate. Nonetheless, to ensure consistency, and in light of the overview trends, the discussion will focus on the $\Delta G^{\ddagger}_{TS-R}$ values.

4.5.2 Separate reactants as reference

Upon analysing the relative energies, we defined a 26 × 26 matrix, $\boldsymbol{\Delta G^{\ddagger}_{TS-R}}$, containing 676 activation energy values. Henceforth, the term $\Delta G^{\ddagger}_{TS-R}$ is simplified to $\Delta G^{\ddagger}$. Initial inspection of the data is shown in Figure 4.6.

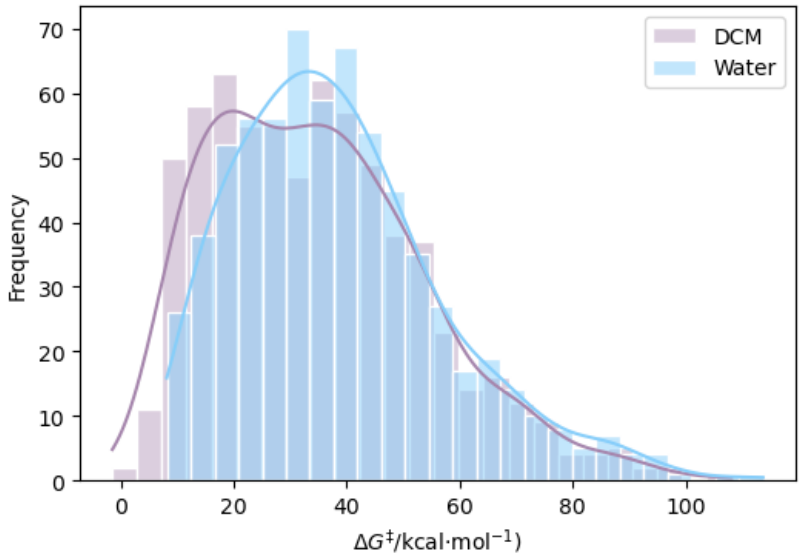


Figure 4.6: Histogram analysis of the free energy barrier (in kcal·mol^{-1}) distribution for water (blue) and dichloromethane (purple). Lines are referred to the KDE distribution plots.

Through the histogram and Kernel Density Estimate (KDE) visualization in Figure 4.6, we can get a grasp on the energy barrier distributions for both solvents. The KDE plot is a method for visualizing the distribution of observations in a dataset using a continuous probability density curve. As expected the absolute values of $\Delta G^{\ddagger}$ are different in each reaction media. Concerning the activation energy values in water, the value range spans from 8.1 to 113.7 kcal·mol^{-1} and in dichloromethane, from 10 to 108.1 kcal·mol^{-1}. Such higher values would not take place in experimental conditions. Nevertheless, the existence of trends in the data suggests that the problem is well-suited for a statistical treatment. The same reactions are more energetically demanding in water than in DCM. When the polarity of the solvent increases, the anionic nucleophile is better solvated than the TS. The Nu is solvated with its charge confined in a small area, but TS contains the separated negative charges along the bigger TS, and thus, the energy

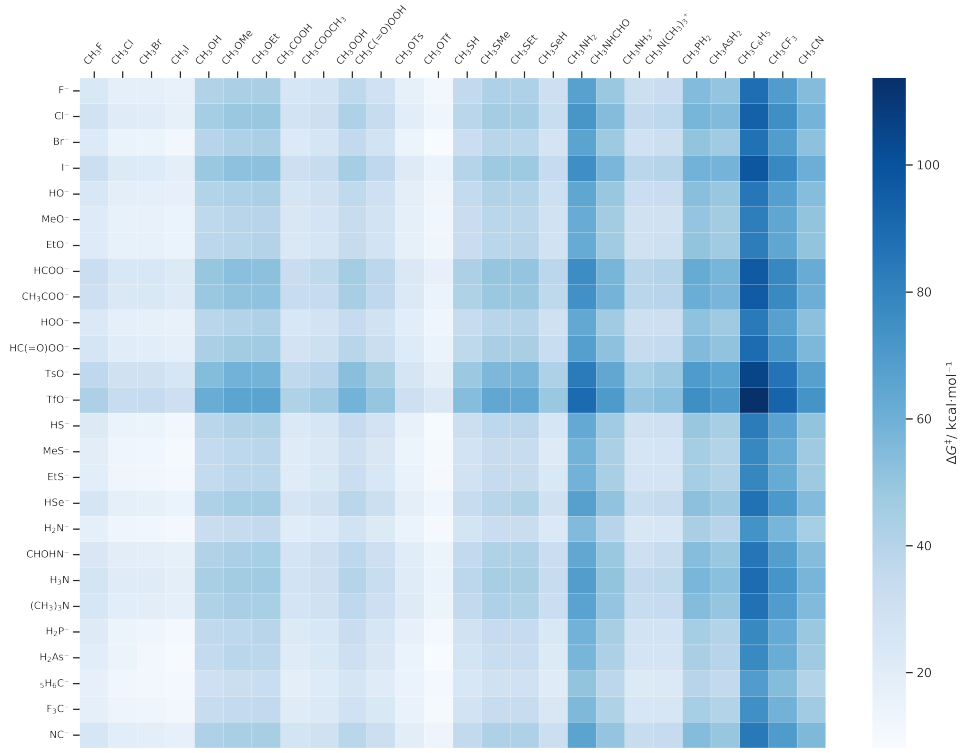


Figure 4.7: Color-coded version of the matrix of the free energy barriers (in kcal·mol^{-1}) from the reactants to the transition state, $\Delta G^{\ddagger}$.

difference between these two stationary points increases in polar solvents.

In addition, Figures 4.7 and 4.8 are color-coded matrices for the activation energy of the reactions in water and DCM, respectively.

We collected the energy barriers between an entering group i and a leaving group j, $\Delta G^{\ddagger}_{i,j}$, as entries of the matrix $\boldsymbol{\Delta G^{\ddagger}}$ where the i and j are the elements in the i^{th} row and the j^{th} column of such matrix, respectively. Therefore, entering groups are arranged in rows, and leaving groups correspond to the columns. Preliminary examination reveals that the color trends are similar in both Figure 4.7 and 4.8 regardless of solvent. In both heatmaps, the $\Delta G^{\ddagger}$ are ruled by the columns rather than the rows, indicating the dominant role of the leaving group in the kinetics of the

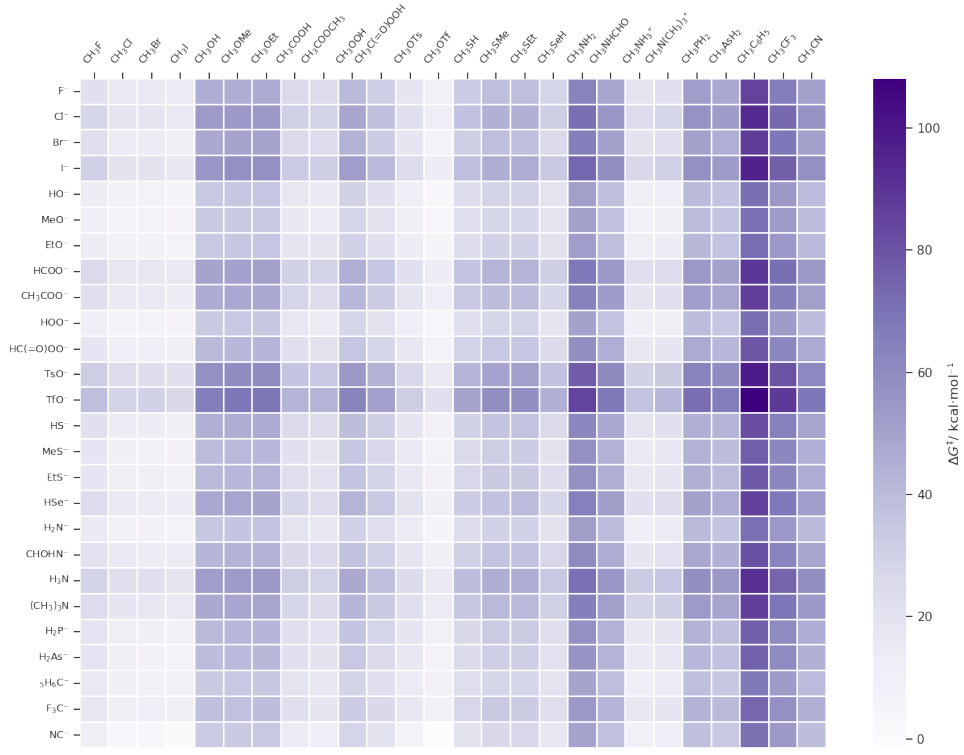


Figure 4.8: Color-coded version of the matrix of the free energy barriers (in kcal·mol^{-1}) from the reactants to the transition state, $\Delta G^{\ddagger}$.

reaction. Darker color in the matrix shows higher $\Delta G^{\ddagger}$, as in the case of the amine, phenyl, and triflate leaving groups. In the opposite behaviour, there are the triflate and tosilate as excellent leaving groups.

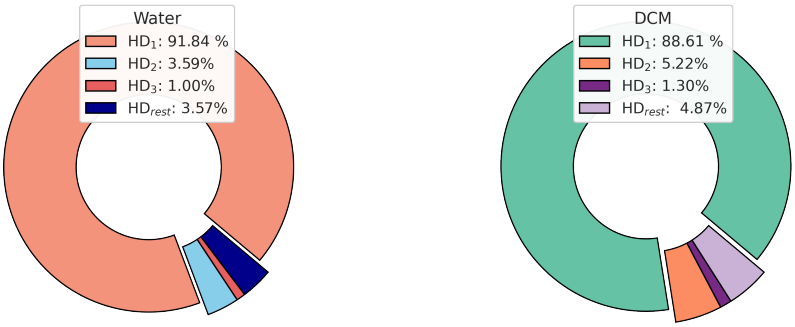
4.6 Calculation of the HDs

Following the procedure outlined in Chapter 2, we carried out the SVD operation over the $\boldsymbol{\Delta G^{\ddagger}}$ matrix. We used NumPy package[78] to decompose $\boldsymbol{\Delta G^{\ddagger}}$ *via* SVD as shown in the equation below 4.6.1,

$$\boldsymbol{\Delta G^{\ddagger} = EG_{n\times n} \cdot W_{n\times n} \cdot LG^{T}_{n\times n}} \tag{4.6.1}$$

where n is the number of nucleophiles under study. The decomposition involved a $n \times n$ unitary matrix ($\boldsymbol{EG}$) which is related to the row variable and thus, to the entering groups; a $n \times n$ diagonal matrix ($\boldsymbol{W}$) representing the positive hidden descriptor weights; and a $n \times n$ matrix ($\boldsymbol{LG^{T}}$) for the column variable, therefore, for the leaving groups. As mentioned in the Chapter 2, the SVD operation is hierarchical, thus, we can reduce the dimensionality of the matrices while maintaining the main chemical information of our dataset. The aim is to choose the smallest number of descriptors that adequately replicate the $\Delta G^{\ddagger}_{i,j}$ matrix's values with an acceptable level of precision. To do so, a measure of the HD weights is conducted. Figure 4.9 depicts the relative percentage of weight (%W) of the first hidden descriptors in the total barrier.

A first inspection of the pie charts shows the very large role of the hidden descriptor 1 (HD_1). HD_1 contains approximately 90% of the overall barrier in both solvents. While the other 25 descriptors do have an impact, we believe that it is so small that it cannot be easily distinguished from statistical fluctuations. After recognizing the main role of HD_1, we used energy criteria to confirm its importance. Thus, the optimal number of hidden descriptors is also elucidated by applying Equation 4.6.2 varying the



(a) Hidden descriptor weights in water.

(b) Hidden descriptor weights in DCM.

Figure 4.9: Pie charts of the Weight % of the hidden descriptors in the S_N2 reaction.

values of hidden descriptors considered, namely k, from 0 to n.

This variation in the number of hidden descriptors, k, included in Equation 4.6.2 refers to the number of vectors maintained in the matrices $\boldsymbol{EG}$, $\boldsymbol{W}$, $\boldsymbol{LG}$. As we want to know which vectors of these matrices are relevant, we refer to these vectors as the hidden descriptors: $HD_{k,EG}$, $HD_{k,W}$, $HD_{k,LG}$.

$$\boldsymbol{\Delta G^{\ddagger}_{pred,k} = HD_{k,EG} \cdot HD_{k,W} \cdot HD_{k,LG}} \tag{4.6.2}$$

Equation 4.6.2 is the truncated form of Equation 4.6.1, where only the first k columns of matrix **EG** and the first k columns of matrix **LG** are maintained. The predicted matrix $\boldsymbol{\Delta G^{\ddagger}_{pred,k}}$ will not reproduce exactly the initial matrix $\boldsymbol{\Delta G^{\ddagger}}$, and the difference between them will provide an evaluation of the accuracy achieved when k descriptors are used. Figure 4.10 presents the maximum error and average error when comparing $\Delta G^{\ddagger}_{pred,k}$ values with DFT-energies using up to six k. The value corresponding $k = 0$ refers to the average and maximum value of the activation energy of the data when 0 descriptors are used.

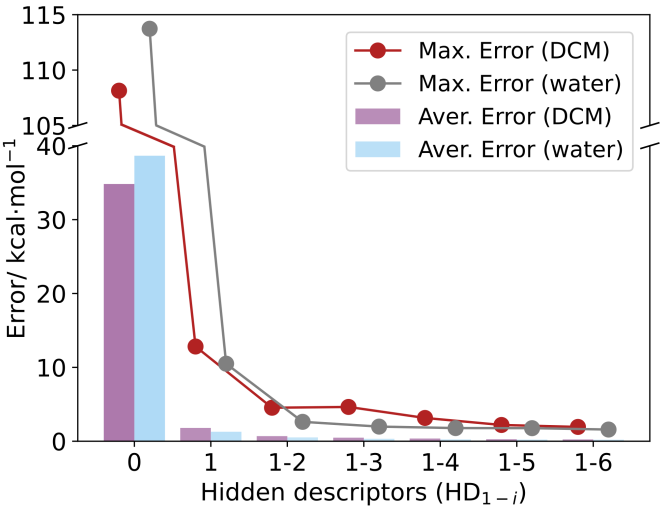


Figure 4.10: Performance measures for the $\Delta G^{\ddagger}_{pred,k}$ values under DCM (purple bar and red dots) and water (blue bar and gray dots) as a function of the number of hidden descriptors considered (HD_{1-i}). Average errors in bars and maximum error in dots (kcal·mol^{-1}).

Despite the differences in the dielectric constants of the solvents, the performance measurements are comparable. The average value of the $\Delta G^{\ddagger}$ is between 30.0 and 40.0 kcal·mol^{-1} for both solvents. Therefore, the average error of 2.0 kcal·mol^{-1} in the prediction with only one HD ($k = 1$) is rather small considering that some DFT functionals can also provide errors with an amount of inaccuracy.[177] Furthermore, the use of only HD_1 results in a maximum error of 12.8 kcal·mol^{-1} for DCM and 10.5 kcal·mol^{-1} for water. Moving forward an increment in the number of descriptors does not significantly improve the prediction performance. Consequently, it is reasonable to claim that a single descriptor sufficiently characterizes this reaction. The aforementioned indicates that only one descriptor allows for assigning two values to each nucleophile: one for its entering ability, HD_{EG}, and another for its leaving capacity, HD_{LG}, during the S_N2 reaction.

We further confirmed the solidness of our data set by carrying out a series of tests to assess its stability. Initially, a new submatrix for the

energy barrier was created by removing one EG(LG) element at a time, which is the leave-one-out (LOO) technique, followed by an SVD analysis. Subsequently, a comparison was made between the new HD_1 values and the directly derived HD_1 values from the computed 26 × 26 $\boldsymbol{\Delta G^{\ddagger}}$. Herein, the differences in the HD_1 were minimal, i.e. lower than 0.009, which confirmed the stability of the data set. In a second analysis, a distinct energy barrier submatrix was designed by removing a set of EG(LG) elements either randomly or selectively *via* cross-validation. This new submatrix was then subjected to SVD analysis, and as well, its HD_1 values were compared to those obtained from the 26 × 26 $\boldsymbol{\Delta G^{\ddagger}}$. As expected the dissimilarities between the original set and the one created increased as the size of the new energy submatrix decreased. The fact that the error increased linearly further proves the stability of the model.

4.7 Analysis of the HD_1

Tables 4.1 and 4.2 show the values of the HD_1 descriptor. Several insights can be gleaned from these two tables. We have established two scales, HD_{1EG} and HD_{1LG}, to characterize the entering and leaving group ability of twenty-six nucleophiles, respectively. Throughout the remaining analysis, HD_{1EG} and HD_{1LG} are also denoted as HD_{EG} and HD_{LG}, respectively. This simplification is made because we focus solely on the first hidden descriptor as the most significant in the investigation of the reaction. Higher values of these variables, HD_{EG} and HD_{LG}, refer to a diminished ability to attack or leave the sp^3 carbon center. Conversely, lower HD values state the opposite trend. In water, the range of values for entering ability is notably narrower (0.135 units ranging from 0.145 to 0.280) than that of leaving ability (0.338 units ranging from 0.060 to 0.398). This pattern is similar in dichloromethane medium: the EG ability spans from 0.146 to 0.289, and the LG ability covers the range from 0.049 to 0.406. Hence, as previously anticipated, the barrier is more sensitive to the nature of the LG.

Table 4.1: Values of the HD_1 in water.

Water			
EG	$\mathbf{HD}_{1EG}$	**LG**	$\mathbf{HD}_{1LG}$
TfO^-	0.280	TfO^-	0.060
TsO^-	0.255	I^-	0.076
$HCOO^-$	0.228	Br^-	0.086
$CH3COO^-$	0.224	Cl^-	0.088
I^-	0.223	TsO^-	0.089
Cl^-	0.212	F^-	0.118
H_3N	0.206	$HCOO^-$	0.122
$HC(=O)OO^-$	0.203	CH_3COO^-	0.136
$(CH_3)_3N$	0.197	HSe^-	0.141
HSe^-	0.196	$HC(=O)OO^-$	0.143
NC^-	0.196	H_3N	0.151
F^-	0.195	$(CH_3)_3N$	0.156
$CHOHN^-$	0.194	HS^-	0.162
HO^-	0.191	HOO^-	0.175
HOO^-	0.186	HO^-	0.192
Br^-	0.186	MeS^-	0.192
EtO^-	0.181	EtS^-	0.193
HS^-	0.180	MeO^-	0.204
MeO^-	0.180	EtO^-	0.206
H_2P^-	0.169	H_2As^-	0.228
EtS^-	0.167	$CHOHN^-$	0.230
MeS^-	0.165	H_2P^-	0.246
H_2As^-	0.164	NC^-	0.249
F_3C^-	0.161	H_2N^-	0.305
H_2N^-	0.156	F_3C^-	0.319
$H_5C_6^-$	0.145	$H_5C_6^-$	0.398

Table 4.2: Values of the HD_1 in dichloromethane.

Dichloromethane			
EG	**HD**$_{1EG}$	**LG**	**HD**$_{1LG}$
TfO^-	0.289	TfO^-	0.049
TsO^-	0.254	I^-	0.062
I^-	0.240	Br^-	0.071
H_3N	0.234	Cl^-	0.072
Cl^-	0.229	TsO^-	0.094
$HCOO^-$	0.219	H_3N	0.098
$(CH_3)_3N$	0.213	F^-	0.108
HSe^-	0.207	$(CH_3)_3N$	0.111
CH_3COO^-	0.206	CH_3COO^-	0.117
Br^-	0.205	$HCOO^-$	0.127
F^-	0.202	HSe^-	0.132
HS^-	0.193	HS^-	0.154
$CHOHN^-$	0.192	$HC(=O)OO^-$	0.155
$HC(=O)OO^-$	0.183	MeS^-	0.185
EtS^-	0.181	EtS^-	0.187
H_2P^-	0.178	HOO^-	0.198
MeS^-	0.176	HO^-	0.220
H_2As^-	0.173	H_2As^-	0.222
F_3C^-	0.172	MeO^-	0.226
H_2N^-	0.158	EtO^-	0.228
EtO^-	0.157	$CHOHN^-$	0.238
HO^-	0.154	H_2P^-	0.243
$H_5C_6^-$	0.153	NC^-	0.244
HOO^-	0.151	H_2N^-	0.306
MeO^-	0.151	F_3C^-	0.320
NC^-	0.146	$H_5C_6^-$	0.406

An initial analysis of the chemical fragment ranking leads to an overall agreement of the trends with expectations. For instance, TfO^- has a HD_{1EG} (water) = 0.280, whereas $H_5C_6^-$ has HD_{1EG} (water) = 0.145. This suggests that the entering group ability of TfO^- is comparatively weaker than that of $H_5C_6^-$, implying a slower reaction rate for the alkoxide compared to the phenyl anion. In the case of the leaving group scale, the pattern is qualitatively the contrary. The worse LG within our chemical space is $H_5C_6^-$ with HD_{1LG} (water) = 0.398, and the best LG is the TfO^- with a HD_{1LG} (water) of 0.060. The HDs and trends are comparable in the dichloromethane solvent, where the best EG is the cyanide (HD_{1EG} (DCM) = 0.146), whereas the worse is the triflate HD_{1EG} (DCM) = 0.280. In the case of the leaving group, TfO^- is endowed with the highest leaving capacity (HD_{1LG} (DCM) = 0.049), and the lowest releasing capacity belongs to the phenyl anion (HD_{1EG} (DCM) = 0.406).

In addition, several interesting trends can be observed in Tables 4.1 and 4.2. The following values are referred to: (HD (water), HD (DCM)). Poor EGs include those where the attacking atom is a carbon, $H_5C_6^-$ (0.146, 0.153), F_3C^- (0.161, 0.172). We also observed that for some good LGs, their interacting atom belong to the right groups of the periodic table. HD_{1LG} values decrease on going down along the periodic table for the halide group: F^- (0.118, 0.108) > Cl^- (0.088, 0.072) > Br^- (0.086, 0.071) > I^- (0.076, 0.062); for the chalcogen group: HO^- (0.192, 0.220) > HS^- (0.162, 0.154) > HSe^- (0.192, 0.132); and for the nitrogen group H_2N^- (0.305, 0.306) > H_2P^- (0.246, 0.243) > H_2As^- (0.228, 0.222). The effect of increasing the length of the carbon chain in the oxygen and sulfur atoms implies a slight worsening in the LG nature going from H– (HO^- (0.192, 0.220)), Me– (MeO^- (0.204, 0.226)) to Et– (EtO^- (0.206, 0.228)) substituted atoms. Besides the alpha-nucleophiles as HOO^- (0.175, 0.198), $HC(=O)OO^-$ (0.143, 0.155) are better LGs than HO^- (0.192, 0.220).

While there is an overall trend of inverse correlation between the two scales, there are numerous variations in the ordering between them. These

variations among the scale groups led us to conduct a more in-depth analysis of the correlation between them.

It is interesting to compare these results with those of our previous study on M–L bonds.[104] When examining the M–L bond, HD_M and HD_L quantify equivalent chemical attributes, meaning that in this problem, HD_1 corresponds to the σ donation capacity, applicable to both the metal fragment HD_{M1}, and the ligand moiety HD_{L1}. However, the HD interpretation depends on the solvent, distinguishing between gas phase and water environments. In this current chapter, our objective was to unveil if this behavior also holds true for EG and LG scales. We further evaluated the correlations among the vectors present in Tables 4.1 and 4.2. For the sake of simplicity, Figure 4.11 provides an overview of the resulting R^2 values.

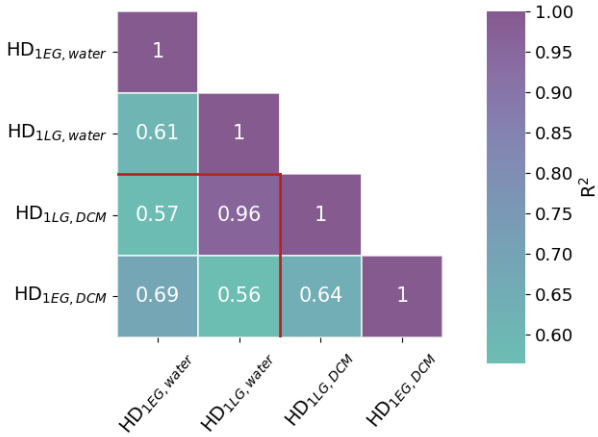


Figure 4.11: Correlation matrix between the HD_1 scales.

The diagonal elements of Figure 4.11 are 1.00 since they represent the correlation of each vector with itself. The other value that comes close to unity is the correlation between the HD_{1LG} (water) and the HD_{1LG} (DCM) vectors. This indicates that the ability of a given nucleophile to act as a leaving group is quite similar in both solvents. Yet, the rest of the correlations are significantly smaller, with R^2 coefficients ranging from 0.56

to 0.69. The lowest values of R^2 were found for the correlations between the H_{1EG} and H_{1LG} which indicates that they correspond to different magnitudes. In contrast with the previous chapter, the solvent nature seems to have little influence on the HD scale.

4.8 Limitations of the calculations

We acknowledge that B3LYP-D3 may not be the most precise method for computing the barriers in these reactions. Several studies have indicated its tendency to underestimate barriers,[178] but the trends are anticipated to be similar. Moreover, we believe it offers a reasonable balance between quality and cost-effectiveness. The ideal scenario would be to perform a benchmarking against experimental data, but the available data are limited in scope.[179–181] It is worth mentioning that the reported patterns are mostly reproduced by our calculations.[182]

The second limitation arises from not explicitly accounting for solvent and counterion effects, a constraint that extends to various other systems. For instance, it is known that in protic solvents like water, the presence of hydrogen bonds between solvent and solute may be significant to consider. Moreover, in a recent report, it was evidenced that the Lewis acid interacting through hydrogen or halogen bonds disrupts the energy of the intermediates and TS in the S_N2 gas phase reaction.[183] However, we do not consider this to be a serious problem, as it points to a problem of interpretation not directly related to our treatment, and this is beyond the scope of the present work.

A nice illustration of seemingly problems really explainable by a careful interpretation concerns the role of iodide additives for nucleophilic catalysis. Iodide is indeed recognized to enhance sluggish transformations. Yet within our HD_{1EG} scale, I^- is categorized as a poor EG. To shed light on this contradiction, we conducted an analysis of the Williamson reaction.[184–186] This reaction serves as one of the standard processes for the synthesis of ethers. The reaction is $RO^- + MeX \rightarrow MeOR + X^-$, where X^- is a Cl^-

or Br^-. This slow process is accelerated when the iodide replaces X^-, and then, I^- is a more efficient LG for the reaction with methoxide. This suggests that iodide should be a good nucleophile, yet attending to our HD_{1EG} parameter, $HD_{1EG}(I^-) = 0.240$, $HD_{1EG}(MeO^-) = 0.151$ in DCM, thus MeO^- should be better EG. Indeed, an examination of our computed energy barriers (Equations 4.8.1 and 4.8.2) suggests that the addition of the iodide anion should not favor the reaction.

$$I^- + MeCl \rightarrow MeI + Cl^- \ \Delta G^{\ddagger} = 22.0 \text{ kcal} \cdot \text{mol}^{-1} \tag{4.8.1}$$

$$MeO^- + MeCl \rightarrow MeOMe + Cl^- \ \Delta G^{\ddagger} = 9.9 \text{ kcal} \cdot \text{mol}^{-1} \tag{4.8.2}$$

In order to clarify the issue, we performed some mechanistic studies considering the presence of counterions in the solution.

The mechanism of Figure 4.12 shows that the presence of the countercation Na^+, which inverts the trends with respect to the reactions computed without it, discussed in Equations 4.8.1 and 4.8.2. Herein, the highest rate-determining step is found in the absence of the iodide (purple path in Figure 4.12). This demonstrates that the addition of the iodide salt kinetically facilitates the reaction (black path). The interaction of the counterion with the chloride promotes the leaving of such anionic fragment and reduces the energy barrier. This result aligns with the experimental evidence. The key to the "catalytic" ability of iodide is thus, its weak binding to the associated counterions.

While these topics hold interest for further investigation, we believe that delving deeply into these matters is not the primary focus of our work. Our aim is to introduce hidden descriptors for understanding reactivity.

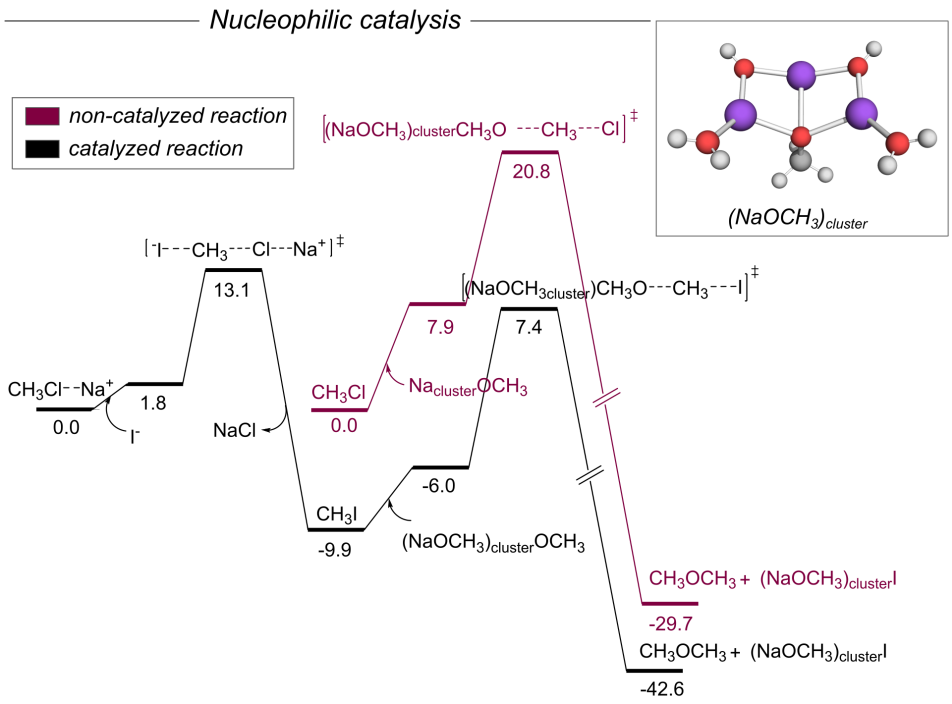


Figure 4.12: 3D structure of the ($NaOCH_3$)cluster, with the following the color-coded for the atoms: • oxygen, • sodium, • carbon and • hydrogen (top). Mechanistic profile of free energy (in kcal·mol^{-1}) for the formation of methoxide with the sodium iodide (black path) and without the catalyst (red path) (bottom).

4.9 Chemical meaning of the the first hidden descriptor

Thus far, we have presented the data concerning HD_1 and conducted a statistical analysis to identify its trends. Our objective was also to comprehend the chemical concepts underlying the reaction. A *priori*, we did not know whether these two entering and leaving scales could be simple sheer mathematical scales, or by contrast, they correspond to well-known chemical driving forces. To shed light on the topic, we used statistical tools including LR, and MLR. A set of 191 quantum-chemically derived conventional descriptors were selected and used to seek significant correlations with HD_1. These descriptors were Frontier Molecular Orbital (FMO) energies and their derived concepts, different atomic-based charges, energetic parameters, solvent terms, geometrical features, and measures of bond order. The full list of chemical descriptors is supplied in the Appendix B. Our approach was to encompass most of the physical organic chemistry properties expected *a priori* to be related to the mechanisms of the nucleophilic attack.[175,187]

Figure 4.13 illustrates the probability density plot of the R^2 values obtained from the linear regression analysis conducted between HD_1 and the ensemble of 191 descriptors.

A general inspection of Figure 4.13 indicates that most of the correlations were not satisfactory with an R^2 lower than 0.500. We only found a few models that fit at higher R^2. Table 4.3 contains the significant results of the statistical analysis.

Considering the mechanism of the reaction, we expected that the FMO energetic descriptors related to the donor-acceptor interactions would correlate the best. However, upon inspecting the correlations with the highest occupied molecular orbital (HOMO) and the lowest occupied molecular orbital (LUMO) energies, we noticed weak correlations between these descriptors and HD_1 (Table 4.3, entries 1 and 8). We explored other descriptors that involve electrodonating and electrodonating power[188] but

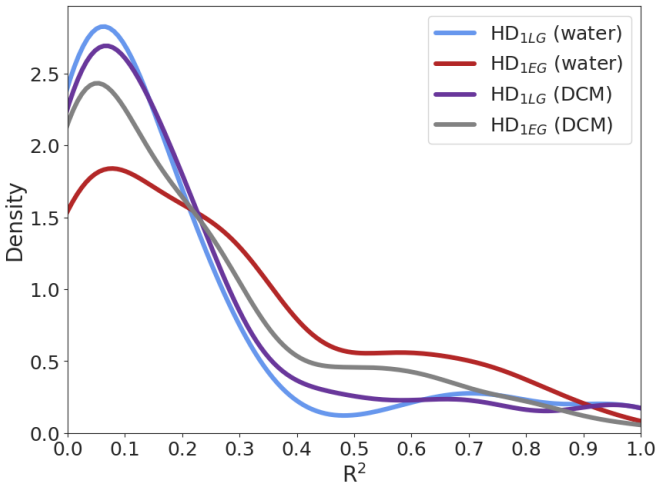


Figure 4.13: Density plot of the square correlation coefficient (R^2) derived from the linear regression between the conventional descriptors and the HD_{1LG} (water) in blue; HD_{1EG} (water) in red; HD_{1LG} (DCM) in purple; and HD_{1EG} (DCM) in gray.

Table 4.3: Square correlation coefficient (R^2) between different HD_1 values and selected conventional descriptors.

No.	**Descriptor**	**Water**		**DCM**	
		$\mathbf{R^2}$ $\mathbf{HD_{1LG}}$	$\mathbf{R^2}$ $\mathbf{HD_{1EG}}$	$\mathbf{R^2}$ $\mathbf{HD_{1LG}}$	$\mathbf{R^2}$ $\mathbf{HD_{1EG}}$
1	HOMO of EG^-	0.287	0.634	0.322	0.483
8	LUMO of CH_3LG	0.183	0.307	0.218	0.412
50	ω^-(I) of EG^-	0.169	0.307	0.083	0.017
60	ω^+(I) of CH_3LG	0.199	0.695	0.088	0.099
124	ΔG_{solv} for EG^-	0.007	0.002	0.063	0.147
128	ΔG_{solv} for CH_3LG	0.021	0.028	0.077	0.087
114	ΔV for (I^- + CH_3LG → CH_3I + LG^-)	**0.975**	0.735	0.969	0.784
146	d(I–C) for TS of (EG^- + CH_3I → CH_3EG + I^-)	0.713	**0.897**	0.672	0.833

none of them provided significant results (Table 4.3, entries 50 and 60). As the reactions were computed with implicit solvation, we thought that some solvent-based descriptors would provide good fits. Nevertheless, none of these descriptors displayed significant correlation values (entries 124 and 128).

Let us first analyse the conventional descriptors closer to HD_{1LG}. Despite the fact the LUMO energy of CH_3LG molecule is believed to be responsible for the reaction, it displays a weak interaction of only R^2 of 0.183 in water and R^2 of 0.218 in DCM (entry 8 in Table 4.3). This can be attributed to the fact that the LUMO in the reactant might not always be the crucial orbital in the TS, as it may be located elsewhere in the molecule. We also expected to identify correlations with alternative FMO attributes, such as the electrophilicity index of CH_3LG, ω, or its electron affinity, EA. The formulas to compute all descriptors are detailed in the Annex B. After exploring the trends, the best correlation with HD_{1LG} is obtained with descriptor number 114 (Table 4.3, entry 7), which is the potential energy value of the reaction, $I^- + CH_3LG \rightarrow CH_3I + LG^-$.

The left plot of Figure 4.14 illustrates the linear regression between the HD_{1LG} and descriptor number 114 for the twenty-six studied chemical fragments. In this context, the effect of the EG in the reaction parameter value is not critical, iodine in this case, is not critical. Other EG groups had similar effects. Iodine was selected as the representative EG during the parameter extraction, ensuring uniform energy influences throughout. Descriptor 114 is a thermodynamic property that measures the strength of the bond between the carbon center and the LG moiety, CH_3–LG. The reaction involves a two-electron transfer mechanism, making ΔV a quantifier of heterolytic bond dissociation. This parameter is crucial in the S_N1 reaction where the initial step involves the formation of the carbocation CH_3^+. This descriptor refers to the DFT energy of the reaction. Unsurprisingly, descriptors incorporating thermodynamic contributions such as ΔH, ΔG and $\Delta V + ZPVE$ (zero point vibrational

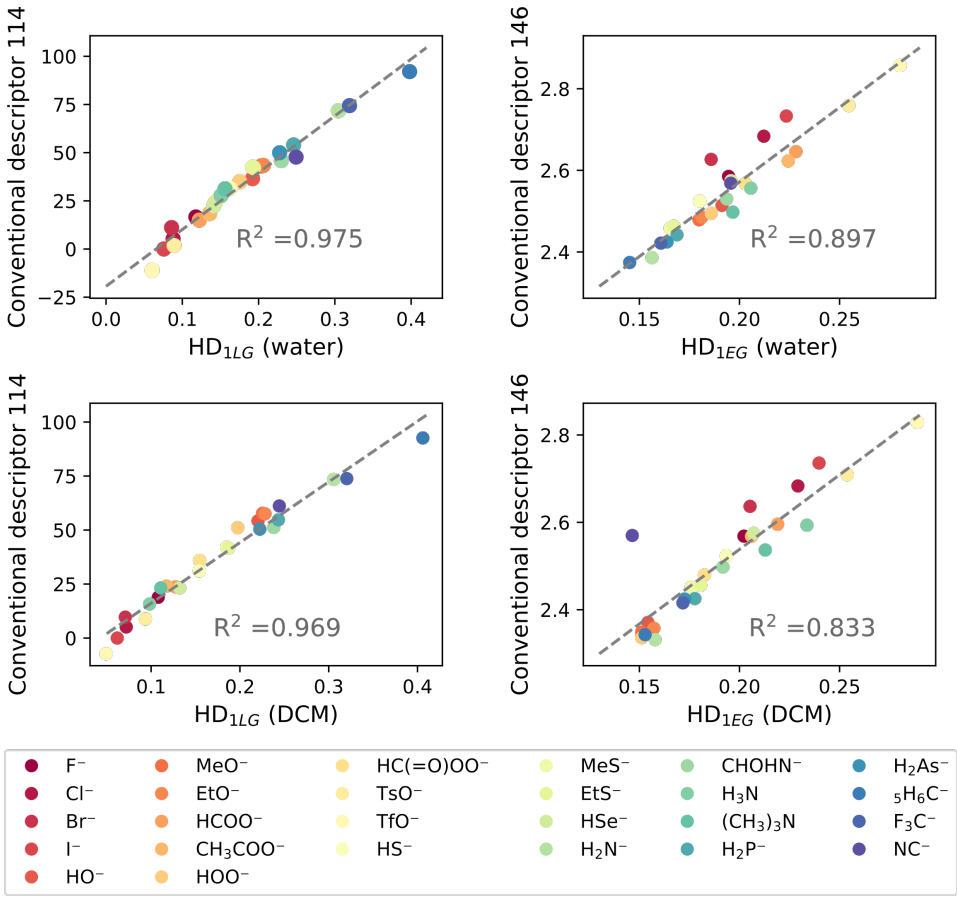


Figure 4.14: Plots of the values of the HD_{1LG} (left plots) and HD_{1EG} (right plots) *vs.* the values of the corresponding conventional descriptor providing the best correlation outcome. Analysis in water media (top plots) and in DCM solvent (bottom plots)

energy), exhibit also good correlations with R^2 higher than 0.964 in both solvents. Therefore, the cleavage of the bond CH_3–LG is pivotal in determining the barrier for the reaction. This observation gains further support from the correlation observed in the breaking of the CH_3–LG bond. The less bounding chemical fragments, such as triflate, tosylate, acetate, and formate are the exceptional LGs, consequently, they yield the lowest $\Delta G^{\ddagger}$. In the opposite side, stronger bonds with LGs like $C_6H_5^-$, and CF_3^- have to overcome high activation energies. This underlines the key role of the cleavage energy of the leaving group in shaping the core characteristics of the S_N2 process.

Our attention now shifts towards the entering group, with the associated HD_{1EG} scale. We anticipated that the energy of the HOMO or the electrodonating power ω^- of the EG, (entries 1 and 50 in Table 4.3, respectively), could be critical. However, none of these descriptors furnished significant correlations. The HOMO energy value achieves a R^2 of 0.634 in water and 0.483 in DCM, but these are relatively moderate values. Descriptor number 146 in Table 4.3 exhibited the best correlation with the HD_{1EG} descriptor with a R^2 of 0.897 in water and 0.833 in DCM. The right plots of the Figure 4.14 displays the LR between that property and the HD_{1EG}. Descriptor 146 comprises a geometrical feature of the transition state structure (TSS). This descriptor has a more kinetic component since it is related to a TS. In particular, it measures the distance between iodine and the carbon center (d(CH_3–I)) within the TSS of the reaction: $EG^- + CH_3I \rightarrow CH_3EG + I^-$. Interestingly, this distance is not about the iodine LG. Instead, it reflects the nucleophilicity ability of the EG chemical fragment. This effect can be viewed as a trans-effect, where the EG pushes the LG away. For the TSs with the shortest distances, d_{min}(CH_3–I), it indicates a lower resistance for the EG to break the CH_3–LG bond. Therefore, this corresponds to an early transition state where the EG is effective. In contrast, when the interaction between the CH_3– and I– fragments is exceptionally strong, a longer elongation of the CH_3–I is required for the

Table 4.4: Square correlation coefficient (R^2) between different HD_1 values and selected conventional descriptors in water.

			Water				
HD	**No.**	**Descriptors**	**R^2**	**No.**	**R^2**	**No.**	**R^2**
HD_{1LG}	56, 117	ω^-(I) of CH_3LG, ΔG (I^- + CH_3LG → CH_3I + LG^-)	**0.990**	56	0.199	117	0.966
HD_{1EG}	56, 117	ω^-(I) of CH_3LG, ΔG (I^- + CH_3LG → CH_3I + LG^-)	**0.933**	56	0.695	117	0.779

bond to be broken. Herein, the EG needs to push the LG further away to attain the TS. This last situation relates to the later TS and involves poorer EGs. Furthermore, it is noteworthy that the impact of the entering group on the barrier appears to be less significant than that of the leaving group, as already discussed earlier. This is evident from the narrower range of values and lower correlation values compared to those of the LGs.

In the case of the HD_{1EG} variable, its correlation with a TS property adds complexity to estimating the EG ability *via* easy calculations. Drawing from these results, we examined the question of whether HD_1 might correlate with two conventional descriptors instead of only one. We conducted multilinear regression analysis aiming to identify correlations that could be easier to interpret. The set of descriptors chosen was the same as for the LR approach, but removing the two highest-performing descriptors (114 and 146), the hidden descriptor values, and any of the transition state-based descriptors. By considering all possible pairs of descriptors without repetition from the remaining 138 conventional descriptors, we had a total of 9453 unique combinations of two variables. Finally, 9453 MLRs were calculated.

The performance of the correlations increases respect to the LR from 0.897 (Table 4.3, number 146) to 0.933 (Table 4.4, numbers 56 and 117) for the entering group and from 0.975 (Table 4.3, entry 114) to 0.990 (Table 4.4, entry 56, 117) in water, and equivalent results for DCM (one descriptor in

Table 4.5: Square correlation coefficient (R^2) between different HD_1 values and selected conventional descriptors in dichloromethane.

			Dichloromethane				
HD	**No.**	**Descriptors**	**R^2**	**No.**	**R^2**	**No.**	**R^2**
HD_{1LG}	26, 117	η^- of X^-, ΔG (I^- + $CH_3LG \rightarrow CH_3I + LG^-$)	**0.990**	26	0.151	117	0.970
HD_{1EG}	26, 115	η^- of X^-, ΔG (I^- + $CH_3LG \rightarrow CH_3I + LG^-$)	**0.933**	26	0.437	115	0.811

Table 4.3, entries 26 and 117 to two descriptors in Table 4.5 entries 56 and 117). Analysing in-depth the results show that the highest correlations include one energetic reaction feature, which also reaches satisfactory outcomes by themselves alone.

To assess the role of the high-order HDs (HD_2, HD_3, and HD_4), we also conducted LRs between them and the reported set of 191 descriptors. None of the resulting LRs yielded a R^2 value above 0.830. This outcome supports the initial hypothesis of section 4.7, that HD_2, HD_3, and HD_4 are mainly affected by numerical noise and do not provide significant insights into the chemical reaction.

4.10 Prediction of HDs

The procedure outlined above has furnished HD values for twenty-six nucleophiles covering a broad chemical space. Our subsequent objective was to be able to derive hidden descriptors for any given nucleophile out of the original dataset without significant computational expense. We could certainly repeat for each new nucleophile the exhaustive methodology detailed throughout this Chapter: the determination of HD values for new target chemical fragments implies the computation of twenty-seven reactions, the subsequent application of the SVD process, and the final analysis of the HDs. Nevertheless, this volume of calculations becomes impractical given the simplified approach we are pursuing. Moreover,

this approach might introduce minor alterations to the already computed descriptors. Thus, inspired by previous studies in our group,[104], we defined a procedure to derive those values with low computational cost and small error. Figure 4.15 summarizes the process.

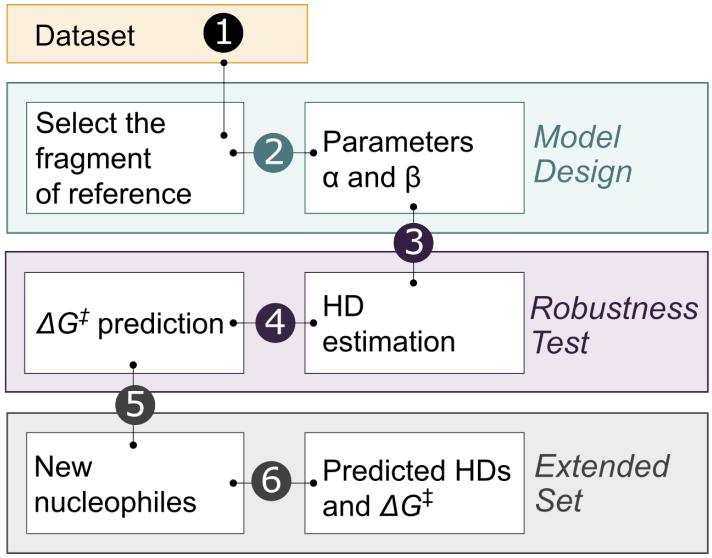


Figure 4.15: Workflow for the design of the HD prediction model for new target entering and leaving groups.

We started by establishing the dataset for the construction of the prediction model. Then, we employed MLR analysis to select the representative chemical fragments for the model. Once the fragments were chosen, we evaluated its performance by comparing the initially computed HD and $\Delta G^{\ddagger}$ values with the predicted counterparts. Finally, we applied the technique on novel target nucleophiles.

4.10.1 Model design

We employed as dataset the collection of 676 DFT activation barrier, along with the SVD-derived matrices: $\boldsymbol{EG}$, $\boldsymbol{LG^T}$, $\boldsymbol{W}$, each within the two solvents. Using this data, we sought to identify the reference set of

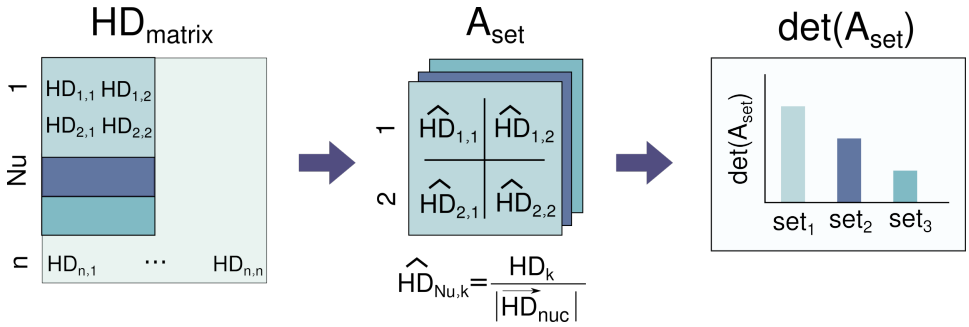


Figure 4.16: Procedure followed for the selection of the reference set of nucleophiles.

nucleophiles that exhibits the greater variability of the entering and leaving group descriptors. Those would be suitable chemical fragment variables for predicting accurate hidden descriptors. The procedure for achieving this is depicted in Figure 4.16.

For the sake of clarity, $\boldsymbol{EG}$ and $\boldsymbol{LG^T}$ matrices are denoted as $\boldsymbol{HD_{matrix}}$ in the Figure 4.16, since they contain the hidden descriptor vectors. Figure 4.16 shows a hidden descriptor matrix $\boldsymbol{HD_{matrix}}$ from which we built square matrices $\boldsymbol{A_{set}}$ with the same number of nucleophiles and HDs. These matrices encompass all possible nucleophile combinations along with their corresponding hidden descriptors, arranged in a hierarchical manner. Thus, the first column vector of $\boldsymbol{A_{set}}$ is for the HD_1, the second column is for the HD_2, and so forth. We formed matrices of the following dimensions 2×2, 3×3, 4×4. In total, we computed 17875 combinations with two, three, and four species each:

$$\binom{26}{2} = 325; \binom{26}{3} = 2600; \binom{26}{4} = 14950 \qquad (4.10.1)$$

The hidden descriptor vectors included in the $\boldsymbol{A_{set}}$ were normalized for each nucleophile $\hat{\mathrm{H}}\mathrm{D}_{Nu,k}$. Then, the reference set of nucleophiles was determined by evaluating the highest determinant of these matrices, $\det(\boldsymbol{A_{set}})$. The more different the hidden descriptors of the nucleophiles

in the subset, the higher the determinant. This concept relates to the fact that the determinant of a matrix provides an idea about how much the multiplication by such matrix stretches or shrinks the space. In our problem, expansion across as much space as possible is desired. We performed this procedure for water and dichloromethane, and for entering groups and leaving groups independently. As a result, the outlined procedure was carried out for four different hidden descriptor matrices: $\boldsymbol{EG_{water}}$, $\boldsymbol{LG^{T}_{water}}$, $\boldsymbol{EG_{DCM}}$ and $\boldsymbol{LG^{T}_{DCM}}$.

At first, we did not know how many nucleophiles were required in the set, thus, $\boldsymbol{A_{set}}$ were formed with 2, 3, and 4 different fragments. It gave rise to several reference sets, and consequently, different model equations.

The R^2 for the HD_1 prediction within the training set is outlined in Figure 4.17. The values of those correlations are always above 0.997 for the first $HD_{EG(LG}$ prediction regardless of the number of parameters included.

We also selected the mean-square error (MSE) metric to assess the appropriateness of the reference set. The following Figure 4.18 illustrates the predictive performance of the models for HD values, based on different quantities of HD variables.

The left plot of Figure 4.18 demonstrates that when predicting HD_{EG}, employing two elements in the reference set results in an MSE slightly above 0.08. As the number of elements increases, the error remains relatively constant. In the case of the leaving group HD, the error is remarkably low from the first attempt (right plot in Figure 4.18). For the sake of simplicity, we decided to adopt the same number of elements as used for the entering group. Consequently, we concluded that an equation model involving two nucleophiles is adequately accurate for predicting $HD_{EG/LG}$.

We then established four sets of references. The most illustrative entering groups are: Br^- and CH_3COO^- for water and NH_3 and $C_6H_5^-$ for dichloromethane, and the sets of leaving groups are: $HC(=O)OO^-$, TfO^- for water and I^-, PH_2^- for dichloromethane.

Upon definition of the reference chemical fragments, we conducted MLR

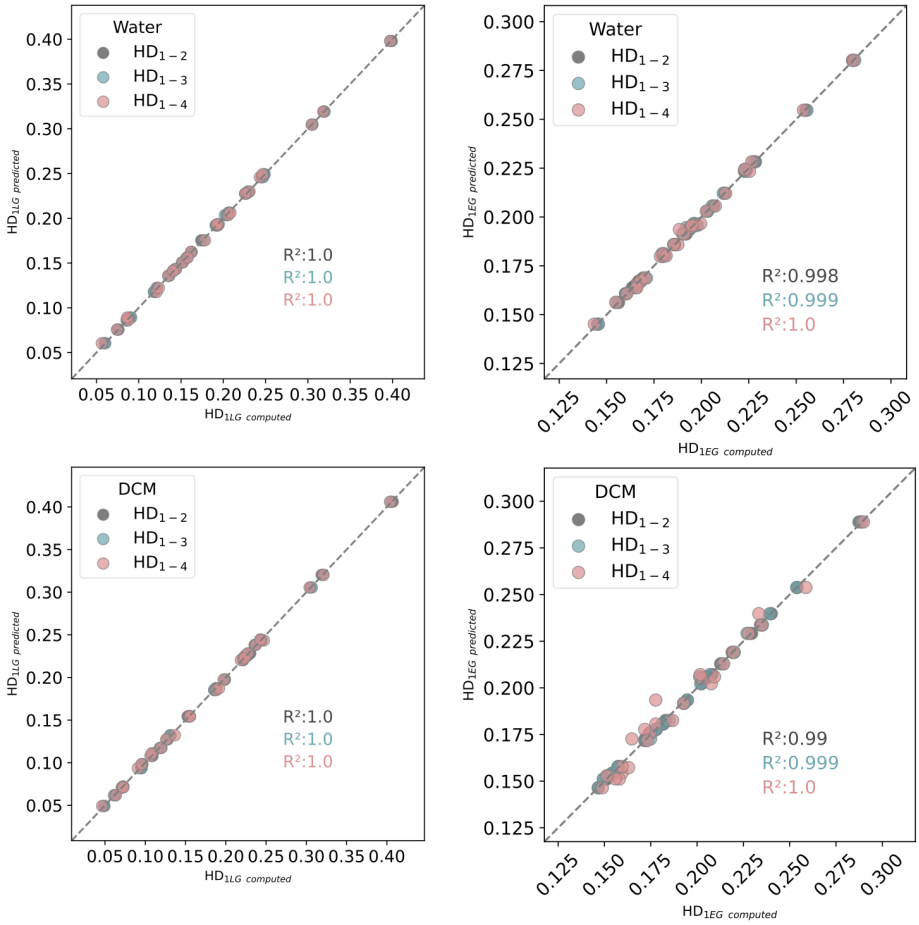


Figure 4.17: Plots of the computed $HD_{1EG/1LG}$ *vs* estimated $HD_{1EG/1LG}$ in water (top) and in dichloromethane (bottom) with different numbers of variables .

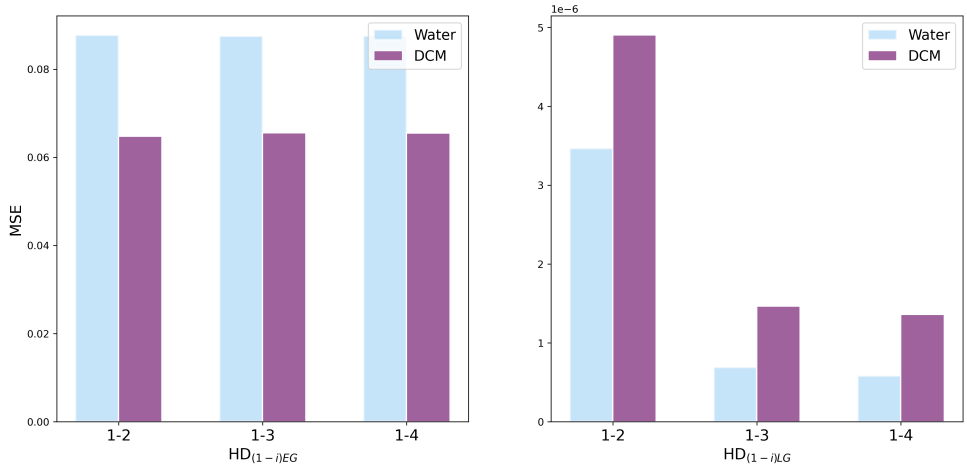


Figure 4.18: Mean-squared error (MSE) for HD_{pred} under water (blue bars) and DCM (purple bars) as a function of the hidden descriptors considered (HD_{1-i}).

analysis to predict HD_1 values. In this context, HD values are the dependent variables related to the independent energy barrier values. We fitted a linear model as Equation 4.10.2 resulting in the parameters $\alpha_{k,ref}$ and β_k.

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \alpha_{k,ref} + \beta_k \tag{4.10.2}$$

$\boldsymbol{Y}$ matrix contains the HD_1 values of the 26 nucleophiles (Nu). The $\boldsymbol{X}$ matrix refers to the energy barrier between the nucleophiles of the reference set and the whole set of nucleophiles $\boldsymbol{\Delta G^{\ddagger}_{ref,Nu}}$. The calculated $\alpha_{k,ref}$ and β_k parameters are depicted in Tables 4.6a and 4.6b.

4.10.2 Consistency of the model

First, we analysed the performance of the training set. The variables derived before, α and β, are employed to predict $HD_{k,i}$ with the Equation 4.10.3, where k is 1, and i represents the target nucleophiles,

$$HD_{k,i\ pred} = \Delta G^{\ddagger}_{ref_1,i} \cdot \alpha_{k,ref_1} + \Delta G^{\ddagger}_{ref_2,i} \cdot \alpha_{k,ref_2} + \beta_k \tag{4.10.3}$$

Table 4.6: Values of the coefficient parameters α and β of the MLR equation for HD_{Nu} prediction.

(a) Values of the coefficient parameters α and β for the HD_{EG} prediction.

EGs of reference

	k	α_{k,Br^-}	α_{k,CH_3COO^-}	β_1
Water	1	0.00294	0.0012	-0.0055
	2	0.0586	-0.0190	-0.2778
	k	α_{k,NH_3}	$\alpha_{k,C_6H_5^-}$	β_2
DCM	1	0.0031	0.0012	0.0138
	2	-0.0434	0.0341	0.0992

(b) Values of the coefficient parameters α and β for the HD_{LG} prediction

LGs of reference

	k	α_{k,NH_3}	$\alpha_{k,C_6H_5^-}$	β_1
Water	1	0.0031	0.0012	0.0138
	2	-0.0434	0.0341	0.0992
	k	α_{k,I^-}	α_{k,PH_2^-}	β_2
DCM	1	0.0034	0.0013	-0.0083
	2	-0.0308	0.0176	0.1186

Since our focus is on calculating HD_1, *i.e.* HD_{1EG} and HD_{1LG}, we set the value 1 to k in Equation 4.10.3. The unknown value $HD_{k,i\ pred}$ is unveiled once we input the energy barrier $\Delta G^{\ddagger}_{ref_1,i}$ and $\Delta G^{\ddagger}_{ref_2,i}$, which represent the activation energy for the reaction between the reference set of nucleophiles *ref* and the new nucleophile *i*, into the equation.

We then validated the consistency of the model by predicting the energy barriers for each entering and leaving group pair with Equation 4.10.4. Subsequently, these predicted values were compared with the corresponding DFT-energy barriers.

$$\Delta G^{\ddagger}_{pred} = HD_{EG_{pred}} \cdot HD_{1,W} \cdot HD_{LG_{pred}} \tag{4.10.4}$$

Plots of Figure 4.19 show such comparison of the response value in Equation 4.10.4 with those explicitly computed.

Using two energy barrier values to calculate the first hidden descriptor,

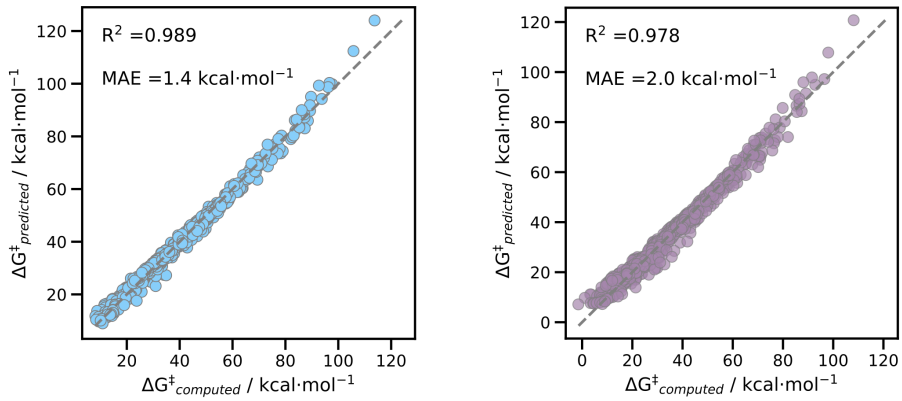


Figure 4.19: Plots of the computed $\Delta G^{\ddagger}$ *vs* predicted $\Delta G^{\ddagger}$ (in kcal·mol^{-1}) using the first HD in water (left plot) and in dichloromethane (right plot).

the mean absolute error (MAE) was 1.4 kcal·mol^{-1} in water and 2.0 kcal·mol^{-1} in DCM; and the maximum error was 10.3 kcal·mol^{-1} in water and 12.6 kcal·mol^{-1} in DCM. In Section 4.7, we provided a detailed analysis of the errors obtained when predicting the energy barrier using HD_{EG} and HD_{LG} directly derived from the SVD. In that scenario, the MAE was 1.3 kcal·mol^{-1} and 1.8 kcal·mol^{-1} with maximum errors of 10.5 kcal·mol^{-1} and 12.8 kcal·mol^{-1} for water and for DCM, respectively. Herein, the errors for the computation of $\Delta \mathrm{G}^{\ddagger}_{pred}$ are nearly equal when the HD_1 is estimated instead of directly derived from the SVD. The correlations between the computed energy barrier and the predicted ones are $R^2 = 0.989$ for water and $R^2 = 0.978$ for DCM, within the set. Therefore, we conclude that our prediction model adequately captures the first hidden descriptor of the nucleophiles and subsequently, the calculation of the $\Delta G^{\ddagger}_{pred}$ exhibits satisfactory accuracy.

4.10.3 Application to extended series of nucleophiles

Aiming to assess the predictive model we extend the use of this model to new chemical candidates. We tested seventeen additional nucleophiles which

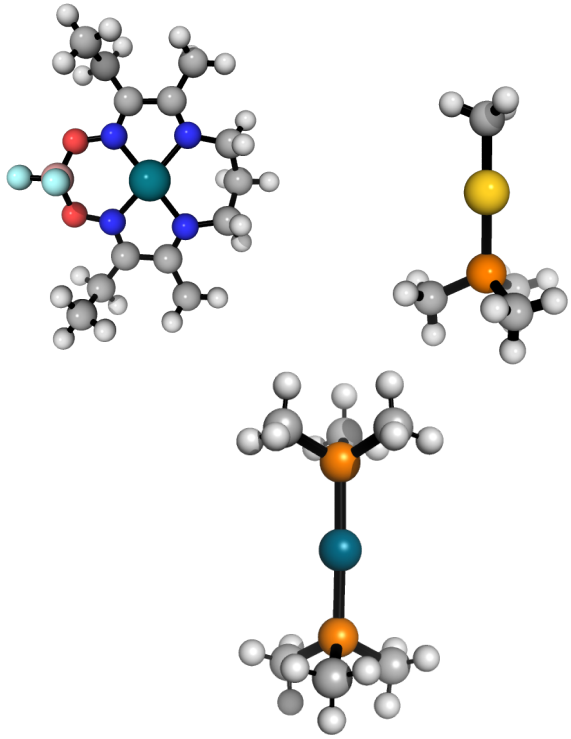


Figure 4.20: 3D Representation of the considered metal fragments. Following the order top, left-to-right, the metal centers of the complexes are • Rh, • Au, • Pd. The color-coded for the atoms is the following: • carbon, • hydrogen, • oxygen, • nitrogen, • phosphorous, • fluorine, and • boron.

act both as entering and leaving groups. In this set, we consider organic, CH_3^-, CO, NCH, NHC, C_5H_5N, $C_6H_4MeO^-$, and inorganic molecules H_2O, H^-, O^{2-}, PH_3, S^{2-}, PCl_3, $SiMe^-$. Moreover, four metal complexes were evaluated (see Figure 4.20) since some oxidative additions of organic molecules have been tested to occur through a concerted S_N2 reaction.[189,190] We selected $AuP(Me_3)Me$, $MP(Me_3)_2$ being M = Pd, Pt and, a Rh complex, which was one of the first examples to participate in this type of reaction.[191]

As mentioned above, we needed to perform a few calculations for each of those chemical species. We computed the activation energy of each target fragment with the reference nucleophiles. Figure 4.21 shows a color-coded

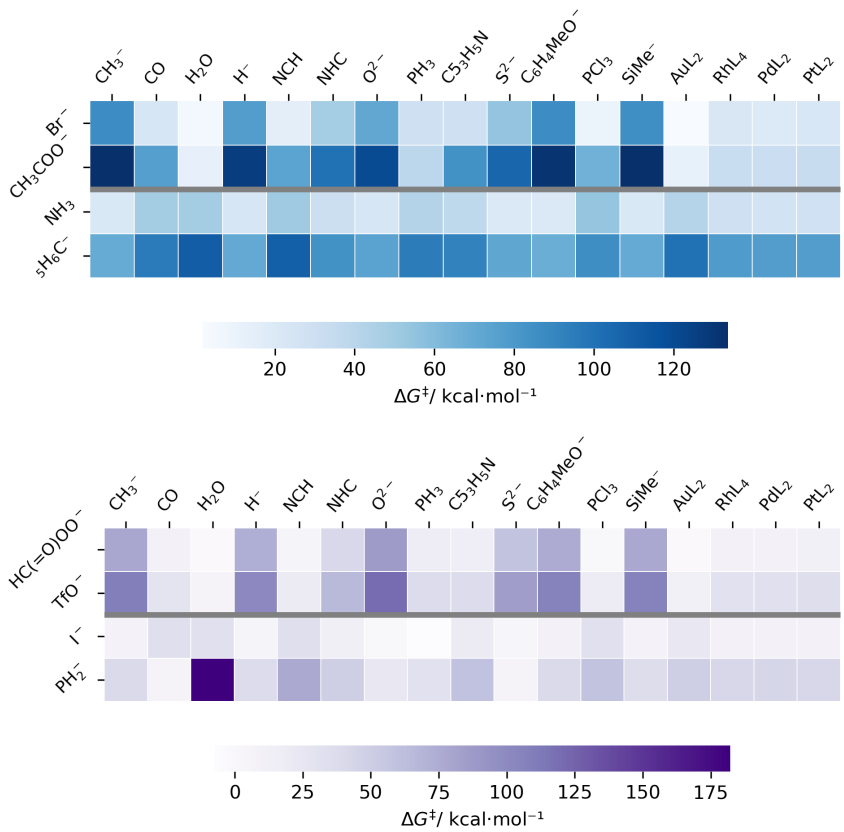


Figure 4.21: Color-coded version of the matrices of the free energy barriers, $\Delta G^{\ddagger}_{R-TS}$, (in kcal·mol^{-1}) in water solvent (top in blue) in dichloromethane (bottom in purple).

representation of the 17 × 2 matrices with the free energy barriers in water and in dichloromethane. Each heatmap (Figure 4.21) similarly organizes the activation energy. In the first two rows, the horizontal axis represents the EG (Br^-, CH_3COO^- in water; $HC(=O)OO^-$, TfO^- in DCM) and in the last two rows, this axis depicts the leaving group (NH_3, $C_6H_5^-$ in water; I^-, PH_2^- in DCM). The horizontal axis is first LG and then the same species act as EGs.

Following the same Equation 4.10.3, we computed the HDs of the

Table 4.7: Predicted values of the HD_1 descriptor.

Water				Dichloromethane			
EG	**HD**$_{1EG}$	**LG**	**HD**$_{1LG}$	**EG**	**HD**$_{1EG}$	**LG**	**HD**$_{1LG}$
$C_6H_4MeO^-$	0.137	CH_3^-	0.439	CO	0.014	O^{2-}	0.447
CH_3^-	0.142	$C_6H_4MeO^-$	0.436	S^{2-}	0.014	CH_3^-	0.407
$SiMe^-$	0.142	$SiMe^-$	0.436	O^{2-}	0.087	$SiMe^-$	0.401
S^{2-}	0.143	H^-	0.404	PH_3	0.121	$C_6H_4MeO^-$	0.392
H^-	0.148	O^{2-}	0.377	$SiMe^-$	0.144	H^-	0.377
O^{2-}	0.154	S^{2-}	0.308	H^-	0.148	S^{2-}	0.305
PdL_2	0.167	NHC	0.281	CH_3^-	0.154	NHC	0.221
PtL_2	0.173	C_5H_5N	0.204	$C_6H_4MeO^-$	0.155	PH_3	0.102
RhL_4	0.176	CO	0.178	PtL_2	0.167	C_5H_5N	0.097
NHC	0.187	PH_3	0.151	RhL_4	0.168	PtL_2	0.076
C_5H_5N	0.217	NCH	0.148	PdL_2	0.172	PdL_2	0.069
PH_3	0.233	PtL_2	0.123	NHC	0.197	RhL_4	0.066
PCl_3	0.237	RhL_4	0.121	AuL_2	0.198	CO	0.063
CO	0.254	PCl_3	0.120	PCl_3	0.237	NCH	0.036
AuL_2	0.237	PdL_2	0.114	C_5H_5N	0.241	PCl_3	0.017
H_2O	0.272	H_2O	0.043	NCH	0.314	AuL_2	-0.001
NCH	0.276	AuL_2	0.034	H_2O	0.760	H_2O	-0.009

fragments in the test set. Table 4.7 presents the values of the hidden descriptor 1 for the test set in water.

An interesting result from Table 4.7 is that most of the estimated hidden descriptor values fall within the range of the SVD operation results (HD_{EG} range: 0.145 to 0.280; HD_{LG} range: 0.060 to 0.398). This confirms the premise that the initial dataset covers most of the chemical space. Among the new nucleophiles, the best EG (0.137) is anisolate, and the worst is NCH (0.276). The best LG in the scale is AuL_2 (0.034) and the worst is CH_3^- (0.439). Similar trends are also found in the dichloromethane solvent along the species where the best EG is CO (0.014) and the poorest entering fragment is H_2O (0.760); the LG ability ranges from the O^{2-} (0.447) to the best nucleofuge H_2O (0.009).

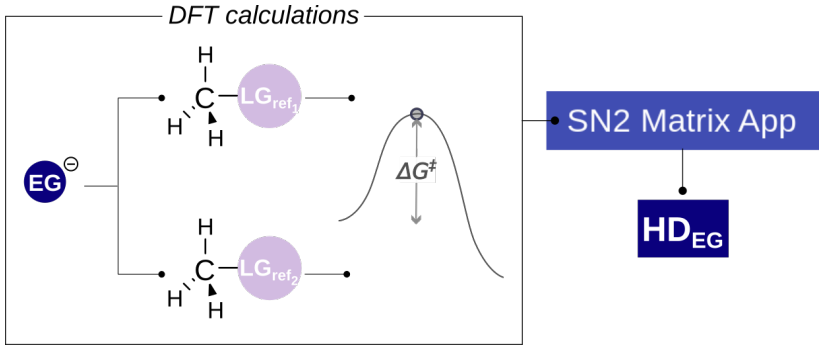


Figure 4.22: Scheme of the procedure to predict HD values for EG using the S_N2 Matrix App.

4.11 S_N2 Matrix App

The HD values are thus employed for both locating the EG and LG ability of nucleophiles in the chemical space and predicting $\Delta G^{\ddagger}$ to the design of chemical reactions. In light of these results, we developed a user-friendly app, that anyone can use to compute HD values. This open-access application is accessible through https://maserasgroup-repo.github.io/sn2app/.[192]

In order to employ it, it is required to compute the DFT energy barriers between a target nucleophile with two selected nucleophiles (left in Figure 4.22), and introduce these values in a system of linear equations (right in Figure 4.22). This enables to spawn the hidden descriptors of any given nucleophile.

4.12 HD for S_N2 *vs.* HD for BDE

In the previous Chapter, we examined the application of the hidden descriptor methodology to unravel the electronic contributions of moieties involved in the metal-ligand bond (M–L) within a water environment. We wonder now if those thermodynamic contributions can be related in some way to our S_N2 hidden descriptors in water.

To investigate this potential relationship, we will use several ligands that in previous work[104] were characterized using the hidden descriptors for ligands, HD_L. Herein, we examined the ability of these ligands to participate in bimolecular nucleophilic substitution processes type using the S_N2 Matrix App to compute their HD-S_N2 values. We are aware that the hidden descriptors for metals were also determined. Nonetheless, because of the difficulty to compute the energy barriers between those exact metal fragments and our nucleophiles of references, we did not predict the HD-S_N2 variables for metal fragments. Therefore, we established correlations between both sets of HDs: $HD_{EG/LG}$ and HD_L.

We firstly performed linear regressions (LR) between the HD_{1EG} and HD_{1LG} and each of the other HD_L. In a first inspection, correlation is at most moderate, being the highest R^2 of 0.782 with HD_{1EG} and R^2 of 0.761 with HD_{1LG}. It is noteworthy, that while we only used the first hidden descriptor for describing the entering and leaving ability of the nucleophiles, HD_{1EG} and HD_{1LG}, respectively, in the M−L analysis we employed five HD_L for characterizing each ligand's behaviour ($HD_{L(1-5)}$). This raised the question of whether the correlation of $HD_{EG/LG}$ might extend to a few hidden descriptors of ligands rather than being solely linked to a single descriptor.

Aiming at finding answers, we explored multilinear regressions between HD_{1EG} and HD_{1LG} and all the possible combination of the five $HD_{L(1-5)}$. We set up a threshold of $R^2 = 0.9$ to analyse the contributions of each hidden descriptor. The best fittings were the following,

$$\boldsymbol{HD_{1EG}} = -0.427\boldsymbol{HDL_1} + 0.042\boldsymbol{HDL_2} - 0.048\boldsymbol{HDL_3} + 0.278$$
$$R^2 = 0.903 \tag{4.12.1}$$

$$\boldsymbol{HD_{1LG}} = 1.284\boldsymbol{HDL_1} - 0.280\boldsymbol{HDL_2} - 0.022 \qquad R^2 = 0.924 \tag{4.12.2}$$

Considering that HDs are normalized, we can assume that the weight of each HD_L over the HD_{1EG} and HD_{1LG} can be measured with their coefficients. Their contributions are for the HD_{1EG} (water): HD_{L1} 82.6%, HD_{L2} 8.1% and HD_{L3} 9.1%; and for HD_{1LG} (water): HD_{L1} 82.1% and HD_{L2} 17.9%.

It is clear that HD_{L1} is the main parameter in both equations 4.12.1 and 4.12.2 as it can be observed by their % contribution. This is not surprising because the σ-orbitals drive the formation and the breaking of bonds in the reaction.

Let us focus now on HD_{LG}. As it is shown in Equation 4.12.2, HD_{1LG} has good correlation with a combination of HD_{L1} and HD_{L2}. This HD_{L2} describes the π effects of the nucleophiles. Therefore, as the first term of the MLR equation 4.12.2 is positive, to get a small value of HD_{1LG} and thus, a good leaving group, the HD_{L2} should be also positive. This suggests that the better leaving group are the ones with a low capability of σ donation and high π donating properties. Figure 4.23 displays the distribution of the chemical species according to their HD_{L1} and HD_{L2} and their color position details their nucleofuge ability.

In the case of the HD_{1EG}, those values correlate twith a combination of HD_{L1}, HD_{L2} and HD_{L3}. The latter is associated with the *cis*-contribution of the chemical species. Considering that all our nucleophiles have a positive HD_{L1} value, to obtain a good entering group, HD_{L2} and HD_{L3} should be negative values, *i.e.* its π acceptor capacity and its *cis*-influence should be small. However, here the correlation is worse than in the case of HD_{1LG} and in addition, the contribution of HD_{L3} is almost negligible. The noteworthy observation is that the descriptors derived from kinetic S_N2@C reactions are correlated with those resulting from the thermodynamic values of the bond dissociation energy, but this correlation is not trivial.

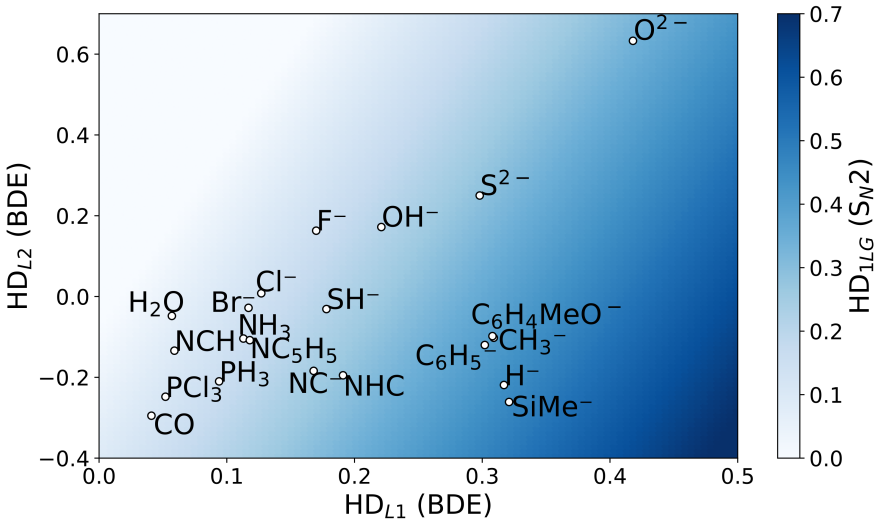


Figure 4.23: Scatter plot of HD_{L1}, HD_{L2} and color-coded grid of the HD_{LG} values in water.

4.13 Conclusions

This chapter delved into the application of the hidden descriptor methodology to bimolecular nucleophilic substitution reactions. In the first part, we characterized the concerted mechanism of 676 reactions encompassing twenty-six nucleophiles in implicit water and dichloromethane solvents. The reactants were categorized either as separated entering groups and electrophiles, or as adducts. We evaluated the minima of the reactants by selecting a subset of reactions. Then, their energy barrier values were computed starting from both the reactants and adducts. The results showed higher stability of the separate reactants across most of the reactions. Next, we conducted SVD on the full DFT-$\boldsymbol{\Delta G^{\ddagger}}$ matrices, followed by dimensionality reduction to derive hidden descriptors. A sufficient number of HDs were chosen based on various factors such as weight, energy, and chemical concepts. The first hidden descriptor rules over the rest of the HDs accounting for %90 of the information of our chemical system. Furthermore,

using only the first HD_1 derived from the matrix decomposition, the prediction of activation energy yielded a MAE of 2.0 kcal·mol^{-1}. Therefore, the HD_1 was defined as a quantitative measure of the intrinsic reactivity of a nucleophile that contains three numerical values: HD_{1EG} referring to the entering group ability, HD_{1LG} denoting the leaving group ability, and HD_{1W} indicating the weight of this hidden descriptor in the matrix decomposition. The study found that the identity of the leaving group had a higher influence on the barrier than the identity of the nucleophile. An exhaustive LR and MLR analysis was performed to unravel the chemical meaning of variables HD_{1EG} and HD_{1LG}. Unexpectedly, this hidden descriptor is poorly correlated with magnitudes intuitively associated with frontier molecular orbitals or solvation descriptors. Instead, it correlates with a thermodynamic property in the case of the HD_{1LG} and with a geometric property for HD_{1EG}. The performance of MLR did not achieve substantial improvement of the correlations.

In the second part of the Chapter, upon statistical treatment of the S_N2-HD data, we constructed a predictive model to derive hidden descriptors of out-of-the-sample nucleophiles in two different solvents. The process began with the selection of the chemical fragments of reference. In water, reference EGs included Br^-, CH_3COO^-, while LGs encompassed $HC(=O)OO^-$, TfO^-. In dichloromethane, the EG selected were NH_3 and $C_6H_5^-$, and the LGs I^-, PH_2^-. Then, we ensured the robustness of the model analysing the training set performance. We compared the computed 676 $\Delta G^{\ddagger}$ values with the predicted $\Delta G^{\ddagger}_{pred}$ values derived from HD_{pred}, and the corresponding equation. The successful comparison yielded R^2 values exceeding 0.978 for both water and DCM, and the MAE was roughly 2.0 kcal·mol^{-1}. To further evaluate the performance of the prediction model, we introduced a test set comprising four metal complexes and several nucleophiles. Their resulting HD values were in line with the expected trends. In order to extend the concept of the HD variable for nucleophiles to the scientific community, we developed an open-access web application. This model can be accessed at

https://maserasgroup-repo.github.io/sn2app/, and the user can employ it to predict HD_{1EG} and HD_{1LG} of any nucleophile.

Finally, we explored correlations between S_N2-HDs and the HD_L to find some relationships between the thermodynamic and the kinetic parameters. Our findings revealed that leaving group ability is strongly influenced by contributions from σ and π orbitals, while entering group ability is predominantly governed by the σ donation.

Overall, the current work furnished HD values that can aid in both understanding and forecasting the kinetics of this organic reaction, ultimately reducing the resources needed for designing chemical processes.

Chapter 5

AABBA graph kernel

> Well, it should be obvious to even the most dim-witted individual who holds an advanced degree in hyperbolic topology, that Homer Simpson has stumbled into... the third dimension
>
> — Professor Frink – The Simpson

5.1 Introduction

This Chapter is part of the project undertaken during my PhD secondment at the Hylleraas Centre for Quantum Molecular Sciences at the University of Oslo, under the supervision of Dr. David Balcells. The graph database used was provided by Hannes Kneiding, along with the preliminary code for running the neural networks. I developed the AABBA kernel, executed the code over the dataset, and provided the results of the DNN, which represent the overall content of this Chapter. The ChemRxiv preprint contains additional work[32] that was not included in this Thesis, which was conducted by Jørn Eirik Betten.

So far in this Thesis, we have utilized a reverse engineering approach to obtain the hidden descriptors. That methodology has leveraged valuable insights into chemical phenomena. However, the initial acquisition of HD

involves the computation of a large chemical space varying two variables, *i.e.* metal fragment and ligand in HD-BDE, and entering and leaving groups in HD-S_N2. As mentioned in the Chapters 1 and 2, consistency between data size, algorithm, and target problem is essential when dealing with data-led approaches. Therefore, in this last Chapter, we have changed the strategy to extract chemical descriptors. We no longer created chemical descriptors from a target property, but we refined a protocol for extracting chemical characteristics.

At the beginning, it was showed the broad number of available chemical descriptors that could be used for statistical approaches. Choosing the suitable ones requires a trade-off between information content, dimensionality, and computational requirements. Depending on the chemical system, we can select precise representations that effectively capture the relevant chemical information. For instance, SMILES[37] encodes a recognized nomenclature that is successfully employed in the organic chemistry field.[193] However, extrapolating this representation to metal complexes, such as the TMCs, has been proved challenging. Descriptors that are effective for organic molecules may result unsuitable for inorganic materials.[70]

This problem appears in complexes such as $Mo(PCy_3)_2(CO)_3$, which displays agostic interactions between the metal centre and a remote C–H bond. When using SMILES notations, this stabilizing interaction will be disregarded, and consequently, it will not have an impact on the SMILES-based models. Furthermore, in the case of TMC complexes, where the bond connectivity is diffuse, the SMILES may not account for the correct bond structure (Figure 5.1). In this context, SMARTS (SMiles ARbitrary Target Specification)[194] or SELFIES (Self-Referencing Embedded String)[195] have emerged to overcome certain limitations of SMILES strings. However, they still remain underdeveloped for TMCs. To fill this void, there has been a growing interest in graph-based representations in chemistry. The architecture of molecular graphs was thoroughly explained in Chapter 2.

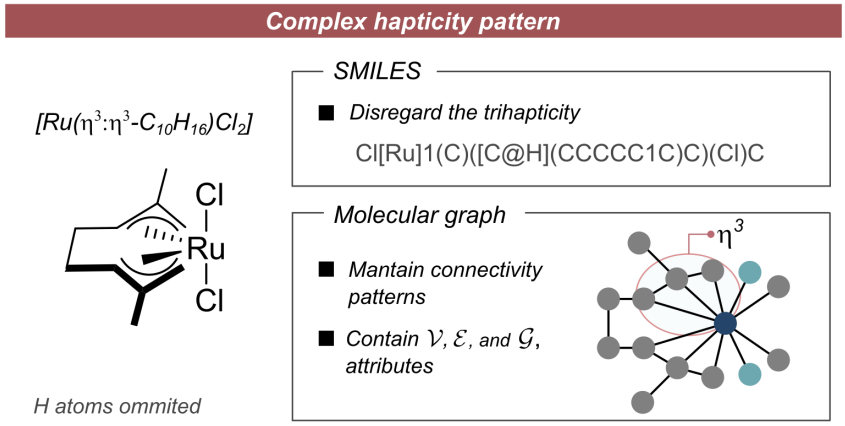


Figure 5.1: Molecular representations of the ruthenium-based catalyst: 2D drawing, SMILE notation and molecular graph. Hydrogen atoms are omitted for clarity.

The graphs consider the topology of a compound through the nodes and edges, further integrating properties through the use of attributes.

Moreover, these representations can be fed directly into Graph Neural Networks. This ML architecture processes graphs as inputs by loading the attributes of the scaffold. The use of molecular graphs in GNN has shown promising success within various chemical applications.[23,72] Nevertheless, this cutting-edge strategy is computationally expensive. Additionally, many ML models cannot handle graphs as input. Alternatively, molecular graphs may be utilized as a tool to incorporate sorted information, that can eventually be processed and served as inputs of a statistical model. To achieve this, *graph kernel* functions are utilised to extract relevant aspects from the graph as a feature engineering tool. This method accounts for similarities and differences, reducing the dimensionality of the representations, while preventing the loss of information. Herein, we have extended and developed a graph kernel to convert molecular graphs into fixed-length vectors. These vectors can be used as fingerprints in data-lead methods, effectively turning 2D graph information into 1D strings.

Among these kernels, we found the Moreau-Broto autocorrelation, which was first introduced for its applicability in organic chemoinformatics.[196] This approach is based on *walking* over the graph to obtain products of atomic properties. These atomic properties refer to five atomic heuristics: the atomic number (Z), the covalent radius (R), the electronegativity (χ), the atomic valence (V), and the identity (I). The algorithm maps the graph up to a certain depth, tracing the shortest connection between two autocorrelated atoms. Each atomic attribute provides a product property, and when aggregated, they constitute the components of a final vector. This final fixed representation is characterized by the user-selected depth and the number of properties per atom, $v(P, d)$. Therefore, the extraction of these product features leverages graph-based descriptors that can be seamlessly integrated into different ML models.

Kulik *et al.* have applied this approach to TMCs. They reformulated the classical autocorrelation function to provide a revised autocorrelation (RAC) formulation.[197] The authors defined new scopes regarding metal and ligand moieties, and incorporated additional mathematical operations into the functions beyond multiplications.

In this Chapter of the Thesis, we aim to reformulate the traditional Moreau-Broto autocorrelations and evaluate their efficacy in a TMC dataset. Our goal is to design autocorrelations that go beyond the classical representation, not only considering atomic products but also bond-bond and bond-atom (AABBA) correlations. Furthermore, we will also extend the five heuristic attributes to include geometric and electronic features. Finally, we will benchmark the performance of this representation within a Vaska's complex dataset.[124] A ML model will be fed using the AABBA-based autocorrelation vectors to predict the energy barriers and bond distances during the oxidative addition of the dihydrogen molecule.

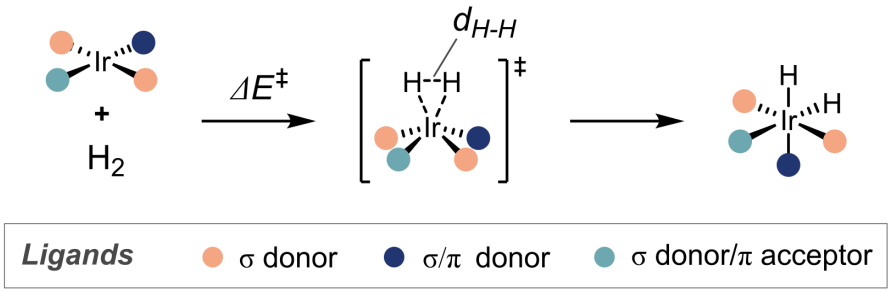


Figure 5.2: Scheme of the hydrogenation step of the Vaska's catalyst.

5.2 Database

Aiming to test the validity of the graph-based kernel, we employed a collection of 1,947 iridium complexes. These were derived from the structure of the Vaska's complex, $[Ir(PPh_3)_2(CO)(Cl)]$, and used in the study of the H_2-activation process (Figure 5.2). The compound scaffold contains one iridium metal with four ligands in a square planar manner. The set of ligands encompasses diverse σ donor, σ/π donor, and σ donor/π acceptor species. Further computational details and curation properties of the set are found in *Chem. Sci.*, **2020**, *11*, 4584-4601.

This transformation, as previously mentioned in Chapter 3, is highly employed in the hydrogenation stage leading to a stable molecular dihydride.[198] Thus, the dataset comprises 1,947 iridium-based compounds, along with computed results for their respective transition states in the oxidative addition of molecular hydrogen. This includes information about the energy barrier and the bond distance for breaking H−H.

In this work, we pursue to evaluate the performance of the vectors derived from the graph kernel, therefore, it was required to build the molecular graphs. These were downstreamed by Kneiding using the Hylleraas deep graph learning (HyDGL) program.[199] Here, the graph skeleton is provided by natural bond orbital theory, and second-order perturbation analysis[200] (SOPA), avoiding the bias of using manual

descriptions. These orbital-based topologies are defined as natural quantum graphs (NatQG). The implementation can account for interactions that may be crucial in catalytic procedures,[201] or complex denticity partners (*vida supra*). Following the drawing of the undirected NatQGs (u-NatQG), they are equipped with $\mathcal{G}$, $\mathcal{V}$, and $\mathcal{E}$ attributes. These graphs fall into two categories based on the origin of their properties: generic or NBO electronic properties.

5.2.1 Generic properties

We already mentioned the atomic generic properties (P_A):

$$P_A = \{Z, I, V, R, \chi\} \tag{5.2.1}$$

They are founded on tabulated constant values and are not derived from electronic calculations. Along this Chapter, we have referred to them as generic or periodic properties interchangeably. They provide information about the uniqueness of the atom (Z), the shape of the TMC (I), the number of bonds connected to an atom (V), the volume (R), and their electronic behaviour (χ). The identity property may be regarded as an obscure parameter, yet, it only takes the values of 0 or 1. It indicates whether a node is present or absent at any given depth. Moreover, bond polarization can be calculated from the difference between the electronegativities of the atoms. Overall, this ensemble of properties offers insight into the chemical constitution and environment.

The following properties correspond to the bond generic attributes (P_B):

$$P_B = \{BO, I, BD\} \tag{5.2.2}$$

where, BO is the Wiberg-based natural bond rounded order, I is the identity, in this case for the bond existence, and BD is the bond distance in Å.

5.2.2 NBO properties

The second category of graphs is characterized by its NBO electronic attributes. This compendium of NBO properties is obtained through NBO analysis, and the complete set of properties can be found in the Appendix C, specifically in Tables C.1 and C.2. This collection of NBO properties encompasses electronic attributes, such as the natural charge of the atom (q_{Nat}), the electronic occupancies of the s, p, d orbitals, (N_s, N_p, N_d, respectively), or the character of the bonding NBOs (BN_s, BN_p, BN_d). The number of properties is higher for the NBO set than for the GP group, $|P_{A,GP}| = 5$, $|P_{B,GP}| = 3$, and $|P_{A,NBO}| = 19$, $|P_{B,NBO}| = 16$.

5.2.3 Whole-graph properties

Besides the periodic and NBO properties, each $\mathcal{G}$ is characterized by four global properties, $P_{\mathcal{G}}$. The set includes the charge of the metal complex (q), its molecular mass (M), and the total number of atoms (N_{At}) and electrons (N_e).

5.3 ABBA Kernel

This work focused on the development of the graph kernel called atom–atom, bond–bond, and bond–atom autocorrelation (AABBA-AC). This autocorrelation function (f_{AC}) yields a fixed-length vector (v_{AC}) for any molecular graph $\mathcal{G}$, regardless of its size. This transformation is considered a compression operation, since the dimension of v_{AC} is generally smaller than the $\mathcal{G}$, given a molecular fingerprint. In this context, two different flavours have been proposed: AABBA(I) and AABBA(II). Further details of each function are detailed below, and the code to perform them is accessible at https://github.com/lmoranglez/AABBA.

Before, delving deep into the formulation, two main concepts must be addressed: the depth and the origin of the AC. The depth (d) refers to

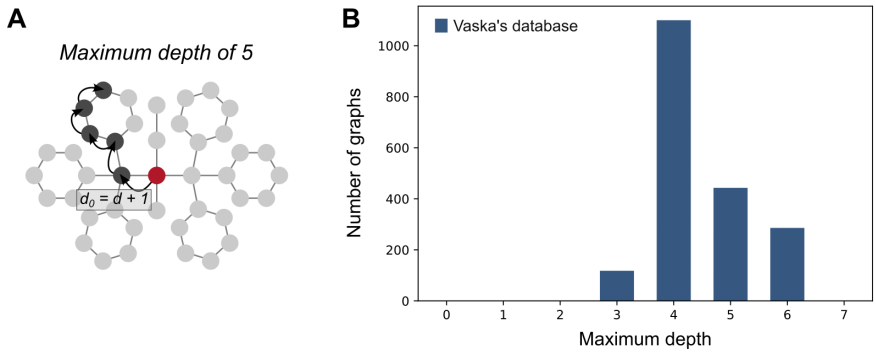


Figure 5.3: Path to reach the maximum metal-centered depth for the molecular graph: the path connects the metal index with the furthest atomic node (A). Histogram plot of the distribution of the maximum depth per molecular graph in the Vaska's dataset (B).

the minimum number of edges connecting two nodes. In the case, of the bond-bond AC (BB-AC), this applies to the minimum number of nodes connecting two edges. In this walk, one of the nodes (or edges), i, is the origin, thus, it is located at $d_i = 0$, and the path finishes in a target node (or edge) set at the specified depth ($d_j = d_{i,j}$). The maximum depth (D) denotes the necessary number of jumps to cover all the nodes (or edges) of the $\mathcal{G}$. Therefore, the path starts in the origin index and goes to a target index.

Figure 5.3 shows a walk over the $\mathcal{G}$. Starting from the red node ($d = 0$) and proceeding stepwise through the nodes to finish in the dot at $d = 5$. In this case, D is equal to 5. When we apply the autocorrelation functions, two scenarios are possible: (i) to fix the origin center in the metal, which is trivial, since the dataset contains mononuclear metal complexes; or (ii) to apply the AC function recursively while fixing all the nodes (or edges) of the $\mathcal{G}$ at $d = 0$ once. The first approach is metal-centered (MC) AC, and the second is referred to as full (F) ACs. From a chemical perspective, the former method is preferable for better interpretability. Initiating the analysis from the metal centre simplifies the extraction of chemical insights

from the prediction models. For further details, see the full set of equations below.

5.3.1 Atom-Atom autocorrelations

The first method encompasses the traditional Moreau-Broto autocorrelation. We denoted this as atom-atom autocorrelation (AA-AC) (Figure 5.4, and to compute v_{AA}, the following autocorrelation function must be applied.

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{i=1}^{N_{\mathcal{G}}} \sum_{j=1}^{N_{\mathcal{G}}} p_i p_j \delta_{d,d_{i,j}} \tag{5.3.1}$$

where $N_{\mathcal{G}}$ is the number of atomic nodes in the molecular graph, p is an atomic property, d is the depth, i and j are the atomic indices, and $d_{i,j}$ is the shortest path connecting the node j from i. The last term, $\delta_{d,d_{i,j}}$, is the Kronecker delta, which is defined as $\delta_{d,d_{i,j}} = 1$ for $d = d_{i,j}$, and 0 for $d \neq d_{i,j}$.

Upon application of Equation 5.3.1, the properties of the atoms i and j, which are separated at a given d, are then correlated by multiplication. The results constitute the elements of the final v_{AA} vector. The f_{AC} is applied to a property and at a specific depth. Yet, it is important to note that the resulting v_{AC} is accumulative, meaning that it incorporates all the depths up to the specified value of d. In the Equation 5.3.2 the d is equal to D.

$$v_{AA}(D) = (f_{AC}(d = 0), f_{AC}(d = 1), ..., f_{AC}(d = D)) \tag{5.3.2}$$

Furthermore, as explained in Section 5.2, each component of the molecular graph contains a group of properties. Thus, the f_{AC} may be executed as many times as, in this case, total atomic properties (K) are. Combining both concepts, the accumulative depth and the different properties, the resulting v_{AA} vector is built as below,

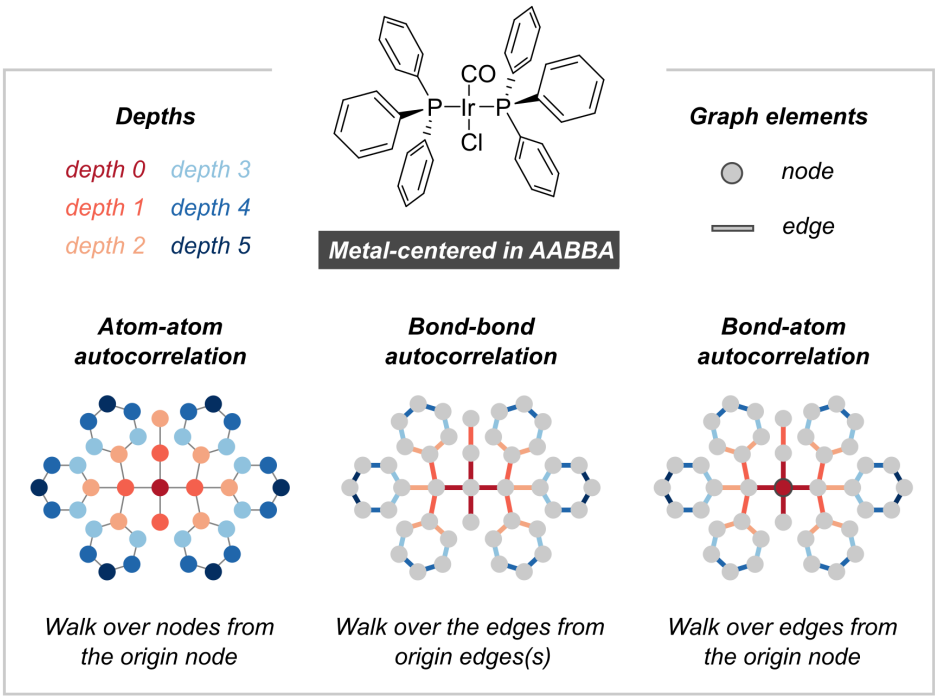


Figure 5.4: Definition of metal-centered depth approaches in AABBA kernel: atom–atom, bond–bond, and bond–atom autocorrelations. Hydrogen atoms and nodes are omitted for clarity.

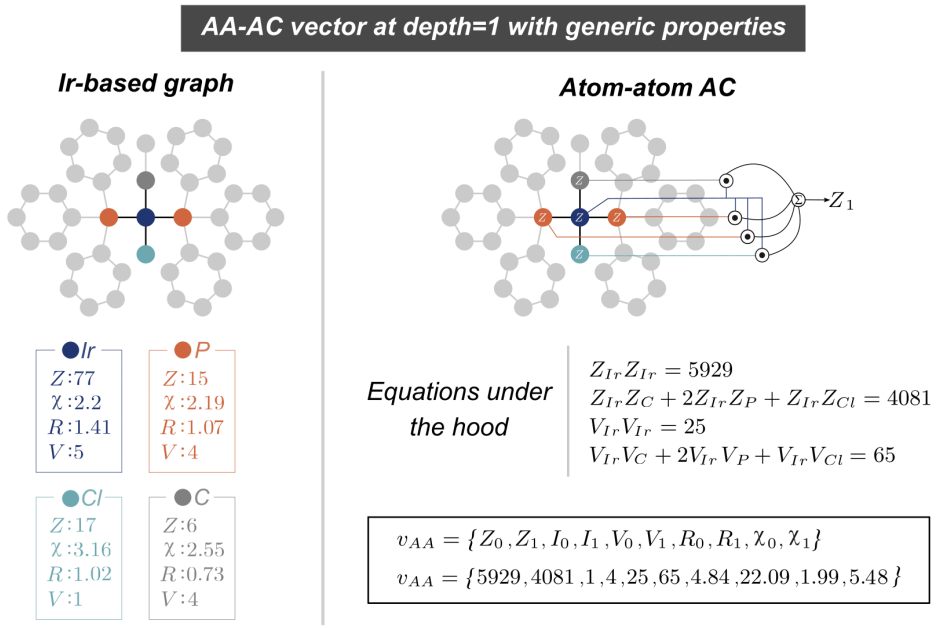


Figure 5.5: Practical example of calculating the metal-centered v_{AA} of the Ir-based graph with depth equal to 1.

$$v_{AA}(P_A, D) = (f_{AC}(P_{A,1}, d = 0), ..., f_{AC}(P_{A,1}, d = D), ..., f_{AC}(P_{A,K}, d = 0), ..., f_{AC}(P_{A,K}, d = D)) \tag{5.3.3}$$

An interesting aspect of the graph kernel is that leverages fixed-length vectors, which can be seamlessly integrated into machine learning models. The following Equation 5.3.4 provides the maximum dimension of the vector v_{AA}.

$$dim(v_{AA}) = (D + 1) \times K \tag{5.3.4}$$

Figure 5.5 provides an illustrative example of generating a v_{AA} with a depth equal to 1, containing all the generic atomic properties. Within it,

the equation below must be solved.

$$\begin{aligned} v_{AA}(P_A, d=1) = (&f_{AC}(Z, d=0), f_{AC}(Z, d=1), f_{AC}(I, d=0) \\ &f_{AC}(I, d=1), f_{AC}(V, d=0), f_{AC}(V, d=1), \\ &f_{AC}(R, d=0), f_{AC}(R, d=1), f_{AC}(\chi, d=0), f_{AC}(\chi, d=1)) \\ = &(Z_0, Z_1, I_0, I_1, V_0, V_1, R_0, R_1, \chi_0, \chi_1) \end{aligned}$$

(5.3.5)

The final vector is formed by five properties at a depth of 1, therefore, $dim(v_{AA}) = (1+1) \times 5 = 10$.

In addition, as mentioned before, the origin of the equations can vary to perform F or MC autocorrelations. In the case of the full autocorrelation (Equation 5.3.1), the vector derived from each origin node i of a given depth, $v_{AA}(i, d)$ is summed with the vectors resulting from other origin nodes i' at the same depth, thus,

$$v_{AA}(d) = v_{AA}(i, d) + v_{AA}(i', d) + ... + v_{AA}(N_i, d) \tag{5.3.6}$$

where N_i denotes the total number of nodes at a determined depth. Whereas in the MC equations, the index i of Equation 5.3.1 is always the metal center M.

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{j=1}^{N_{\mathcal{G}}} p_M p_j \delta_{d, d_{M,j}} \tag{5.3.7}$$

The final aspect to consider in the AC functions is the mathematical operation. In the first implementation of the autocorrelations, the properties were only correlating via a multiplication ($\odot$).[196] Herein, we expanded the equations to consider different arithmetic procedures: division as ratiometric function, $\oslash$ (Equation 5.3.8), summation as summetric function, $\oplus$ (Equation 5.3.9), and subtraction as deltametric function, $\ominus$

(Equation 5.3.10) autocorrelations.

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{i=1}^{N_{\mathcal{G}}} \sum_{j=1}^{N_{\mathcal{G}}} \frac{p_i}{p_j} \delta_{d,d_{i,j}} \tag{5.3.8}$$

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{i=1}^{N_{\mathcal{G}}} \sum_{j=1}^{N_{\mathcal{G}}} (p_i + p_j) \delta_{d,d_{i,j}} \tag{5.3.9}$$

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{i=1}^{N_{\mathcal{G}}} \sum_{j=1}^{N_{\mathcal{G}}} (p_i - p_j) \delta_{d,d_{i,j}}, \tag{5.3.10}$$

5.3.2 Bond-Bond autocorrelations

The reactivity pattern is primarily governed by the atomic nature of the compounds. As a result, atomic properties will be closely linked to these transformations. Nevertheless, it is important not to overlook bond properties, as they can also furnish valuable insights for predicting reactions. Inspired by this argument and the donor-acceptor interactions between bond orbitals in NBO, we extended the classical AC concept to develop the bond-bond autocorrelations (BB-AC) (Figure 5.4). The autocorrelation function can be performed in the same fashion as for the AA-AC (Equation 5.3.1). Nevertheless, in this context, the correlated features are bond properties providing the v_{BB} vector. Additionally, the $N_{\mathcal{G}}$ is redefined to the number of bonds in the molecular graph, p is referred to a bond property, and i and j are bond indices. The depth is referred to as the shortest distance, in number of atoms, connecting two bonds i and j.

As the number of properties is different in atoms or bonds, the Equation 5.3.11 to determine the vector dimensionality is also different.

$$dim(v_{AC}^{BB}) = (D + 1) \times L \tag{5.3.11}$$

being L the number of bond properties. Thus, to derive the v_{BB} with the same conditions as before ($d = 1$) and GP properties as in Equation

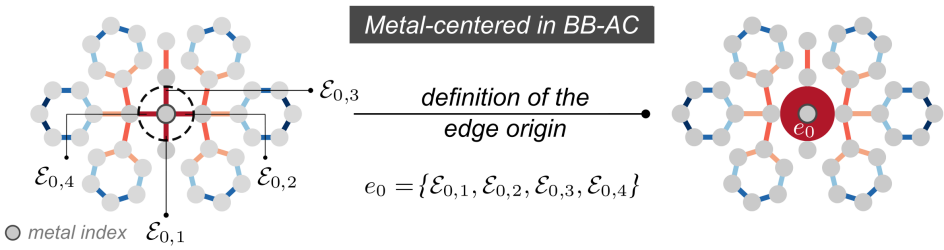


Figure 5.6: Definition of the edge origin e_0 (red ellipse) in the metal-centered BB-AC autocorrelation. The dotted-line black circle comprises the metal center and the bonds connected to it.

5.3.5), the formula to be applied is the following,

$$\begin{aligned} v_{BB}(P_B, d=1) = (&f_{AC}(BO, d=0), f_{AC}(BO, d=1), f_{AC}(I, d=0) \\ &f_{AC}(I, d=1), f_{AC}(BD, d=0), f_{AC}(BD, d=1), \\ = (&B_0, B_1, I_0, I_1, BD_0, BD_1) \end{aligned}$$

The resulting v_{BB} formed with three properties at $d = 1$ has a dimension of 6.

In the full BB-AC, the reinterpreted Equation 5.3.1 is conducted defining all bonds at the origin once. However, it is difficult to transfer the metal-center approach into the BB-AC. Here, there is not a unique bond that dominates over the rest. Thus, we need to reformulate the strategy in the metal-centered BB-AC. In this implementation, the origin is defined by the several bonds connecting the metal center to the ligands, the metal-ligand bonds.

Figure 5.6 illustrates the definition of the edge origin (e_0). To do so, the index of the metal serves as a means to identify the bonds associated with it. These labelled bonds are the edges at $d = 0$, and as a whole, they form the edge origin e_0, *i.e.* $e_0 = \{\mathcal{E}_{0,1}, \mathcal{E}_{0,2}, ..., \mathcal{E}_{0,CN}\}$, where CN is the coordination number of the metal center. The next step to be addressed is the definition of

the properties associated with the edge origin, p_{e_0}. Two different strategies have been implemented: either (i) averaging the properties of the edge set,

$$\overline{p}_{e_0}(CN, p) = \frac{\sum_{i=1}^{CN} p_{e_{0,i}}}{CN} \tag{5.3.12}$$

(ii) or summing them up to create a *super bond*:

$$p_{e_0}(CN, p) = \sum_{i=1}^{CN} p_{e_{0,i}} \tag{5.3.13}$$

Once the edge origin and its properties are defined, the metal-centered bond-bond autocorrelation is calculated with this function:

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{j=1}^{N_{\mathcal{G}}} \mathcal{P}_{e_0} p_j \delta_{d,d_{e_0,j}} \tag{5.3.14}$$

This AC function is denoted as $\overline{\text{BB}}$-AC or BB-AC, depending on whether $\mathcal{P}_{e_0}$ is equal to $\overline{p}_{e_0}$ or p_{e_0}, respectively. It is remarkable that, although the approach to handling the properties of the first bond sphere varies, the properties themselves remain constant. Consequently, substantial changes are not expected.

5.3.3 Bond-Atom autocorrelations

The last implementation within the AABBA(I) model is the bond-atom autocorrelation (BA-AC). This method relies on the concepts of the message-passing in GNNs, where neighbouring nodes or edges exchange information and influence each other's updated embeddings. It also draws from NBO analysis, where localized and bond orbitals interact with each other. In this context, an atom is set as the origin $d = 0$, and the AABBA graph walks from there to over the edges of the $\mathcal{G}$ (Figure 5.4). Equation 5.3.15 defines the BA-AC function.

$$f_{AC}(N_{\mathcal{G},\mathcal{V}}, N_{\mathcal{G},\mathcal{E}}, p, d) = \sum_{i=1}^{N_{\mathcal{G},\mathcal{V}}} \sum_{j=1}^{N_{\mathcal{G},\mathcal{E}}} p_i p_j \delta_{d,d_{i,j}} \tag{5.3.15}$$

where $N_{\mathcal{G},\mathcal{V}}$ refers to atomic nodes and $N_{\mathcal{G},\mathcal{E}}$ to bond edges, i and j correspond to the indices of the atom and bond, respectively, ad thus, the p denotes the corresponding properties of each. The term $d_{i,j}$ is the shortest path in nodes between the node i and the edge j. For the metal-centered BA-AC, the equation that leverages the v_{BA} is analogous to Equation 5.3.15. However, in this case, the node i always corresponds to the metal center M.

At first glance, it may seem challenging to correlate these two sets of properties, given their different dimensionality. We designed Equation 5.3.16, with the specified arithmetic operation, to obtain the autocorrelations between bonds and atoms.

$$p_i p_j = \sum_{l=1}^{L} p_i p_{j,l} \tag{5.3.16}$$

In this context, i and j are the indices of the origin node and the edges, respectively, and L is the number of bond properties. To account for the shape of the v_{BA} is necessary to apply again Equation 5.3.4. In this method, to obtain the v_{BA} at depth 0, involves computing the following array,

$$v_{BA}(d=0) = Z'_0 + T'_0 + I'_0 + S'_0 + \chi'_0 \tag{5.3.17}$$

The derivation of this $v_{BA}(d = 0)$ is explained below, where Each element in this array has undergone *prior* processing.

$$
\begin{aligned}
f_{AC}(Z, d=0) &= Z_0B_0 + Z_0I_0 + Z_0BD_0 = Z'_0 \\
f_{AC}(T, d=0) &= T_0B_0 + T_0I_0 + T_0BD_0 = T'_0 \\
f_{AC}(I, d=0) &= I_0B_0 + I_0I_0 + I_0BD_0 = I'_0 \\
f_{AC}(S, d=0) &= S_0B_0 + S_0I_0 + S_0BD_0 = S'_0 \\
f_{AC}(\chi, d=0) &= \chi_0B_0 + \chi_0I_0 + \chi_0BD_0 = \chi'_0
\end{aligned}
\tag{5.3.18}
$$

$$
\begin{aligned}
v_{BA} &= \{Z'_0, Z'_1, I'_0, I'_1, V'_0, V'_1, R'_0, R'_1, \chi'_0, \chi'_1\} \\
v_{BA} &= \{5929, 4081, 1, 4, 25, 65, 4.84, 22.09, 1.99, 5, 48\}
\end{aligned}
\tag{5.3.19}
$$

5.3.4 Autocorrelation ensemble - AABBA(I) and AABBA(II)

At that point, we developed autocorrelation functions to relate atomic and bond properties, both independently and crossing the terms. Aiming to obtain a comprehensive representation of the molecular graph, and thus, of the metal complex, we formulated a generalized feature vector, denoted as v_{AABBA}. We yielded this vector following two strategies: AABBA(I)-AC and AABBA(II)-AC. The former compilates the previously defined autocorrelations, including AA-AC, $\overline{\text{BB}}$-AC and BA-AC, join their outcomes and generate the v^I_{AABBA} (Equation 5.3.20).

$$
v^I_{AABBA} = v_{AA} \oplus v_{\overline{BB}} \oplus v_{BA} \tag{5.3.20}
$$

This novel vector has significantly increased its dimensionality with respect to the previous implementations, as demonstrated in Equation 5.3.21.

$$
dim(v^I_{AABBA}) = (D+1) \times (2K+L) \tag{5.3.21}
$$

Here, D is maximum depth, K is the number of atomic properties, L, the number of bond properties. As discussed thoroughly in the Thesis, a higher amount of dimensionality leads to a greater characterization of the molecule. However, it is worth mentioning that this practice comes at a higher computational cost and may potentially saturate the models in which these vectors are employed.

The second strategy, AABBA(II), relies on a sort of manually *data comprehension*. In this context, we selected the edges of the graph to extract information about both the bond and the atoms that constitute them. Thus, we obtain a combination of atomic and bond properties, P_{AB} in the feature vector. In total, new five sets of P_{AB} were obtained to describe each edge of the molecular graph. Within the generic properties, we found $P_{AB,1}$, $P_{AB,2}$ and $P_{AB,3}$ (Equation 5.3.22).

$$\begin{aligned} P_{AB,1} &= \{Z_i, Z_j, V_i, V_j, \chi_i, \chi_j, BD, BO, I\}; M = 9 \\ P_{AB,2} &= \{Z_i, Z_j, V_i, V_j, \chi_i - \chi_j, BD, BO, I\}; M = 8 \\ P_{AB,3} &= \{Z_i, Z_j, V_i, V_j, \chi_i - \chi_j, R_i, R_j, BO, I\}; M = 9 \end{aligned} \tag{5.3.22}$$

The indices i and j denote each atomic index forming the edge $\mathcal{E}_{i,j}$. The three groups of properties contain the atomic number, and the valence of each of the atoms belonging to the bonds, together with the bond distance and the bond identity. Furthermore, the first set, $P_{AB,1}$, also includes the electronegativity of each of the atoms of the bond, χ_i and χ_j, and the bond order. Within the $P_{AB,2}$ and $P_{AB,3}$, we included the bond polarity via the difference in electronegativity of the atoms, $\chi_i - \chi_j$. In addition, $P_{AB,2}$ contains the bond distance parameter, while the $P_{AB,3}$ accounts for the dimensionality of the bond by including the covalent radius of each of the binding atoms. It is worth mentioning that the last group of GP properties does not hold any computed property; all of the parameters are tabulated constants.

We have additionally defined two sets of P_{AB} based on NBO properties.

$$P_{AB,4} = \{q_{Nat,i}, q_{Nat,j}, V_{Nat,i}, V_{Nat,j}, N_{s,i}, N_{s,j}, N_{p,i}, N_{p,j}, N_{d,i}, N_{d,j}, N_{LP,i}, N_{LP,j}, N_{LV,i}, N_{LV,j}, BD, BO_{Nat}, N_{BN}, BN_s, BN_p, BN_d, N_{BN^*}, BN_s^*, BN_p^*, BN_d^*, I\}; M = 25$$

$$P_{AB,5} = \{q_{Nat,i}, q_{Nat,j}, V_{Nat,i}, V_{Nat,j}, N_{LP,i}, N_{LP,j}, LP_{E,i}, LP_{E,j}, LP_{\Delta E,i}, LP_{\Delta E,j}, N_{LV,i}, N_{LV,j}, LV_{E,i}, LV_{E,j}, LV_{\Delta E,i}, LV_{\Delta E,j}, BD, BO_{Nat}, N_{BN}, BN_E, BN_{\Delta E}, N_{BN^*}, BN_E^*, BN_{\Delta E}^*, I\}; M = 25$$

Concerning the $P_{AB,4}$ and $P_{AB,5}$ collections, they shared the natural charge and valence properties. For the rest of the properties, whereas $P_{AB,4}$ primarily encompasses properties related to orbital symmetry behaviour, $P_{AB,5}$ mainly characterizes the orbital energy information.

These groups of properties characterize each bond, thus, these will be the autocorrelated properties during the execution of AABBA(II) kernel. To clarify, when applying AABBA(II), we focus on the bond-bond autocorrelation mapping. When applying metal-centered approach $\mathcal{P}_{e_0}$ is equal to $\overline{p}_{e_0}$. Lastly, the dimensionality of the resulting vector, v_{AABBA}^{II} is

$$dim(v_{AABBA}^{II}) = (D+1) \times M \tag{5.3.23}$$

In this Equation, M is the number of properties of each $P_{AB,n}$ set and $n = 1, 2, 3, 4, 5$. In this variation of AABBA, the global information of the molecule is condensed in a reduced v_{AABBA}^{II}. For instance, if we set a $d = 3$, using generic properties, and we employed the AABBA(I) version, the $dim(v_{AABBA}^{I})$ is equal to $(3+1)\times(2\cdot5+3) = 52$. Instead, using AABBA(II), the total number of elements in v_{AABBA}^{II} is 36 for $P_{AB,1}$ and $P_{AB,3}$; and 32 for $P_{AB,2}$. Therefore, AABBA kernel offers two distinct approaches for

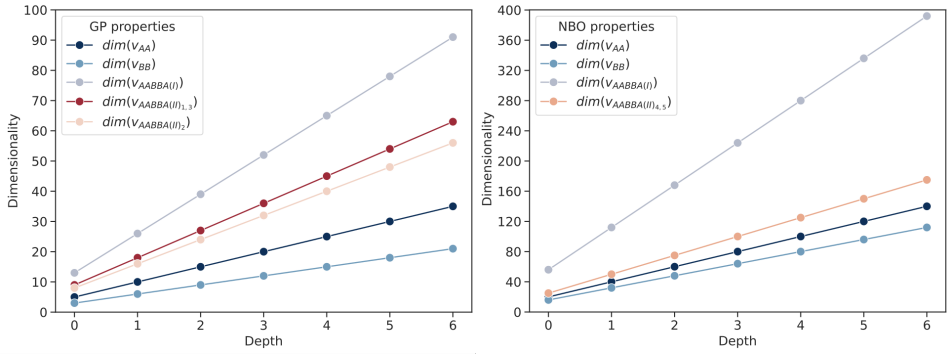


Figure 5.7: Scatter plots of the dimensionality of the vectors derived from the autocorrelation functions for the generic properties (left) and for the NBO-based properties (right).

combining the autocorrelated properties of atoms and bonds, known as the explicit, AABBA(I), and implicit modes, AABBA(II).

5.3.5 Dimensionality overview

In previous sections, we detailed the equations to calculate the dimensionality of the autocorrelation vectors. Figure 5.7 depicts a visual summary of how the number of autocorrelated features increases linearly as we delve deeper into the walk. In addition to the vector size, the time required to generate these vectors may vary significantly among the different AC functions, as well as the dataset's regime.

5.3.6 Computational implementation

After thoroughly discussed the theoretical framework of the AABBA graph kernel, some additional details will be provided on how the model is coded. This tool is an open-source code written in Python, which is capable of processing u-NatQGs to get the corresponding fixed-length autocorrelation vector. The code is built using the NetworkX[202] and Numpy[78] libraries. It can be found at https://github.com/lmoranglez/AABBA, and it is

organized into the following scripts:

- `graph_info.py` - read the chemical graph and extract a dictionary with the node and edge indices at specified depths. It also contains the property labels.
- `ac_functions.py` - perform autocorrelation functions.
- `utilities.py` - tools for manipulating and saving the data.
- `ac_multithread.py` - parallel implementation to perform the autocorrelation functions.

The code is designed in a functional manner, thus allowing the user to use the code according to the given problem. Revised graphs with other non-defined properties are easily implemented to address autocorrelation functions in other mononuclear TMCs. It can be transferred easily.

5.4 Neural network models

Upon performing the autocorrelations on the Vaska's dataset, we leveraged a set of vectors for each sample containing both periodic and NBO autocorrelated features. This process was executed for each of the mathematical operations mentioned previously (as outlined in Equations 5.3.7, 5.3.8, 5.3.9, and 5.3.10). To evaluate the effectiveness of these new chemical descriptors, we conducted predictive experiments with neural networks. The target properties were the energy barrier and H−H distance during the oxidative addition of H_2 within Ir-based complexes. The NNs contain a multilayer perception architecture where the hyperparameters were set manually based on previous work. We defined two hidden layers consisting of 128 nodes each, ReLU as the activation function, and Adam optimizer to minimize the MSE loss function. The dataset was randomly split into train, validation, and test sets at a ratio of 80:10:10 using a seed

of 2022. This seed is maintained during the evaluation process. However, the randomness of the model comes from the weight initialization. Thus, we repeated each training ten times providing the average mean absolute errors, and the 95% of confidence level. Moreover, we also highlighted the best outcome within each training set. We employed batch training, with a batch size of 32 for 200 epochs. The learning rate (lr) was initialized at 0.01, and it dynamically changes according to `ReduceLROnPlateau` scheduler of PyTorch library.[76] Here, the validation error is assessed, and in case it does not change during five consecutive epochs (*patience*), the lr is reduced by a *factor* of 0.7, as $lr_{updated} = 0.7lr$. Therefore, after each optimization, the training with the highest validation accuracy was evaluated over the test set. The performance metrics provided in the Thesis are the errors during the test stage. It is worth mentioning that upon the NN training model, data was subjected to standardization. To do so, we computed the mean, μ_{train}, and the standard deviation, σ_{train}, of the training set.

$$X_{std,train} = (X_{train} - \mu_{train})/\sigma_{train} \tag{5.4.1}$$

To prevent data leakage, we transferred the μ_{train} and σ_{train} to the validation and test sets. Furthermore, if a property has the same value across all samples, it was removed from the feature vector as it did not contribute meaningful information to the model. We conducted numerous NN trainings, but for the sake of simplicity and considering the benchmark aim of this project, we only highlighted some relevant outcomes. The results in the following sections correspond to inputs derived from metal-centered functions with a depth of 3, unless otherwise stated.

5.4.1 Energy barrier prediction

In the first attempt, we explored the regression prediction ability of the vectors with the energy barrier. Figure 5.8 shows the distribution of the energy barrier in our dataset. The range of values is distributed among the

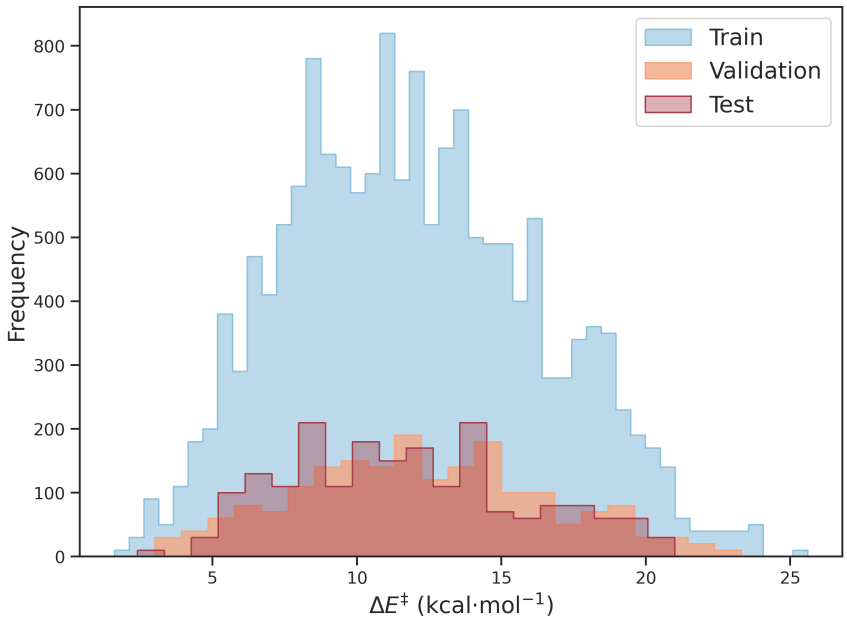


Figure 5.8: Distributions of energy barriers (in kcal·mol^{-1}) during the neural network training process.

three data clusters.

All results are collected in the Appendix C. In previous publication, authors employed the Moreau-Broto AA-AC, with different hyperparameters and another Python library, achieving after an in-depth search, a minimum MAE of 1.12 kcal·mol^{-1}.[124] In this study, we conducted a hierarchical testing of the input vectors, progressively increasing their complexity in terms of information and depth. Thus, we started by defining a metal-centered origin vector with a maximum depth of three. A general overview of the outcomes is depicted in Figure 5.9.

We started our analysis with the traditional autocorrelation with generic properties, AA-AC, where the best-reached outcome was a MAE of 1.16 kcal·mol^{-1} (entry 1 in Table C.3). Then, we tested the novel autocorrelations that include bond properties. Figure 5.9 shows a poorer prediction when using BB-, $\overline{\text{BB}}$-, and BA-AC approaches, being BA-AC

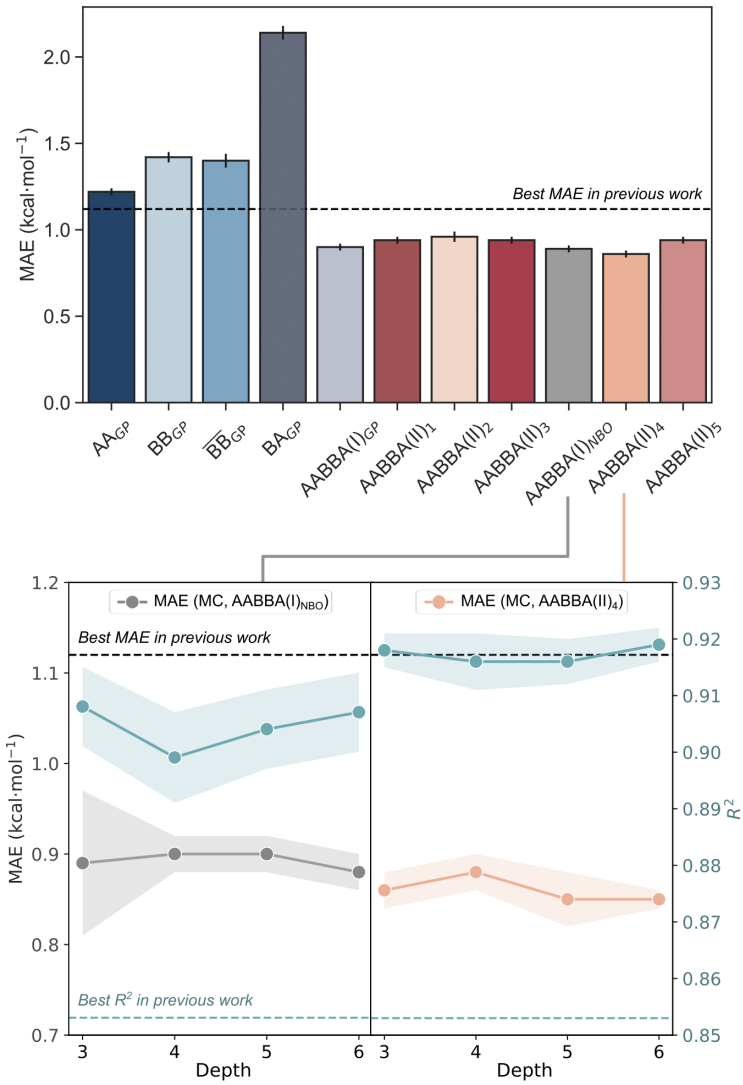


Figure 5.9: Mean absolute error (MAE) (in kcal·mol^{-1}) in the prediction of the energy barrier using vectors derived from autocorrelation functions (top). Mean absolute error (in kcal·mol^{-1}) and R^2 in function of a given depth using the ABBA(I)$_{NBO}$ (left-bottom) and ABBA(II)$_4$ autocorrelations (right-bottom).

the worst of all the methods tested so far. The results suggest that autocorrelated bond properties are less effective molecular representations than classical AC in this task. Furthermore, a decrease in vector dimensionality, such as $dim(v_{AA}) = 18$ to $dim(v_{\overline{BB}}) = 10$, can lead to reduced prediction accuracy. Furthermore, the average bond, $\overline{BB}$-AC, outperformed the superbond approach, and thus, the former will be selected for concatenation purposes in metal-centered ACs. To sum up, none of the isolated kernels achieved better results than the previously attained ones. Later, we concatenated all the derived vectors, v_{AA}, $v_{\overline{BB}}$, v_{BA}, within the defined AABBA(I)-AC. This global molecular representation yielded a minimum MAE of 0.86 kcal·mol^{-1} (fifth bar in Figure 5.9). This result was a hit as its MAE is below 1 kcal·mol^{-1}, and outperformed the reported attempts. This implementation represents a 26 % decrease in error compared to the first model (AA-AC).

We mainly focus on the product autocorrelation, but there are alternative arithmetic operations. In this regard, we thought about properties that represent key chemical concepts. For instance, the polarity of the bond is the difference in electronegativities between the binding atoms, thus, a deltametric function will be suitable for the χ property. Moreover, we also proposed the ratiometric function for the covalent radius, since it is related to the relative atomic size. Models 6 and 7 of Table C.3 show that these fine-grained settings do not improve the forecast performance.

The subsequent refinement in the complexity of the model is the application of NBO properties. We exploited the AABBA(I) kernel with P_{NBO} leveraging results (Figure 5.9) that outperformed the GP-based predictions reaching a maximum accuracy of 0.85 kcal·mol^{-1}, in the prediction, and R^2 of 0.913. This outcome is enhanced with the replacement of the metal-center functions by the full autocorrelations achieving a MAE of 0.76 kcal·mol^{-1} (model 9 in Table C.3). In light of the results, we attempted to enhance the accuracy of the full autocorrelation at $d = 3$ by

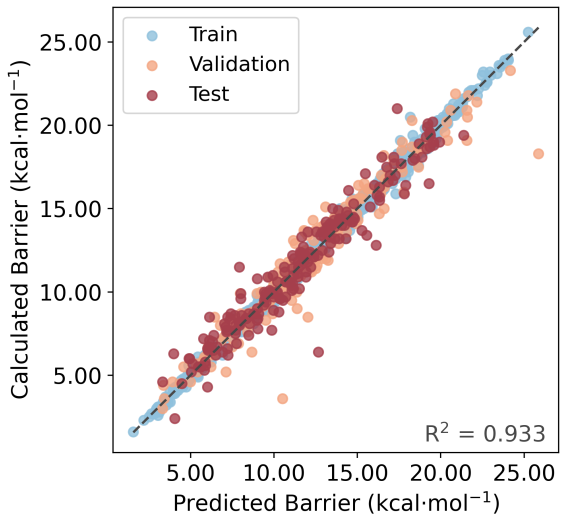


Figure 5.10: Pair plots depicting the correlation between the DFT-calculated and NN-predicted energy barrier ($\Delta E^{\ddagger}$) within the Vaska's dataset. Data points follow the color-coded: • training, • validation, and • test sets.

concatenation the whole-graph properties. The outcome was only slightly improved (model 13 in Table C.3). This model achieved the best outcome within all the numerical experiments and its prediction ability is shown in Figure 5.10. We examined the robustness of the most accurate ML model. To do so, we repeated ten times the training of model using different random training:validation:test split. The results were aligned with the expectations to further confirm their reliability. At that point, it is important to remark that MC approaches are preferred for the sake of interpretability. It is easier to rationalize the impact of the properties according to the distance to the metal index, *e.g.* proximal, middle, or distance effects.

Therefore, we focused on the impact of increasing the depth of the molecular fingerprint from 3 to 6 in the metal-centered AABBA(I)$_{NBO}$ implementation. The left-bottom plot in Figure 5.9 provided an overview of the results, where all the models achieved averaged results of roughly 0.90 kcal·mol^{-1}. The best performance was found at a depth equal to 6

yielding a MAE of 0.84 kcal·mol^{-1} and a R^2 of 0.918 (model 12 in Table C.3). However, this increment of the depth did not significantly improve the prediction performance.

Lastly, we turned our attention to the AABBA(II) kernel. As mentioned in Section 5.3.4, each of the five sets encompassed either generic or NBO features. The MAE is less than 1.00 kcal·mol^{-1} when applying all these strategies (Figure 5.9). In the case of the GP, the concatenation-based AABBA(I) remained giving the lowest MAE compared to AABBA(II)$_1$, AABBA(II)$_2$ and AABBA(II)$_3$. In contrast, within the NBO frame, the compressed global-NBO metal-centered AABBA(II)$_4$ attained the best results within the set. In particular, the lowest average MAE was reached at $d = 5$ and $d = 6$ (right-bottom plot of Figure 5.9). The highest performance of this collection of metal-centered numerical experiments was attained at a depth equal to 5 with a MAE of 0.81 kcal·mol^{-1} (entry 21 in Table C.3).

Upon a thorough examination of Table C.3, we observed that the best accuracy was attained with the full-AABBA(I)$_{NBO}$-AC. Its derived input vector contains a $dim(v^I_{AABBA_{NBO}})$ of 227 elements. On the other hand, a slightly reduced accuracy was observed with MC-AABBA(II)$_4$-AC with a $dim(v^{II}_{AABBA_4})$ of 129. This implies a reduction of 42 % of the features. In addition, this latter approach is beneficial in terms of elapsed time and comprehensibility.

We could conclude after an in-depth analysis, that the $\Delta E^{\ddagger}$ parameter is mostly related and thus, accurately reproduced with global electronic features.

5.4.2 H-H bond distance prediction

In exploring the H−H distance during the transition state, we also plotted the dataset based on the training split, as shown in Figure 5.11.

We started to run numerical experiments with the baseline approach: metal-centered at a depth of 3 with generic properties. We also depicted the average prediction ability using different inputs with the earlier mentioned

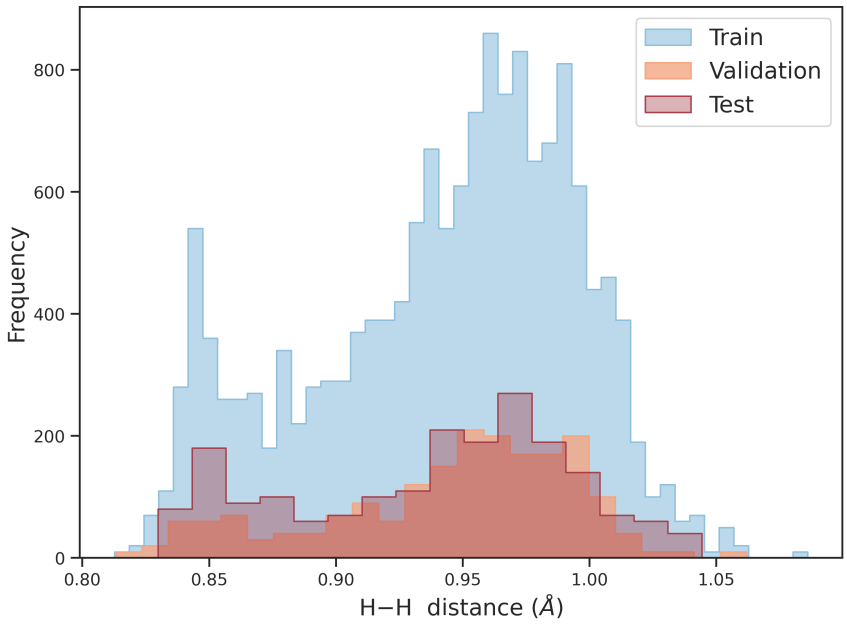


Figure 5.11: Distributions of $H - H$ distance in Å during the neural network training process.

parameters and diverse properties, accordingly (Figure 5.12).

Regarding the simplified approaches, AA-, $\overline{\text{BB}}$-, BB-, BA-AC, we noticed that the first three functions yielded comparable accuracies, *i.e.* their MAEs range from $2.30{\cdot}10^{-2}$ to $2.38{\cdot}10^{-2}$ Å (Table C.4). Unlike the barrier prediction (Table C.3), where the classical AA-AC exhibited the best outcome among these kernels, here, the $\overline{\text{BB}}$-AC obtained the highest performance within this set achieving a MAE of $2.09{\cdot}10^{-2}$ Å. The BA-AC was also tested but displayed poorer efficiency. Consequently, we confirmed the expected hypothesis that bond properties are crucial in the prediction model of the bond distances. It is worth mentioning that the models exhibit lower accuracy in terms of R^2 for the $d(H - H)$ values when compared to the predictions for the energy barriers (Tables C.3 and C.4).

Following the same procedure as in the energy barrier experiment, we built the comprehensive vector by concatenating all the individual

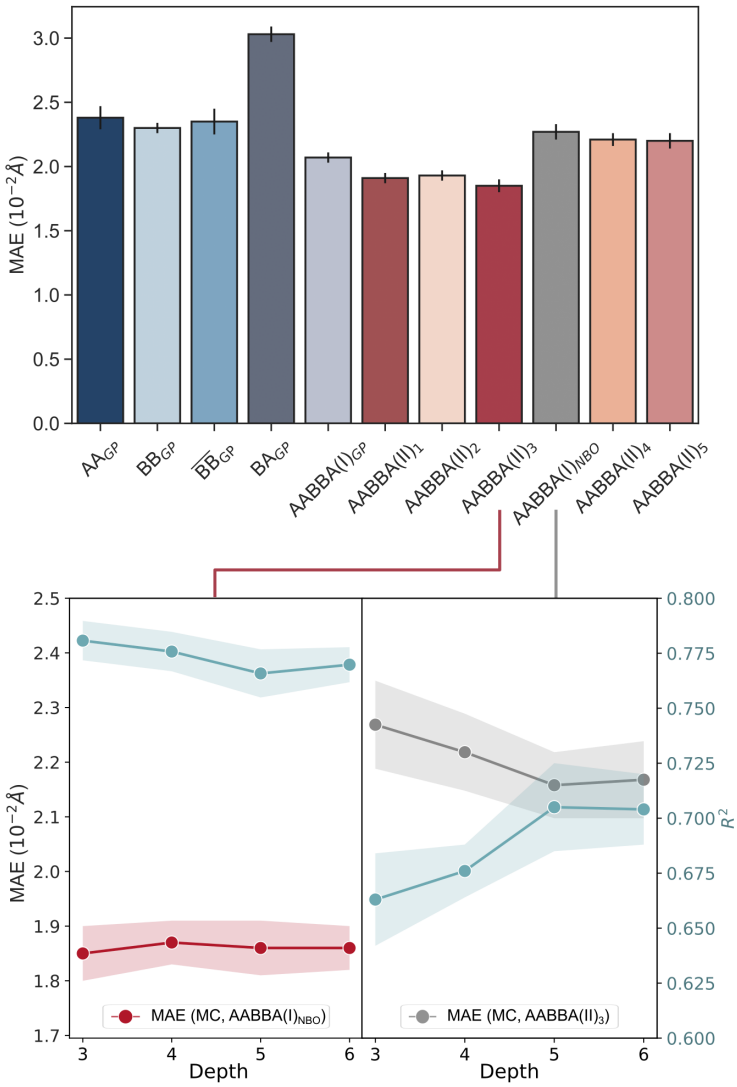


Figure 5.12: Mean absolute error (MAE) in 10^{-2}Å in the prediction of the H−H distance using different vectors derived from autocorrelation functions (top). Mean absolute error and R^2 in function of the given depth applying the ABBA(II)$_3$ (left-bottom) and ABBA(I)$_{NBO}$ autocorrelations (right-bottom).

approaches. This implementation involves a reduction of the MAE with respect to the $\overline{\text{BB}}$-AC by 12 % (model 27 in Table C.4). The performance with the full AC for AABBA(I) was unsuccessful, leading to a poorer prediction ability compared to the metal-centered approach (model 28 in Table C.4). Inspired by earlier experiments, we also fined-tuning the product of the electronegativity and the covalent radius by their deltametric and ratiometric functions, respectively. For the first replacement, we achieved a prediction with a MAE of $1.87{\cdot}10^{-2}$ Å. Therefore, the bond polarity plays a crucial part in predicting the bond distance. While for the relative size, the performance was not significantly improved compared to the original AABBA(I)-AC.

To further analyse the impact of the properties, numerical experiments were conducted using electronic features. We observed that the models' accuracy diminished with the use of NBO in comparison to periodic properties. This behaviour is opposed to the one found in the $\Delta E^{\ddagger}$ prediction. For instance, the ABBA(I)-AC approach with NBO properties provided a MAE 10.4 % higher compared to its periodic analogous. The bar plot presented in Figure 5.12 shows a lower MAE for the GP-based approaches, contrasted with the last three bars founded on NBO properties. Aiming to enhance the NBO-informed forecasting, we investigated NNs by shifting the depth from 3 to 6 in ABBA(I)_{NBO} (as depicted in the left-bottom plot of Figure 5.12). Despite the rise in vector dimensionality, the performance remained almost unchanged across the different depths.

Delving into the customized atomic-bond property vectors, we found that GP features outperformed electronic ones, as shown by the AABBA(II)_n bars in Figure 5.12. Figure 5.13 displays the most accurate model driven by the AABBA(II)_3-AC with a MAE of $1.72{\cdot}10^{-2}$ Å, which encompasses generic properties (entry 40 in Table C.4). Interestingly, the $P_{AB,3}$ set contains the polarity of the bond and the covalent radius of each of the atoms forming the bond. Therefore, the $v^{II}_{ABBA,3}$ lacks the bond distance feature, being a geometric-agnostic approach.

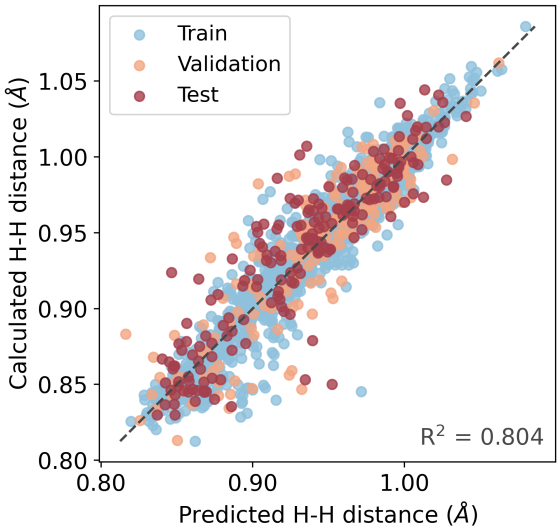


Figure 5.13: Pair plots depicting the correlation between the DFT-calculated and NN-predicted distance (d(H−H)) within the Vaska's dataset. Data points follow the color-coded: • training, • validation, and • test sets.

Endeavours to further minimize the MAE were collected in the left-bottom plot of Figure 5.12 increasing the depth of the input vectors. In this regard, the profoundness increment did not yield more accurate results. Finally, we repeated ten times the training process of the most accurate model using random splits. The model's robustness was confirmed as the results align with those previously obtained.

To conclude, the distance between the hydrogen atoms is largely affected by the generic properties in a local environment rather than global electronic effects.

5.5 Conclusions

This Chapter focused on the development and application of the graph kernel. This tool extracts information from molecular graphs to derive molecular vector representations. Our approach involves the introduction

of the atom–atom, bond–bond and bond-atom autocorrelations. These functions gather generic and electronic properties of both atoms and bonds, and encode them in autocorrelation vectors. The concatenation of these autocorrelations establishes the AABBA(I)-kernel. Our implementation leverages vectors containing correlated properties of a chemical compound. On the other hand, we also created the AABBA(II) variant. It involves a fusion of atomic and bond characteristics for correlation. Here, tracing a path over the edges of the molecular graph serves to derive the atom-bond features. This approach diminishes the vector size by concentrating on a limited number of properties. The code for employing the AABBA kernel is available on GitHub,[203]. It is organized into individual functions, which ensures modularity and customization.

The AABBA kernel was applied in the generation of the vectors for Vaska's database, which contains 1,947 molecular graphs. Thus, we derived all possible vectors for each of the iridium-based graphs. The prediction ability of the vectors was evaluated in the prediction of the energy barrier and H–H bond distance during the catalysed oxidative addition of H_2. To carry out this task, a multilayer perceptron with two hidden layers was employed. We provided a review of the performance of various autocorrelation vectors in a consistent manner: starting from the simplest approaches to the most complex vectors. The predictions with the extended BB- and BA-AC autocorrelations yielded outcomes that align with the ones provided by the traditional AA-AC. This indicates that the introduction of bond properties is beneficial for regression prediction tasks. In addition, the models' accuracy was significantly enhanced by introducing a comprehensive molecular graph descriptor using either AABBA(I) or AABBA(II) strategies, resulting in the best performance to date. Interestingly, the two AABBA flavours provide similar errors. In the case of the energy barrier predictions the v^{I}_{AABBA} is slightly superior than v^{II}_{AABBA}, whereas for the distance bond forecast,s the v^{II}_{AABBA} representation outperformed the v^{I}_{AABBA}. This finding is unexpected as the

dimensionality of the v^{I}_{AABBA} is much higher than the v^{II}_{AABBA}, showing the potential benefits of dimensionality reduction in this context. Particularly, in the prediction of the energy barrier, global electronic properties highly influenced the accuracy of the design. The best model within the set is achieved by using the full AABBA(I) with NBO properties combined with the entire properties of the graph. On the other hand, the bond distance between the hydrogen atoms is effectively predicted using generic properties. Specifically, the metal-centered $AABBA_3$(II) with a depth of 3, provided the most successful outcome. Notably, this last implementation does not contain any geometrical features.

In summary, this work presents a novel methodology for extracting molecular graph features, and utilizing the resulting vector in machine learning models, without losing information that may impact prediction accuracy.

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

Conclusions

Everything's coming up Milhouse!

— Milhouse Van Houten – The Simpson

This Thesis introduces novel descriptors aimed at enhancing our understanding of chemical processes.

In Chapter 3, we set up a free-to-access web application, *BDE Matrix App* and we used it to study two classes of ligands: N-heterocyclic carbenes (NHC) and monohapto dihydrogen. In the first example, we calculated the BDE between NHCs and five metal fragments of reference to derive the HDs of the NHC ligands. We observed that these descriptors identify electronic patterns within the NHC families. The HD_{L1} was claimed as a suitable descriptor to account for σ donation, and a global picture of the NHC compared with other ligands was provided. In the second example, the formation of stable $L_nM(\eta^1\text{-}H_2)$ complexes with various metal fragments was investigated. The generation of the HDs using the *BDE Matrix App*, and its subsequent evaluation, revealed that none of the metal fragments proposed are suitable for monohapto binding.

Along Chapter 4, we delved into the study of the bimolecular nucleophilic substitution reaction. We characterized hundreds of reactions involving nucleophiles acting as entering and leaving groups. Singular value decomposition was applied to the data to derive the importance of the first hidden descriptor (HD_1), which encapsulates 90 % of the chemical

information of the fragments. This descriptor was used to build a prediction model accessible through the *S_N2 Matrix App* to extend the analysis beyond the initial nucleophiles. Additional investigations found correlations between these HD-S_N2 descriptors and those HDs-BDE from the previous Chapter. This approach established a workflow for extracting nucleophile data and predicting energy barriers for reaction design.

In Chapter 5, we shifted the paradigm of the hidden descriptors, designing a tool that generates well-suited molecular descriptors for ML models. We developed the AABBA graph kernel which extracts molecular representations using both periodic and electronic properties. We assessed these descriptors in two regression tasks using a data set of the oxidative addition of H_2 within Vaska's complexes. We observed low mean absolute errors (MAEs) in the prediction of the energy barriers and bond distances. Furthermore, we attributed the most effective properties to precisely estimate the selected targeted properties. Our modular Python code for the AABBA kernel is publicly available, offering a valuable tool for generating molecular descriptors in chemical research.

Overall, an adequate representation of chemical compounds is essential. We have sought accurate and interpretable chemical descriptors, harnessing the power of data-driven approaches. These data-driven strategies have emerged as alternatives to traditional methods. They require less time and fewer resources, and are capable of solving chemical problems. Achieving a balance between time and effectiveness is crucial in the design of chemical descriptors. When dealing with a small dataset and intricate mechanisms, the search for optimal descriptors should be accomplished. Strategies such as hidden descriptors can untangle chemical problems, pinpoint their main driving forces, and accurately predict their target properties. However, when treating databases that contain thousands of molecules, the use of fast-to-calculate descriptors that can offer significant information are required. Along this Thesis we proposed the AABBA kernel to accomplish this task.

We have provided an in-depth explanation of the hidden descriptor

methodology. We exhibited their applications in two different approaches: thermal and kinetic scenarios. Particularly, we employed bond dissociation energy as a key parameter to unravel electronic interactions in metal–ligand bonds, and the energy barrier of bimolecular nucleophilic substitution reaction for tacking kinetic problems. This new approach contributes to the generation of specific chemical descriptors, thus avoiding the design of suboptimal models, and assisting in the targeted design of experiments. Furthermore, the design of the AABBA graph kernel enables the generation of molecular representations for large data sets, which can be seamlessly integrated into ML models. These data-driven strategies show a promising future for generating more robust models, promoting sustainability, understanding, and effectiveness in research approaches. We conclude that the applications proposed in this Thesis have the potential to be generalized across a wider range of chemical systems.

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

Appendix A

Bond dissociation energies

In this Appendix, we collect all bond dissociation energies that we characterized between NHCs, cyclopropylidenes, amines, and phosphine ligands and the metal fragments of reference. Energies are expressed in kcal·mol^{-1}.

Table A.1: Bond dissociation energies (in kcal·mol^{-1}) between NHC ligands and each of the metal fragments of reference.

Ligands	**Metal fragments of reference**					
	$\mathbf{OsO_3^{2+}}$	**PdH-**$\mathbf{(PH_3)_2^+}$	$\mathbf{PdPH_3}$	$\mathbf{ZrCl_5^-}$	$\mathbf{InCl_2^+}$	$\overline{BDE}$
Im*N*H$_2$	-112.18	-39.77	-38.20	-28.98	-75.73	-58.97
Im*N*Me$_2$	-119.62	-42.84	-40.05	-23.73	-79.16	-61.08
Im*N*Et$_2$	-119.13	-44.46	-40.98	-24.37	-80.75	-61.94
Im*N*^{i}Pr$_2$	-119.33	-45.14	-41.27	-25.57	-81.56	-62.57
Im*N*Ph$_2$	-124.12	-45.58	-40.15	-19.32	-79.74	-61.78
Im(NO$_2$)$_2$-*N*Me$_2$	-80.82	-35.74	-37.31	-15.51	-59.46	-45.77
ImCN$_2$*N*Me$_2$	-89.17	-37.31	-37.49	-17.66	-63.71	-49.07
ImF$_2$*N*Me$_2$	-110.17	-39.91	-38.75	-21.32	-72.35	-56.50
ImMe$_2$*N*Me$_2$	-129.73	-44.00	-40.60	-24.39	-82.27	-64.20

ImNMe$_2$-*N*Me$_2$	-158.02	-44.33	-40.47	-24.86	-84.18	-70.37
sIm*N*H$_2$	-105.09	-40.94	-38.30	-28.47	-74.73	-57.51
sIm*N*Me$_2$	-110.90	-43.58	-39.64	-21.82	-76.48	-58.48
sIm*N*Et$_2$	-110.02	-45.06	-40.62	-22.09	-78.47	-59.25
PyrazC3*N*H$_2$	-122.25	-43.40	-38.78	-30.24	-83.89	-63.71
PyrazC3-*N*Me$_2$	-127.55	-45.31	-40.44	-30.08	-88.45	-66.36
sPm*N*Me$_2$	-112.50	-43.92	-38.65	-14.34	-78.37	-57.56
PyC4*N*H	-139.85	-48.97	-42.75	-35.16	-95.11	-72.37
PyC4-3,5-Me$_2$*N*H	-143.02	-52.23	-44.16	-24.29	-97.20	-72.18
BIm*N*Me$_2$	-115.27	-42.10	-39.80	-21.59	-75.48	-58.85
DPyIm	-149.70	-42.18	-39.88	-27.2	-81.22	-68.03
aIm*N*Me$_2$	-138.12	-46.15	-41.45	-32.11	-93.11	-70.19
1,2,4-Triaz*N*Me$_2$	-100.49	-39.90	-38.22	-21.84	-71.22	-54.33

Table A.2: Bond dissociation energies (in kcal·mol^{-1}) between ligands and each of the metal fragments of reference.

Ligands	**Metal fragments of reference**					
	OsO$_3^{2+}$	**PdH-(PH$_3$)$_2^+$**	**PdPH$_3$**	**ZrCl$_5^-$**	**InCl$_2^+$**	$\overline{BDE}$
CP$_1$	-84.79	-35.05	-36.58	-20.29	-62.21	-47.78
CP$_2$	-102.69	-38.54	-37.95	-24.19	-72.09	-55.09
CP$_3$	-116.20	-40.36	-39.46	-27.33	-75.99	-59.87
CP$_4$	-140.29	-42.40	-39.53	-30.31	-87.16	-67.94
CP$_5$	-116.03	-38.74	-37.54	-25.77	-77.75	-59.17
CP$_6$	-120.97	-39.83	-38.20	-27.54	-79.40	-61.19
CP$_7$	-102.16	-37.32	-37.09	-23.79	-71.59	-54.39
CP$_8$	-144.41	-47.40	-42.54	-25.98	-91.33	-70.33
PH$_3$	-60.74	-19.02	-24.71	-2.37	-36.39	-28.64
PMe$_3$	-102.70	-31.33	-32.05	-15.04	-59.68	-48.16
PPh$_3$	-108.71	-31.94	-33.26	-10.92	-56.54	-48.27

PF_3	-7.37	-11.72	-29.14	4.50	-6.77	-10.10
PHF_2	-33.39	-17.40	-30.25	-0.42	-19.36	-20.16
PH_2F	-49.08	-19.45	-28.25	-2.59	-28.56	-25.59
NH_3	-59.84	-20.81	-19.94	-16.06	-49.43	-33.22
NMe_3	-66.18	-20.71	-19.93	-11.63	-52.94	-34.28

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

Appendix B

Computed energy barriers and chemical descriptors

In this Appendix, we collect the computed energy barriers and the chemical descriptors that we have employed for the LR and MLR analysis in Chapter 4.

Frontier Molecular Orbital descriptors

1	Energy of the HOMO orbital of the EG^-.
2	Energy of the HOMO-1 orbital of the EG^-.
3	Energy of the LUMO orbital of the EG^-.
4	Energy of the LUMO+1 orbital of the EG^-.
5	E_{HOMO} – E_{LUMO} of the EG^-.
6	Energy of the HOMO orbital of the CH_3LG.
7	Energy of the HOMO-1 orbital of the CH_3LG.
8	Energy of the LUMO orbital of the CH_3LG.
9	Energy of the LUMO+1 orbital of the CH_3LG.
10	E_{HOMO} – E_{LUMO} of the CH_3LG.
11	Energy of the HOMO orbital in the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$.
12	Energy of the HOMO-1 orbital in the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$.
13	Energy of the LUMO orbital in the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$.

14	Energy of the LUMO-1 orbital in the transition state of the reaction EG^- + $CH_3I \rightarrow CH_3EG + I^-$.
15	E_{HOMO} – E_{LUMO} of the transition state of the reaction EG^- + $CH_3I \rightarrow CH_3EG + I^-$.
16	Ionization potential of the EG^-, IP, computed as Energy of the cation-Energy of the nucleophile: E(EG)-E(EG^-).[204]
17	Electron affinity of the EG^-, EA, computed as Energy of the nucleophile-Energy of the anion: E(EG^-)-E(X^{2-}).[204]
18	Chemical potential of the EG^-, μ, computed as –(IP+EA)/2 = (E(X^{2-})-E(EG))/2.[205]
19	Electronegativity of the EG^-, $\chi\chi$, μ = - χ.[206]
20	Hardness of the EG^-, η, computed as (IP-EA)/2 = (E(X^{2-})+E(EG)-2xE(EG^-))/2.[207]
21	Softness of the EG^-, S, computed as 1/(IP-EA) = 1/(E(X^{2-})+E(EG)-2xE(EG^-)).[208]
22	Electrophilicity index of the EG^-, ω, computed as $\mu 2/(2\eta)$ = (E(X^{2-})-E(EG))2/(4xE(X^{2-})+E(EG)-2xE(EG^-)).[209,210]
23	Nucleophilicity of the EG^- computed as $1/\omega$ = $(2\eta)/\ \mu 2$ = (4xE(X^{2-})+E(EG)-2xE(EG^-))/(E(X^{2-})-E(EG))^2.[209]
24	Chemical potential of the EG^-, μ, computed as –((-E_{HOMO})+E_{LUMO})/2.[205]
25	Electronegativity of the EG^-, χ, computed as -μ = (-E_{HOMO}+E_{LUMO})/2.[206]
26	Hardness of the EG^-, η, computed as (-E_{HOMO}-E_{LUMO})/2.[207]
27	Softness of the EG^-, S, computed as 1/(-E_{HOMO}-E_{LUMO}).[208]
28	Electrophilicity index of the EG^-, ω, computed as $\mu 2/(2\eta)$ = ((E_{HOMO}-E_{LUMO})/2)2/((-E_{HOMO}-E_{LUMO})/2).[209,210]
29	Nucleophilicity of the EG^- computed as $1/\omega$ = $(2\eta)/\ \mu 2$ = (-E_{HOMO}+E_{LUMO})/((E_{HOMO}+E_{LUMO})/2)^2.[209]
30	Ionization potential of the CH_3LG, IP, computed as Energy of the cation-Energy of the nucleophile: E(CH_3LG^+)-E(CH_3LG).[204]
31	Electron affinity of the CH_3LG, EA, computed as Energy of the nucleophile-Energy of the anion: E(CH_3LG)-E(CH_3LG^-).[204]
32	Chemical potential of the CH_3LG, μ, computed as –(IP+EA)/ $\rightarrow$ 2 = (E(CH_3LG^-)- E(CH_3LG))/2.[205]
33	Electronegativity of the CH_3LG, χ, μ = - χ.[206]
34	Hardness of the CH_3LG, η, computed as (IP-EA)/2 = (E(CH_3LG^-) +E(CH_3LG)- 2xE(CH_3LG))/2 .[207]
35	Softness of the CH_3LG, $\mathcal{S}$, computed as 1/(IP-EA) = 1/(E(CH_3LG^-) +E(CH_3LG^+)- 2xE(CH_3LG))/2.[208]

36	Electrophilicity index of the CH_3LG, ω, computed as $\mu 2/(2\eta)$ = (E(CH_3LG^-)-(E(CH_3LG^+))2)/(4xE(CH_3LG^-)+E(CH_3LG^+)-2xE(CH_3LG)).[209,210]
37	Nucleophilicity of the CH_3LG computed as $1/\omega$ = $(2\eta)/\ \mu 2$ = (4xE(CH_3LG^-)+E(CH_3LG^+)-2xE(CH_3LG))/(E(CH_3LG^-)-(E(CH_3LG^+)))2.[209]
38	Chemical potential of the CH_3LG, μ, computed as –((-E_{HOMO})+E_{LUMO})/2.[205]
39	Electronegativity of the CH_3LG, χ, computed as -μ = (-E_{HOMO}+E_{LUMO})/2.[206]
40	Hardness of the CH_3LG, η, computed as (-E_{HOMO}-E_{LUMO})/2.[207]
41	Softness of the CH_3LG, $\mathcal{S}$, computed as 1/(-E_{HOMO}-E_{LUMO}).[208]
42	Electrophilicity index of the CH_3LG, ω, computed as $\mu 2/(2\eta)$ = (($E_{HOMO}E_{LUMO}$)/2)2/((-E_{HOMO}-E_{LUMO})/2).[209,210]
43	Nucleophilicity of the CH_3LG computed as $1/\omega$ = $(2\eta)/\ \mu 2$ = (-E_{HOMO}+E_{LUMO})/((E_{HOMO}+E_{LUMO})/2).[209]

Alternatives to the classical FMO descriptors

44	Electrodonating power I of the EG^-, ω^-, computed as ω^- = (IP)2/2(IP-AE).[188]
45	Electrodonating power II of the EG^-, ω^-, computed as ω^- = (3IP+AE)2/16(IP-AE).[188]
46	Nucleophilicity I of the EG^-, computed as $10/\omega^-$ = 20(IP-AE)/(IP).[211]
47	Nucleophilicity II of the EG^-, computed as $10/\omega^-$ = 160(IP-AE)/(3IP+AE).[211]
48	Electroaccepting power I of the EG^-, ω^+, computed as ω^+ = AE2/2(IP-AE).[188]
49	Electroaccepting power II of the EG^-, ω^+, computed as ω^+ = (IP+3AE)2/16(IPAE).[188]
50	Electrodonating power I of the EG^-, ω^-, computed as ω^- = (-E_{HOMO})2/2(-$E_{HOMO}E_{LUMO}$).[188]
51	Electrodonating power II of the EG^-, ω^-, computed as ω^- = (-3E_{HOMO}-E_{LUMO})2/16(-E_{HOMO}-E_{LUMO}).[188]
52	Nucleophilicity I of the EG^-, computed as $10/\omega^-$ = 20(-E_{HOMO}-E_{LUMO})/(-E_{HOMO})2.[211]
53	Nucleophilicity II of the EG^-, computed as $10/\omega^-$ = 160(-E_{HOMO}-E_{LUMO})/(-3$E_{HOMO}E_{LUMO}$)2.[211]

54	Electroaccepting power I of the EG^-, ω^+, computed as $\omega^+ = E_{LUMO}$ 2/2(-E_{HOMO} - E_{LUMO}).[188]
55	Electroaccepting power II of the EG^-, ω^+, computed as ω^+ = (-E_{HOMO} + 3E_{LUMO})2/16(-E_{HOMO} - E_{HOMO}).[188]
56	Electrodonating power I of the CH_3LG, ω^-, computed as ω^- = (IP)2/2(IP-AE).[188]
57	Electrodonating power II of the CH_3LG, ω^-, computed as ω^- = (3IP+AE)2/16(IPAE).[188]
58	Nucleophilicity I of the CH_3LG, computed as $10/\omega^-$ = 20(IP-AE)/(IP).[211]
59	Nucleophilicity II of the CH_3LG, computed as $10/\omega^-$ = 160(IP-AE)/(3IP+AE).[211]
60	Electroaccepting power I of the CH_3LG, ω^+, computed as ω^+ = AE2/2(IP-AE).[188]
61	Electroaccepting power II of the CH_3LG, ω^+, computed as ω^+ = (IP+3AE)2/16(IPAE).[188]
62	Electrodonating power I of the CH_3LG, ω^-, computed as ω^-= (-E_{HOMO})2/2(-$E_{HOMO}E_{LUMO}$).[188]
63	Electrodonating power II of the CH_3LG, ω^-, computed as ω^- = (-3$E_{HOMO}E_{LUMO}$)2/16(-E_{HOMO}-E_{LUMO}).[188]
64	Nucleophilicity I of the CH_3LG, computed as $10/\omega^-$ = 20(-E_{HOMO}-E_{LUMO})/(- E_{HOMO})2.[211]
65	Nucleophilicity II of the CH_3LG, computed as $10/\omega^-$ = 160(-E_{HOMO}-E_{LUMO})/(- 3E_{HOMO}-E_{LUMO})2.[211]
66	Electroaccepting power I of the CH_3LG, ω^+, computed as ω^+ = (-E_{HOMO})2/2(-$E_{HOMO}E_{LUMO}$).[188]
67	Electroaccepting power II of the CH_3LG, ω^+, computed as ω^+ = (-E_{HOMO} + 3E_{LUMO})2/16(-E_{LUMO} - E_{HOMO}).10
68	Nucleophilicity of the EG^-, N, computed as N = (3IP-AE)2/8(IP-AE).[211]
69	Electrophilicity of the EG^-, E, computed as E = (IP+AE)2/8(IP-AE).[211]
70	Nucleophilicity of the EG^-, N, computed as N = (-3E_{HOMO} – E_{LUMO})2/8(-E_{HOMO} – E_{LUMO}).[211]
71	Electrophilicity of the EG^-, E, computed as E = (-E_{HOMO} + E_{LUMO})2/8(-E_{HOMO} – E_{LUMO}).[211]
72	Nucleophilicity of the CH_3LG , N, computed as N = (3IP-AE)2/8(IP-AE).[211]
73	Electrophilicity of the CH_3LG , E, computed as E = (IP+AE)2/8(IP-AE).[211]
74	Nucleophilicity of the CH_3LG, N, computed as N = (-3E_{HOMO} – E_{LUMO})2/8(-E_{HOMO} – E_{LUMO}).[211]

75	Electrophilicity of the CH_3LG, E, computed as $E = (-E_{HOMO} + E_{LUMO})2/8(-E_{HOMO} - E_{LUMO})$.[211]
76	$\Delta E_{nucleofuge}$ of the EG^-, computed as $\Delta E_{nucleofuge} = (IP-3AE)2/8(IP-AE)$.[212,213]
77	$\Delta E_{electrofuge}$ of the EG^-, computed as $\Delta E_{electrofuge} = (3IP-AE)2/8(IP-AE)$.[212,213]
78	Nucleofugality of the EG^-, λN, computed as $\lambda N = e^{-\beta E \Delta E nucleofuge}$, $\beta N = 67.6556$.[212,213]
79	ln(Nucleofugality) of the EG^-, $\ln(\lambda N)$, being $\lambda N = e^{-\beta E \Delta E nucleofuge}$, $\beta N = 67.6556$.[212,213]
80	Electrofugality of the EG^-, λE, computed as $\lambda E = e^{-\beta E \Delta E electrofuge}$, $\beta E = 5.4389$.[212,213]
81	ln(Electrofugality) of the EG^-, $\ln(\lambda E)$ being $\lambda E = e^{-\beta E \Delta E electrofuge}$, $\beta E = 5.4389$.[212,213]
82	$\Delta E_{nucleofuge}$ of the EG^-, computed as $\Delta E_{nucleofuge} = (-E_{HOMO}-3E_{LUMO})2/8(-E_{HOMO}E_{LUMO})$.[212,213]
83	$\Delta E_{electrofuge}$ of the EG^-, computed as $\Delta E_{electrofuge} = (-3E_{HOMO}-E_{LUMO})2/8(-E_{HOMO}E_{LUMO})$.[212,213]
84	Nucleofugality of the EG^-, λN, computed as $\lambda N = e^{-\beta E \Delta E nucleofuge}$, $\beta N = 67.6556$.[212,213]
85	ln(Nucleofugality) of the EG^-, $\ln(\lambda N)$, being $\lambda N = e^{-\beta E \Delta E nucleofuge}$, $\beta N = 67.6556$.[212,213]
86	Electrofugality of the EG^-, λE, computed as $\lambda E = e^{-\beta E \Delta E electrofuge}$, $\beta E = 5.4389$.[212,213]
87	ln(Electrofugality) of the EG^-, $\ln(\lambda E)$ being $\lambda E = e^{-\beta E \Delta E electrofuge}$, $\beta E = 5.4389$.[212,213]
88	$\Delta E_{nucleofuge}$ of the CH_3LG, computed as $\Delta E_{nucleofuge} = (IP-3AE)2/8(IP-AE)$.[212,213]
89	$\Delta E_{electrofuge}$ of the CH_3LG, computed as $\Delta E_{electrofuge} = (3IP-AE)2/8(IP-AE)$.[212,213]
90	Nucleofugality of the CH_3LG, λN, computed as $\lambda N = e^{-\beta E \Delta E nucleofuge}$, $\beta N = 67.6556$.[212,213]
91	ln(Nucleofugality) of the CH_3LG, $\ln(\lambda N)$ being $\lambda N = e^{-\beta E \Delta E nucleofuge}$, $\beta N = 67.6556$.[212,213]
92	Electrofugality of the CH_3LG, λE, computed as $\lambda E = e^{-\beta E \Delta E electrofuge}$, $\beta E = 5.4389$.[212,213]
93	ln(Electrofugality) of the CH_3LG, $\ln(\lambda E)$ being $\lambda E = e^{-\beta E \Delta E electrofuge}$, $\beta E = 5.4389$.[212,213]

94	$\Delta E_{nucleofuge}$ of the CH_3LG, computed as $\Delta E_{nucleofuge}$ = (-E_{HOMO}-3E_{LUMO})2/8(-$E_{HOMO}E_{LUMO}$).[212,213]
95	$\Delta E_{electrofuge}$ of the CH_3LG, computed as $\Delta E_{electrofuge}$ = (-3E_{HOMO}–E_{LUMO})2/8($E_{HOMO}E_{LUMO}$).[212,213]
96	Nucleofugality of the CH_3LG, λN, computed as $\lambda N = e^{-\beta E \Delta E nucleofuge}$, βN = 67.6556.[212,213]
97	ln(Nucleofugality) of the CH_3LG, ln(λN) being $\lambda N = e^{-\beta E \Delta E nucleofuge}$, βN = 67.6556.[212,213]
98	Electrofugality of the CH_3LG, λE, computed as $\lambda E = e^{-\beta E \Delta E electrofuge}$, βE = 5.4389.[212,213]
99	ln(Electrofugality) of the CH_3LG, ln(λE) being $\lambda E = e^{-\beta E \Delta E electrofuge}$, βE = 5.4389.[212,213]

Atomic Charges

100	Mulliken charge on the attacking atom of the EG^-.[214,215]
101	APT charge on the attacking atom of the attacking atom of the EG^-.[216]
102	NPA charges of the attacking atom of the EG^-.[217]
103	Hirshfeld charges of the attacking atom of the EG^-.[218–220]
104	Charge Model 5 of the attacking atom of the EG^-.[221]
105	Electric potential of the attacking atom of EG^-.[222,223]
106	Merz-Kollman charge of the attacking atom of EG^-.[222,223]
107	Mulliken charge on the electrophilic carbon of CH_3LG.[214,215]
108	APT charge on the electrophilic carbon of CH_3LG.[216]
109	NPA charges of the electrophilic carbon of CH_3LG.[217]
110	Hirshfeld charges of the electrophilic carbon of CH_3LG.[218–220]
111	Charge Model 5 of the electrophilic carbon of CH_3LG.[221]
112	Electric potential, of the electrophilic carbon of CH_3LG.[222,223]
113	Merz-Kollman charges of the electrophilic carbon of CH_3LG.[222,223]

Energetic parameters

114	Potential energy of the reaction I^- + $CH_3LG \rightarrow CH_3I + LG^-$. (kcal/mol).
115	Potential + ZPVE energy of the reaction I^- + $CH_3LG \rightarrow CH_3I + LG^-$ (kcal/mol).
116	Enthalpy of the reaction I^- + $CH_3LG \rightarrow CH_3I + LG^-$ (kcal/mol).
117	Free energy of the reaction I^- + $CH_3LG \rightarrow CH_3I + LG^-$(kcal/mol).
118	Potential energy of the Proton Affinity of the EG^- (kcal/mol).

119	Potential + ZPVE energy of the Proton Affinity of the EG^- (kcal/mol).
120	Enthalpy of the Proton Affinity of the EG^- (kcal/mol).
121	Free energy of the Proton Affinity of the EG^- (kcal/mol).
122	Imaginary frequency of the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$ (cm^{-1}).
123	Imaginary frequency of the transition state of the reaction $EG^- + CH_3C_6H_5 \rightarrow CH_3EG + C_6H_5^-$ (cm^{-1})

Solvent features

124	ΔG(solvation) in water of EG^- (kcal/mol).
125	ΔG(solvation) in DCM of EG^- (kcal/mol).39
126	ΔG(solvation) in DMSO of EG^- (kcal/mol).
127	ΔG(solvation) in cyclohexane of EG^- (kcal/mol).
128	ΔG(solvation) in water of CH_3LG (kcal/mol).
129	ΔG(solvation) in DCM of CH_3LG (kcal/mol).
130	ΔG(solvation) in DMSO of CH_3LG (kcal/mol).
131	ΔG(solvation) in cyclohexane of CH_3LG (kcal/mol).
132	ΔG(solvation) in water of the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$ (kcal/mol).
133	ΔG(solvation) in DCM of the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$ (kcal/mol).
134	Cavity volume of the EG^- in water with SMD according to Gaussian defaults.
135	Cavity volume of the EG^- in DCM with SMD according to Gaussian defaults.
136	Cavity volume of the EG^- in DMSO with SMD according to Gaussian defaults.
137	Cavity volume of the EG^- in cyclohexane with SMD according to Gaussian defaults.
138	Cavity volume of the CH_3LG in water with SMD according to Gaussian defaults.
139	Cavity volume of the CH_3LG in DCM with SMD according to Gaussian defaults.
140	Cavity volume of the CH_3LG in DMSO with SMD according to Gaussian defaults.
141	Cavity volume of the CH_3LG in cyclohexane with SMD according to Gaussian defaults.

	Geometrical features parameters
142	Box volume of the EG^-.
143	Molar volume of the EG^-computed according to Gaussian's keyword Volume(cm^3/mol).
144	Box volume of the CH_3LG.
145	Molar volume of the CH_3LG computed according to Gaussian's keyword Volume(cm^3/mol).
146	Distance between the electrophilic carbon and I^- in the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$, d($I^-$–C), ($\mathring{A}$). 40
147	Standard deviation of d(I^-–C) in the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$
148	Distance between the electrophilic carbon and EG^- in the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$, d(C–X), ($\mathring{A}$).
149	Standard deviation of d(C-X) in the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$
150	Angle between the EG^-, the electrophilic carbon and I^-in the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$ (^{o}).
151	Average of the distance between the EG^- and the three H of the methyl group in the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$($\mathring{A}$).
152	Average of the distance between the I^- and the three H of the methyl group in the transition state of the reaction $EG^- + CH_3I \rightarrow CH_3EG + I^-$($\mathring{A}$).
153	Distance between the electrophilic carbon and TsO^- in the transition state of the reaction $EG^- + CH_3TsO \rightarrow CH_3EG + TsO^- + EG^-$, d($TsO^-$–C), ($\mathring{A}$).
154	Standard deviation of d(TsO^-–C) in the transition state of the reaction EG^-+$CH_3TsO \rightarrow CH_3EG + TsO^- + EG^-$.
155	Distance between the electrophilic carbon and EG^- in the transition state of thereaction $EG^- + CH_3TsO \rightarrow CH_3EG + TsO^-$, d(C–X), ($\mathring{A}$).
156	Standard deviation of d(C–X) in the transition state of the reaction $EG^- + CH_3TsO \rightarrow CH_3EG + TsO^-$
157	Angle between the EG^-, the electrophilic carbon and TsO^- in the transition state of the reaction $EG^- + CH_3TsO \rightarrow CH_3EG + TsO^-$ (^{o}).
158	Average of the distance between the EG^- and the three H of the methyl group in the transition state of the reaction $EG^- + CH_3TsO \rightarrow CH_3EG + TsO^-$ ($\mathring{A}$).
159	Average of the distance between the TsO^- and the three H of the methyl group in the transition state of the reaction $EG^- + CH_3TsO \rightarrow CH_3EG + TsO^-$ ($\mathring{A}$).

160	Distance between the electrophilic carbon and TfO- in the transition state of the reaction EG^- + $CH_3TfO \rightarrow CH_3EG$ + TfO- + EG^-, d(TsO^-–C), ($\mathring{A}$).
161	Sterimol L parameter (a1=C and a2=X) for the transition geometry I^- + $CH_3LG \rightarrow CH_3I + LG^-$.[224]
162	Sterimol B1 parameter (a1=C and a2=X) for the transition geometry I^- + $CH_3LG \rightarrow CH_3I + LG^-$.[224]
163	Sterimol B5 parameter (a1=C and a2=X) for the transition geometry I^- + $CH_3LG \rightarrow CH_3I + LG^-$.[224]
164	Sterimol L parameter (a1=C and a2=I) for the transition geometry I^- + $CH_3LG \rightarrow CH_3I + LG^-$.[224]
165	Sterimol B1 parameter (a1=C and a2=I) for the transition geometry I^- + $CH_3LG \rightarrow CH_3I + LG^-$.[224]
166	Sterimol B5 parameter (a1=C and a2=I) for the transition geometry I^- + $CH_3LG \rightarrow CH_3I + LG^-$.[224]
167	Molecular weight of EG^-.
168	Molecular weight of CH_3LG.
169	Molecular weight of the transition state of the reaction EG^- + $CH_3EG \rightarrow CH_3EG$ + X

Bond order parameters

170	Pauling Bond Order, n≠, of the bond between the electrophilic carbon and I^- in the transition state of the reaction EG^- + $CH_3I \rightarrow CH_3EG + I^-$, calculated as n≠ =exp[(r-r≠)/0.6], being r and r≠ the bond lengths at the reactant and at the TS, respectively.[225]
171	Pauling Bond Order, n≠, of the bond between the electrophilic carbon and EG^- in the transition state of the reaction I^- + $CH_3LG \rightarrow CH_3I + LG^-$, calculated as n≠ =exp[(r-r≠)/0.6], being r and r≠ the bond lengths at the reactant and at the TS, respectively.[225]

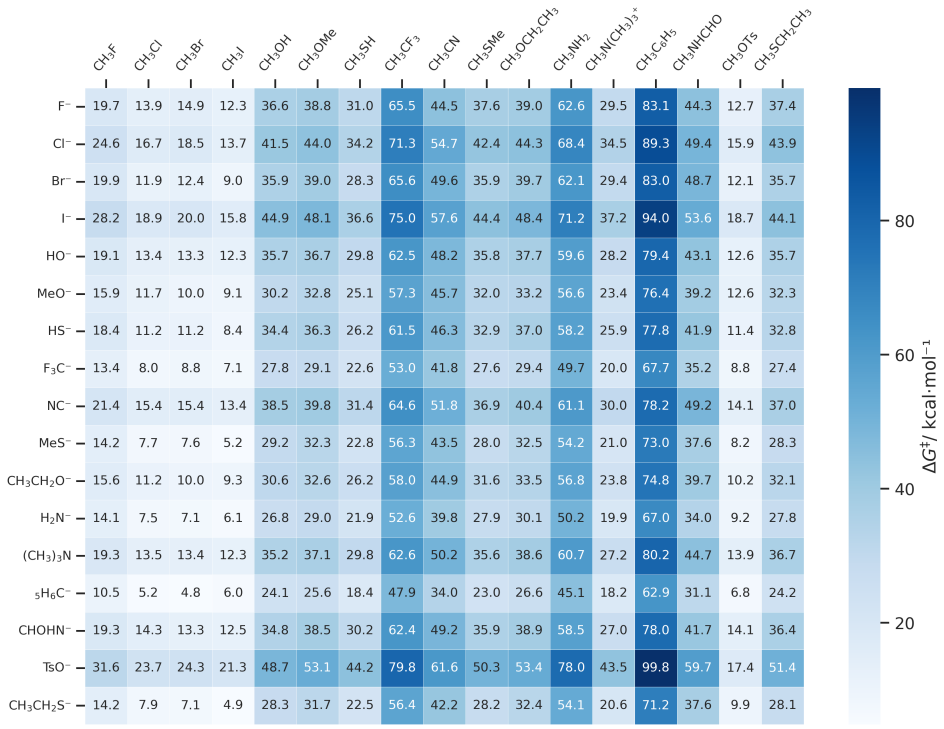


Figure B.1: Matrix $17x17$ of the free energy barriers (in kcal·mol^{-1}) from the adducts to the transition states in water ($\Delta G^{\ddagger}_{TS-Adduct}$)

Figure B.2: Matrix 17x17 of the free energy barriers (in kcal·mol^{-1}) from the adducts to the transition states in dichloromethane ($\Delta G^{\ddagger}_{TS-Adduct}$).

UNIVERSITAT ROVIRA I VIRGILI
DECODING CHEMICAL PROCESSES: THE POWER OF DATA-DRIVEN DESCRIPTORS
Lucía Morán González

Appendix C

NBO properties and neural network models

Table C.1: NBO properties included in the atomic ($P_{A,NBO}$) property sets.[a]

	$P_{A,NBO}$
Z	Atomic number
q_{Nat}	Natural charge (e)
V_{Nat}	Natural valence index
N_s	# s electrons in nat. config.
N_p	# p electrons in nat. config.
N_d	# d electrons in nat. config.
N_{LP}	# Lone pairs
LP_E	E of highest-lying LP (Ha)
$LP_{\Delta E}$	Lowest/highest-lying LP E gap (Ha)
LP_{Occ}	Electron occupancy of highest-E LP
LP_s	s-character of highest-E LP (%)
LP_p	p-character of highest-E LP (%)
LP_d	d-character of highest-E LP (%)
N_{LV}	# Lone vacancies
LV_E	E of lowest-lying LV (Ha)
$LV_{\Delta E}$	Lowest/highest-lying LV E gap LV (Ha)
LV_{Occ}	Electron occupancy of lowest-E LV
LV_s	s-character of lowest-E LV (%)
LV_p	p-character of lowest-E LV (%)
LV_d	d-character of lowest-E LV (%)

[a]Abbreviations: # = Number of; E = Energy; Nat. = Natural; LP = Lone Pair; LV = Lone Vacancy; Config. = Configuration.

Table C.2: NBO properties included in the bond ($P_{B,NBO}$) property sets.[a]

	$P_{B,NBO}$
BD	Bond distance ($\mathring{A}$)
BO_{Nat}	Natural Wiberg bond order
N_{BN}	# bonding NBOs
BN_E	E of highest-lying BN (Ha)
$BN_{\Delta E}$	Lowest/highest-lying BN E gap (Ha)
BN_{Occ}	Electron occupancy of highest-E BN
BN_s	s-character of highest-E BN (%)
BN_p	p-character of highest-E BN (%)
BN_d	d-character of highest-E BN (%)
N_{BN^*}	# non- & anti-bonding NBOs
BN^*_E	E of lowest-lying BN* (Ha)
$BN^*_{\Delta E}$	Lowest/highest-lying BN* E gap (Ha)
BN^*_{Occ}	Electron occupancy of lowest-E BN*
BN^*_s	s-character of lowest-E BN* (%)
BN^*_p	p-character of lowest-E BN* (%)
BN^*_d	d-character of lowest-E BN* (%)

[a]Abbreviations: # = Number of; E = Energy; Nat. = Natural; BO = Bond Order; NBOs = Natural Bond Orbitals; BN = Bonding NBO; BN* = Non- and anti-bonding NBOs.

Table C.3: Average and lowest test errors in the prediction of the Vaska's dataset energy barriers with neural networks. The inputs passed to the models were vectors defined with different graph kernels ($\mathcal{G}_\mathcal{K}$), property types (Prop), operators ($\hat{\text{Op}}$), origins ($\emptyset$), and maximum depths (D), yielding different dimensionality (dim). The mean absolute error (MAE) is given in kcal/mol.

Model	Input						Average Error[a]		Lowest Error[a]	
	$\mathcal{G}_\mathcal{K}$	Prop	$\hat{\text{Op}}$	$\emptyset$[b]	D	dim[c]	MAE	r^2	MAE	r^2
1	**AA**	**P**[d]	$\odot$[e]	**MC**	**3**	**18**	**1.22 ± 0.02**	**0.844 ± 0.007**	**1.16**	**0.850**
2	BB	P	$\odot$	MC	3	12	1.42 ± 0.03	0.801 ± 0.005	1.37	0.803
3	**BB**	**P**	$\odot$	**MC**	**3**	**10**	**1.40 ± 0.04**	**0.765 ± 0.016**	**1.26**	**0.791**
4	BA	P	$\odot$	MC	3	20	2.14 ± 0.04	0.571 ± 0.013	2.05	0.595
5	**I**[f]	**P**	$\odot$	**MC**	**3**	**48**	**0.90 ± 0.02**	**0.914 ± 0.003**	**0.86**	**0.916**
6	I	P	$\ominus_\chi$[e]	MC	3	47	0.91 ± 0.01	0.913 ± 0.003	0.89	0.919
7	I	P	$\oslash_R$[e]	MC	3	47	0.92 ± 0.02	0.911 ± 0.004	0.89	0.919
8	I	NBO[g]	$\odot$	MC	3	212	0.89 ± 0.02	0.908 ± 0.007	0.85	0.913
9	**I**	**NBO**	$\odot$	**F**	**3**	**220**	**0.79 ± 0.02**	**0.927 ± 0.002**	**0.76**	**0.928**
10	I	NBO	$\odot$	MC	4	263	0.90 ± 0.02	0.899 ± 0.008	0.84	0.925
11	I	NBO	$\odot$	MC	5	298	0.90 ± 0.02	0.904 ± 0.007	0.87	0.899
12	**I**	**NBO**	$\odot$	**MC**	**6**	**303**	**0.88 ± 0.02**	**0.907 ± 0.007**	**0.84**	**0.918**
13	I[h]	NBO	$\odot$	F	3	223	0.78 ± 0.02	0.928 ± 0.004	0.73	0.933
14	**II$_1$**[f]	**P**	$\odot$	**MC**	**3**	**33**	**0.94 ± 0.02**	**0.906 ± 0.002**	**0.89**	**0.913**
15	II$_2$	P	$\odot$	MC	3	29	0.96 ± 0.03	0.904 ± 0.005	0.86	0.917
16	II$_3$	P	$\odot$	MC	3	33	0.94 ± 0.02	0.908 ± 0.004	0.90	0.913
17	II$_4$	NBO	$\odot$	MC	3	92	0.86 ± 0.02	0.918 ± 0.003	0.82	0.926
18	II$_5$	NBO	$\odot$	MC	3	80	0.94 ± 0.02	0.907 ± 0.004	0.88	0.915
19	II$_4$	NBO	$\odot$	F	3	98	1.15 ± 0.03	0.850 ± 0.007	1.05	0.862
20	II$_4$	NBO	$\odot$	MC	4	111	0.88 ± 0.02	0.916 ± 0.005	0.85	0.914
21	**II$_4$**	**NBO**	$\odot$	**MC**	**5**	**129**	**0.85 ± 0.03**	**0.916 ± 0.004**	**0.81**	**0.923**
22	II$_4$	NBO	$\odot$	MC	6	129	0.85 ± 0.01	0.919 ± 0.003	0.83	0.921

Table C.4: Average and lowest test errors in the prediction of the Vaska's dataset H$\cdots$H distance with neural networks. The inputs passed to the models were vectors defined with different graph kernels ($\mathcal{G}_\mathcal{K}$), property types (Prop), operators ($\hat{\text{Op}}$), origins (Ø), and maximum depths (D), yielding different dimensionality (*dim*). The mean absolute error (MAE) is given in $\mathring{A}$.

Model	Input						Average Error[a]		Lowest Error[a]	
	$\mathcal{G}_\mathcal{K}$	Prop	$\hat{\text{Op}}$	Ø[b]	D	*dim*[c]	MAE	r^2	MAE	r^2
23	**AA**	**P**[d]	$\odot$[e]	**MC**	**3**	**18**	$\mathbf{2.38{\cdot}10^{-2} \pm 9{\cdot}10^{-4}}$	**0.687 ± 0.012**	$\mathbf{2.11{\cdot}10^{-2}}$	**0.727**
24	BB	P	$\odot$	MC	3	10	$2.30{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.706 ± 0.007	$2.21{\cdot}10^{-2}$	0.714
25	$\overline{\textbf{BB}}$	**P**	$\odot$	**MC**	**3**	**12**	$\mathbf{2.35{\cdot}10^{-2} \pm 1.0{\cdot}10^{-3}}$	**0.673 ± 0.020**	$\mathbf{2.09{\cdot}10^{-2}}$	**0.729**
26	BA	P	$\odot$	MC	3	20	$3.03{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$	0.537 ± 0.012	$2.91{\cdot}10^{-2}$	0.551
27	**I**[f]	**P**	$\odot$	**MC**	**3**	**48**	$\mathbf{2.07{\cdot}10^{-2} \pm 4{\cdot}10^{-4}}$	**0.747 ± 0.006**	$\mathbf{1.93{\cdot}10^{-2}}$	**0.767**
28	I	P	$\odot$	F	3	220	$2.12{\cdot}10^{-2} \pm 7{\cdot}10^{-4}$	0.716 ± 0.015	$1.98{\cdot}10^{-2}$	0.739
29	I	P	$\odot$	F[j]	3	220	$2.54{\cdot}10^{-2} \pm 1.0{\cdot}10^{-3}$	0.632 ± 0.019	$2.37{\cdot}10^{-2}$	0.669
30	**I**	**P**	$\ominus_\chi$[e]	**MC**	**3**	**47**	$\mathbf{1.96{\cdot}10^{-2} \pm 5{\cdot}10^{-4}}$	**0.767 ± 0.007**	$\mathbf{1.87{\cdot}10^{-2}}$	**0.769**
31	I	P	$\oslash_R$[e]	MC	3	47	$2.04{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$	0.764 ± 0.004	$1.96{\cdot}10^{-2}$	0.761
32	I	NBO[g]	$\odot$	MC	3	212	$2.27{\cdot}10^{-2} \pm 8{\cdot}10^{-4}$	0.663 ± 0.021	$2.13{\cdot}10^{-2}$	0.702
33	I	NBO	$\odot$	MC	4	263	$2.22{\cdot}10^{-2} \pm 7{\cdot}10^{-4}$	0.676 ± 0.012	$2.12{\cdot}10^{-2}$	0.690
34	I	NBO	$\odot$	MC	5	298	$2.16{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$	0.705 ± 0.020	$2.04{\cdot}10^{-2}$	0.706
35	**I**	**NBO**	$\odot$	**F**	**5**	**316**	$\mathbf{2.10{\cdot}10^{-2} \pm 1.0{\cdot}10^{-3}}$	**0.718 ± 0.021**	$\mathbf{1.89{\cdot}10^{-2}}$	**0.748**
36	I	NBO	$\odot$	F[h]	5	316	$2.60{\cdot}10^{-2} \pm 5.2{\cdot}10^{-3}$	0.586 ± 0.153	$2.06{\cdot}10^{-2}$	0.727
37	I	NBO	$\odot$	MC	6	303	$2.17{\cdot}10^{-2} \pm 7{\cdot}10^{-4}$	0.704 ± 0.016	$1.99{\cdot}10^{-2}$	0.745
38	II_1[f]	P	$\odot$	MC	3	33	$1.91{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.771 ± 0.009	$1.81{\cdot}10^{-2}$	0.788
39	II_2	P	$\odot$	MC	3	29	$1.93{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.768 ± 0.006	$1.81{\cdot}10^{-2}$	0.772
40	$\mathbf{II_3}$	**P**	$\odot$	**MC**	**3**	**33**	$\mathbf{1.85{\cdot}10^{-2} \pm 5{\cdot}10^{-4}}$	**0.781 ± 0.009**	$\mathbf{1.72{\cdot}10^{-2}}$	**0.804**
41	II_3	P	$\odot$	F	3	54	$1.97{\cdot}10^{-2} \pm 7{\cdot}10^{-4}$	0.747 ± 0.013	$1.87{\cdot}10^{-2}$	0.765
42	II_4	NBO	$\odot$	MC	3	92	$2.21{\cdot}10^{-2} \pm 5{\cdot}10^{-4}$	0.685 ± 0.013	$2.13{\cdot}10^{-2}$	0.704
43	II_5	NBO	$\odot$	MC	3	80	$2.20{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$	0.687 ± 0.024	$2.05{\cdot}10^{-2}$	0.749
44	II_3	P	$\odot$	MC	4	42	$1.87{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.776 ± 0.009	$1.82{\cdot}10^{-2}$	0.785
45	II_3	P	$\odot$	MC	5	51	$1.86{\cdot}10^{-2} \pm 5{\cdot}10^{-4}$	0.766 ± 0.011	$1.78{\cdot}10^{-2}$	0.795
46	II_3	P	$\odot$	MC	6	51	$1.86{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.770 ± 0.008	$1.74{\cdot}10^{-2}$	0.792
47	II_3[i]	P	$\odot$	MC	3	36	$1.86{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.777 ± 0.008	$1.74{\cdot}10^{-2}$	0.799

[a]From ten repetitions with a training:validation:test split of 80:10:10; [b]Metal-centered (MC) or full (F); [c]After removing redundant dimensions; [d]*I.e.* P_A, P_B, and P_{AB} periodic and generic property sets; [e]All properties correlated by product ($\odot$), except the electronegativity in entry 6 in Table C.3, and in entry 8 in Table C.4 (subtracted, $\ominus_\chi$), and the covalent radius in entry 7 in Table C.3, and in entry 9 in Table C.4 (divided, $\oslash_R$); [f]Entries 5-13 and 14-22 correspond to the AABBA(I) and AABBA(II) kernels, respectively, in Table C.3 and entries 5-15 and 16-24 correspond to the AABBA(I) and AABBA(II) kernels, respectively in Table C.4;[g]*I.e.* $P_{A,NBO}$, $P_{B,NBO}$, and $P_{AB,NBO}$ NBO property sets. [h]Including whole-graph properties; [i]Also including whole-graph properties; [j]From an extended neural network of 3 hidden layers with 256 nodes each.

Bibliography

[1] Mannhold, R., Kubinyi, H. and Folkers, G. 'Front Matter'. In: *Molecular Descriptors for Chemoinformatics*. John Wiley & Sons, Ltd, **2009**, I–XLI. DOI: https://doi.org/10.1002/9783527628766.fmatter1 (cit. on p. 1).

[2] Cassier, E. 'The Philosophy of Symbolic Forms'. In: *Volume 4: The Metaphysics of Symbolic Forms*. Ed. by John Michael Krois and Donald Phillip Verene. Yale University Press, **1953** (cit. on p. 1).

[3] Garay-Ruiz, D. and Bo, C. 'Chemical reaction network knowledge graphs: the OntoRXN ontology'. In: *J Cheminform* **2022**, *14*. DOI: 10.1186/s13321-022-00610-x (cit. on p. 1).

[4] Wilkinson, M. D. et al. 'Comment: The FAIR Guiding Principles for scientific data management and stewardship'. In: *Sci. Data* **2016**, *3*. DOI: 10.1038/sdata.2016.18 (cit. on p. 2).

[5] *World Bank Open Data*. URL: https://data.worldbank.org/ (cit. on p. 2).

[6] Mendeléeff. 'LXIII.—The Periodic Law of the Chemical Elements'. In: *J. Chem. Soc., Trans.* **1889**, *55*, 634–656. DOI: 10.1039/CT8895500634 (cit. on p. 2).

[7] Linstrom, P. J. and Mallard, W. G. 'The NIST Chemistry WebBook: A chemical data resource on the Internet'. In: *J. Chem. Eng. Data* **2001**, *46*, 1059–1063. DOI: 10.1021/je000236i (cit. on p. 2).

[8] Gibb, B. C. 'Big (chemistry) data'. In: *Nat. Chem* **2013**, *5*, 248–249. DOI: 10.1038/nchem.1604 (cit. on p. 2).

Bibliography

[9] Goodman, J. 'Computer Software Review: Reaxys'. In: *J. Chem. Inf. Model.* **2009**, *49*, 2897–2898. DOI: 10.1021/ci900437n (cit. on p. 2).

[10] Milo, A. 'Democratizing synthesis by automation'. In: *Science* **2019**, *363*, 122–123. DOI: 10.1126/science.aav8816 (cit. on p. 2).

[11] Weber, L., Illgen, K. and Almstetter, M. 'Discovery of New Multi Component Reactions with Combinatorial Methods'. In: *Synlett* **1999**, *3*, 366–374 (cit. on p. 2).

[12] Collins, K. D., Gensch, T. and Glorius, F. 'Contemporary screening approaches to reaction discovery and development'. In: *Nat. Chem.* **2014**, *6*, 859–871. DOI: 10.1038/nchem.2062 (cit. on p. 2).

[13] *Web of Science*. URL: https://www.webofscience.com/wos/woscc/basic-search (cit. on p. 2).

[14] Toyao, T., Maeno, Z., Takakusagi, S., Kamachi, T., Takigawa, I. and Shimizu, K. I. 'Machine Learning for Catalysis Informatics: Recent Applications and Prospects'. In: *ACS Catal* **2020**, *10*, 2260–2297. DOI: 10.1021/acscatal.9b04186 (cit. on p. 2).

[15] Wiest, O. et al. 'On the Use of Real-World Datasets for Reaction Yield Prediction'. In: *Chem. Sci.* **2023**, *14*, 4997–5005. DOI: 10.1039/d2sc06041h (cit. on pp. 3, 6).

[16] Bo, C., Maseras, F. and López, N. 'The role of computational results databases in accelerating the discovery of catalysts'. In: *Nat. Catal.* **2018**, *1*, 809–810. DOI: 10.1038/s41929-018-0176-4 (cit. on pp. 3, 23).

[17] Grzybowski, B. A., Bishop, K. J. M., Kowalczyk, B. and Wilmer, C. E. 'The 'wired' universe of organic chemistry'. In: *Nat. Chem.* **2009**, *1*, 31–36. DOI: 10.1038/nchem.136 (cit. on p. 3).

[18] Ramakrishnan, R., Dral, P. O., Rupp, M. and Lilienfeld, O. A. V. 'Quantum chemistry structures and properties of 134 kilo molecules'. In: *Sci. Data* **2014**, *1*. DOI: 10.1038/sdata.2014.22 (cit. on p. 3).

[19] Álvarez-Moreno, M., Graaf, C. D., López, N., Maseras, F., Poblet, J. M. and Bo, C. 'Managing the computational chemistry big data problem: The ioChem-BD platform'. In: *J. Chem. Inf. Model.* **2015**, *55*, 95–103. DOI: 10.1021/ci500593j (cit. on pp. 3, 59, 96).

[20] Nakata, M. and Shimazaki, T. 'PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry'. In: *J. Chem. Inf. Model.* **2017**, *57*, 1300–1308. DOI: 10.1021/acs.jcim.7b00083 (cit. on p. 3).

[21] Jain, A. et al. 'Commentary: The materials project: A materials genome approach to accelerating materials innovation'. In: *APL Mater.* **2013**, *1*, 011002. DOI: 10.1063/1.4812323 (cit. on p. 3).

[22] Balcells, D. and Skjelstad, B. B. 'The tmQM Dataset - Quantum Geometries and Properties of 86k Transition Metal Complexes'. In: *J. Chem. Inf. Model* **2020**, *60*, 6135–6146. DOI: 10.1021/acs.jcim.0c01041 (cit. on p. 3).

[23] Kneiding, H., Lukin, R., Lang, L., Reine, S., Pedersen, T. B., Bin, R. D. and Balcells, D. 'Deep learning metal complex properties with natural quantum graphs'. In: *Digital Discovery* **2023**, *2*, 618–633. DOI: 10.1039/D2DD00129B (cit. on pp. 3, 8, 32, 139).

[24] Fey, N., Tsipis, A. C., Harris, S. E., Harvey, J. N., Orpen, A. G. and Mansson, R. A. 'Development of a ligand knowledge base, Part 1: Computational descriptors for phosphorus donor ligands'. In: *Chem. Eur. J.* **2006**, *12*, 291–302. DOI: 10.1002/chem.200500891 (cit. on pp. 3, 17).

[25] Jover, J., Fey, N., Harvey, J. N., Lloyd-jones, G. C., Orpen, A. G. and Owensmith, G. J. J. 'Expansion of the ligand knowledge base for monodentate P-donor ligands (LKB-P)'. In: *Organometallics* **2010**, *29*, 6245–6258. DOI: 10.1021/om100648v (cit. on pp. 3, 17).

[26] Fey, N., Haddow, M. F., Harvey, J. N., McMullin, C. L. and Orpen, A. G. 'A ligand knowledge base for carbenes (LKB-C): Maps of ligand space'. In: *Dalton Trans.* **2009**, 8183–8196. DOI: 10.1039/b909229c (cit. on pp. 3, 18, 33, 39).

[27] Gensch, T. et al. 'A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis'. In: *J. Am. Chem. Soc.* **2022**, *144*, 1205–1217. DOI: 10.1021/jacs.1c09718 (cit. on p. 3).

[28] Brown, A. C. 'On the theory of chemical combination'. In: **1861** (cit. on p. 4).

[29] Kekulé, A. 'Sur la constitution des substances aromatiques'. In: *Bulletin mensuel de la Société Chimique de Paris* **1865**, *3*, 98 (cit. on p. 4).

[30] Hammett, L. P. 'The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives'. In: *J. Am. Chem. Soc.* **1937**, *59*, 96–103. DOI: 10.1021/ja01280a022 (cit. on p. 5).

[31] Martinez-Cuezva, A., Pastor, A., Marin-Luna, M., Diaz-Marin, C., Bautista, D., Alajarin, M. and Berna, J. 'Cyclization of interlocked fumaramides into β-lactams: experimental and computational mechanistic assessment of the key intercomponent proton transfer and the stereocontrolling active pocket'. In: *Chem. Sci.* **2021**, *12*, 747–756. DOI: 10.1039/d0sc05757f (cit. on p. 6).

[32] Morán-González, L., Betten, J. E., Kneiding, H. and Balcells, D. 'AABBA: Atom-Atom Bond-Bond Bond-Atom Graph Kernel for Machine Learning on Molecules and Materials'. In: *ChemRxiv* **2023**, 1–43. DOI: 10.26434/chemrxiv-2023-5wbkr (cit. on pp. 6, 18, 23, 137).

[33] Tang, T., Hazra, A., Min, D. S., Williams, W. L., Jones, E., Doyle, A. G. and Sigman, M. S. 'Interrogating the Mechanistic Features of Ni(I)-Mediated Aryl Iodide Oxidative Addition Using Electroanalytical and Statistical Modeling Techniques'. In: *J. Am. Chem. Soc.* **2023**. DOI: 10.1021/jacs.3c01726 (cit. on pp. 6, 14).

[34] Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. and Glorius, F. 'A Structure-Based Platform for Predicting Chemical Reactivity'. In: *Chem* **2020**, *6*, 1379–1390. DOI: 10.1016/j.chempr.2020.02.017 (cit. on p. 8).

[35] Young, T. A., Silcock, J. J., Sterling, A. J. and Duarte, F. 'autodE: Automated Calculation of Reaction Energy Profiles— Application to Organic and Organometallic Reactions'. In: *Angew. Chem. Int. Ed.* **2021**, *60*, 4266–4274. DOI: 10.1002/anie.202011941 (cit. on p. 8).

[36] Sabadell-Rendón, A., Kaźmierczak, K., Morandi, S., Euzenat, F., Curulla-Ferré, D. and López, N. 'Automated Multiscale Universal Simulation Environment'. In: *ChemRxiv* **2023**, 1–26 (cit. on p. 8).

[37] Weininger, D. 'SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules'. In: *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. DOI: 10.1021/ci00057a005 (cit. on pp. 8, 138).

[38] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P. and Aspuru-Guzik, A. 'Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules'. In: *ACS Cent. Sci.* **2018**, *4*, 268–276. DOI: 10.1021/acscentsci.7b00572 (cit. on p. 8).

[39] Niemeyer, Z. L., Milo, A., Hickey, D. P. and Sigman, M. S. 'Parameterization of phosphine ligands reveals mechanistic pathways and predicts reaction outcomes'. In: *Nat. Chem* **2016**, *8*, 610–617. DOI: 10.1038/nchem.2501 (cit. on p. 8).

[40] Cramer, R. D., Patterson, D. E. and Bunce, J. D. 'Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins'. In: *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967. DOI: 10.1021/ja00226a005 (cit. on p. 9).

[41] Dijk, L. van, Ardkhean, R., Sidera, M., Karabiyikoglu, S., Sari, Ö., Claridge, T. D. W., Lloyd-Jones, G. C., Paton, R. S. and Fletcher, S. P. 'Mechanistic investigation of Rh(i)-catalysed asymmetric Suzuki–Miyaura coupling with racemic allyl halides'. In: *Nat. Catal.* **2021**, *4*, 284–292. DOI: 10.1038/s41929-021-00589-y (cit. on p. 9).

[42] Li, X., Zhang, S. Q., Xu, L.-C. and Hong, X. 'Predicting Regioselectivity in Radical C–H Functionalization of Heterocycles through Machine Learning'. In: *Angew. Chem. Int. Ed.* **2020**, *59*, 13253–13259. DOI: 10.1002/anie.202000959 (cit. on p. 10).

[43] Poater, A., Cosenza, B., Correa, A., Giudice, S., Ragone, F., Scarano, V. and Cavallo, L. 'SambVca: A Web Application for the Calculation of the Buried Volume of N-Heterocyclic Carbene Ligands'. In: *Eur. J. Inorg. Chem.* **2009**, *2009*, 1759–1766. DOI: 10.1002/ejic.200801160 (cit. on p. 10).

[44] III, R. D. C. 'Chapter 31: Quantitative Drug Design'. In: *Annual Reports in Medicinal Chemistry*. Ed. by Frank H. Clarke. *Vol. 11*. Academic Press, **1976**, 301–310. DOI: 10.1016/S0065-7743(08)61415-3 (cit. on p. 10).

[45] Zahrt, A. F., Athavale, S. V. and Denmark, S. E. 'Quantitative Structure-Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future'. In: *Chem. Rev.* **2020**, *120*, 1620–1689. DOI: 10.1021/acs.chemrev.9b00425 (cit. on p. 10).

[46] Vermeire, F. H., Chung, Y. and Green, W. H. 'Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures'. In: *J. Am. Chem. Soc.* **2022**, *144*, 10785–10797. DOI: 10.1021/jacs.2c01768 (cit. on p. 10).

[47] Yang, K. et al. 'Analyzing Learned Molecular Representations for Property Prediction'. In: *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388. DOI: 10.1021/acs.jcim.9b00237 (cit. on p. 10).

[48] Singh, S. and Sunoj, R. B. 'Molecular Machine Learning for Chemical Catalysis: Prospects and Challenges'. In: *Acc. Chem. Res.* **2023**, *56*, 402–412. DOI: 10.1021/acs.accounts.2c00801 (cit. on p. 12).

[49] Schwaller, P., Vaucher, A. C., Laino, T. and Reymond, J. L. 'Prediction of chemical reaction yields using deep learning'. In: *Mach. Learn.: Sci. Technol.* **2021**, *2*, 015016. DOI: 10.1088/2632-2153/abc81d (cit. on p. 12).

[50] Santiago, C. B., Guo, J. Y. and Sigman, M. S. 'Predictive and mechanistic multivariate linear regression models for reaction development'. In: *Chem. Sci.* **2018**, *9*, 2398–2412. DOI: 10.1039/c7sc04679k (cit. on p. 12).

[51] Gallegos, L. C., Luchini, G., John, P. C. S., Kim, S. and Paton, R. S. 'Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties'. In: *Acc. Chem. Res.* **2021**, *54*, 827–836. DOI: 10.1021/acs.accounts.0c00745 (cit. on p. 12).

[52] Williams, W. L., Zeng, L., Gensch, T., Sigman, M. S., Doyle, A. G. and Anslyn, E. V. 'The Evolution of Data-Driven Modeling in Organic Chemistry'. In: *ACS Cent. Sci.* **2021**, *7*, 1622–1637. DOI: 10.1021/acscentsci.1c00535 (cit. on pp. 13, 17).

[53] Luchini, G. and Paton, R. S. 'Bottom-up Atomistic Descriptions of Top-Down Macroscopic Measurements: Computational Benchmarks for Hammett Electronic Parameters'. In: *ChemRxiv* **2023**, 1–16. DOI: 10.26434/chemrxiv-2023-n8jsm-v2 (cit. on p. 13).

[54] Brønsted, J. N. and Pedersen, K. 'Die katalytische Zersetzung des Nitramids und ihre physikalisch-chemische Bedeutung'. In: *Zeitschrift für Physikalische Chemie* **1924**, *108U*, 185–235. DOI: 10.1515/zpch-1924-10814 (cit. on p. 13).

[55] Gasteiger, J. 'Chemoinformatics: Achievements and challenges, a personal view'. In: *Molecules* **2016**, *21*, 151. DOI: 10.3390/molecules21020151 (cit. on p. 13).

[56] Wodrich, M. D., Sawatlon, B., Busch, M. and Corminboeuf, C. 'The Genesis of Molecular Volcano Plots'. In: *Acc. Chem. Res.* **2021**, *54*, 1107–1117. DOI: 10.1021/acs.accounts.0c00857 (cit. on p. 13).

[57] Kalkman, E. D., Qiu, Y. and Hartwig, J. F. 'Transition-State Stabilization by Secondary Orbital Interactions between Fluoroalkyl Ligands and Palladium During Reductive Elimination from Palladium(aryl)(fluoroalkyl) Complexes'. In: *ACS Catal.* **2023**, 12810–12825. DOI: 10.1021/acscatal.3c02648 (cit. on p. 14).

[58] Sigman, M. S. and Miller, J. J. 'Examination of the role of Taft-type steric parameters in asymmetric catalysis'. In: *J. Org. Chem.* **2009**, *74*, 7633–7643. DOI: 10.1021/jo901698t (cit. on p. 14).

[59] Solé-Daura, A., Poblet, J. M. and Carbó, J. J. 'Structure–Activity Relationships for the Affinity of Chaotropic Polyoxometalate Anions towards Proteins'. In: *Chem. Eur. J.* **2020**, *26*, 5799–5809. DOI: 10.1002/chem.201905533 (cit. on p. 14).

[60] Taft, R. W. J. 'Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters1'. In: *J. Am. Chem. Soc.* **1952**, *74*, 3120–3128. DOI: 10.1021/ja01132a049 (cit. on p. 14).

[61] Besora, M., Olmos, A., Gava, R., Noverges, B., Asensio, G., Caballero, A., Maseras, F. and Pérez, P. J. 'A Quantitative Model for Alkane Nucleophilicity Based on C-H Bond Structural/Topological Descriptors'. In: *Angew. Chem. Int. Ed.* **2020**, *59*, 3112–3116. DOI: 10.1002/anie.201914386 (cit. on p. 14).

[62] Engel, T. 'Basic Overview of Chemoinformatics'. In: *J. Chem. Inf. Model.* **2006**, *46*, 2267–2277. DOI: 10.1021/ci600234z (cit. on p. 16).

[63] Appel, R., Hochstrasser, D., Roch, C., Funk, M., Muller, A. F. and Pellegrini, C. 'Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: A step toward

machine learning'. In: *ELECTROPHORESIS* **1988**, *9*, 136–142. DOI: 10.1002/elps.1150090307 (cit. on p. 16).

[64] Coley, C. W. et al. 'A robotic platform for flow synthesis of organic compounds informed by AI planning'. In: *Science* **2019**, *365*, eaax1566. DOI: 10.1126/science.aax1566 (cit. on p. 17).

[65] Mishra, C., Wolff, N. von, Tripathi, A., Brodie, C. N., Lawrence, N. D., Ravuri, A., Brémond, É., Preiss, A. and Kumar, A. 'Predicting ruthenium catalysed hydrogenation of esters using machine learning'. In: *Digital Discovery* **2023**, *2*, 819–827. DOI: 10.1039/D3DD00029J (cit. on p. 17).

[66] Pollice, R. et al. 'Data-Driven Strategies for Accelerated Materials Design'. In: *Acc. Chem. Res.* **2021**, *54*, 849–860. DOI: 10.1021/acs.accounts.0c00785 (cit. on p. 17).

[67] Jumper, J. et al. 'Highly accurate protein structure prediction with AlphaFold'. In: *Nature* **2021**, *596*, 583–589. DOI: 10.1038/s41586-021-03819-2 (cit. on p. 17).

[68] Rahman, T., Petrus, E., Segado, M., Martin, N. P., Palys, L. N., Rambaran, M. A., Ohlin, C. A., Bo, C. and Nyman, M. 'Predicting the Solubility of Inorganic Ion Pairs in Water'. In: *Angew. Chem. Int. Ed.* **2022**, *61*. DOI: 10.1002/anie.202117839 (cit. on p. 17).

[69] Hueffel, J. A., Sperger, T., Funes-Ardoiz, I., Ward, J. S., Rissanen, K. and Schoenebeck, F. 'Accelerated dinuclear palladium catalyst identification through unsupervised machine learning'. In: *Science* **2021**, *374*, 1134–1140. DOI: 10.1126/science.abj0999 (cit. on pp. 17, 39).

[70] Janet, J. P. and Kulik, H. J. 'Predicting electronic structure properties of transition metal complexes with neural networks'. In: *Chem. Sci.* **2017**, *8*, 5137–5152. DOI: 10.1039/c7sc01247k (cit. on pp. 18, 138).

[71] Saadun, A. J., Pablo-García, S., Paunović, V., Li, Q., Sabadell-Rendón, A., Kleemann, K., Krumeich, F., López, N. and Pérez-Ramírez, J. 'Performance of Metal-Catalyzed Hydrodebromination of Dibromomethane Analyzed by Descriptors Derived from Statistical Learning'. In: *ACS Catal.* **2020**, *10*, 6129–6143. DOI: 10.1021/acscatal.0c00679 (cit. on pp. 18, 38).

[72] Pablo-García, S., Morandi, S., Vargas-Hernández, R. A., Jorner, K., Ivković, Ž., López, N. and Aspuru-Guzik, A. 'Fast evaluation of the adsorption energy of organic molecules on metals via graph neural networks'. In: *Nat. Comput. Sci* **2023**, *3*, 433–442. DOI: 10.1038/s43588-023-00437-y (cit. on pp. 18, 139).

[73] Skoraczyński, G., Dittwald, P., Miasojedow, B., Szymkuć, S., Gajewska, E. P., Grzybowski, B. A. and Gambin, A. 'Predicting the outcomes of organic reactions via machine learning: Are current descriptors sufficient?' In: *Sci. Rep.* **2017**, *7*. DOI: 10.1038/s41598-017-02303-0 (cit. on p. 18).

[74] Ahneman, D. T., F., E. J., Shishi, L., D., D. S. and G., D. A. 'Predicting reaction performance in C-N cross-coupling using machine learning'. In: *Science* **2018**, *360*, 186–190. DOI: 10.1126/science.aar5169 (cit. on p. 18).

[75] Chuang, K. V. and Keiser, M. J. 'Comment on "Predicting reaction performance in C–N cross-coupling using machine learning"'. In: *Science* **2018**, *362*, eaat8603. DOI: 10.1126/science.aat8603 (cit. on p. 18).

[76] Paszke, A. et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates, Inc., **2019**, 8024–8035 (cit. on pp. 19, 51, 158).

[77] Pedregosa, F. et al. 'Scikit-learn: Machine Learning in Python'. In: *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830 (cit. on p. 19).

[78] Harris, C. R. et al. 'Array programming with NumPy'. In: *Nature* **2020**, *585*, 357–362. DOI: 10.1038/s41586-020-2649-2 (cit. on pp. 19, 103, 156).

[79] Tzaguy, A., Masip-Sánchez, A., Avram, L., Solé-Daura, A., López, X., Poblet, J. M. and Neumann, R. 'Electrocatalytic Reduction of Dinitrogen to Ammonia with Water as Proton and Electron Donor Catalyzed by a Combination of a Tri-ironoxotungstate and an Alkali Metal Cation'. In: *J. Am. Chem. Soc.* **2023**, *145*, 19912–19924. DOI: 10.1021/jacs.3c06167 (cit. on p. 23).

[80] Jensen, F. *Introduction to Computational Chemistry*. Third Edition. Wiley, **2016**, 0-664 (cit. on pp. 23, 28).

[81] Dirac, P. A. M. 'Quantum Mechanics of Many-Electron Systems'. In: *Proc. R. Soc. A: Math. Phys. Eng. Sci.* **1929**, *123*, 714–733. DOI: 10.1098/rspa.1929.0094 (cit. on p. 24).

[82] Eyring, H. 'The Activated Complex in Chemical Reactions'. In: *J. Chem. Phys.* **1935**, *3*, 107–115. DOI: 10.1063/1.1749604 (cit. on p. 26).

[83] Keith, J. A., Vassilev-Galindo, V., Cheng, B., Chmiela, S., Gastegger, M., Müller, K.-R. and Tkatchenko, A. 'Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems'. In: *Chem. Rev.* **2021**, *121*, 9816–9872. DOI: 10.1021/acs.chemrev.1c00107 (cit. on p. 28).

[84] Řezáč, J. and Hobza, P. 'Describing noncovalent interactions beyond the common approximations: How accurate is the "gold standard," CCSD(T) at the complete basis set limit?' In: *J. Chem. Theory Comput.* **2013**, *9*, 2151–2155. DOI: 10.1021/ct400057w (cit. on p. 28).

[85] Hohenberg, P. 'Inhomogeneous Electron Gas'. In: *Phys. Rev.* **1964**, *136*, B864–B871. DOI: 10.1103/PhysRev.136.B864 (cit. on p. 29).

[86] Kohn, W. and Sham, L. J. 'Self-Consistent Equations Including Exchange and Correlation Effects'. In: *Phys. Rev.* **1965**, *140*, A1133–A1138. DOI: 10.1103/PhysRev.140.A1133 (cit. on p. 29).

[87] Foundation, N. *The Nobel Prize in Chemistry 1998*. **2023**. URL: https://www.nobelprize.org/prizes/chemistry/1998/summary/ (cit. on p. 29).

[88] Perdew, J. P. and Schmidt, K. 'Jacob's ladder of density functional approximations for the exchange-correlation energy'. In: *AIP Conf. Proc.* **2001**, *577*, 1–20. DOI: 10.1063/1.1390175 (cit. on p. 29).

[89] Bursch, M., Mewes, J.-M., Hansen, A. and Grimme, S. 'Best-Practice DFT Protocols for Basic Molecular Computational Chemistry'. In: *Angew. Chem. Int. Ed.* **2022**, *61*, e202205735. DOI: 10.1002/ange.202205735 (cit. on p. 31).

[90] West, D. B. et al. *Introduction to graph theory. Vol. 2*. Prentice hall Upper Saddle River, **2001** (cit. on p. 31).

[91] Petrus, E., Segado, M. and Bo, C. 'Nucleation mechanisms and speciation of metal oxide clusters'. In: *Chem. Sci.* **2020**, *11*, 8448–8456. DOI: 10.1039/d0sc03530k (cit. on p. 32).

[92] Morán-González, L., Besora, M. and Maseras, F. 'Seeking the Optimal Descriptor for SN2 Reactions through Statistical Analysis of Density Functional Theory Results'. In: *J. Org. Chem.* **2022**, *87*, 363–372. DOI: 10.1021/acs.joc.1c02387 (cit. on p. 33).

[93] Alegre-Requena, J. V. and Dalmau, D. *ROBERT v1.0.* **2023**. URL: https://github.com/jvalegre/robert (cit. on p. 34).

[94] Stewartt, G. W. 'On the Early History of the Singular Value Decomposition'. In: *SIAM Review* **1993**, *35*, 1876–1959. URL: http://www.siam.org/journals/ojsa.php (cit. on p. 35).

[95] Harris, C. R. et al. 'Array programming with NumPy'. In: *Nature* **2020**, *585*, 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2 (cit. on p. 38).

[96] Pablo-García, S., García-Muelas, R., Sabadell-Rendón, A. and López, N. *Dimensionality reduction of complex reaction networks in heterogeneous catalysis: From linear-scaling relationships to statistical learning techniques.* **2021**. DOI: 10.1002/wcms.1540 (cit. on p. 39).

[97] Okuwaki, K., Akisawa, K., Hatada, R., Mochizuki, Y., Fukuzawa, K., Komeiji, Y. and Tanaka, S. 'Collective residue interactions in trimer complexes of SARS-CoV-2 spike proteins analyzed by fragment molecular orbital method'. In: *Appl. Phys. Express* **2022**, *15*, 017001. DOI: 10.35848/1882-0786/ac4300 (cit. on p. 43).

[98] Blavier, M., Levine, R. D. and Remacle, F. 'Time evolution of entanglement of electrons and nuclei and partial traces in ultrafast photochemistry'. In: *Phys. Chem. Chem. Phys.* **2022**, *24*, 17516–17525. DOI: 10.1039/d2cp01440h (cit. on p. 43).

[99] Schoendorff, G., West, A. C., Schmidt, M. W., Ruedenberg, K., Wilson, A. K. and Gordon, M. S. 'Relativistic ab Initio Accurate Atomic Minimal Basis Sets: Quantitative LUMOs and Oriented Quasi-Atomic Orbitals for the Elements Li-Xe'. In: *J. Phys. Chem. A* **2017**, *121*, 3588–3597. DOI: 10.1021/acs.jpca.7b01916 (cit. on p. 43).

[100] Maruyama, K., Sheng, Y., Watanabe, H., Fukuzawa, K. and Tanaka, S. 'Application of singular value decomposition to the inter-fragment interaction energy analysis for ligand screening'. In: *Comput. Theor. Chem.* **2018**, *1132*, 23–34. DOI: 10.1016/j.comptc.2018.04.001 (cit. on p. 43).

[101] Cooke, S. A. and Minei, A. J. 'The pure rotational spectrum of 1,1,2,2,3-pentafluorocyclobutane and applications of singular value decomposition signal processing'. In: *J. Mol. Spectrosc.* **2014**, *306*, 37–41. DOI: 10.1016/j.jms.2014.10.007 (cit. on p. 43).

[102] Navarro-Vázquez, A. 'MSpin-RDC. A program for the use of residual dipolar couplings for structure elucidation of small molecules'. In: *Magn. Reson. Chem* **2012**, *50*, S73–S79. DOI: 10.1002/mrc.3905 (cit. on p. 43).

[103] Gomez, J. C., Moreno, J., Ibarra-Manzano, M.-A. and Almanza-Ojeda, D.-L. 'Reconstructive Classification for Age and Gender Identification in Social Networks'. In: *IEEE Trans. Comput. Soc.* **2023**, 1–11. DOI: 10.1109/TCSS.2023.3267766 (cit. on p. 43).

[104] Lakuntza, O., Besora, M. and Maseras, F. 'Searching for Hidden Descriptors in the Metal-Ligand Bond through Statistical Analysis of Density Functional Theory (DFT) Results'. In: *Inorg. Chem.* **2018**, *57*, 14660–14670. DOI: 10.1021/acs.inorgchem.8b02372 (cit. on pp. 45, 53, 58, 59, 65, 73–75, 77, 80, 110, 121, 132).

[105] Chollet, F. et al. *Keras*. **2015**. URL: https://github.com/fchollet/keras (cit. on p. 51).

[106] Arduengo, A. J., Davidson, F., Dias, H. V. R., Goerlich, J. R., Khasnis, D., Marshall, W. J. and Prakasha, T. K. 'An Air Stable Carbene and Mixed Carbene "Dimers"'. In: *J. Am. Chem. Soc.* **1997**, *119*, 12742–12749. DOI: 10.1021/ja973241o (cit. on p. 55).

[107] Hsu, Y. C. et al. 'One-Pot Tandem Photoredox and Cross-Coupling Catalysis with a Single Palladium Carbodicarbene Complex'. In: *Angew. Chem. Int. Ed.* **2018**, *57*, 4622–4626. DOI: 10.1002/anie.201800951 (cit. on p. 55).

[108] Ohmiya, H. 'N-Heterocyclic Carbene-Based Catalysis Enabling Cross-Coupling Reactions'. In: *ACS Catal.* **2020**, *10*, 6862–6869. DOI: 10.1021/acscatal.0c01795 (cit. on p. 55).

[109] Vougioukalakis, G. C. and Grubbs, R. H. 'Ruthenium-Based Heterocyclic Carbene-Coordinated Olefin Metathesis'. In: *Chem. Rev.* **2010**, *110*, 1746–1787. DOI: doi.org/10.1021/cr9002424 (cit. on p. 55).

[110] Ogba, O. M., Warner, N. C., O'Leary, D. J. and Grubbs, R. H. 'Recent advances in ruthenium-based olefin metathesis'. In: *Chem. Soc. Rev.* **2018**, *47*, 4510–4544. DOI: 10.1039/c8cs00027a (cit. on p. 55).

[111] Khan, R. K. M., Torker, S. and Hoveyda, A. H. 'Readily Accessible and Easily Modifiable Ru-Based Catalysts for Efficient and Z-Selective Ring-Opening Metathesis Polymerization and Ring-Opening/Cross-Metathesis'. In: *J. Am. Chem. Soc.* **2013**, *135*, 10258–10261. DOI: doi.org/10.1021/ja404208a (cit. on p. 55).

[112] Zhao, K. and Enders, D. 'Merging N-Heterocyclic Carbene Catalysis and Single Electron Transfer: A New Strategy for Asymmetric Transformations'. In: *Angew. Chem. Int. Ed.* **2017**, *56*, 3754–3756. DOI: 10.1002/anie.201700370 (cit. on p. 55).

[113] Peris, E. 'Smart N-Heterocyclic Carbene Ligands in Catalysis'. In: *Chem. Rev.* **2018**, *118*, 9988–10031. DOI: 10.1021/acs.chemrev.6b00695 (cit. on pp. 55, 70).

[114] Scholl, M., Ding, S., Lee, C. W. and Grubbs, R. H. 'Synthesis and activity of a new generation of ruthenium-based olefin metathesis catalysts coordinated with 1,3-dimesityl-4,5-dihydroimidazol-2-ylidene ligands'. In: *Org. Lett.* **1999**, *1*, 953–956. DOI: 10.1021/ol990909q (cit. on p. 55).

[115] Huang, J., Stevens, E. D., Nolan, S. P. and Petersen, J. L. 'Olefin metathesis-active ruthenium complexes bearing a nucleophilic carbene ligand'. In: *J. Am. Chem. Soc.* **1999**, *121*, 2674–2678. DOI: 10.1021/ja9831352 (cit. on p. 55).

[116] Bernhammer, J. C., Frison, G. and Huynh, H. V. 'Electronic structure trends in N-heterocyclic carbenes (NHCs) with varying number of nitrogen atoms and NHC-transition-metal bond properties'. In: *Chem. Eur. J.* **2013**, *19*, 12892–12905. DOI: 10.1002/chem.201301093 (cit. on p. 55).

[117] Heydenrych, G., Hopffgarten, M. von, Stander, E., Schuster, O., G., R. H. and Frenking, G. 'The Nature of the Metal–Carbene Bond in Normal and Abnormal Pyridylidene, Quinolylidene and Isoquinolylidene Complexes'. In: *Eur. J. Inorg. Chem.* **2009**, *2009*, 1892–1904. DOI: 10.1002/ejic.200801244 (cit. on p. 55).

[118] Gusev, D. G. 'Electronic and steric parameters of 76 N-heterocyclic carbenes in Ni(CO)3(NHC)'. In: *Organometallics* **2009**, *28*, 6458–6461. DOI: 10.1021/om900654g (cit. on pp. 55, 60).

[119] Antonova, N. S., Carbó, J. J. and Poblet, J. M. 'Quantifying the donor-acceptor properties of phosphine and N-heterocyclic carbene ligands in grubbs' catalysts using a modified EDA procedure based on orbital deletion'. In: *Organometallics* **2009**, *28*, 4283–4287. DOI: 10.1021/om900180m (cit. on p. 55).

[120] Azofra, L. M., Vummaleti, S. V. C., Zhang, Z., Poater, A. and Cavallo, L. 'σ/π Plasticity of NHCs on the Ruthenium–Phosphine and Ruthenium-Ylidene Bonds in Olefin Metathesis Catalysts'. In: *Organometallics* **2020**, 3972–3982. DOI: 10.1021/acs.organomet.0c00536 (cit. on p. 55).

[121] Tonner, R., Heydenrych, G. and Frenking, G. 'Bonding analysis of N-heterocyclic carbene tautomers and phosphine ligands in transition-metal complexes: A theoretical study'. In: *Chem. Asian J.* **2007**, *2*, 1555–1567. DOI: 10.1002/asia.200700235 (cit. on p. 55).

[122] Hopkinson, M. N., Richter, C., Schedler, M. and Glorius, F. 'An overview of N-heterocyclic carbenes'. In: *Nature* **2014**, *510*, 485–496. DOI: 10.1038/nature13384 (cit. on p. 55).

[123] Balaraman, E., Gunanathan, C., Zhang, J., Shimon, L. J. and Milstein, D. 'Efficient hydrogenation of organic carbonates, carbamates and formates indicates alternative routes to methanol based on CO2 and CO'. In: *Nat. Chem.* **2011**, *3*, 609–614. DOI: 10.1038/nchem.1089 (cit. on p. 55).

[124] Friederich, P., Gomes, G. D. P., Bin, R. D., Aspuru-Guzik, A. and Balcells, D. 'Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex'. In: *Chem. Sci.* **2020**, *11*, 4584–4601. DOI: 10.1039/d0sc00445f (cit. on pp. 55, 140, 159).

[125] Besora, M. and Maseras, F. 'Chapter Six - Computational insights into metal-catalyzed asymmetric hydrogenation'. In: ed. by Montserrat Diéguez and Antonio Pizzano. *Vol. 68*. Academic Press, **2021**, 385–426. DOI: https://doi.org/10.1016/bs.acat.2021.08.006 (cit. on p. 55).

[126] Riehl, J. F., Pelissier, M. and Eisenstein, O. 'Influence of a cis hydride on a coordinated molecular hydrogen ligand cis hydride, Ab initio calculations'. In: *Inorg. Chem.* **1992**, *31*, 3344–3345. DOI: 10.1021/ic00042a003 (cit. on p. 55).

[127] Kubas, G. J. 'Metal – dihydrogen and s-bond coordination: the consummate extension of the Dewar–Chatt–Duncanson model for metal–olefin p bonding'. In: *J. Organomet. Chem.* **2001**, *635*. DOI: 10.1016/S0022-328X(01)01066-X (cit. on p. 55).

[128] Heinekey, M. D., Lledós, A. and Lluch, J. M. 'Elongated dihydrogen complexes: what remains of the H–H Bond?' In: *Chem. Soc. Rev.* **2004**, *33*, 175–182. DOI: 10.1039/B304879A (cit. on p. 55).

[129] Kubas, G. J., Ryan, R. R., Swanson, B. I., Vergamini, P. J. and Wasserman, H. J. 'Characterization of the first examples of isolable molecular hydrogen complexes, M(CO)3(PR3)2(H2) (M = molybdenum or tungsten; R = Cy or isopropyl). Evidence for a side-on bonded dihydrogen ligand'. In: *J. Am. Chem. Soc.* **1984**, *106*, 451–452. DOI: 10.1021/ja00314a049 (cit. on p. 55).

[130] Maseras, F., Duran, M., Lledós, A. and Bertrán, J. 'Molecular Hydrogen Complexes with a Hydride Ligand. An ab Initio Study on the [Fe(PR3)4H(H2)]+ System'. In: *J. Am. Chem. Soc.* **1991**, *113*, 2879–2884. DOI: 10.1021/ja00008a014 (cit. on p. 57).

[131] Maseras, F., Duran, M., Lledós, A. and Bertrán, J. 'Intramolecular atom exchange between molecular hydrogen and hydride ligands in cis-[Fe(PR3)4H(H2)]+ complexes. An ab initio theoretical study'. In: *J. Am. Chem. Soc.* **1992**, *114*, 2922–2928. DOI: 10.1021/ja00034a025 (cit. on p. 57).

[132] Maseras, F., Lledós, A., Clot, E. and Eisenstein, O. 'Transition Metal Polyhydrides: From Qualitative Ideas to Reliable Computational Studies'. In: *Chem. Rev.* **2000**, *100*, 601–636. DOI: 10.1021/cr980397d (cit. on p. 57).

[133] Besora, M., Lledós, A. and Maseras, F. 'Protonation of transition-metal hydrides: a not so simple process'. In: *Chem. Soc. Rev.* **2009**, *38*, 957. DOI: 10.1039/b608404b (cit. on p. 57).

[134] Ortuño, M. A. and Lledós, A. 'How acid can become a dihydrogen complex in water? A DFT study'. In: *J. Org. Chem.* **2021**, *949*, 121957. DOI: 10.1016/j.jorganchem.2021.121957 (cit. on p. 57).

[135] Ozin, G. A. and Garcia-Prieto, J. 'Pd(n1-H2) and Pd(n2-H2): Ligand-Free End-on and Side-on Bonded Molecular Dihydrogen Complexes'. In: *J. Am. Chem. Soc.* **1986**, *108*, 3100–3102. DOI: 10.1021/ja00271a047 (cit. on p. 57).

[136] Musaev, D. G. and Charkin, O. P. 'Theoretical study of the structure and stability of complexes of molecular hydrogen with K+, Cu+, Be2+, and Zn2+'. In: *J. Struct. Chem.* **1989**, *30*, 365–372. DOI: 10.1007/BF00751892 (cit. on p. 57).

[137] Morán-González, L., Pedregal, J. R.-G., Besora, M. and Maseras, F. 'Understanding the Binding Properties of N-heterocyclic Carbenes through BDE Matrix App'. In: *Eur. J. Inorg. Chem.* **2021**, e202100932. DOI: 10.1002/ejic.202100932 (cit. on p. 58).

[138] Pedregal, J. R.-G. *BDE Matrix App*. Version 1.0.0. **2018**. URL: https://maserasgroup-repo.github.io/bdeapp/ (cit. on p. 58).

[139] Frisch, M. J. et al. *Gaussian 09*. Gaussian Inc. Wallingford CT. **2009** (cit. on p. 59).

[140] Becke, A. D. 'Density-functional thermochemistry. III. The role of exact exchange'. In: *J. Chem. Phys.* **1993**, *98*, 5648–5652. DOI: 10.1063/1.464913 (cit. on pp. 59, 95).

[141] Lee, C., Yang, W. and Parr, R. G. 'Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density'. In: *Phys. Rev. B* **1988**, *37*, 785–789. DOI: 10.1103/PhysRevB.37.785 (cit. on pp. 59, 95).

[142] Grimme, S., Antony, J., Ehrlich, S. and Krieg, H. 'A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu'. In: *J. Chem. Phys.* **2010**, *132*, 1–19. DOI: 10.1063/1.3382344 (cit. on p. 59).

[143] Petersson, G. A., Bennett, A., Tensfeldt, T. G., Al-Laham, M. A., Shirley, W. A. and Mantzaris, J. 'A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements'. In: *J. Chem. Phys.* **1988**, *89*, 2193–2218. DOI: 10.1063/1.455064 (cit. on p. 59).

[144] Petersson, G. A. and Al-Laham, M. A. 'A Complete Basis Set Model Chemistry. II. The Total Energies of open-Shell Atoms and Hydrides of the First-Row Atoms'. In: *J. Chem. Phys.* **1991**, *9*, 6081–6090. DOI: 10.1063/1.455064 (cit. on p. 59).

[145] Dolg, H., Wedig, U., Stoll, H. and Preuss, H. 'Energy-adjusted ab initio pseudopotentials for the first row transition elements'. In: *J. Chem. Phys.* **1987**, *86*. DOI: 10.1063/1.452288 (cit. on p. 59).

[146] Igel-Mann, G., Stoll, H. and Preuss, H. 'Pseudopotentials for main group elements (IIIa through VIIa)'. In: *Mol. Phys.* **1988**, *65*, 1321–1328. DOI: 10.1080/00268978800101811 (cit. on p. 59).

[147] Tomasi, J., Mennucci, B. and Cammi, R. 'Quantum Mechanical Continuum Solvation Models'. In: *Chem. Rev.* **2005**, *105*, 2999–3094. DOI: 10.1021/cr9904009 (cit. on p. 59).

[148] Scalmani, G. and Frisch, M. J. 'Continuous surface charge polarizable continuum models of solvation. I. General formalism'. In: *J. Chem. Phys.* **2010**, *132*. DOI: 10.1063/1.3359469 (cit. on p. 59).

[149] Pérez-Soto, R., Besora, M. and Maseras, F. *pyssian v1.0.2, Maseras Lab.* **2021**. DOI: 10.5281/ZENODO.5055860 (cit. on pp. 59, 96).

[150] III, R. A. K., Clavier, H., Giudice, S., Scott, N. M., Stevens, E. D., Bordner, J., Samardjiev, I., Hoff, C. D., Cavallo, L. and Nolan, S. P. 'Determination of N-heterocyclic carbene (NHC) steric and electronic parameters using the system'. In: *Organometallics* **2008**, *27*, 202–210. DOI: 10.1021/om701001g (cit. on p. 67).

[151] Fürstner, A., Alcarazo, M., Radkowski, K. and Lehmann, C. W. 'Carbenes stabilized by ylides: Pushing the limits'. In: *Angew. Chem. Int. Ed.* **2008**, *47*, 8302–8306. DOI: 10.1002/anie.200803200 (cit. on p. 67).

[152] Schuster, O., Yang, L., Raubenheimer, H. G. and Albrecht, M. 'Beyond conventional N-heterocyclic carbenes: Abnormal, remote, and other classes of NHC ligands with reduced heteroatom stabilization'. In: *Chem. Rev.* **2009**, *109*, 3445–3478. DOI: 10.1021/cr8005087 (cit. on p. 67).

[153] Huang, H., Strater, Z. M. and Lambert, T. H. 'Electrophotocatalytic C-H Functionalization of Ethers with High Regioselectivity'. In: *J. Am. Chem. Soc.* **2020**, *142*, 1698–1703. DOI: 10.1021/jacs.9b11472 (cit. on p. 70).

[154] Tu, H. F., Jeandin, A. and Suero, M. G. 'Catalytic Synthesis of Cyclopropenium Cations with Rh-Carbynoids'. In: *J. Am. Chem. Soc.* **2022**, *144*, 16737–16743. DOI: 10.1021/jacs.2c07769 (cit. on p. 70).

[155] Lavallo, V., Canac, Y., Donnadieu, B., Schoeller, W. W. and Bertrand, G. 'Cyclopropenylidenes: From Interstellar Space to an Isolated Derivative in the Laboratory'. In: *Science* **2006**, *312*, 722–724. DOI: 10.1126/science.1126675 (cit. on p. 70).

[156] Morán-González, L. and Maseras, F. 'A computational search of the ideal metal fragment for monohapto coordination of dihydrogen'. In: *Aust. J. Chem.* **2023**. DOI: 10.1071/CH23121 (cit. on p. 73).

[157] Bento, A. P. and Bickelhaupt, F. M. 'Nucleophilicity and leaving-group ability in frontside and backside S N2 reactions'. In: *J. Org. Chem.* **2008**, *73*, 7290–7299. DOI: 10.1021/jo801215z (cit. on p. 94).

[158] Bickelhaupt, F. M., Swart, M. and Sola, M. 'Energy Landscapes of Nucleophilic Substitution Reactions : A Comparison of Density Functional Theory and Coupled Cluster Methods'. In: *J. Comput. Chem* **2007**, *28*, 1551–1560. DOI: 10.1002/jcc.20653 (cit. on p. 94).

[159] Kubelka, J. and Bickelhaupt, F. M. 'Activation strain analysis of SN2 reactions at C, N, O, and F centers'. In: *J. Phys. Chem. A* **2017**, *121*, 885–891. DOI: 10.1021/acs.jpca.6b12240 (cit. on p. 94).

[160] Keil, F. and Ahlrichs, R. 'Theoretical study of SN2 reactions. Ab initio computations on HF and CI level'. In: *J. Am. Chem. Soc.* **1976**, *98*, 4787–4793. DOI: 10.1021/ja00432a017 (cit. on p. 94).

[161] Alemán, C., Maseras, F., Lledós, A., Duran, M. and Bertrán, J. 'Analysis of solvent effect on SN2 reactions by different theoretical models'. In: *J. Phys. Org. Chem.* **1989**, *2*, 611–622. DOI: 10.1002/poc.610020804 (cit. on p. 94).

[162] Deng, L., Branchadell, V. and Ziegler, T. 'Potential Energy Surfaces of the Gas-Phase SN2 Reactions X-+ CH3X= XCH3+ X-(X= F, Cl, Br, I): A Comparative Study by Density Functional Theory and ab Initio Methods'. In: *J. Am. Chem. Soc.* **1994**, *116*, 10645–10656. DOI: 10.1021/ja00102a034 (cit. on p. 94).

[163] Swart, M., Solà, M. and Bickelhaupt, F. M. 'Energy landscapes of nucleophilic substitution reactions: A comparison of density functional theory and coupled cluster methods'. In: *J. Comput. Chem.* **2007**, *28*, 1551–1560. DOI: 10.1002/jcc.20653 (cit. on p. 94).

[164] Fernández, I., Frenking, G. and Uggerud, E. 'The interplay between steric and electronic effects in SN2 reactions'. In: *Chem. Eur. J.* **2009**, *15*, 2166–2175. DOI: 10.1002/chem.200801833 (cit. on p. 94).

[165] Gimadiev, T., Igor Tetko, M., Nugmanov, R., Casciuc, I., Klimchuk, O., Bodrov, A., Polishchuk, P., Antipin, I. and Varnek, A. 'Bimolecular Nucleophilic Substitution Reactions: Predictive Models for Rate Constants and Molecular Reaction Pairs Analysis'. In: *Mol. Inf.* **2019**, *38*, 1800104. DOI: 10.1002/minf.201800104 (cit. on p. 94).

[166] Fernández, I. and Bickelhaupt, F. M. 'The activation strain model and molecular orbital theory: understanding and designing chemical reactions'. In: *Chem. Soc. Rev.* **2014**, *43*, 4953–4967. DOI: 10.1039/C4CS00055B (cit. on p. 94).

[167] Frisch, M. J. et al. *Gaussian˜16 Revision C.01*. Gaussian Inc. Wallingford CT. **2016** (cit. on p. 95).

[168] McLean, A. D. and Chandler, G. S. 'Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11-18'. In: *J. Chem. Phys.* **1980**, *72*, 5639–5648. DOI: 10.1063/1.438980 (cit. on p. 95).

[169] Krishnan, R., Binkley, J. S., Seeger, R. and Pople, J. A. 'Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions'. In: *J. Chem. Phys.* **1980**, *72*, 650–654. DOI: 10.1063/1.438955 (cit. on p. 95).

[170] Clark, T., Chandrasekhar, J., Spitznagel, G. W. and Schleyer, P. V. R. 'Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li–F'. In: *J. Comput. Chem.* **1983**, *4*, 294–301. DOI: 10.1002/jcc.540040303 (cit. on p. 95).

[171] Wadt, W. R. and Hay, P. J. 'Ab initio effective core potentials for molecular calculations. Potentials for main group elements Na to Bi'. In: *J. Chem. Phys.* **1985**, *82*, 284–298. DOI: 10.1063/1.448800 (cit. on p. 95).

[172] Check, C. E., Faust, T. O., Bailey, J. M., Wright, B. J., Gilbert, T. M. and Sunderlin, L. S. 'Addition of polarization and diffuse functions to the LANL2DZ basis set for P-block elements'. In: *J. Phys. Chem. A* **2001**, *105*, 8111–8116. DOI: 10.1021/jp0119451 (cit. on p. 95).

[173] Marenich, A. V., Cramer, C. J. and Truhlar, D. G. 'Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions'. In: *J. Phys. Chem. B* **2009**, *113*, 6378–6396. DOI: 10.1021/jp810292n (cit. on p. 96).

[174] Gonzales, J. M., Cox, R. S., Brown, S. T., Allen, W. D. and Schaefer, H. F. 'Assessment of Density Functional Theory for Model S N 2 Reactions : CH3X + F- (X = F , Cl, CN, OH, SH, NH2, PH2)'. In: *J. Phys. Chem. A* **2001**, *105*, 11327–11346. DOI: 10.1021/jp012892a (cit. on p. 97).

[175] Uggerud, E. 'Nucleophilicity - Periodic trends and connection to basicity'. In: *Chem. Eur. J.* **2006**, *12*, 1127–1136. DOI: 10.1002/chem.200500639 (cit. on pp. 97, 114).

[176] Alkorta, I., Thacker, J. C. R. and Popelier, P. L. A. 'An Interacting Quantum Atom Study of Model SN2 Reactions (X-···CH3X, X = F, Cl, Br and I)'. In: *J. Comput. Chem* **2018**, *39*, 546–556. DOI: 10.1002/jcc.25098 (cit. on p. 97).

[177] Peverati, R. and Truhlar, D. G. 'Quest for a universal density functional: The accuracy of density functionals across a broad spectrum of databases in chemistry and physics'. In: *Phil. Trans. R. Soc. A.* **2014**, *372*. DOI: 10.1098/rsta.2012.0476 (cit. on p. 105).

[178] Zhao, Y. and Truhlar, D. G. 'The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals'. In: *Theor. Chem. Acc.* **2008**, *120*, 215–241. DOI: 10.1007/s00214-007-0310-x (cit. on p. 111).

[179] Bathgate, R. H. and Moelwyn-Hughes, E. A. '530. The kinetics of certain ionic exchange reactions of the four methyl halides in aqueous solution'. In: *J. Chem. Soc.* **1959**, 2642–2648. DOI: 10.1039/JR9590002642 (cit. on p. 111).

[180] Vlasov, V. M. 'Energetics of bimolecular nucleophilic reactions in solution'. In: *Russ. Chem. Rev.* **2006**, *75*, 765–796. DOI: 10.1070/rc2006v075n09abeh003614 (cit. on p. 111).

[181] Vlasov, V. M. 'Effects of substituents on activation parameters in sn2 reactions at aliphatic carbon in solution'. In: *J. Phys. Org. Chem.* **2010**, *23*, 468–476. DOI: 10.1002/poc.1634 (cit. on p. 111).

[182] Hamlin, T. A., Swart, M. and Bickelhaupt, F. M. 'Nucleophilic Substitution (SN2): Dependence on Nucleophile, Leaving Group, Central Atom, Substituents, and Solvent'. In: *ChemPhysChem* **2018**, *19*, 1315–1330. DOI: 10.1002/cphc.201701363 (cit. on p. 111).

[183] Iribarren, I., Trujillo, C., Sánchez-Sanz, G., Hénon, E., Elguero, J. and Alkorta, I. 'Influence of Lewis acids on the symmetric SN2 reaction'. In: *Theor. Chem. Acc.* **2023**, *142*. DOI: 10.1007/s00214-023-03013-9 (cit. on p. 111).

[184] Shibatomi, K., Kotozaki, M., Sasaki, N., Fujisawa, I. and Iwasa, S. 'Williamson Ether Synthesis with Phenols at a Tertiary Stereogenic Carbon: Formal Enantioselective Phenoxylation of β-Keto Esters'. In: *Chem. Eur. J.* **2015**, *21*, 14095–14098. DOI: 10.1002/chem.201502042 (cit. on p. 111).

[185] Mandal, S., Mandal, S., Ghosh, S. K., Sar, P., Ghosh, A., Saha, R. and Saha, B. 'A review on the advancement of ether synthesis from organic solvent to water'. In: *RSC Adv.* **2016**, *6*, 69605–69614. DOI: 10.1039/c6ra12914e (cit. on p. 111).

[186] Diamanti, A., Ganase, Z., Grant, E., Armstrong, A., Piccione, P. M., Rea, A. M., Richardson, J., Galindo, A. and Adjiman, C. S. 'Mechanism, kinetics and selectivity of a Williamson ether synthesis: Elucidation under different reaction conditions'. In: *React. Chem. Eng.* **2021**, *6*, 1195–1211. DOI: 10.1039/d0re00437e (cit. on p. 111).

[187] Jorner, K., Brinck, T., Norrby, P.-O. and Buttar, D. 'Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies'. In: *Chem. Sci.* **2021**, *12*, 1163–1175. DOI: 10.1039/D0SC04896H (cit. on p. 114).

[188] Gázquez, J. L., Cedillo, A. and Vela, A. 'Electrodonating and electroaccepting powers'. In: *J. Phys. Chem. A* **2007**, *111*, 1966–1970. ISSN: 10895639. DOI: 10.1021/jp065459f (cit. on pp. 114, 181, 182).

[189] Crespo, M., Martínez, M., Nabavizadeh, S. M. and Rashidi, M. 'Kinetico-mechanistic studies on CX (X=H, F, Cl, Br, I) bond activation reactions on organoplatinum(II) complexes'. In: *Coord. Chem. Rev.* **2014**, *279*, 115–140. DOI: 10.1016/j.ccr.2014.06.010 (cit. on p. 128).

[190] Kaneko, M. and Nakashima, S. 'Density Functional Theory Study on the ^{193}Ir Mössbauer Spectroscopic Parameters of Vaska's Complexes and Their Oxidative Adducts'. In: *Inorg. Chem.* **2021**, *60*, 12740–12752. DOI: 10.1021/acs.inorgchem.1c00239 (cit. on p. 128).

[191] Collman, J. P. and MacLaury, M. R. 'Neighboring group effect during oxidative addition'. In: *J. Am. Chem. Soc.* **1974**, *96*, 3019–3020. DOI: 10.1021/ja00816a073 (cit. on p. 128).

[192] Morán-González, L. *SN2 Matrix App.* **2023**. URL: https://maserasgroup-repo.github.io/sn2app/ (cit. on p. 131).

[193] Gao, K., Nguyen, D. D., Tu, M. and Wei, G.-W. 'Generative Network Complex for the Automated Generation of Drug-like Molecules'. In: *J. Chem. Inf. Model.* **2020**, *60*, 5682–5698. DOI: 10.1021/acs.jcim.0c00599 (cit. on p. 138).

[194] *SMARTS*. URL: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (cit. on p. 138).

[195] Krenn M Häse, F., Nigam, A., Friederich, P. and Aspuru-Guzik, A. 'Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation'. In: *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024. DOI: 10.1088/2632-2153/aba947 (cit. on p. 138).

[196] Moreau, G. and Broto, P. 'The autocorrelation of a topological structure: A new molecular descriptor'. In: *Nouv. J. Chim.* **1980**, 359–360 (cit. on pp. 140, 148).

[197] Janet, J. P. and Kulik, H. J. 'Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships'. In: *J. Phys. Chem. A* **2017**, *121*, 8939–8954. DOI: 10.1021/acs.jpca.7b08750 (cit. on p. 140).

[198] Knowles, W. S. 'Asymmetric Hydrogenations (Nobel Lecture)'. In: *Angew. Chem. Int. Ed.* **2002**, *41*, 1998–2007. DOI: https://doi.org/10.1002/1521-3773(20020617)41:12<1998::AID-ANIE1998>3.0.CO;2-8 (cit. on p. 141).

[199] Kneiding, H. *HyDGL*. URL: https://github.com/hkneiding/HyDGL (cit. on p. 141).

[200] Glendening, E. D., Landis, C. R. and Weinhold, F. 'Natural bond orbital methods'. In: *WIREs Comput Mol Sci* **2012**, *2*, 1–42. DOI: 10.1002/wcms.51 (cit. on p. 141).

[201] Cao, H. J., Zhao, Q., Zhang, Q. F., Li, J., Hamilton, E. J., Zhang, J., Wang, L. S. and Chen, X. 'Catalyst design based on agostic interactions: Synthesis, characterization, and catalytic activity of bis(pyrazolyl)borate copper complexes'. In: *Dalton Trans.* **2016**, *45*, 10194–10199. DOI: 10.1039/c6dt01272h (cit. on p. 142).

[202] Hagberg, A. A., Schult, D. A. and Swart, P. J. 'Exploring Network Structure, Dynamics, and Function using NetworkX'. In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught and Jarrod Millman. Pasadena, CA USA, **2008**, 11–15 (cit. on p. 156).

[203] Morán-González, L., Betten, J. E., Kneiding, H. and Balcells, D. *AABBA*. **2023**. URL: https://github.com/lmoranglez/AABBA (cit. on p. 168).

[204] Hackett, J. C. 'Chemical Reactivity Theory: A Density Functional View'. In: *J. Am. Chem. Soc.* **2010**, *132*, 7558–7558. DOI: 10.1021/ja1030744 (cit. on p. 180).

[205] Parr, R. G., Donnelly, R. A., Levy, M. and Palke, W. E. 'Electronegativity: the density functional viewpoint'. In: *J. Chem. Phys.* **1978**, *68*, 3801–3807. DOI: 10.1063/1.436185 (cit. on pp. 180, 181).

[206] Mulliken, R. S. 'A new electroaffinity scale; together with data on valence states and on valence ionization potentials and electron affinities'. In: *J. Chem. Phys.* **1934**, *2*, 782–793. DOI: 10.1063/1.1749394 (cit. on pp. 180, 181).

[207] Parr, R. G. and Pearson, R. G. 'Absolute hardness: companion parameter to absolute electronegativity'. In: *J. Am. Chem. Soc.* **1983**, *105*, 7512–7516. DOI: 10.1021/ja00364a005 (cit. on pp. 180, 181).

[208] Yang, W. and Parr, R. G. 'Hardness, softness, and the fukui function in the electronic theory of metals and catalysis.' In: *Proc. Natl. Acad. Sci. U. S. A.* **1985**, *82*, 6723–6726. DOI: 10.1073/pnas.82.20.6723 (cit. on pp. 180, 181).

[209] Maynard, A., Huang, M., Rice, W. and Covell, D. 'Reactivity of the HIV-1 nucleocapsid protein p7 zinc finger domains from the perspective of density-functional theory'. In: *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 11578–11583. DOI: 10.1073/pnas.95.20.11578 (cit. on pp. 180, 181).

[210] Parr, R. G., Szentpály, L. v. and Liu, S. 'Electrophilicity Index'. In: *J. Am. Chem. Soc.* **1999**, *121*, 1922–1924. DOI: 10.1021/ja983494x (cit. on pp. 180, 181).

[211] Pratihar, S. 'Electrophilicity and nucleophilicity of commonly used aldehydes'. In: *Org. Biomol. Chem.* **2014**, *12*, 5781–5788. DOI: 10.1039/c4ob00555d (cit. on pp. 181–183).

[212] Ayers, P. W., Anderson, J. S., Rodriguez, J. I. and Jawed, Z. 'Indices for predicting the quality of leaving groups'. In: *Phys. Chem. Chem. Phys* **2005**, *7*, 1918–1925. DOI: 10.1039/b500996k (cit. on pp. 183, 184).

[213] Ayers, P. W., Anderson, J. S. and Bartolotti, L. J. 'Perturbative perspectives on the chemical reaction prediction problem'. In: *Int. J. Quantum Chemistry* **2005**, *101*, 520–534. DOI: 10.1002/qua.20307 (cit. on pp. 183, 184).

[214] Montgomery Jr, J. A., Frisch, M. J., Ochterski, J. W. and Petersson, G. A. 'A complete basis set model chemistry. VI. Use of density functional geometries and frequencies'. In: *J. Chem. Phys.* **1999**, *110*, 2822–2827. DOI: 10.1063/1.477924 (cit. on p. 184).

[215] Montgomery Jr, J. A., Frisch, M. J., Ochterski, J. W. and Petersson, G. A. 'A complete basis set model chemistry. VII. Use of the minimum population localization method'. In: *J. Chem. Phys.* **2000**, *112*, 6532–6542. DOI: 10.1063/1.481224 (cit. on p. 184).

[216] Stephens, P., Jalkanen, K. and Kawiecki, R. 'Theory of vibrational rotational strengths: comparison of a priori theory and approximate models'. In: *J. Am. Chem. Soc.* **1990**, *112*, 6518–6529. DOI: 10.1021/ja00174a011 (cit. on p. 184).

[217] Reed, A. E., Weinstock, R. B. and Weinhold, F. 'Natural population analysis'. In: *J. Chem. Phys.* **1985**, *83*, 735–746. DOI: 10.1063/1.449486 (cit. on p. 184).

[218] Hirshfeld, F. L. 'Bonded-atom fragments for describing molecular charge densities'. In: *Theor. Chim. Acta* **1977**, *44*, 129–138. DOI: 10.1007/BF00549096 (cit. on p. 184).

[219] Ritchie, J. P. 'Electron density distribution analysis for nitromethane, nitromethide, and nitramide'. In: *J. Am. Chem. Soc.* **1985**, *107*, 1829–1837. DOI: 10.1021/ja00293a005 (cit. on p. 184).

[220] Ritchie, J. P. and Bachrach, S. M. 'Some methods and applications of electron density distribution analysis'. In: *J. Comput. Chem.* **1987**, *8*, 499–509. DOI: 10.1002/jcc.540080430 (cit. on p. 184).

[221] Marenich, A. V., Jerome, S. V., Cramer, C. J. and Truhlar, D. G. 'Charge model 5: An extension of Hirshfeld population analysis for the accurate description of molecular interactions in gaseous and condensed phases'. In: *J. Chem. Theory Comput.* **2012**, *8*, 527–541. DOI: 10.1021/ct200866d (cit. on p. 184).

[222] Singh, U. C. and Kollman, P. A. 'An approach to computing electrostatic charges for molecules'. In: *J. Comp. Chem.* **1984**, *5*, 129–145. DOI: 10.1002/jcc.540050204 (cit. on p. 184).

[223] Besler, B. H., Merz Jr, K. M. and Kollman, P. A. 'Atomic charges derived from semiempirical methods'. In: *J. Comp. Chem.* **1990**, *11*, 431–439. DOI: 10.1002/jcc.540110404 (cit. on p. 184).

[224] Brethome, A. V., Fletcher, S. P. and Paton, R. S. 'Conformational effects on physical-organic descriptors: the case of sterimol steric parameters'. In: *ACS Catal.* **2019**, *9*, 2313–2323. DOI: 10.1021/acscatal.8b04043 (cit. on p. 187).

[225] Wu, X.-P., Sun, X.-M., Wei, X.-G., Ren, Y., Wong, N.-B. and Li, W.-K. 'Exploring the reactivity trends in the E2 and SN2 reactions of X-+ CH3CH2Cl (X= F, Cl, Br, HO, HS, HSe, NH2 PH2, AsH2, CH3, SiH3, and GeH3)'. In: *J. Chem. Theory Comput.* **2009**, *5*, 1597–1606. DOI: 10.1021/ct900041y (cit. on p. 187).

UNIVERSITAT
ROVIRA i VIRGILI